関原　2023-9-25

Variational Autoencoders (VAE)

1　はじめに

Variational Autoencoders について解説する．

2　変分ベイズ法

汎関数（自由エネルギー）の最大化による事後分布の導出（ベイズ信号処理より）

まず，次のような評価関数 $\mathcal{F}[q,\boldsymbol{\theta}]$ を定義する．

$$\mathcal{F}[q,\boldsymbol{\theta}]=\int d\boldsymbol{x}q(\boldsymbol{x})[\log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})-\log q(\boldsymbol{x})] \tag{1}$$

この $\mathcal{F}[q,\boldsymbol{\theta}]$ は 2 つの量，ハイパーパラメータ $\boldsymbol{\theta}$ と任意の確率分布 $q(\boldsymbol{x})$ の関数である．この $\mathcal{F}[q,\boldsymbol{\theta}]$ は統計力学の用語を使い自由エネルギー（free energy）と呼ばれる．ここで，$\boldsymbol{\theta}$ は通常の変数であるが，$q(\boldsymbol{x})$ は確率分布であり，変数 $\boldsymbol{x}$ の関数である．したがって，$\mathcal{F}[q,\boldsymbol{\theta}]$ は関数 $q(\boldsymbol{x})$ の関数である．このように関数の関数であるものを汎関数と呼ぶ．

自由エネルギー $\mathcal{F}[q,\boldsymbol{\theta}]$ を最大にする確率分布 $q(\boldsymbol{x})$ を求めてみよう．汎関数 $\mathcal{F}[q,\boldsymbol{\theta}]$ を関数 $q(\boldsymbol{x})$ に関して最大とするのだが，関数 $q(\boldsymbol{x})$ はそもそも確率分布であるため制約条件，$\int_{-\infty}^{\infty}q(\boldsymbol{x})d\boldsymbol{x}=1$ が存在する．したがって，この最適化問題は

$$\widehat{q}(\boldsymbol{x})=\underset{q(\boldsymbol{x})}{\operatorname{argmax}}\,\mathcal{F}[q,\boldsymbol{\theta}],\quad \text{subject to}\quad \int_{-\infty}^{\infty}q(\boldsymbol{x})d\boldsymbol{x}=1 \tag{2}$$

となる．制約付き最適化問題であるので，ラグランジェ未定定数法を用いて無制約最適化問題に置き換えて解く．つまり，ラグランジェ未定定数を γ として，ラグランジアンを

$$\mathbb{L}[q,\gamma]=\mathcal{F}[q,\boldsymbol{\theta}]+\gamma\left[\int_{-\infty}^{\infty}q(\boldsymbol{x})d\boldsymbol{x}-1\right]=\int_{-\infty}^{\infty}d\boldsymbol{x}q(\boldsymbol{x})[\log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})-\log q(\boldsymbol{x})]+\gamma\left[\int_{-\infty}^{\infty}q(\boldsymbol{x})d\boldsymbol{x}-1\right] \tag{3}$$

と定義する．そして，$\mathbb{L}[q,\gamma]$ を $q(\boldsymbol{x})$ と γ について微分しゼロと置くことにより最適解を求める．

まず，$q(\boldsymbol{x})$ について汎関数 $\mathbb{L}[q,\gamma]$ の微分を行い，その微係数をゼロとすれば，

$$\frac{\partial\mathbb{L}[q(\boldsymbol{x}),\gamma]}{\partial q(\boldsymbol{x})}=\log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})-\log q(\boldsymbol{x})-1+\gamma=0 \tag{4}$$

を得る．汎関数の関数での微分の説明と式 (4) の導出がベイズ信号処理の A.4 節にある．

また，$\mathbb{L}[q,\gamma]$ を γ について微分し，その微係数をゼロとすれば，

$$\frac{\partial\mathbb{L}[q(\boldsymbol{x}),\gamma]}{\partial\gamma}=\int_{-\infty}^{\infty}q(\boldsymbol{x})d\boldsymbol{x}-1=0 \tag{5}$$

を得る．式 (4) から，

$$\widehat{q}(\boldsymbol{x}) = e^{\gamma-1} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \tag{6}$$

を得，また式 (5) から，

$$\int_{-\infty}^{\infty} q(\boldsymbol{x}) d\boldsymbol{x} = e^{\gamma-1} \int_{-\infty}^{\infty} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{x} = e^{\gamma-1} p(\boldsymbol{y}|\boldsymbol{\theta}) = 1 \tag{7}$$

となるので，したがって，

$$e^{\gamma-1} = \frac{1}{p(\boldsymbol{y}|\boldsymbol{\theta})} \tag{8}$$

を得る．式 (6) に代入すれば，結局，

$$\widehat{q}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})} = p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \tag{9}$$

となる．

上式は，自由エネルギー $\mathcal{F}[q, \boldsymbol{\theta}]$ を最大とする確率分布は未知量 $\boldsymbol{x}$ の事後分布 $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ であることを示している．すなわち，自由エネルギー $\mathcal{F}[q, \boldsymbol{\theta}]$ を確率分布 $q(\boldsymbol{x})$ について最大化することで事後分布を導くことができる．この事後分布の導出はベイズの定理を用いずに行われたことに注意いただきたい．

つまり，事後分布を求める手段はベイズの定理のみではなく，自由エネルギーを最大とする確率分布を求めることによっても行うことができるのである．自由エネルギーと言う汎関数を最大にする確率分布として事後分布を求める方法を変分ベイズ法と呼ぶ．

ここで，自由エネルギー $\mathcal{F}[q, \boldsymbol{\theta}]$ を確率分布 $q(\boldsymbol{x})$ について最大化すれば，$q(\boldsymbol{x})$ は事後分布 $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ に等しくなるため，自由エネルギーは

$$\begin{aligned}
\mathcal{F}[p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta}] &= \int d\boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) [\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})] \\
&= \int d\boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})} = \int d\boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \log p(\boldsymbol{y}|\boldsymbol{\theta}) \\
&= \log p(\boldsymbol{y}|\boldsymbol{\theta}) \int d\boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})
\end{aligned} \tag{10}$$

となる．すなわち，自由エネルギー $\mathcal{F}[q, \boldsymbol{\theta}]$ を確率分布 $q(\boldsymbol{x})$ について最大化すれば，自由エネルギーの値は周辺尤度 $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ に等しくなっている．

自由エネルギーの定義式を

$$\begin{aligned}
\mathcal{F}[q, \boldsymbol{\theta}] &= \int d\boldsymbol{x} q(\boldsymbol{x}) [\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) - \log q(\boldsymbol{x})] \\
&= \int d\boldsymbol{x} q(\boldsymbol{x}) [\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) - \log q(\boldsymbol{x}) - \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})] \\
&= \int d\boldsymbol{x} q(\boldsymbol{x}) \underbrace{[\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})]}_{= \log p(\boldsymbol{y}|\boldsymbol{\theta})} \\
&\quad - \int d\boldsymbol{x} q(\boldsymbol{x}) [\log q(\boldsymbol{x}) - \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})]
\end{aligned} \tag{11}$$

