目次

1 近似推論法

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

変分推論

- 変分推論が必要だった理由
  - 潜在変数に関する事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の計算は、困難であることが多い
  - どのような場合に困難になるのか、次の図 1 に示す
  - $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の厳密な計算は諦める代わりに、別の確率分布で近似したい
  - 別の確率分布で近似するとき、単純な項の積として表現できるといった、何らかの仮定を置く

変分推論

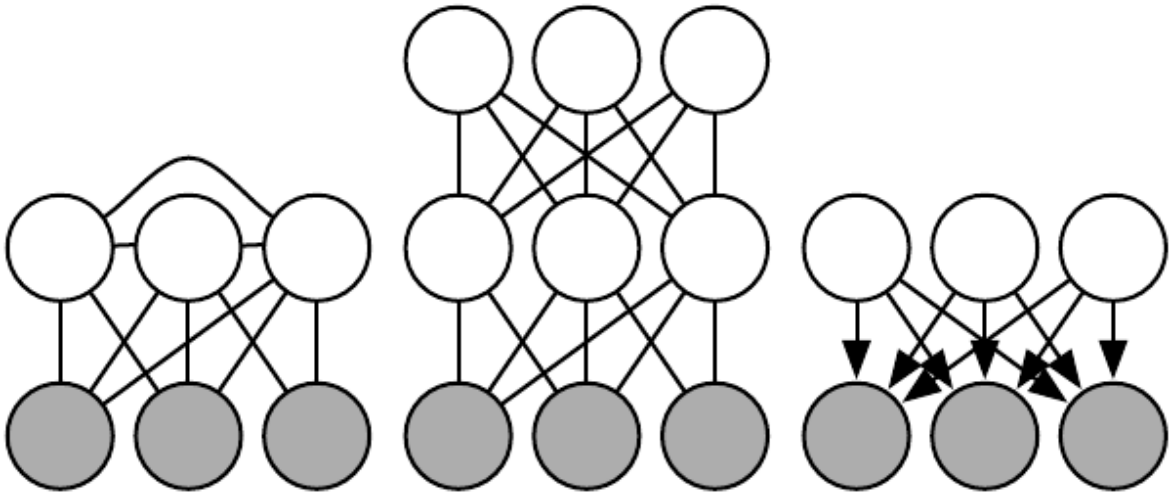

図 1: 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の計算が困難な場合

Figure 19.1: Intractable inference problems in deep learning are usually the result of interactions between latent variables in a structured graphical model. These can be due to edges directly connecting one latent variable to another, or due to longer paths that are activated when the child of a V-structure is observed. *(Left)*A **semi-restricted Boltzmann machine** (Osindero and Hinton, 2008) with connections between hidden units. These direct connections between latent variables make the posterior distribution intractable due to large cliques of latent variables. *(Center)*A deep Boltzmann machine, organized into layers of variables without intra-layer connections, still has an intractable posterior distribution due to the connections between layers. *(Right)*This directed model has interactions between latent variables when the visible variables are observed, because every two latent variables are co-parents. Some probabilistic models are able to provide tractable inference over the latent variables despite having one of the graph structures depicted above. This is possible if the conditional probability distributions are chosen to introduce additional independences beyond those described by the graph. For example, probabilistic PCA has the graph structure shown in the right, yet still has simple inference

図 2: 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の計算が困難な場合

変分推論

- 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の計算が困難な場合 1
  - グラフィカルモデルにおいて、潜在変数間の相互作用がある場合、計算が困難になる
  - 左側の半制限付きボルツマンマシンでは、全ての潜在変数の組み合わせ間で、接続がある
  - 従って、潜在変数間に依存関係が存在し、事後分布の計算が手に負えない
  - 白丸で描かれた潜在変数を $\boldsymbol{Z} = \{z_1, z_2, z_3\}$、灰色で描かれた観測データを $\boldsymbol{X}$ とおくと、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ は、例えば次のようになる

$$
\begin{aligned}
& p(z_1, z_2, z_3|\boldsymbol{X}) \\
= \quad & p(z_1|\boldsymbol{X}, z_2, z_3)p(z_2|\boldsymbol{X}, z_3)p(z_3|\boldsymbol{X})
\end{aligned} \tag{1}
$$

