

## 1 变分自己符号化器

## 1 変分自己符号化器

- 生成モデル
  - 変分自己符号化器 (VAE) の概要
  - 変分自己符号化器 (VAE) の理論

- 生成モデルの目的

- データ  $x$  に関する分布  $p(x)$  を推定する

- データ生成過程

- データ  $x$  は一般に高次元である
- 但し、実際にデータが分布しているのは、ごく限られた一部の低次元の領域であると考えられる (多様体仮説)
- データ  $x$  自体は高次元だが、本質的には低次元の情報しか持たないと考えられる
- データ  $x$  を、より低次元なベクトル  $z$  を使って、表現することを考える
- データに関する分布  $p(x)$  を、潜在変数  $z$  に関する分布と、うまく組み合わせる
- 潜在変数からデータが生成されるまでの過程を組み込んで、 $p(x)$  を記述する

## 1 変分自己符号化器

- 生成モデル
- 変分自己符号化器 (VAE) の概要
- 変分自己符号化器 (VAE) の理論

# 変分自己符号化器 (VAE) の概要

- 深層学習における生成モデル
  - 主に以下の 2 つの手法が存在する
    - 敵対的生成ネットワーク (Generative Adversarial Networks, GAN)
    - 変分自己符号化器 (Variational Auto Encoders, VAE)
  - ここでは変分自己符号化器 (VAE) について扱う
  - VAE を、異常検知 (不良品の検出など) に使った例がある

# 変分自己符号化器 (VAE) の概要

- VAE におけるグラフィカルモデル

- 図 1 のような、潜在変数を含んだグラフィカルモデルを考える
- データ  $x$  について、ある一つの潜在変数  $z$  が対応しているとする
- 各データ  $x$  は、分布  $p(x)$  から独立にサンプルされるとする
- 従って、データ  $\{x_1, \dots, x_N\}$  は独立同分布標本とする
- $\theta$  は、潜在変数  $z$  からデータ  $x$  を取得する際に使用されるパラメータ
- $\phi$  は、データ  $x$  から潜在変数  $z$  を生成する際に使用されるパラメータ
- $N$  は、データ数である

# 変分自己符号化器 (VAE) の概要

- データ  $x$  の生成過程

- データ  $x$  の生成過程は、次のように考える
  - 分布  $p(z|\theta)$  から、潜在変数  $z_i$  がサンプルされる
  - 分布  $p(x|z_i, \theta)$  から、データ  $x_i$  がサンプルされる
- これより、データ  $x$  の分布を次のように表現できる

$$p(x|\theta) = \int p(x|z, \theta)p(z|\theta)dz \quad (1)$$

- 潜在変数  $z$  をデータ  $x$  から取得する過程

- 潜在変数  $z_i$  をデータ  $x_i$  から得る過程は、次のように考える
  - 分布  $q(z|x_i, \phi)$  から、潜在変数  $z_i$  がサンプルされる

# 変分自己符号化器 (VAE) の概要

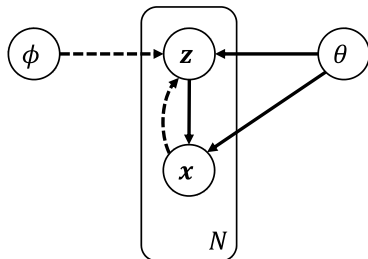


図 1: 変分自己符号化器 (VAE) におけるグラフィカルモデル



# 変分自己符号化器 (VAE) の概要

- 確率分布のニューラルネットワークによる表現
  - 潜在変数を含む確率モデルについて、パラメータの最尤解を求めるために、EM アルゴリズムを導出した
  - EM アルゴリズムでは、潜在変数に関する事後分布  $p(z|x, \theta)$  を計算する必要があった
  - この事後分布  $p(z|x, \theta)$  の計算が困難であるとき、 $p(z|x, \theta)$  を別の分布  $q(z|x, \phi)$  で近似し、変分推論によって  $q(z|\phi)$  の最適解を求めた
- VAE は変分推論の変種であり、近似事後分布  $q(z|x, \phi)$  と、 $p(x|z, \theta)$  の2つをニューラルネットワークで表現する
- データ  $x$  を潜在変数  $z$  に対応付けるニューラルネットワークを、Encoder という
- 潜在変数  $z$  からデータ  $x$  を復元するニューラルネットワークを、Decoder という
- 分布  $q(z|x, \phi)$  は Encoder、分布  $p(x|z, \theta)$  は Decoder に相当する