と変形すれば，最右辺の第 1 項は式 (10) より，周辺尤度 $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ に等しい．第 2 項は任意の確率分布 $q(\boldsymbol{x})$ と事後分布 $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ との KL ダイバージェンスは

$$\mathcal{K}_L[q(\boldsymbol{x}) \| p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})] = \int d\boldsymbol{x} q(\boldsymbol{x}) [\log q(\boldsymbol{x}) - \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})] \tag{12}$$

と定義される．したがって，自由エネルギーは

$$\mathcal{F}[q,\boldsymbol{\theta}] = \log p(\boldsymbol{y}|\boldsymbol{\theta}) - \mathcal{K}_L[q(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})] \tag{13}$$

と表すことができる．KL ダイバージェンスは非負，すなわち，必ず $\mathcal{K}_L[q(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})] \geq 0$ を満たし，等号成立は $q(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ の場合のみである．したがって，

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) \geq \mathcal{F}[q,\boldsymbol{\theta}] \tag{14}$$

の関係が必ず成立する．すなわち，自由エネルギー $\mathcal{F}[q,\boldsymbol{\theta}]$ は周辺尤度 $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ の下限 (lower bound) を構成している．周辺尤度 $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ の下限は，DNN の分野では Evidence lower bound，略して ELBO と呼ばれている．（こんな言葉知らなかった.）つまり，ELBO の名でしばしば登場する量は，われわれが自由エネルギーと呼んでいる量である．

3 Variational Autoencoders(VAE)

3.1 DNN の言葉による問題の書き直し

MEG における信号源推定問題においては，データ $\boldsymbol{y}$ からソースの分布 $\boldsymbol{x}$ を推定する．ここで，ソースの確率分布に仮定したハイパーパラメータを $\boldsymbol{\theta}$ としている．生成 DNN の議論では $\boldsymbol{x}$ がデータであり，データは潜在変数 $\boldsymbol{z}$ から生成されるとする．

$\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ はネットワークのパラメータ（全ユニットの接続重みとバイアスの値）に対応する．$\boldsymbol{\phi}$ は近似事後分布の推定に係わるネットワークパラメータであり，近似事後分布を $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ と書く．$\boldsymbol{x}$ を与えて $\boldsymbol{z}$ を近似事後分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ にしたがって生成するニューラルネットワークをエンコーダー（Encoders）と呼ぶ.

$\boldsymbol{\theta}$ は確率分布 $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ を実現するニューラルネットワークのパラメータである．潜在変数 $\boldsymbol{z}$ からデータ $\boldsymbol{x}$ を，確率分布 $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ にしたがって生成するネットワークをデコーダー（Decoders）と呼ぶ.

自由エネルギーの式 (1) をこれらを用いて書き換えれば，ELBO は

$$\mathcal{F}[q] = \int d\boldsymbol{x} q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})[\log p(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right]$$

と表される．式 (13) を DNN の対応する表現に置き換えれば,

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] = \log p(\boldsymbol{x}) - \mathcal{K}_L[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{x})]$$

つまり,

$$\log p(\boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathcal{K}_L[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{x})] \tag{15}$$

となる．式 (15) において，左辺は，$\boldsymbol{\phi}$ には依存しない量である．したがって，右辺の ELBO（右辺第 1 項）をなるべく大きくなるようにネットワークパラメータを選べば，必然的に右辺第 2 項の KL ダイバージェンスが小さくなり，近似事後分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ は真の事後分布 $p(\boldsymbol{z}|\boldsymbol{x})$ にできるだけ近いものとなる．

3.2 ELBO の最小化

3.2.1 ELBO の分解

さて，ここで，$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ であるので，ニューラルネットで近似した $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ を用いて，

$$p(\boldsymbol{x}, \boldsymbol{z}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$$

として，ELBO をさらに分解すれば，

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] &\approx \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \mathcal{K}_L\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right] \quad (16)\end{aligned}$$

上式の最右辺第 1 項は，潜在変数からデータを復元する際の尤度の近似事後分布での期待値であり，平均データ尤度と呼ばれる量である．(EM アルゴリズムでお馴染みの量である．)

第 2 項は，事後分布の変分近似 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ が $\boldsymbol{z}$ の事前分布 $p(\boldsymbol{z})$ になるべく近くなるようにするための項で，ELBO の最大化により，第 1 項を最大化し第 2 項を最小化する $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ を選ぶことになる．

3.2.2 KL ダイバージェンス

式 (16) の右辺第 2 項（KL ダイバージェンスの項）については，更に以下のような仮定を置いて積分を計算する．

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}^2)) \quad (17)$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}) \quad (18)$$

すなわち，事後分布の変分近似は平均 $\boldsymbol{\mu}_{\boldsymbol{\phi}}$，共分散行列 $\mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}^2)$ の多変量正規分布として，$\boldsymbol{z}$ の事前確率は標準正規分布とする．

確率分布に対して式 (17) と (18) の仮定を置くと，式 (16) における最右辺の KL ダイバージェンス項は，解析的に計算できる．すなわち，

$$\mathcal{K}_L\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right] = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\left(\log p(\boldsymbol{z}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\right)d\boldsymbol{z} = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log p(\boldsymbol{z})d\boldsymbol{z} - \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}$$

である．ここで，$\boldsymbol{z}$ が J 次元のベクトル（$\boldsymbol{z}$ は J 個の要素を持っている）として，

$$\begin{aligned}\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log p(\boldsymbol{z})d\boldsymbol{z} &= \int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}^2))\log\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})d\boldsymbol{z} \\ &= -\frac{1}{2}\int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}^2))\left(J\log(2\pi) + \boldsymbol{z}^T\boldsymbol{z}\right)d\boldsymbol{z} \\ &= -\frac{1}{2}J\log(2\pi) - \frac{1}{2}\int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}^2))\left(z_1^2 + z_2^2 + \cdots + z_J^2\right)dz_1dz_2\cdots dz_J \\ &= -\frac{1}{2}J\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}\int z_j^2\mathcal{N}(z_j|\mu_j, \sigma_j^2)dz_j = -\frac{1}{2}J\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}(\mu_j^2 + \sigma_j^2)\end{aligned}$$

である．さらに，$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}^2)$ と書くことにして，

$$\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})] = -\frac{1}{2} J \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})]$$

となるが，右辺最後の項は

$$\begin{aligned}\frac{1}{2} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})] &= \frac{1}{2} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\mathrm{trace}\left[\boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})^T (\boldsymbol{z}-\boldsymbol{\mu})\right]] \\ &= \frac{1}{2} \mathrm{trace}\left[\boldsymbol{\Sigma}^{-1} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[(\boldsymbol{z}-\boldsymbol{\mu})^T (\boldsymbol{z}-\boldsymbol{\mu})]\right] = \frac{1}{2} \mathrm{trace}\left[\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}\right] = \frac{J}{2}\end{aligned}$$