変分推論

- 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の計算が困難な場合 2
  - 中央は、層間の結合がない潜在変数の層で構成される、深層ボルツマンマシンを表す
  - 潜在変数の層間の結合があるため、事後分布の計算が手に負えない

  - 上の層の潜在変数を $\boldsymbol{Z}_1 = \{z_{11}, z_{12}, z_{13}\}$、中間層の潜在変数を $\boldsymbol{Z}_2 = \{z_{21}, z_{22}, z_{23}\}$、灰色で描かれた観測データを $\boldsymbol{X}$ とおくと、事後分布は、例えば次のようになる

$$
\begin{aligned}
& p(\boldsymbol{Z}_1, \boldsymbol{Z}_2|\boldsymbol{X}) \\
= \; & p(\boldsymbol{Z}_2|\boldsymbol{X}) p(\boldsymbol{Z}_1|\boldsymbol{Z}_2) \qquad (2) \\
= \; & p(z_{21}, z_{22}, z_{23}|\boldsymbol{X}) p(z_{11}, z_{12}, z_{13}|\boldsymbol{Z}_2) \\
= \; & p(z_{21}|\boldsymbol{X}) p(z_{22}|\boldsymbol{X}) p(z_{23}|\boldsymbol{X}) \\
& \quad p(z_{11}|\boldsymbol{Z}_2) p(z_{12}|\boldsymbol{Z}_2) p(z_{13}|\boldsymbol{Z}_2) \qquad (3)
\end{aligned}
$$

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

変分推論

- 変分推論の目的
  - 同時分布 $p(\boldsymbol{X}, \boldsymbol{Z})$ が求まっているときに、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ と、エビデンス $p(\boldsymbol{X})$ の近似を求める

- 注意点
  - 観測変数と潜在変数をそれぞれ、$\boldsymbol{X}, \boldsymbol{Z}$ とおく
  - $p(\boldsymbol{X})$ は、確率モデルからデータ $\boldsymbol{X}$ が生起する確率である
  - データからみたモデルの好みと解釈できるから、$p(\boldsymbol{X})$ をモデルエビデンスという

変分推論

- 周辺分布の対数 $\ln p(\boldsymbol{X})$ の分解
  - EM アルゴリズムのときと同様であり、次のように分解できる

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{4}$$

  - 但し、$\mathcal{L}(q)$ と $\mathrm{KL}(q||p)$ は次のように定義した

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{5}$$

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{6}$$

- パラメータ $\boldsymbol{\theta}$ の扱い
  - EM アルゴリズムとは異なり、パラメータ $\boldsymbol{\theta}$ がどこにも現れていない
  - パラメータも潜在変数として扱っているので、パラメータベクトルは明示的には書かない

変分推論

- ここでは、パラメータ $\boldsymbol{\theta}$ については、あまり気にしない
- ここでは連続潜在変数について考えるが、離散潜在変数であれば、積分を $\boldsymbol{Z}$ に関する総和に置き換えればよい

- 下界 $\mathcal{L}(q)$ を最適化する動機
  - $\mathcal{L}(q)$ はエビデンスの対数 $\ln p(\boldsymbol{X})$ の下界であるから、エビデンス下界 (Evidence lower bound, ELBO) ともいう
  - または、負の変分自由エネルギー (Variational free energy) という
  - $\ln p(\boldsymbol{X})$ は q には依存しないため、定数項とみなせる
  - 従って、$\mathcal{L}(q)$ を q について最大化することは、$\mathrm{KL}(q||p)$ の最小化に相当
  - このとき、分布 $q(\boldsymbol{Z})$ を真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ に近づけられる
  - $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ が分かれば、データ $\boldsymbol{X}$ から、潜在変数やパラメータ $\boldsymbol{Z}$ が得られる