## 1 変分自己符号化器

- 生成モデル
- 変分自己符号化器 (VAE) の概要
- 変分自己符号化器 (VAE) の理論

# 変分自己符号化器 (VAE) の理論

- 変分自己符号化器 (VAE) の理論
  - 変分下界  $\mathcal{L}(q)$  は次のようであった

$$\mathcal{L}(q) = \int q(z|\mathbf{x}) \ln \frac{p(\mathbf{x}, z)}{q(z|\mathbf{x})} dz \quad (2)$$

$$= \int q(z|\mathbf{x}) \ln \frac{p(\mathbf{x}|z)p(z)}{q(z|\mathbf{x})} dz \quad (3)$$

$$= \int q(z|\mathbf{x}) \ln p(\mathbf{x}|z) dz + \int q(z|\mathbf{x}) \ln \frac{p(z)}{q(z|\mathbf{x})} dz \quad (4)$$

$$= \int q(z|\mathbf{x}) \ln p(\mathbf{x}|z) dz - \text{KL}(q(z|\mathbf{x})||p(z)) \quad (5)$$

$$= \mathbb{E}_{z \sim q(z|\mathbf{x})} [\ln p(\mathbf{x}|z)] - \text{KL}(q(z|\mathbf{x})||p(z)) \quad (6)$$

- ここでは、単一のデータ  $\mathbf{x}$  と、それに対応する潜在変数  $z$  を考えている
- また、パラメータ  $\theta, \phi$  は省略している

# 変分自己符号化器 (VAE) の理論

- 第1項  $\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [\ln p(\mathbf{x}|\mathbf{z})]$  を大きく、また第2項  $\text{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$  を小さくすることで、変分下界  $\mathcal{L}(q)$  を大きくできる
- VAE では、変分下界  $\mathcal{L}(q)$  を最大化するパラメータ  $\theta, \phi$  を求めるために、ニューラルネットを使用する (変分推論にニューラルネットをねじ込んだもの)
- KL ダイバージェンスの項は、後ほど求めることにする (解析的に求められる)
- 第1項は、分布  $q(\mathbf{z}|\mathbf{x})$  に関する期待値であり、VAE ではサンプリングで近似する

$$\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [\ln p(\mathbf{x}|\mathbf{z})] \simeq \frac{1}{L} \sum_{l=1}^L \ln p(\mathbf{x}_i | \mathbf{z}_{i,l}) \quad (7)$$

# 変分自己符号化器 (VAE) の理論

- これより、変分下界  $\mathcal{L}(q)$  は以下のように書ける

$$\mathcal{L}(q) \simeq -\text{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \frac{1}{L} \sum_{i=1}^L \ln p(\mathbf{x}_i|\mathbf{z}_{i,l}) \quad (8)$$

- パラメータ  $\theta, \phi$  を含めれば、次のように書ける

$$\mathcal{L}(q) \simeq -\text{KL}(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}|\theta)) + \frac{1}{L} \sum_{i=1}^L \ln p(\mathbf{x}_i|\mathbf{z}_{i,l}, \theta) \quad (9)$$

# 変分自己符号化器 (VAE) の理論

- Encoder のニューラルネットの入出力
  - Encoder は、分布  $q(z|x, \phi)$  を表現するニューラルネット
  - 入力  $x$  に対応する潜在変数  $z$  を得る
- Encoder の入力、明らかにデータ  $x$  である
- 変分下界  $\mathcal{L}(q)$  において、項  $\mathbb{E}_z [\ln p(x|z)]$  は近似する必要があった
- $z$  を、分布  $q(z|x, \phi)$  から  $L$  回サンプリングした
- ニューラルネットで、 $x$  から  $z$  を直接サンプリングするのは困難
- そこで、Encoder では、サンプルされたデータは出力しないことにする
- その代わりに、サンプルする分布のパラメータを出力する
- 例えば、サンプルする分布がガウス分布であれば、平均と分散の2つのパラメータを出力する

# 変分自己符号化器 (VAE) の理論

- 後述のように、分布  $q(z|\mathbf{x}, \phi)$  はガウス分布になるので、Encoder は平均ベクトル  $\mu$  と共分散行列  $\Sigma$  を出力する
- 共分散行列は、実際には対角行列であるため、実際には行列ではなく、行列の対角成分を要素にもつベクトルを出力する