である．したがって，

$$\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} = -\frac{1}{2} J \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} \log(\sigma_j^2) - \frac{J}{2} = -\frac{1}{2} J \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log(\sigma_j^2)\right)$$

を得る．したがって，最終的に，

$$\mathcal{K}_L \left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})\right] = \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right) \tag{19}$$

を得る．

3.2.3 平均データ尤度

式 (16) の右辺第 1 項は，潜在変数からデータを復元するための平均データ尤度の項である．すなわち，

$$\text{潜在変数に対するデータ尤度の期待値} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]$$

である．上の期待値（積分計算）は，モンテカルロ推定値に置き換える．つまり，$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ から発生した M 個のランダムサンプル z^m（$m = 1, \ldots, M$）を用いて

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] \approx \frac{1}{M} \sum_{m=1}^{M} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)}) \tag{20}$$

と，期待値をモンテカルロ平均に置き換える．ここで，

$$\boldsymbol{z}^{(m)} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \qquad (m = 1, \ldots, M)$$

である．(つまり，確率分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ からの M 個のサンプルである．) ただし，式 (20) はこのままでは勾配が計算できないため，Reparameterization trick と言う手法を使って変更する．これについては，後ほど述べる．

3.3 ニューラルネットワークによる実現

3.3.1 ネットワークの概略

VAE を実現するニューラルネットワークの例を図 1 に示す．VAE を実現するニューラルネットワークは，入力層と出力層を含めで 6 つの層からなる．左から層に番号付けをする．入力層が第 1 層であり，出力層が第 6 層とする．第 ℓ 層へ入力する重みを $\boldsymbol{w}^{(\ell)}$，バイアスを $\boldsymbol{b}^{(\ell)}$ と書く．ℓ 層の出力に用いられる活性化関数を $f^{(\ell)}(\cdot)$ と書く．

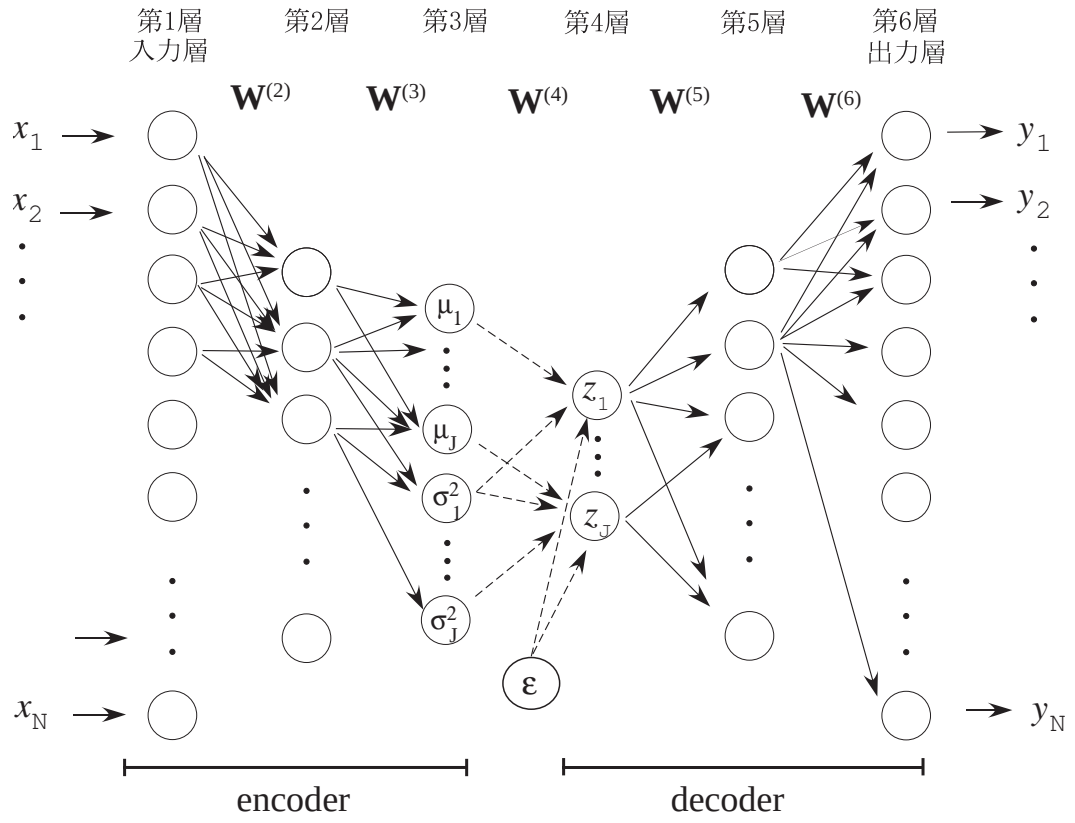

図 1: VAE を実現するネットワーク．左から層に番号付けをする．入力層が第 1 層であり，出力層が第 6 層である．第 1 層から第 3 層までをエンコーダー，第 4 層から第 6 層までをデコーダーと呼ぶ．第 ℓ 層のユニットへの入力に対する重みを $\boldsymbol{w}^{(\ell)} = \{w_{ji}^{(\ell)}\}$，バイアスを $\boldsymbol{b}^{(\ell)}$ と書く．これらをまとめて $\boldsymbol{W}^{(\ell)} = [\boldsymbol{w}^{(\ell)}, \boldsymbol{b}^{(\ell)}]$ と表す．$\boldsymbol{W}^{(2)}$ と $\boldsymbol{W}^{(3)}$ がエンコーダーのパラメータで，これらをまとめて $\boldsymbol{\phi} = (\boldsymbol{W}^{(2)}, \boldsymbol{W}^{(3)})$ と表す．$\boldsymbol{W}^{(5)}$ と $\boldsymbol{W}^{(6)}$ がデコーダーのパラメータで，これらをまとめて $\boldsymbol{\theta} = (\boldsymbol{W}^{(5)}, \boldsymbol{W}^{(6)})$ と表す．$\boldsymbol{W}^{(4)}$ はエンコーダーとデコーダー，つまり第 3 層と第 4 層をつなぐ部分で，図で点線で表されており，固定の重みを持つ．

デコーダは（エンコーダーとも）複数の全結合層から構成される.（1 層の隠れ層を持つ全結合型のニューラルネットを multi-layered perceptrons(MLP) と呼ぶ．デコーダーとエンコーダーは MLP で構成されている.）エンコーダー（左から 3 つの層）は，N 個のデータ $\boldsymbol{x} = [x_1, \ldots, x_N]$ を入力し，J 個の潜在（確率）変数 $\boldsymbol{z} = [z_1, \ldots, z_J]$ を出力する．デコーダー（右側の 3 つの層）は，エンコーダーが出力した J 個の潜在変数から入力を再構成するネットワークであり，再構成結果は $\boldsymbol{y} = [y_1, \ldots, y_N]$ と書かれる．ここで，第 3 層と第 4 層間は固定重みを持ち，エンコーダーとデコーダーをつなぐ働きをする．