変分推論

- 下界 $\mathcal{L}(q)$ の最適化
  - EM アルゴリズムのときと同じように、下界 $\mathcal{L}(q)$ を、分布 $q(\boldsymbol{Z})$ について最大化する
  - これは、KL ダイバージェンス $\mathrm{KL}(q||p)$ を最小化することと等価である
  - 従って、もし $q(\boldsymbol{Z})$ を任意の分布にしてよければ、$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ とおいて、KL ダイバージェンスを 0 にすればよい
  - しかしここでは、真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を求めることは不可能と仮定する
- 分布 $q(\boldsymbol{Z})$ の近似
  - 計算コストを削減するために、$q(\boldsymbol{Z})$ の形をある程度制限する
  - 制限したクラスの $q(\boldsymbol{Z})$ の中で、KL ダイバージェンス $\mathrm{KL}(q||p)$ を最小化するものを探す

変分推論

- 変分推論の目的
  - 分布のクラスを制限することで、$q(\boldsymbol{Z})$ を計算可能にすること
  - 表現力が豊かなクラスを使うことで、真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を良く近似する

  - 計算可能な分布のクラスの中で、可能な限り豊かな表現力を持つものを選びたい
  - 表現力が豊かな分布を使うことは、真の事後分布を、精度良く近似することにつながるのであって、従って過学習は発生しない

  - 分布 $q(\boldsymbol{Z})$ は、$q(\boldsymbol{Z}|\boldsymbol{X})$ と書くこともある

変分推論

- 分布 $q(\boldsymbol{Z})$ のクラスを制限する方法
  - 例えば分布 $q(\boldsymbol{Z})$ を、パラメトリックな分布に限定することができる
  - 即ち、パラメータベクトル $\boldsymbol{\omega}$ によって $q(\boldsymbol{Z}|\boldsymbol{\omega})$ と記述されるような、分布に制限する

  - 分布 $q(\boldsymbol{Z})$ を、ガウス分布などの、何らかの特別なパラメトリックな分布と仮定することに相当

変分推論

- クラスを制限する別の方法 (平均場近似)
  - 分布 $q(\boldsymbol{Z})$ のクラスを制限する別の方法として、平均場近似がある
  - 潜在変数 $\boldsymbol{Z}$ を、M 個の互いに排反なグループ $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M\}$ に分割
  - 分布 $q(\boldsymbol{Z})$ が、これらのグループによって分解されると仮定

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i) \tag{7}$$

  - 分布 q について、これ以上の仮定はしない
  - 従って、各因子 $q_i(\boldsymbol{Z}_i)$ の関数形については、何の制限も課さない
  - 平均場近似とは、元々は物理学における用語である

変分推論

- 下界 $\mathcal{L}(q)$ の最大化
  - $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ と分解できるような分布 $q(\boldsymbol{Z})$ の中で、下界 $\mathcal{L}(q)$ を最大にするものを探す
  - $\mathcal{L}(q)$ を $q(\boldsymbol{Z})$ について最大化するために、$\mathcal{L}(q)$ を各因子 $q_i(\boldsymbol{Z}_i)$ について順番に最大化していく
  - $\mathcal{L}(q)$ に $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ を代入して、因子の一つ $q_j(\boldsymbol{Z}_j)$ に関する依存項を抜き出してみよう

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{8}$$

$$= \int \left(\prod_i q_i(\boldsymbol{Z}_i) \right) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{9}$$

$$= \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \tag{10}$$

$$= \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z} - \int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i)\right) d\boldsymbol{Z} \tag{11}$$

ここで第 1 項は

$$\int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z}$$

$$= \int q_j(\boldsymbol{Z}_j) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i)\right) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z} \tag{12}$$