# 変分自己符号化器 (VAE) の理論

## ● Encoder の損失関数

- VAE では、事前分布として、平均ベクトル  $\mathbf{0}$ 、共分散行列  $\mathbf{I}$  のガウス分布  $\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$  を仮定する
- データ  $\mathbf{x}$  は高次元だが、実際にはそのうちの低次元な領域にまとまって存在する (多様体仮説)
- 従って、データ  $\mathbf{x}$  の構造を、より低次元な潜在変数  $\mathbf{z}$  の空間  $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$  に押し込めることができる
- Encoder の損失関数は、KL ダイバージェンス  $\text{KL}(q(\mathbf{z}|\mathbf{x}, \phi) || p(\mathbf{z}|\theta))$  で定義できる
- この KL ダイバージェンスを最小化することは、Encoder の分布  $q(\mathbf{z}|\mathbf{x}, \phi)$  を、 $p(\mathbf{z}|\theta) \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$  に近づける制約に相当する
- $p(\mathbf{z}|\theta)$  がガウス分布であれば、事後分布  $p(\mathbf{z}|\mathbf{x}, \theta)$  もガウス分布であり、従って  $q(\mathbf{z}|\mathbf{x}, \phi)$  もガウス分布となる



# 変分自己符号化器 (VAE) の理論

- よって  $\text{KL}(q(z|\mathbf{x}, \phi) || p(z|\theta))$  は、2つのガウス分布間の KL ダイバージェンスである
- 一般に、2つのガウス分布  $p(z) = \mathcal{N}(z|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ 、 $q(z) = \mathcal{N}(z|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  間の KL ダイバージェンスは、解析的に計算できる
- KL ダイバージェンス  $\text{KL}(p(z) || q(z))$  を順番に求めてみよう

$$\begin{aligned} & \text{KL}(p(z) || q(z)) \\ &= \int p(z) \ln \frac{p(z)}{q(z)} dz \end{aligned} \quad (10)$$

$$= \int p(z) (\ln p(z) - \ln q(z)) dz \quad (11)$$

$$= \int p(z) (\ln \mathcal{N}(z|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ln \mathcal{N}(z|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) dz \quad (12)$$

$$= \mathbb{E}_{z \sim p(z)} [\ln \mathcal{N}(z|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ln \mathcal{N}(z|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)] \quad (13)$$

- データを  $D$  次元、潜在変数を  $K$  次元とする

# 変分自己符号化器 (VAE) の理論

- ここで

$$\begin{aligned} & \ln \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ = & \ln \left( \frac{1}{(2\pi)^{\frac{K}{2}}} \frac{1}{|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{z} - \boldsymbol{\mu}_0) \right) \right) \\ = & -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_0| - \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{z} - \boldsymbol{\mu}_0) \quad (14) \end{aligned}$$

$$\begin{aligned} & \ln \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ = & -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{z} - \boldsymbol{\mu}_1) \quad (15) \end{aligned}$$

であるので

$$\begin{aligned} & \ln \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ln \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ = & -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_0| - \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{z} - \boldsymbol{\mu}_0) - \end{aligned}$$

# 変分自己符号化器 (VAE) の理論

$$\begin{aligned} & \left( -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_1| - \frac{1}{2} (z - \mu_1)^T \Sigma_1^{-1} (z - \mu_1) \right) \\ = & \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2} (z - \mu_0)^T \Sigma_0^{-1} (z - \mu_0) + \\ & \frac{1}{2} (z - \mu_1)^T \Sigma_1^{-1} (z - \mu_1) \end{aligned} \quad (16)$$

- よって

$$\begin{aligned} & \text{KL}(p(z) \| q(z)) \\ = & \mathbb{E}_{z \sim p(z)} [\ln \mathcal{N}(z | \mu_0, \Sigma_0) - \ln \mathcal{N}(z | \mu_1, \Sigma_1)] \\ = & \mathbb{E}_{p(z)} \left[ \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2} (z - \mu_0)^T \Sigma_0^{-1} (z - \mu_0) + \right. \\ & \left. \frac{1}{2} (z - \mu_1)^T \Sigma_1^{-1} (z - \mu_1) \right] \quad (17) \\ = & \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2} \mathbb{E}_{p(z)} \left[ (z - \mu_0)^T \Sigma_0^{-1} (z - \mu_0) \right] + \end{aligned}$$

# 変分自己符号化器 (VAE) の理論

$$\frac{1}{2} \mathbb{E}_{p(z)} \left[ (z - \mu_1)^T \Sigma_1^{-1} (z - \mu_1) \right] \quad (18)$$