3.3.2 デコーダー：入力が 2 値データの場合

デコーダーは，エンコーダーの出力した J 個の潜在変数 $\boldsymbol{z} = [z_1, \ldots, z_J]$ を入力し，データ $\boldsymbol{x}$ の再構成結果 $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ を出力する．

2 値データの場合，データ尤度 $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ は，クロスエントロピー

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \sum_{i=1}^{N} (x_i \log y_i + (1 - x_i) \log(1 - y_i))$$

と与えられる．ここで，$\boldsymbol{y}$ は潜在変数 $\boldsymbol{z}$ から再構成された信号（すなわち $\boldsymbol{z}$ をデコードした結果）である．したがって，尤度は，期待値をモンテカルロ期待値に置き換えて,（式 (20)）

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] \approx \frac{1}{M}\sum_{m=1}^{M}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\left(x_i \log y_i(\boldsymbol{z}^{(m)}) + (1-x_i)\log(1-y_i(\boldsymbol{z}^{(m)})\right)$$

で与えられる．ここで，$\boldsymbol{z}^{(m)} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$，すなわち，確率分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ の実現値である．実際には，$M=1$ として計算する事も多い．この場合，この項は

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] \approx \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \sum_{i=1}^{N}\left(x_i \log y_i(\boldsymbol{z}) + (1-x_i)\log(1-y_i(\boldsymbol{z})\right)$$

となる．

デコーダー部分のニューラルネットは以下のように働く．潜在変数 $\boldsymbol{z}$ の第 5 層への入力 $\boldsymbol{u}^{(5)}$ と第 5 層からの出力 $\boldsymbol{h}^{(5)}$ は

$$\boldsymbol{u}^{(5)} = \boldsymbol{w}^{(5)}\boldsymbol{z} + \boldsymbol{b}^{(5)}, \qquad \boldsymbol{h}^{(5)} = f^{(5)}(\boldsymbol{u}^{(5)}) = \tanh(\boldsymbol{u}^{(5)})$$

である．ここでは，[1] にしたがって活性化関数 $f^{(5)}(\cdot)$ として $\tanh(\cdot)$ を用いた．最終層（第 6 層）への入力は

$$\boldsymbol{u}^{(6)} = \boldsymbol{w}^{(6)}\boldsymbol{h}^{(5)} + \boldsymbol{b}^{(6)}$$

であり，最終的なネットワークの出力はシグモイド関数 $\sigma(\cdot)$ を用いて

$$\boldsymbol{y} = \sigma(\boldsymbol{u}^{(6)})$$

と表される．ここで，デコーダーのネットワークパラメータ $\boldsymbol{\theta}$ は

$$\boldsymbol{\theta} = \{\boldsymbol{W}^{(5)}, \boldsymbol{W}^{(6)}\} = \{\boldsymbol{w}^{(5)}, \boldsymbol{w}^{(6)}, \boldsymbol{b}^{(5)}, \boldsymbol{b}^{(6)}\}$$

である．

3.3.3 デコーダー：入力データが連続実数値の場合

平均データ尤度は（モンテカルロ期待値を $M=1$ として求めて）

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \log \mathcal{N}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{I}) = -\|\boldsymbol{x} - \boldsymbol{y}(\boldsymbol{z})\|^2 = -\sum_{i=1}^{N}(x_i - y_i(\boldsymbol{z}))^2$$

と与えられる．ここで，$\boldsymbol{y}$ はデコーダーによる（潜在変数 $\boldsymbol{z}$ に対応した）$\boldsymbol{x}$ の再構成である．この場合も，第 5 層への入力 $\boldsymbol{u}^{(5)}$ と第 5 層からの出力 $\boldsymbol{h}^{(5)}$ は

$$\boldsymbol{u}^{(5)} = \boldsymbol{w}^{(5)}\boldsymbol{z} + \boldsymbol{b}^{(5)}, \qquad \boldsymbol{h}^{(5)} = f^{(5)}(\boldsymbol{u}^{(5)}) = \tanh(\boldsymbol{u}^{(5)})$$

となる.（活性化関数の選択は $ReLU$ などもあり得るか.）したがって，第 6 層（出力層）への入力は

$$\boldsymbol{u}^{(6)} = \boldsymbol{w}^{(6)}\boldsymbol{h}^{(5)} + \boldsymbol{b}^{(6)}$$

であり，最終的なネットワークの出力は $f^{(6)}(\cdot)$ に恒等写像を用いて

$$\boldsymbol{y} = \boldsymbol{u}^{(6)}$$

とする．

3.3.4 エンコーダー

エンコーダーは観測データ $\boldsymbol{x}$（$\in \mathbb{R}^N$）から，観測データを表す潜在変数の従う確率分布を推定する.（実際には多数の $\boldsymbol{x}$ から推定するのであるが，誤差関数は単純に和になるだけなので，観測データは $\boldsymbol{x}$ として定式化を行う.）

具体的には，この確率分布を正規分布として，ニューラルネットワークにより，その期待値と分散 μ_j，σ_j^2（$j = 1, 2, \cdots, J$）を求める.

第 2 層への入力と出力は

$$\boldsymbol{u}^{(2)} = \boldsymbol{w}^{(2)}\boldsymbol{x} + \boldsymbol{b}^{(2)} \qquad \boldsymbol{h}^{(2)} = f^{(2)}(\boldsymbol{u}^{(2)})$$

第 3 層への入力と出力は，$\boldsymbol{\mu}$ の推定と $\boldsymbol{\sigma}^2$ の推定に分けて，A の下付きが $\boldsymbol{\mu}$ の推定，B の下付きが $\boldsymbol{\sigma}^2$ の推定に係わるパラメータとすれば,

$$\boldsymbol{u}_A^{(3)} = \boldsymbol{w}_A^{(3)}\boldsymbol{h}_{(2)} + \boldsymbol{b}_A^{(3)} \qquad \boldsymbol{\mu} = f_A^{(3)}(\boldsymbol{u}_A^{(3)})$$

さらに,

$$\boldsymbol{u}_B^{(3)} = \boldsymbol{w}_B^{(3)}\boldsymbol{h}_{(2)} + \boldsymbol{b}_B^{(3)} \qquad \boldsymbol{\sigma}^2 = f_B^{(3)}(\boldsymbol{u}_B^{(3)})$$

である.

$f^{(2)}(\,\cdot\,)$, $f_A^{(3)}(\,\cdot\,)$, $f_B^{(3)}(\,\cdot\,)$ は活性化関数である.

$$f^{(2)} = \tanh(\,\cdot\,) \qquad f_A^{(3)} = \text{恒等写像} \qquad f_B^{(3)} = \text{ソフトマックス関数}$$

とする．$f_B^{(3)} =$ ソフトマックス関数 とするのは σ_j^2 に対する非負制約を組み込むためである.