$d\boldsymbol{Z} = d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M$ であるから

$$= \int q_j(\boldsymbol{Z}_j) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i)\right) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \tag{13}$$

$$= \int q_j(\boldsymbol{Z}_j) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i)\right) \left(\prod_{i \neq j} d\boldsymbol{Z}_i\right) d\boldsymbol{Z}_j \quad (14)$$

$$= \int q_j(\boldsymbol{Z}_j) \left(\int \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i\right) d\boldsymbol{Z}_j \quad (15)$$

$$= \int q_j(\boldsymbol{Z}_j) \ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j \quad (16)$$

- 但し、新しい分布 $\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ は以下の式で定義した (積分の結果であるため、定数項が出現する)

$$\ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \quad (17)$$

変分推論

- 記法 $\mathbb{E}_{i \neq j}$ は、$i \neq j$ をみたす全ての分布 $q_i(\boldsymbol{Z}_i)$ による、期待値を取ることを表す

$$\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int (\ln p(\boldsymbol{X}, \boldsymbol{Z})) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{18}$$

- 第 2 項は

$$\begin{aligned}
& \int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \\
= & \sum_i \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} && (19) \\
= & \sum_i \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M && (20) \\
= & \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_j(\boldsymbol{Z}_j) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M + &
\end{aligned}$$

$$\sum_{i \neq j} \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i)\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \tag{21}$$

但し

$$\int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_j(\boldsymbol{Z}_j)\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \tag{22}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \left(\prod_{k \neq j} q_k(\boldsymbol{Z}_k) \right) \left(\prod_{k \neq j} d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_j \tag{23}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \left(\int \prod_{k \neq j} q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_j \tag{24}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \prod_{k \neq j} \underbrace{\left(\int q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)}_{=1} d\boldsymbol{Z}_j \tag{25}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \tag{26}$$

であるほか

$$\sum_{i \neq j} \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M$$

$$= \sum_{i \neq j} \int q_i(\boldsymbol{Z}_i) q_j(\boldsymbol{Z}_j) \left(\prod_{k \neq i,j} \ln q_k(\boldsymbol{Z}_k) \right)$$

$$\left(\ln q_i(\boldsymbol{Z}_i) \right) \left(\prod_{k \neq i,j} d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_i d\boldsymbol{Z}_j \tag{27}$$

$$= \sum_{i \neq j} \underbrace{\left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \right)}_{=1} \left(\int \prod_{k \neq i,j} \ln q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)$$

$$\int \ln q_i(\boldsymbol{Z}_i) q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{28}$$

$$= \sum_{i \neq j} \prod_{k \neq i,j} \underbrace{\left(\int \ln q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)}_{=\text{Const.}} \underbrace{\int \ln q_i(\boldsymbol{Z}_i) q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i}_{=\text{Const.}} \tag{29}$$

$$= \sum_{i \neq j} \text{Const.} = \text{Const.} \tag{30}$$

となるから結局

$$\int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \text{Const.} \tag{31}$$

変分推論

- 従って、下界 $\mathcal{L}(q)$ から $q_j(\boldsymbol{Z}_j)$ に依存する項を取り出すと

$$\mathcal{L}(q) = \int q_j(\boldsymbol{Z}_j) \ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j - \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \text{Const.} \tag{32}$$

但し

$$\ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{33}$$

- $\mathcal{L}(q)$ を、$i \neq j$ である全ての $q_i(\boldsymbol{Z}_i)$ について固定した上で、$q_j(\boldsymbol{Z}_j)$ について最大化することになる
- $q_j(\boldsymbol{Z}_j)$ について可能な全ての分布の中で、$\mathcal{L}(q)$ を最大にするようなものを選ぶ

変分推論

- $\mathcal{L}(q)$ は次のように変形できる

$$\mathcal{L}(q) = \int q_j(\boldsymbol{Z}_j) \ln \frac{\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)}{q_j(\boldsymbol{Z}_j)} d\boldsymbol{Z}_j + \text{Const.} \tag{34}$$

$$= -\text{KL}\left(q_j(\boldsymbol{Z}_j) || \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)\right) + \text{Const.} \tag{35}$$