- ここで、期待値についての式を導出しておく
- $\mathbb{E}[z] = \mu$ 、 $\mathbb{E}[(z - \mu)(z - \mu)^T] = \Sigma$  とする
- $z$  の  $i$  成分を  $z_i$ 、 $\mu$  の  $i$  成分を  $\mu_i$ 、 $\Sigma$  の  $i, j$  成分を  $\Sigma_{ij}$  とする
- このとき

$$\begin{aligned} \Sigma_{ij} &= \mathbb{E}[(z_i - \mu_i)(z_j - \mu_j)] \\ &= \mathbb{E}[z_i z_j - z_i \mu_j - z_j \mu_i + \mu_i \mu_j] \\ &= \mathbb{E}[z_i z_j] - \mu_j \mathbb{E}[z_i] - \mu_i \mathbb{E}[z_j] + \mu_i \mu_j \\ &= \mathbb{E}[z_i z_j] - \mu_j \mu_i - \mu_i \mu_j + \mu_i \mu_j \\ &= \mathbb{E}[z_i z_j] + \mu_i \mu_j \end{aligned} \quad (19)$$

であるから

$$\mathbb{E}[z_i z_j] = \Sigma_{ij} - \mu_i \mu_j \quad (20)$$

# 変分自己符号化器 (VAE) の理論

- そして、行列  $A$  の  $i, j$  成分を  $A_{ij}$  とすれば、以下を得る

$$\begin{aligned}\mathbb{E}[z^T A z] &= \mathbb{E}\left[\sum_i \sum_j z_i A_{ij} z_j\right] \\&= \sum_i \sum_j A_{ij} \mathbb{E}[z_i z_j] \\&= \sum_i \sum_j A_{ij} (\Sigma_{ij} + \mu_i \mu_j) \\&= \sum_i \sum_j A_{ij} \Sigma_{ij} + \sum_i \sum_j A_{ij} \mu_i \mu_j \\&= \sum_i \sum_j A_{ij} \Sigma_{ji} + \sum_i \sum_j \mu_i A_{ij} \mu_j \\&= \sum_i (A \Sigma)_{ii} + \mu^T A \mu \\&= \text{Tr}(A \Sigma) + \mu^T A \mu\end{aligned}\tag{21}$$

# 変分自己符号化器 (VAE) の理論

- 上式の変形では、共分散行列  $\Sigma$  が対称行列ゆえ、 $\Sigma_{ij} = \Sigma_{ji}$  が成立することを用いた
- また、ベクトル  $\mathbf{a}$  の  $i$  成分を  $a_i$  とすれば、以下を得る

$$\begin{aligned}\mathbb{E}[\mathbf{a}^T \mathbf{z}] &= \mathbb{E}[\mathbf{z}^T \mathbf{a}] \\ &= \mathbb{E}\left[\sum_i z_i a_i\right] \\ &= \sum_i a_i \mathbb{E}[z_i] \\ &= \sum_i a_i \mu_i \\ &= \mathbf{a}^T \boldsymbol{\mu} = \boldsymbol{\mu}^T \mathbf{a}\end{aligned}\tag{22}$$

# 変分自己符号化器 (VAE) の理論

- これより、 $\mathbf{a}, \mathbf{B}$  をそれぞれ適当なベクトル、行列とすれば、以下を得る

$$\begin{aligned} & \mathbb{E} \left[ (\mathbf{z} - \mathbf{a})^T \mathbf{B} (\mathbf{z} - \mathbf{a}) \right] \\ = & \mathbb{E} \left[ \mathbf{z}^T \mathbf{B} \mathbf{z} - \mathbf{z}^T \mathbf{B} \mathbf{a} - \mathbf{a}^T \mathbf{B} \mathbf{z} + \mathbf{a}^T \mathbf{B} \mathbf{a} \right] \\ = & \mathbb{E} \left[ \mathbf{z}^T \mathbf{B} \mathbf{z} \right] - \mathbb{E} \left[ \mathbf{z}^T \mathbf{B} \mathbf{a} \right] - \mathbb{E} \left[ \mathbf{a}^T \mathbf{B} \mathbf{z} \right] + \mathbf{a}^T \mathbf{B} \mathbf{a} \\ = & \left( \text{Tr}(\mathbf{B} \Sigma) + \boldsymbol{\mu}^T \mathbf{B} \boldsymbol{\mu} \right) - \boldsymbol{\mu}^T \mathbf{B} \mathbf{a} - \mathbf{a}^T \mathbf{B} \boldsymbol{\mu} + \mathbf{a}^T \mathbf{B} \mathbf{a} \\ = & \text{Tr}(\mathbf{B} \Sigma) + \boldsymbol{\mu}^T \mathbf{B} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \mathbf{B} \mathbf{a} + \mathbf{a}^T \mathbf{B} \mathbf{a} \end{aligned} \quad (23)$$