ただし，ソフトマックス関数を使った構成は若干複雑になるので，第 3 層の「B」ユニットは σ_j^2 ではなく，$\varphi_j = \log(\sigma_j^2)$ を出力するよう変更する事も行われる.（こうすれば，非負制約の必要がなくなり，活性化関数にソフトマックス関数を使う必要はなくなる.）そのかわり，KL ダイバージェンスの項は,

$$\mathcal{K}_L\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right] = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right)$$

を，$\varphi_j = \log(\sigma_j^2)$ の置き換えに対応した,

$$\mathcal{K}_L\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right] = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \varphi_j - \mu_j^2 - e^{\varphi_j}\right)$$

とする必要がある.

エンコーダーのパラメータは $\boldsymbol{\phi} = \{\boldsymbol{W}^{(2)}, \boldsymbol{W}^{(3)}\}$ である．ここで，$\boldsymbol{W}^{(3)} = \{\boldsymbol{w}_A^{(3)}, \boldsymbol{w}_B^{(3)}, \boldsymbol{b}_A^{(3)}, \boldsymbol{b}_B^{(3)}\}$ である.

3.3.5 第 4 層について

第 4 層では，第 3 層の出力である μ_j と σ_j^2 を用いて潜在変数 z_j を計算する．z_j は正規分布 $\mathcal{N}(z|\mu_j, \sigma_j^2)$ からのサンプルである．実際の計算は，Reparametrization trick と呼ばれる

$$z_j = \mu_j + \sigma_j\varepsilon_j \qquad (j = 1, \ldots, J) \qquad \varepsilon_j \sim \mathcal{N}(\varepsilon|0, 1)$$

として行う.

第 3 層の $j = J+1, \ldots, 2J$ に対応したユニットの出力が σ_j^2 の場合：

$$u_j^{(3)} = \begin{cases} \mu_j & j = 1, \ldots, J \\ \sigma_j^2 & j = J+1, \ldots, 2J \end{cases}$$

であるので，第 4 層の入力と出力は，$j = 1, \ldots, J$ に対して，

$$\begin{aligned} u_j^{(4)} &= u_j^{(3)} + \sqrt{u_{j+J}^{(3)}}\varepsilon_j \\ z_j &= u_j^{(4)} \end{aligned}$$

となる．

第 3 層の $j = J+1, \ldots, 2J$ 番目ユニットの出力が $\log(\sigma_j^2)$ の場合：

$$u_j^{(3)} = \begin{cases} \mu_j & j = 1, \ldots, J \\ \log(\sigma_j^2) & j = J+1, \ldots, 2J \end{cases}$$

であるので，第 4 層の入力と出力は，$j = 1, \ldots, J$ に対して，

$$\begin{aligned} u_j^{(4)} &= u_j^{(3)} + \sqrt{\exp(u_{j+J}^{(3)})}\varepsilon_j = u_j^{(3)} + \exp(\frac{1}{2}u_{j+J}^{(3)})\varepsilon_j \\ z_j &= u_j^{(4)} \end{aligned}$$

となる．

4 ELBO（変分下限）の勾配計算

4.1 勾配計算の問題点

パラメター $\boldsymbol{\theta}$ を変数に持つあるスカラー関数 $f(\boldsymbol{x}, \boldsymbol{\theta})$ の，確率分布 $p(\boldsymbol{x})$ についての期待値：

$$\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x}, \boldsymbol{\theta})]$$

を考え，この期待値の $\boldsymbol{\theta}$ に関する微分（勾配）を計算することを考える．すると，

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x}, \boldsymbol{\theta})] = \nabla_{\boldsymbol{\theta}} \int p(\boldsymbol{x}) f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta})]$$

である．つまり，勾配計算と期待値計算は可換であり，期待値の勾配は勾配の期待値に等しい．ただし，この場合，確率分布 $p(\boldsymbol{x})$ はパラメータ $\boldsymbol{\theta}$ とは無関係である事が仮定されている．

もし，$p(\boldsymbol{x})$ がパラメータ $\boldsymbol{\theta}$ に依存するなら，

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x})}[f(\boldsymbol{x}, \boldsymbol{\theta})] = \nabla_{\boldsymbol{\theta}} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}) f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} + \int (\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x}))\, f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} \\ &= \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta})] + \int (\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x}))\, f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} \quad (21) \end{aligned}$$

となる．すなわち，勾配計算と期待値計算は可換とはならず，余分な項（右辺第 2 項）が出てきてしまう．この余分な項は一般的には計算困難である．

ELBO の勾配計算の場合：ELBO は式 (16) で与えられる．もう 1 度この式を書くと，

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \mathcal{K}_L\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})\right]$$

である．上式をパラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ で微分するのである．右辺第 2 項は，式 (19) に示すように，正規分布の仮定により解析的に解が得られるので，微分計算に問題は生じない．しかし右辺第一項は，近似事後分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ による期待値を計算する．

この項の計算は，期待値をモンテカルロ平均に置き換え，

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] \approx \frac{1}{M}\sum_{m=1}^{M} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)}) \quad \text{ただし} \quad \boldsymbol{z}^{(m)} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$$

を計算する．しかしながら，先に述べたことにより

$$\nabla_{\boldsymbol{\phi}} \frac{1}{M}\sum_{m=1}^{M} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)}) = \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)})$$

が成り立たないので勾配計算ができない．この問題を回避するため，次に述べる Reparametrization trick が提案されている．

4.2 Reparametrization trick

この問題に対する解決策は，$\boldsymbol{\phi}$ に依存しない独立な確率分布 $p(\boldsymbol{\varepsilon})$ からのサンプル $\boldsymbol{\varepsilon}$ を用いて，$\boldsymbol{z}$ を

$$\boldsymbol{z} = g(\boldsymbol{\varepsilon}, \boldsymbol{x}, \boldsymbol{\phi}) = g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \boldsymbol{x})$$

と表せるような関数 $g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \boldsymbol{x})$ を見出すことである[1]．この $g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \boldsymbol{x})$ の簡単な例としては，VAE で実際に用いられている

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$

と仮定した場合における，

$$\boldsymbol{z} = \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{x}) + \boldsymbol{\sigma}(\boldsymbol{\theta}, \boldsymbol{x}) \odot \boldsymbol{\varepsilon}$$

が挙げられる．ここで，$\odot$ は要素ごとのかけ算を表す．先ほどの式 (21) の一般的な例をもう 1 度考えてみると，

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x})}[f(\boldsymbol{z}, \boldsymbol{\theta})] = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{\varepsilon})}[f(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \boldsymbol{x}))] = \mathbb{E}_{p(\boldsymbol{\varepsilon})}[\nabla_{\boldsymbol{\theta}} f(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \boldsymbol{x}))] = \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\theta}} f(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(m)}, \boldsymbol{x}))$$

として計算可能な量となる．

ELBO の第 1 項

$$\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]$$

についても，

$$\begin{aligned}\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] &= \nabla_{\boldsymbol{\phi}} \mathbb{E}_{q(\boldsymbol{\varepsilon})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \boldsymbol{x}))\right] \\ &= \nabla_{\boldsymbol{\phi}} \frac{1}{M}\sum_{m=1}^{M} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)}) = \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(m)})\end{aligned}$$

[1] この分野の文献はパラメータを下付きで表すことが多い.「関数 g は $\boldsymbol{\phi}$ にも依存する」の意味なので $g(\boldsymbol{\varepsilon}, \boldsymbol{x}, \boldsymbol{\phi})$ と書く方がわかりやすいと思うが.