- これより、$\mathcal{L}(q)$ は $q_j(\boldsymbol{Z}_j)$ と $\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ の間の、負の KL ダイバージェンスとなっている
- そして、$\mathcal{L}(q)$ の $q_j(\boldsymbol{Z}_j)$ に関する最大化は、KL ダイバージェンスの最小化と等価

- KL ダイバージェンスを最小にするためには、$q_j(\boldsymbol{Z}_j) = \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ とすればよい

- 従って、$q_j(\boldsymbol{Z}_j)$ の最適解は、次のように書ける

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{36}$$

変分推論

- 下界 $\mathcal{L}(q)$ を最大化する $\ln q_j(\boldsymbol{Z}_j)$ の解
  - $q_j(\boldsymbol{Z}_j)$ の最適解は次のように書けた

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{37}$$

  - 上式は次のことを意味している
  - 因子 $q_j(\boldsymbol{Z}_j)$ の最適解の対数 $\ln q_j^*(\boldsymbol{Z}_j)$ は、観測データ $\boldsymbol{X}$ と潜在変数 $\boldsymbol{Z}$ の同時分布の対数 $\ln p(\boldsymbol{X}, \boldsymbol{Z})$ を考え、$i \neq j$ である他の因子 $q_i(\boldsymbol{Z}_i)$ について期待値を取ったものである

  - 定数項は、正規化することで得られるので、結局次のようになる

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]) d\boldsymbol{Z}_j} \tag{38}$$

  - 正規化定数は必要に応じて計算すればよいので、取り敢えず無視できる

変分推論

- 最適解の式は、$q(\boldsymbol{Z})$ の分解の数だけ得られるので、$\{q_i(\boldsymbol{Z}_i)\}$ に関する M 本の連立方程式となる
- この方程式は、分布 $q(\boldsymbol{Z})$ が M 個の因子に分解されるという仮定の下で、下界 $\mathcal{L}(q)$ の最大値が満たすべき条件である

- $\ln q_j^*(\boldsymbol{Z}_j)$ の右辺は、$i \neq j$ である $q_i(\boldsymbol{Z}_i)$ の期待値に依存するため、$q_j^*(\boldsymbol{Z}_j)$ を陽に求めることができない
- そこで、下界 $\mathcal{L}(q)$ は次のように最適化される (重要)

- $i \neq j$ である全ての $q_i(\boldsymbol{Z}_i)$ を固定した状態で、$q_j(\boldsymbol{Z}_j)$ を最適化することを、全ての $j = 1, \ldots, M$ について繰り返す手続きを、座標降下法という

変分推論

- 下界 $\mathcal{L}(q)$ の最適化 (座標降下法)
  1. 全ての因子 $q_j(\boldsymbol{Z}_j)$ を適当に初期化する
  2. 各因子を、以下の式を使って更新する

$$\ln q_j(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{39}$$

$$\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{40}$$

     即ち、因子 $q_j(\boldsymbol{Z}_j)$ を、他の全ての因子の現在の値 $q_i(\boldsymbol{Z}_i)$ を使って改良する

  3. (2) を、下界 $\mathcal{L}(q)$ が収束するまで繰り返す

変分推論

- ここまでの話の流れ
  1. $\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}))$ であるから、エビデンス下界 $\mathcal{L}(q)$ を q について最大化すれば、$\mathrm{KL}(q||p) = 0$ とでき、従って $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ を得る
  2. $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ が分かれば、データ $\boldsymbol{X}$ から、潜在変数やパラメータ $\boldsymbol{Z}$ が得られる (パラメータは潜在変数 $\boldsymbol{Z}$ に含まれている)
  3. しかし、事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ は計算不可能なので、何らかの方法で近似するしかない
  4. 近似するといっても、計算可能でなければならないので、$q(\boldsymbol{Z})$ の形には、通常何らかの制限を課す
  5. $q(\boldsymbol{Z})$ を、パラメトリックな分布 $q(\boldsymbol{Z}|\boldsymbol{\omega})$ と仮定することがある

変分推論

6. または、$q(\boldsymbol{Z})$ を、$\prod_i q_i(\boldsymbol{Z}_i)$ のように分解できるとする (平均場近似)