特に、 $\mathbf{a} = \boldsymbol{\mu}, \mathbf{B} = \Sigma^{-1}$  とすれば

$$\begin{aligned} & \mathbb{E} \left[ (\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right] \\ = & \text{Tr}(\Sigma^{-1} \Sigma) + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} \\ = & \text{Tr}(\Sigma^{-1} \Sigma) \\ = & \text{Tr}(\mathbf{I}) = K \end{aligned} \quad (24)$$

# 変分自己符号化器 (VAE) の理論

- これを用いれば、KL ダイバージェンスは次のようになる

$$\begin{aligned} & \text{KL}(p(\mathbf{z})||q(\mathbf{z})) \\ = & \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2} \mathbb{E}_{p(\mathbf{z})} \left[ (\mathbf{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{z} - \boldsymbol{\mu}_0) \right] + \\ & \frac{1}{2} \mathbb{E}_{p(\mathbf{z})} \left[ (\mathbf{z} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{z} - \boldsymbol{\mu}_1) \right] \\ = & \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2} K + \frac{1}{2} \left( \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_0 - \right. \\ & \left. 2\boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 \right) \\ = & \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2} K + \frac{1}{2} \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + \\ & \frac{1}{2} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\ = & \frac{1}{2} \left( \ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - K + \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + \right. \end{aligned}$$



# 変分自己符号化器 (VAE) の理論

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \Big) \quad (25)$$

- これで、2つのガウス分布  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ 、 $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  間の KL ダイバージェンスが、次のようになることが分かった

$$\begin{aligned} & \text{KL}(p(\mathbf{z})||q(\mathbf{z})) \\ &= \frac{1}{2} \left( \ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - K + \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + \right. \\ & \quad \left. (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \right) \end{aligned} \quad (26)$$

- ここでは、 $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  は、Encoder のニューラルネットが表現する分布  $q(\mathbf{z}|\mathbf{x}, \phi)$  である
- そして、分布  $q(\mathbf{z}|\mathbf{x}, \phi)$  はガウス分布であったので、ここでは平均  $\boldsymbol{\mu}_0$  と共分散行列  $\boldsymbol{\Sigma}_0$  を使って、 $q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  と表すことにする

# 変分自己符号化器 (VAE) の理論

- また  $q(z) = \mathcal{N}(z|\mu_1, \Sigma_1)$  は、潜在変数に関する事後分布  $p(z|\theta) = \mathcal{N}(z|\mathbf{0}, I)$  である
- 結局、KL ダイバージェンス  $\text{KL}(q(z|x, \phi)||p(z|\theta))$  は、先程の式に  $\mu_1 = \mathbf{0}$  と  $\Sigma_1 = I$  を代入すれば得られる

$$\begin{aligned} & \text{KL}(q(z|x, \phi)||p(z|\theta)) \\ &= \text{KL}(\mathcal{N}(z|\mu_0, \Sigma_0)||\mathcal{N}(z|\mathbf{0}, I)) \\ &= \frac{1}{2} (-\ln |\Sigma_0| - K + \text{Tr}(\Sigma_0) + \mu_0^T \mu_0) \end{aligned} \quad (27)$$

- この KL ダイバージェンスが、Encoder の損失関数として定義される
- Encoder のニューラルネットは、入力としてデータ  $x$  を取り、平均  $\mu_0$  と共分散行列  $\Sigma_0$  を出力する
- 従って、Encoder の出力と、潜在変数の次元  $K$  を上式に代入すれば、損失関数を容易に計算できる

# 変分自己符号化器 (VAE) の理論

- $\Sigma_0$  は対称行列であるため、実際に出力されるのは、 $\Sigma_0$  の対角成分を並べたベクトルである
- 一般的な VAE の Encoder は次の図 2 のように表せる