として計算できる．ただし，

$$\boldsymbol{z}^{(m)} = g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(m)}, \boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{\phi}, \boldsymbol{x}) + \boldsymbol{\sigma}(\boldsymbol{\phi}, \boldsymbol{x}) \odot \boldsymbol{\varepsilon}^{(m)} \quad \text{および} \quad \boldsymbol{\varepsilon}^{(m)} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{I})$$

である．

以上述べた考え方は，reparametrization trick と呼ばれている．reparametrization trick は，あるパラメータを持つ確率分布の期待値の（そのパラメータでの）勾配を計算するときに，期待値計算をそのパラメータを含まない確率分布で行うよう，確率変数の変換を行うことがポイントであろう．

4.3 誤差逆伝搬の計算式

4.3.1 誤差関数

データが連続値を取る場合を考える．重みを計算するための評価関数（誤差関数）は尤度の項と KL ダイバージェンスの和であり，これらは，(モンテカルロ期待値において $M = 1$ として）

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{N}(x_i - y_i(\boldsymbol{z}(\boldsymbol{\phi}), \boldsymbol{\theta}))^2 - \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j^2(\boldsymbol{\phi})) - \mu_j^2(\boldsymbol{\phi}) - \sigma_j^2(\boldsymbol{\phi})\right)$$

である．ここで，

$$z_j(\boldsymbol{\phi}) = \mu_j(\boldsymbol{\phi}) + \varepsilon\sigma_j(\boldsymbol{\phi})$$

である．z_j, μ_j, σ_j^2 は $\boldsymbol{x}, \boldsymbol{\phi}$ の関数である．($\boldsymbol{x}$ への依存性は，自明であるので省略した.）$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ である．ここで，

$$\mathcal{E}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{Q}(x_i - y_i(\boldsymbol{z}(\boldsymbol{\phi}), \boldsymbol{\theta}))^2 \tag{22}$$

$$\mathcal{K}(\boldsymbol{\phi}; , \boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j^2(\boldsymbol{\phi})) - \mu_j^2(\boldsymbol{\phi}) - \sigma_j^2(\boldsymbol{\phi})\right) \tag{23}$$

として，誤差関数を 2 つの項の和で表して，

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathcal{E}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) + \mathcal{K}(\boldsymbol{\phi}; , \boldsymbol{x})$$

と書く．右辺第 1 項は平均データ尤度に由来する項であるが，パラメータ $\boldsymbol{\theta} = (\boldsymbol{W}^{(5)}, \boldsymbol{W}^{(6)})$ と $\boldsymbol{\phi} = (\boldsymbol{W}^{(2)}, \boldsymbol{W}^{(3)})$ の両方に依存する．一方，第 2 項は KL ダイバージェンスの項で，パラメータ $\boldsymbol{\phi} = (\boldsymbol{W}^{(2)}, \boldsymbol{W}^{(3)})$ に依存するが，$\boldsymbol{\theta} = (\boldsymbol{W}^{(5)}, \boldsymbol{W}^{(6)})$ には依存しない．

4.3.2 デコーダー層における誤差勾配の計算

• **第 6 層の重み勾配計算：**図 1 に示す VAE では，$L = 6$ である．$w_{ji}^{(6)} \in \boldsymbol{\theta}$ であり，$\mathcal{K}(\boldsymbol{\phi}; , \boldsymbol{x})$ の項はこの重みには依存しないので誤差関数を $\mathcal{E}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$ として，計算を進める．

$$\frac{\partial \mathcal{E}}{\partial w_{ji}^{(6)}} = \frac{\partial \mathcal{E}}{\partial u_j^{(6)}}\frac{\partial u_j^{(6)}}{\partial w_{ji}^{(6)}} = \widetilde{\delta}_j^{(6)}\frac{\partial u_j^{(6)}}{\partial w_{ji}^{(6)}}$$

であるが，まず，

$$\frac{\partial u_j^{(6)}}{\partial w_{ji}^{(6)}} = h_i^{(5)}$$

であり，

$$\widetilde{\delta}_j^{(6)} = \frac{\partial \mathcal{E}}{\partial u_j^{(6)}} = \frac{\partial \mathcal{E}}{\partial y_j}\frac{\partial y_j}{\partial u_j^{(6)}} = -(x_j - y_j(\boldsymbol{z}))\frac{\partial y_j}{\partial u_j^{(6)}}$$

である．第 6 層の活性化関数が恒等写像ならば，$y_j = u_j^{(6)}$ なので，$\partial y_j/\partial u_j^{(6)} = 1$ であり，

$$\widetilde{\delta}_j^{(6)} = -(x_j - y_j(\boldsymbol{z}))$$

を得る．また，

$$u_{ji}^{(6)} = w_{ji}^{(6)} h_i^{(5)} + b_j^{(6)}$$

であるので，$\partial u_j^{(6)}/\partial w_{ji}^{(6)} = h_i^{(5)}$ であり，したがって，

$$\frac{\partial \mathcal{E}}{\partial w_{ji}^{(6)}} = \widetilde{\delta}_j^{(6)}\frac{\partial u_j^{(6)}}{\partial w_{ji}^{(6)}} = (y_j(\boldsymbol{z}) - x_i)h_i^{(5)}$$

である．ちなみに，上式は式 (30) から直接導かれる．

- **第 5 層の重み勾配計算：**式 (30) と式 (32) を用いれば，

$$\frac{\partial \mathcal{E}}{\partial w_{ji}^{(5)}} = \widetilde{\delta}_j^{(5)} h_i^{(4)}$$

および

$$\widetilde{\delta}_j^{(5)} = \sum_k \widetilde{\delta}_k^{(6)} w_{kj}^{(6)} \dot{f}^{(5)}(u_j^5)$$

である．したがって，

$$\frac{\partial \mathcal{E}}{\partial w_{ji}^{(5)}} = \left(\sum_k \delta_k^{(6)} w_{kj}^{(6)} \dot{f}^{(5)}(u_j^5)\right) h_i^{(4)} = \left(\sum_k \widetilde{\delta}_k^{(6)} w_{kj}^{(6)} \dot{f}^{(5)}(u_j^5)\right) z_i$$

である．($h_i^{(4)} = z_i$ であるため．)

- **第 4 層の $\widetilde{\delta}_j^{(4)}$ 計算：**第 4 層と第 3 層間のエンコーダー，デコーダー接続部分の重みは固定であり，更新しない．しかし，エンコーダー層における誤差勾配計算のため，第 4 層のデルタ $\widetilde{\delta}_j^{(4)}$ を計算しておく．まず，

$$\widetilde{\delta}_j^{(4)} = \frac{\partial \mathcal{E}}{\partial u_j^{(4)}} = \sum_k \frac{\partial \mathcal{E}}{\partial u_k^{(5)}}\frac{\partial u_k^{(5)}}{\partial u_j^{(4)}} = \sum_k \widetilde{\delta}_k^{(5)}\frac{\partial u_k^{(5)}}{\partial u_j^{(4)}}$$

である．右辺の積における 2 番目の項は，

$$u_k^{(5)} = \sum_j w_{kj}^{(5)} h_j^{(4)} = \sum_j w_{kj}^{(5)} u_j^{(4)}$$

を用いて，

$$\frac{\partial u_k^{(5)}}{\partial u_j^{(4)}} = w_{kj}^{(5)}$$

であるので,

$$\widetilde{\delta}_j^{(4)} = \sum_k \widetilde{\delta}_k^{(5)} w_{kj}^{(5)}$$

を得る.