7. 平均場近似を行うとき、各因子 $q_j(\boldsymbol{Z}_j)$ の最適解は、
$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.}$ であった

8. 全ての因子 $\{q_j(\boldsymbol{Z}_j)\}$ を同時に最適化することはできない

9. 下界 $\mathcal{L}(q)$ を、各因子 $q_j(\boldsymbol{Z}_j)$ について順番に最適化することはできる

- これからの話の流れ
  - MAP 推定と最尤推定は、変分推論の特殊な場合であることを確認する

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

変分推論

- MAP 推定および最尤推定の変分推論からの導出
  - 変分推論で得たいのは、潜在変数に関する事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ である
  - この変分推論の特殊ケースが、MAP 推定や最尤推定であることを導く

  - 変分推論では、エビデンス下界 $\mathcal{L}(q)$ を q について最大化し、従って $\mathrm{KL}(q||p)$ を最小化する
  - いま、分布 $q(\boldsymbol{Z})$ が、次のデルタ関数であるとする
  - 特定の値 $\boldsymbol{Z} = \widehat{\boldsymbol{Z}}$ についてのみ確率が非零になる、無限に鋭い分布

$$q(\boldsymbol{Z}) = \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \tag{41}$$

変分推論

- このとき、KL ダイバージェンス $\mathrm{KL}(q||p)$ は次のようになる

$$
\begin{aligned}
& \mathrm{KL}(q||p) \\
\equiv\ & \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X})) \\
=\ & -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
=\ & -\int q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} && (42) \\
=\ & -\int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z} + \\
& \qquad \int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) d\boldsymbol{Z} && (43) \\
=\ & -\ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) + \ln \delta(\widehat{\boldsymbol{Z}} - \widehat{\boldsymbol{Z}}) && (44) \\
=\ & -\ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) + \mathrm{Const.} && (45)
\end{aligned}
$$

変分推論

- これを $q(\boldsymbol{Z})$ について最小化することは、$\widehat{\boldsymbol{Z}}$ について最小化することに相当する
- よって $\widehat{\boldsymbol{Z}}$ の最適解 $\widehat{\boldsymbol{Z}}^*$ は

$$\begin{aligned}
\widehat{\boldsymbol{Z}}^* &= \arg\min_{\widehat{\boldsymbol{Z}}} \Big(-\ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) \Big) && (46)\\
&= \arg\max_{\widehat{\boldsymbol{Z}}} \ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) && (47)\\
&= \arg\max_{\widehat{\boldsymbol{Z}}} p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) && (48)\\
&= \arg\max_{\widehat{\boldsymbol{Z}}} p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) p(\widehat{\boldsymbol{Z}}) && (49)
\end{aligned}$$

  となるから、MAP 推定 (最大事後確率推定) の式と一致する

- これより、MAP 推定は、KL ダイバージェンス $\mathrm{KL}(q||p)$ の最小化、従って、エビデンス下界 $\mathcal{L}(q)$ の最大化と等価である

変分推論

- 更に、$p(\widehat{\boldsymbol{Z}}) = \text{Const.}$、即ち $\widehat{\boldsymbol{Z}}$ に関する事前の情報がないとすると

$$\begin{aligned} \widehat{\boldsymbol{Z}}^* &= \arg\max_{\widehat{\boldsymbol{z}}} p(\boldsymbol{X}|\widehat{\boldsymbol{Z}})p(\widehat{\boldsymbol{Z}}) \\ &= \arg\max_{\widehat{\boldsymbol{z}}} p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) \end{aligned} \tag{50}$$

となるから、これは最尤推定の式に一致する

変分推論

- 下界 $\mathcal{L}(q)$ の式から導くことも可能である

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
&= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} \qquad (51) \\
&= \int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} - \\
&\qquad \int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) d\boldsymbol{Z} \qquad (52) \\
&= \ln p(\boldsymbol{X}, \widehat{\boldsymbol{Z}}) - \ln \delta(\widehat{\boldsymbol{Z}} - \widehat{\boldsymbol{Z}}) \qquad (53) \\
&= \ln p(\boldsymbol{X}, \widehat{\boldsymbol{Z}}) + \text{Const.} \qquad (54) \\
&= \ln p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) p(\widehat{\boldsymbol{Z}}) + \text{Const.} \qquad (55)
\end{aligned}
$$