# 変分自己符号化器 (VAE) の理論

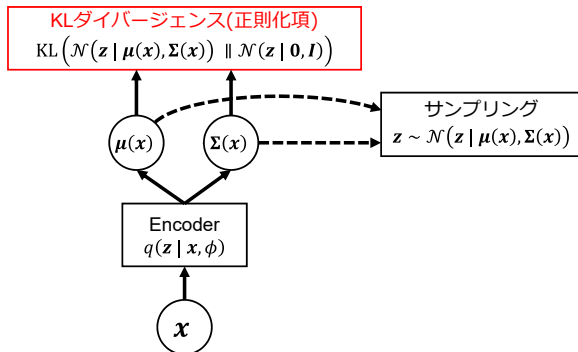


図 2: VAE の Encoder の概要

# 変分自己符号化器 (VAE) の理論

- Encoder のニューラルネットの構造

- VAE が最初に提案された論文では、隠れ層は 1 層となっている
- 最終層は、 $\mu_0$  と  $\sqrt{\Sigma_0}$  を出力する 2 つのユニットから成る
- $\sqrt{\Sigma_0}$  とは、行列  $\Sigma_0$  の各要素の平方根を取った行列である
- 隠れ層の重みを  $W_h$ 、バイアスを  $b_h$ 、活性化関数を  $f(\cdot)$ 、層の出力を  $h$  とする
- 隠れ層で行う処理は、次の式で表される

$$h = f(W_h x + b_h) \quad (28)$$

- $\mu_0$  は、重み  $W_m$  とバイアス  $b_m$  を使って、以下のように計算される

$$\mu_0 = W_m h + b_m \quad (29)$$

- $\sqrt{\Sigma_0}$  は、重み  $W_s$  とバイアス  $b_s$  から、以下のように計算される

$$\sqrt{\Sigma_0} = W_s h + b_s \quad (30)$$

# 変分自己符号化器 (VAE) の理論

- Decoder のニューラルネットの入出力
  - Decoder は、分布  $p(x|z, \theta)$  を表現するニューラルネット
  - 潜在変数  $z$  から元のデータ  $x$  を復元する
- Decoder の入力は、Encoder によってサンプリングされた  $z$  となる
- もう少し正確に表現すると、Encoder から出力されるのは、分布のパラメータ  $\mu, \Sigma$  である
- そして、そのパラメータを使って分布  $\mathcal{N}(z|\mu, \Sigma)$  を構成し、分布から  $z$  をサンプリングする
- Decoder の出力は、復元されたデータ  $y$  である

# 変分自己符号化器 (VAE) の理論

- Decoder の損失関数

- 画像データでは通常、各ピクセルの値が 0 から 1 までになるようにスケーリングされている
- このとき、 $p(x|z)$  はベルヌーイ分布と仮定していることになる
- 出力層のユニット  $j$  は、 $y_j = p(x_j = 1|z)$  を出力しているとみなせる
- $y_j$  は、再構成  $y$  の  $j$  番目の要素であり、元のデータ  $x$  の  $j$  番目の要素  $x_j$  と対応する
- VAE が最初に提案された論文では、隠れ層は 1 層のみである
- 再構成  $y$  は、次のように計算される

$$y = f_{\sigma}(W_o \tanh(W_h z + b_h) + b_o) \quad (31)$$

- $f_{\sigma}(\cdot)$  は、行列の各要素にシグモイド関数  $\sigma(\cdot)$  を適用する活性化関数
- $W_h, b_h$  は隠れ層の重みとバイアス、 $W_o, b_o$  は出力層の重みとバイアス

# 変分自己符号化器 (VAE) の理論

- このとき  $\ln p(\mathbf{x}|\mathbf{z})$  は次のように記述できる

$$\begin{aligned}\ln p(\mathbf{x}|\mathbf{z}) &= \ln \prod_{j=1}^D (p(x_j = 1|\mathbf{z}))^{x_j} (p(x_j = 0|\mathbf{z}))^{1-x_j} \\&= \ln \prod_j (p(x_j = 1|\mathbf{z}))^{x_j} (1 - p(x_j = 1|\mathbf{z}))^{1-x_j} \\&= \ln \prod_j y_j^{x_j} (1 - y_j)^{1-x_j} \\&= \sum_j (x_j \ln y_j + (1 - x_j) \ln (1 - y_j))\end{aligned}\tag{32}$$

- $\mathbf{z}$  は、分布  $q(\mathbf{z}|\mathbf{x}, \phi)$  からサンプリングされている
- 従って、上記を最大化することは、 $\mathbb{E}_{q(\mathbf{z})} [\ln p(\mathbf{x}|\mathbf{z})]$  を最大化することに等しい



# 変分自己符号化器 (VAE) の理論

- 上記は、 $x_j$  と  $y_j$  のいずれもベルヌーイ分布に従う (二値変数) ときの、**負の交差エントロピー**となっていることが分かる
- 従って、 $\mathbb{E}_{q(z)} [\ln p(x|z)]$  を最大化することは、**交差エントロピーを最小化**することに相当
- Decoder の損失関数は、以下の**交差エントロピー誤差**として定義できる

$$E = - \sum_j (x_j \ln y_j + (1 - x_j) \ln (1 - y_j)) \quad (33)$$

- 元データ  $x_j$  と、その再構成  $y_j$  との**差が大きければ大きいほど、上記の誤差は増大**する
- これより、上記の誤差は**再構成誤差** (Reconstruction Error) とよばれる
- これより、VAE の Encoder と Decoder は次の図 3 のように表せる