4.3.3 エンコーダー層における誤差勾配の計算

Case 1 第 3 層の出力が第 3 層の出力が

$$u_j^{(3)} = \begin{cases} \mu_j & j = 1, \dots J \\ \sigma_j^2 & j = J+1, \dots, 2J \end{cases}$$

の場合を考える.

• **第 3 層の重み勾配計算：**第 3 層の重み $w_{ji}^{(3)}$ は，$w_{ji}^{(3)} \in \boldsymbol{\phi}$ であるので，第 2 項も考慮した誤差関数を用いて新たに誤差を計算する.

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^{(3)}} = \frac{\partial \mathcal{L}}{\partial u_j^{(3)}} \frac{\partial u_j^{(3)}}{\partial w_{ji}^{(3)}} = \delta_j^{(3)} \frac{\partial u_j^{(3)}}{\partial w_{ji}^{(3)}}$$

であるが，まず,

$$\frac{\partial u_j^{(3)}}{\partial w_{ji}^{(3)}} = h_i^{(2)}$$

であり,

$$\delta_j^{(3)} = \frac{\partial \mathcal{L}}{\partial u_j^{(3)}} = \frac{\partial \mathcal{E}}{\partial u_j^{(3)}} + \frac{\partial \mathcal{K}}{\partial u_j^{(3)}}$$

である．右辺第一項は $\widetilde{\delta}_j^{(3)} = \partial \mathcal{E} / \partial u_j^{(3)}$ であるので，デコーダー層の誤差逆伝搬を引き継ぐことで求められる．すなわち,

$$\frac{\partial \mathcal{E}}{\partial u_j^{(3)}} = \widetilde{\delta}_j^{(3)} = \sum_k \widetilde{\delta}_k^{(4)} \frac{\partial u_k^{(4)}}{\partial u_j^{(3)}}$$

である．右辺の $\partial u_k^{(4)} / \partial u_j^{(3)}$ の計算を行う.

第 3 層と第 4 層の入力である $u_j^{(3)}$ と $u_j^{(4)}$ の関係は,

$$u_j^{(4)} = u_j^{(3)} + \sqrt{u_{j+J}^{(3)}} \varepsilon_j \qquad (j = 1, \dots, J)$$

であるので,

$$\frac{\partial u_k^{(4)}}{\partial u_j^{(3)}} = \begin{cases} 1 & j = k \\ \varepsilon_j / \left(2\sqrt{u_j^{(3)}} \right) & j = k + J \end{cases}$$

である．したがって,

$$\widetilde{\delta}_j^{(3)} = \begin{cases} \widetilde{\delta}_j^{(4)} & (j = 1, \dots, J) \\ \left[\varepsilon_j / (2\sqrt{u_j^{(3)}}) \right] \widetilde{\delta}_{j-J}^{(4)} & (j = J+1, \dots, 2J) \end{cases} \tag{24}$$

となる．

次に，$\partial\mathcal{K}/\partial u_j^{(3)}$ の計算であるが，

$$\mathcal{K}(\boldsymbol{\phi};,\boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{J}\left(1+\log(\sigma_j^2(\boldsymbol{\phi})) - \mu_j^2(\boldsymbol{\phi}) - \sigma_j^2(\boldsymbol{\phi})\right)$$

であり，

$$u_j^{(3)} = \begin{cases} \mu_j & j = 1,\dots J \\ \sigma_j^2 & j = J+1,\dots,2J \end{cases}$$

であるので，

$$\mathcal{K}(\boldsymbol{\phi};,\boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{J}\left(1+\log(u_{j+J}^{(3)}) - (u_j^{(3)})^2 - u_{j+J}^{(3)}\right)$$

と書ける．したがって，

$$\frac{\partial\mathcal{K}}{\partial u_j^{(3)}} = \begin{cases} u_j^{(3)} & (j = 1,\dots J) \\ \frac{1}{2}\left(1 - 1/(u_j^{(3)})\right) & (j = J+1,\dots,2J) \end{cases} \tag{25}$$

である．したがって，式 (24) と (25) より，

$$\delta_j^{(3)} = \frac{\partial\mathcal{E}}{\partial u_j^{(3)}} + \frac{\partial\mathcal{K}}{\partial u_j^{(3)}} = \widetilde{\delta}_j^{(3)} + \frac{\partial\mathcal{K}}{\partial u_j^{(3)}} = \begin{cases} \widetilde{\delta}_j^{(4)} + u_j^{(3)} & (j = 1,\dots,J) \\ \left(\varepsilon_j/(2\sqrt{u_j^{(3)}})\right)\widetilde{\delta}_{j-J}^{(4)} + \frac{1}{2}\left(1 - 1/(u_j^{(3)})\right) & (j = J+1,\dots,2J) \end{cases} \tag{26}$$

を得る．さらに，これを用いて，誤差関数の勾配

$$\frac{\partial\mathcal{L}}{\partial w_{ji}^{(3)}} = \delta_j^{(3)} h_i^{(2)}$$

を得る．

Case 2 第 3 層の出力が

$$u_j^{(3)} = \begin{cases} \mu_j & j = 1,\dots J \\ \log(\sigma_j^2) & j = J+1,\dots,2J \end{cases}$$

の場合を考える．

- **第 3 層の重み勾配計算：**以下の関係

$$\frac{\partial\mathcal{E}}{\partial u_j^{(3)}} = \widetilde{\delta}_j^{(3)} = \sum_k \widetilde{\delta}_k^{(4)}\frac{\partial u_k^{(4)}}{\partial u_j^{(3)}}$$

において．右辺の $\partial u_k^{(4)}/\partial u_j^{(3)}$ の計算を行う．第 3 層と第 4 層の入力である $u_j^{(3)}$ と $u_j^{(4)}$ の関係は，

$$u_j^{(4)} = u_j^{(3)} + \exp(\frac{1}{2}u_{j+J}^{(3)})\varepsilon_j \qquad (j = 1,\dots,J)$$