- これより、下界 $\mathcal{L}(q)$ の最大化は、$\widehat{\boldsymbol{Z}}$ に関する $p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) p(\widehat{\boldsymbol{Z}})$ の最大化、即ち MAP 推定と等価である

MAP 推定の例

- スパース符号化
  - ここでは MAP 推定の例として、スパース符号化を扱う (変分推論の例ではない)
  - データ $\boldsymbol{x}$ に対する潜在変数 $\boldsymbol{z}$ に、スパース性を持たせる
  - そのために、スパース性を導く事前分布 (ラプラス事前分布) を、潜在変数に用いる
  - データ $\boldsymbol{x}$ を D 次元、また潜在変数 $\boldsymbol{z}$ を K 次元とする
  - データ $\boldsymbol{x}$ に対応する潜在変数を $\boldsymbol{z}_i$ と表し、$\boldsymbol{z}_i$ の第 k 成分を z_{ik} とする

$$p(z_{ik}|\lambda) = \frac{\lambda}{2}\exp\left(-\lambda|z_{ik}|\right) \tag{56}$$

$$p(\boldsymbol{z}_i|\lambda) = \prod_k p(z_{ik}) = \prod_k \frac{\lambda}{2}\exp\left(-\lambda|z_{ik}|\right) \tag{57}$$

MAP 推定の例

- データに関する事後分布 $p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}, \boldsymbol{b}, \beta)$ を、次で定義する

$$
\begin{aligned}
& p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}, \boldsymbol{b}, \beta) \\
= \quad & \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}, \beta^{-1}\boldsymbol{I}) && (58) \\
= \quad & \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\beta^{-1}\boldsymbol{I}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))^T \right. \\
& \left. \left(\beta^{-1}\boldsymbol{I}\right)^{-1}(\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))\right) && (59) \\
= \quad & \frac{1}{(2\pi\beta^{-1})^{\frac{D}{2}}} \exp\left(-\frac{\beta}{2}(\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))^T \right. \\
& \left. (\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))\right) && (60)
\end{aligned}
$$

- データ $\boldsymbol{x}_i$ は、対応する潜在変数 $\boldsymbol{z}_i$ に、線形変換 $\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}$ を施し、更に分散 $\beta^{-1}\boldsymbol{I}$ のガウスノイズを足すことで、生成されると考える

MAP 推定の例

- パラメータについての補足
  - 今回は、$\boldsymbol{b} = 0$ とおいてバイアスを無視したものを考える
  - λ と、精度 β はハイパーパラメータであり、予め決められているとする
  - そこでこれ以降、次のように確率分布を表現する

$$p(\boldsymbol{z}_i|\lambda) = p(\boldsymbol{z}_i) \tag{61}$$

$$p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}, \boldsymbol{b}, \beta) = p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}) \tag{62}$$

  - $\boldsymbol{X}$ を、全てのデータ $\{\boldsymbol{x}_i\}$ を集めた行列とする (第 i 行ベクトルが $\boldsymbol{x}_i^T$)
  - $\boldsymbol{Z}$ についても同様に、全ての潜在変数 $\{\boldsymbol{z}_i\}$ を集めた行列として定める (第 i 行ベクトルが $\boldsymbol{z}_i^T$)

- スパース符号化の学習
  - 事後分布 $p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ は、表現することすら困難であるため、最尤推定による手法 (EM アルゴリズムなど) は利用できない