# 変分自己符号化器 (VAE) の理論

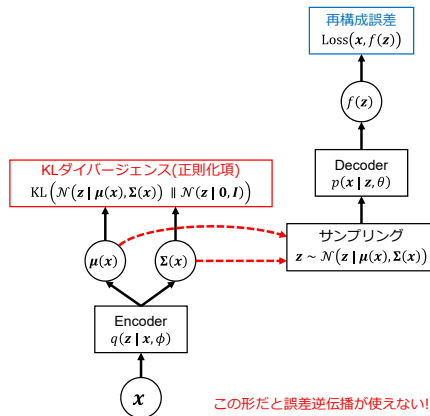


図 3: VAE の Encoder と Decoder の概要

# 変分自己符号化器 (VAE) の理論

- Reparameterization Trick

- VAE が先程の図 3 のようであるとき、大きな問題が生じる
- サンプルを行うと、計算グラフが途中で途切れるため、誤差逆伝播法を実行できない
- そこで、次の図 4 のように構成する
- $z \sim (z|\mu, \Sigma)$  として、 $z$  を分布から直接サンプリングするのではない
- $z$  を、決定論的な関数  $g(\epsilon, x|\phi)$  から決定する
- 但し、 $\epsilon$  は、分布  $p(\epsilon)$  からサンプリングされる
- ニューラルネットの最適化には無関係な項  $\epsilon$  と、Encoder のパラメータ  $\phi$  で  $z$  を表現することで、誤差逆伝播法を実行可能にする
- このテクニックを、Reparameterization Trick という
- $\epsilon \sim (\epsilon|0, I)$  とすれば、 $z$  は次のように計算できる

$$z = g(\epsilon, x|\phi) = \mu + \Sigma^{\frac{1}{2}} \epsilon \quad (34)$$

# 変分自己符号化器 (VAE) の理論

- 共分散行列  $\Sigma$  が対角行列であれば、上記の  $\Sigma^{\frac{1}{2}}\epsilon$  は、単なる要素ごとの積 (対角行列の各要素を並べたベクトルと、 $\epsilon$  の要素ごとの積) として書ける
- $z$  の式は以下のように導出できる
- 確率変数  $z, \epsilon$  間の関係が、次のようになっているとする

$$z = \mu + U\Lambda^{\frac{1}{2}}\epsilon \quad (35)$$

- 但し、正定値対称行列  $\Sigma$  が、固有値分解によって  $\Sigma = U\Lambda U^T = U\Lambda^{\frac{1}{2}}(U\Lambda)^T$  と表せるとする
- $U$  は固有ベクトルを並べた行列、 $\Lambda$  は対角成分に固有値をもつ対角行列とする

# 変分自己符号化器 (VAE) の理論

- 確率分布  $p(\mathbf{z})$  と  $p(\boldsymbol{\epsilon})$  との関係は、ヤコビ行列  $\mathbf{J} = \partial \boldsymbol{\epsilon} / \partial \mathbf{z}$  により次のように記述できる

$$p(\mathbf{z}) = p(\boldsymbol{\epsilon}) |\det(\mathbf{J})| = p(\boldsymbol{\epsilon}) \left| \det \left( \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} \right) \right| \quad (36)$$

- ヤコビ行列  $\mathbf{J}$  を計算すると次のようになる

$$\mathbf{J} = \frac{\partial \boldsymbol{\epsilon}}{\partial \mathbf{y}} = \frac{\partial}{\partial \mathbf{z}} \left( \left( \mathbf{U} \boldsymbol{\Lambda}^{\frac{1}{2}} \right)^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right) = \left( \left( \mathbf{U} \boldsymbol{\Lambda}^{\frac{1}{2}} \right)^{-1} \right)^T \quad (37)$$

# 変分自己符号化器 (VAE) の理論

- ヤコビ行列  $J$  の行列式は次のようになる

$$\begin{aligned}\det(J) &= \left| \left( (U\Lambda^{\frac{1}{2}})^{-1} \right)^T \right| \\ &= \left| (U\Lambda^{\frac{1}{2}})^{-1} \right| \quad (\because |A^T| = |A|) \\ &= \frac{1}{|U\Lambda^{\frac{1}{2}}|} \quad \left( \because |A^{-1}| = \frac{1}{|A|} \right)\end{aligned}\tag{38}$$