であるので，

$$\frac{\partial u_k^{(4)}}{\partial u_j^{(3)}} = \begin{cases} 1 & j = k \\ \frac{1}{2}\varepsilon_j \exp(\frac{1}{2}u_j^{(3)}) & j = k+J \end{cases}$$

である．したがって，

$$\widetilde{\delta}_j^{(3)} = \begin{cases} \widetilde{\delta}_j^{(4)} & (j = 1, \ldots, J) \\ \left[\frac{1}{2}\varepsilon_j \exp(\frac{1}{2}u_j^{(3)})\right] \widetilde{\delta}_{j-J}^{(4)} & (j = J+1, \ldots, 2J) \end{cases} \tag{27}$$

となる．

次に，$\partial\mathcal{K}/\partial u_j^{(3)}$ の計算であるが，

$$\mathcal{K}(\boldsymbol{\phi};,\boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \varphi_j - \mu_j^2 - e^{\varphi_j}\right)$$

であり，

$$u_j^{(3)} = \begin{cases} \mu_j & j = 1, \ldots J \\ \varphi_{j-J} & j = J+1, \ldots, 2J \end{cases}$$

であるので，

$$\mathcal{K}(\boldsymbol{\phi};,\boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{J}\left(1 + u_{j+J}^{(3)} - (u_j^{(3)})^2 - e^{u_{j+J}^{(3)}}\right)$$

と書ける．したがって，

$$\frac{\partial\mathcal{K}}{\partial u_j^{(3)}} = \begin{cases} u_j^{(3)} & (j = 1, \ldots J) \\ -\frac{1}{2}\left(1 - \exp[(u_j^{(3)}])\right) & (j = J+1, \ldots, 2J) \end{cases} \tag{28}$$

である．したがって，式 (27) と (28) より，

$$\delta_j^{(3)} = \widetilde{\delta}_j^{(3)} + \frac{\partial\mathcal{K}}{\partial u_j^{(3)}} = \begin{cases} \widetilde{\delta}_j^{(4)} + u_j^{(3)} & (j = 1, \ldots, J) \\ \left(\frac{1}{2}\varepsilon_j \exp(\frac{1}{2}u_j^{(3)})\right) \widetilde{\delta}_{j-J}^{(4)} - \frac{1}{2}\left(1 - \exp[(u_j^{(3)}])\right) & (j = J+1, \ldots, 2J) \end{cases} \tag{29}$$

を得る．さらに，これを用いて，誤差関数の勾配

$$\frac{\partial\mathcal{L}}{\partial w_{ji}^{(3)}} = \delta_j^{(3)} h_i^{(2)}$$

を得る．

- **第 2 層の重み勾配計算：**第 2 層の重み勾配は $\delta_j^{(3)}$ を用いて，逆誤差伝搬を行えばよい．すなわち，式 (32) より，

$$\delta_j^{(2)} = \sum_k \delta_k^{(3)} w_{kj}^{(3)} \dot{f}^{(2)}(u_j^2)$$

として，$\delta_j^{(2)}$ が求まるので，式 (30) より，

$$\frac{\partial\mathcal{L}}{\partial w_{ji}^{(2)}} = \delta_j^{(2)} h_i^{(1)} = \delta_j^{(2)} x_i$$

として，勾配が求まる．

4.4 Appendix：誤差逆伝搬（漸化式の導出）

第 ℓ 層のユニット j に対する入力を $u_j^{(\ell)}$ として，このユニットの出力を $h_j^{(\ell)} = f^{(\ell)}(u_j^{(\ell)})$ と表す．$f^{(\ell)}$ は活性化関数であり，

$$u_j^{(\ell)} = \sum_i w_{ji}^{(\ell)} h_i^{(\ell-1)}$$

の関係がある．誤差関数 E に対する重み $w_{ji}^{(\ell)}$ の勾配を計算する．誤差関数 E は，$u_j^{(\ell)}$ を通じて（のみ）重み $w_{ji}^{(\ell)}$ に依存している．すなわち，

$$\frac{\partial E}{\partial w_{ji}^{(\ell)}} = \frac{\partial \mathcal{E}}{\partial u_j^{(\ell)}} \frac{\partial u_j^{(\ell)}}{\partial w_{ji}^{(\ell)}}$$

が成り立つ．ここで，

$$\delta_j^{(\ell)} = \frac{\partial E}{\partial u_j^{(\ell)}}$$

と定義する．これは，ℓ 層のユニット j への入力（の総和）がどれだけ最終誤差 E に効いているのかを表す量である．

また，$u_j^{(\ell)}$ と $w_{ji}^{(\ell)}$ は線形な関係なので，

$$\frac{\partial u_j^{(\ell)}}{\partial w_{ji}^{(\ell)}} = h_i^{(\ell-1)}$$

が成り立つので，結局，

$$\frac{\partial E}{\partial w_{ji}^{(\ell)}} = \delta_j^{(\ell)} h_i^{(\ell-1)} \tag{30}$$

が成り立つ．

ここで，ポイントは $\delta_j^{(\ell)}$ をどのように求めるかである．次に $\delta_j^{(\ell)}$ に対する漸化式を導く．まず，

$$\delta_j^{(\ell)} = \frac{\partial E}{\partial u_j^{(\ell)}} = \sum_k \frac{\partial E}{\partial u_k^{(\ell+1)}} \frac{\partial u_k^{(\ell+1)}}{\partial u_j^{(\ell)}} = \sum_k \delta_k^{(\ell+1)} \frac{\partial u_k^{(\ell+1)}}{\partial u_j^{(\ell)}} \tag{31}$$

である．ここで，$\delta_k^{(\ell+1)} = \partial E / \partial u_k^{(\ell+1)}$ を用いた．式 (31) は，δ_j^{ℓ} が $\delta_j^{\ell+1}$ の和で表されるためには，$u_j^{(\ell+1)}$ が $u_k^{(\ell)}$ の微分可能な関数として表されている必要があることを示している．

ただしこのことは，通常のネットワークでは問題とならず，

$$u_k^{(\ell+1)} = \sum_j w_{kj}^{(\ell+1)} h_j^{\ell} = \sum_j w_{kj}^{(\ell+1)} f^{(\ell)}(u_j^{\ell})$$

と表される．したがって，

$$\frac{\partial u_k^{(\ell+1)}}{\partial u_j^{(\ell)}} = w_{kj}^{(\ell+1)} \dot{f}^{(\ell)}(u_j^{\ell})$$

であるので，結局，

$$\delta_j^{(\ell)} = \sum_k \delta_k^{(\ell+1)} w_{kj}^{(\ell+1)} \dot{f}^{(\ell)}(u_j^{\ell}) \tag{32}$$

を得る．この式が $\delta_j^{(\ell)}$ を計算する漸化式である．

参考文献

以下の文献を参考にした.

1. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
2. Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. Foundations and TrendsR in Machine Learning, 12(4), 307-392.
3. Anonymous author, The Reparameterization Trick,

   `https://gregorygundersen.com/blog/2018/04/29/reparameterization/`