- $\Sigma$  の行列式は次のように表せる

$$|\Sigma| = \left| U\Lambda^{\frac{1}{2}} (U\Lambda^{\frac{1}{2}})^T \right| = |U\Lambda^{\frac{1}{2}}| \left| (U\Lambda^{\frac{1}{2}})^T \right| = |U\Lambda^{\frac{1}{2}}|^2 \tag{39}$$

- 従って  $|U\Lambda^{\frac{1}{2}}| = |\Sigma|^{\frac{1}{2}}$  である

# 変分自己符号化器 (VAE) の理論

- $\Sigma$  は正定値である ( $|\Sigma| > 0$ ) から、 $|\Sigma|^{\frac{1}{2}} = |U\Lambda^{\frac{1}{2}}| > 0$  が成立し、 $U\Lambda^{\frac{1}{2}}$  も正定値行列となる
- これより、逆行列  $(U\Lambda^{\frac{1}{2}})^{-1}$  が存在するので、ヤコビ行列  $J$  は計算できることが確認される
- ヤコビ行列  $J$  の行列式の絶対値は、次のようになる

$$|\det(J)| = \left| \frac{1}{|U\Lambda^{\frac{1}{2}}|} \right| = \left| \frac{1}{|\Sigma|^{\frac{1}{2}}} \right| = \frac{1}{|\Sigma|^{\frac{1}{2}}} \quad (40)$$

- $p(\epsilon)$  がガウス分布  $\mathcal{N}(\epsilon|\mathbf{0}, I)$  であるとする、 $p(z)$  は次のようになる

$$\begin{aligned} p(z) &= p(\epsilon) \left| \det \left( \frac{\partial z}{\partial \epsilon} \right) \right| \\ &= \frac{1}{(2\pi)^{\frac{K}{2}}} \exp \left( -\frac{1}{2} \epsilon^T \epsilon \right) \frac{1}{|\Sigma|^{\frac{1}{2}}} \end{aligned} \quad (41)$$

# 変分自己符号化器 (VAE) の理論

- ここで、指数関数の中身は次のように書ける

$$\begin{aligned} & (z - \mu)^T \Sigma^{-1} (z - \mu) \\ = & \left( U \Lambda^{\frac{1}{2}} \epsilon \right)^T (U \Lambda U^T)^{-1} \left( U \Lambda^{\frac{1}{2}} \epsilon \right) \\ = & \epsilon^T \left( \Lambda^{\frac{1}{2}} \right)^T U^T (U^T)^{-1} \Lambda^{-1} U^{-1} U \Lambda^{\frac{1}{2}} \epsilon \\ = & \epsilon^T \left( \Lambda^{\frac{1}{2}} \right)^T \Lambda^{-1} \Lambda^{\frac{1}{2}} \epsilon \\ = & \epsilon^T \Lambda^{\frac{1}{2}} \Lambda^{-1} \Lambda^{\frac{1}{2}} \epsilon \\ = & \epsilon^T \epsilon \end{aligned} \tag{42}$$



# 変分自己符号化器 (VAE) の理論

- これより、 $p(\mathbf{z})$  は平均  $\boldsymbol{\mu}$ 、共分散行列  $\boldsymbol{\Sigma}$  のガウス分布である

$$\begin{aligned} p(\mathbf{z}) &= \frac{1}{(2\pi)^{\frac{K}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}\right) \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \\ &= \frac{1}{(2\pi)^{\frac{K}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right) \end{aligned} \quad (43)$$

$$= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (44)$$

- $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$  について、共分散行列  $\boldsymbol{\Sigma}$  が既に対角行列であれば、 $\mathbf{U} = \mathbf{I}$ 、 $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}$  であるので、結局以下が言える

$$\mathbf{z} = \boldsymbol{\mu} + \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon} \quad (45)$$

# 変分自己符号化器 (VAE) の理論

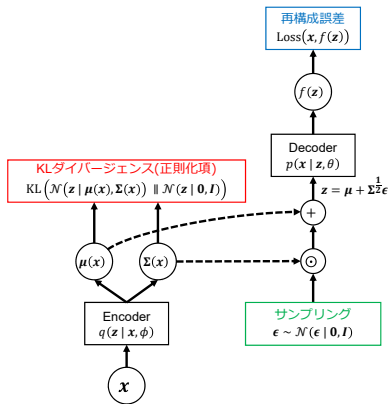


図 4: VAE の構造