MAP 推定の例

- そこで、最尤推定の代わりに MAP 推定を利用することで、最適なパラメータが得られる
- 最大化するのは、以下の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ である

$$
\begin{aligned}
p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W}) &= \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{W})}{p(\boldsymbol{X})} && (63)\\
&\propto p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{W}) && (64)\\
&= p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{W})p(\boldsymbol{Z}) && (65)\\
&= \left(\prod_i p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W})\right)\left(\prod_i p(\boldsymbol{z}_i)\right) && (66)\\
&= \prod_i p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W})p(\boldsymbol{z}_i) && (67)
\end{aligned}
$$

MAP 推定の例

- 対数を取って最大化してもよいので、最大化する量は結局

$$\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{W}) = \ln\left(\prod_i p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W})p(\boldsymbol{z}_i)\right) \tag{68}$$

$$= \sum_i \left(\ln p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W}) + \ln p(\boldsymbol{z}_i)\right) \tag{69}$$

$$= \sum_i \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W}) + \sum_i \ln p(\boldsymbol{z}_i) \tag{70}$$

各項を求めると

$$\sum_i \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W})$$

$$= \sum_i \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i,\beta^{-1}\boldsymbol{I})$$

$$= \sum_i \ln\left(\frac{1}{(2\pi\beta^{-1})^{\frac{D}{2}}}\exp\left(-\frac{\beta}{2}\left(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{z}_i\right)^T\left(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{z}_i\right)\right)\right)$$

MAP 推定の例

$$
\begin{aligned}
&= \sum_i \left(-\frac{D}{2} \ln 2\pi - \frac{D}{2} \ln \beta^{-1} - \right. \\
&\qquad \left. \frac{\beta}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right)^T \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right) \right) \\
&= \sum_i \left(-\frac{\beta}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right)^T \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right) \right) + \text{Const.} \\
&= -\frac{\beta}{2} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right)^T \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right) + \text{Const.} \\
&= -\frac{\beta}{2} \sum_i \left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right)_{i:} \left(\left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right)_{i:} \right)^T + \text{Const.} \qquad (71) \\
&= -\frac{\beta}{2} \sum_{i,j} \left(\left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right) \odot \left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right) \right)_{i,j} + \text{Const.} \qquad (72) \\
&= -\frac{\beta}{2} \sum_{i,j} \left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right)^2_{i,j} + \text{Const.} \qquad (73)
\end{aligned}
$$

$$= \quad -\frac{\beta}{2}||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 + \text{Const.} \tag{74}$$

また

$$\begin{aligned}
& \sum_i \ln p(\boldsymbol{z}_i) \\
= \quad & \sum_i \ln \prod_j \left(\frac{\lambda}{2} \exp\left(-\lambda |z_{ij}|\right) \right) \\
= \quad & \sum_i \sum_j \left(\ln \frac{\lambda}{2} - \lambda |z_{ij}| \right) \\
= \quad & -\lambda \sum_{i,j} |z_{ij}| + \text{Const.} \qquad (75) \\
= \quad & -\lambda \sum_{i,j} |\boldsymbol{Z}_{i,j}| + \text{Const.} \qquad (76) \\
= \quad & -\lambda ||\boldsymbol{Z}||_1 + \text{Const.} \qquad (77)
\end{aligned}$$

MAP 推定の例

のようになる

- $\odot$ はアダマール積、$\boldsymbol{A}_{i:}$ は行列 $\boldsymbol{A}$ の第 i 行目のベクトルを表す
- また、行列のノルム $||\boldsymbol{A}||_p$ は次で定義される

$$||\boldsymbol{A}||_p = \left(\sum_{i,j} |A_{i,j}|^p \right)^{\frac{1}{p}} \tag{78}$$

$p = 1, p = 2$ のときは次のようになる

$$||\boldsymbol{A}||_1 = \sum_{i,j} |A_{i,j}|, \quad ||\boldsymbol{A}||_2 = \sqrt{\sum_{i,j} A_{i,j}^2} \tag{79}$$

- これより、$||\boldsymbol{A}||_1$ は、行列 $\boldsymbol{A}$ の各要素の絶対値の和を表す
- また $||\boldsymbol{A}||_2$ は、行列 $\boldsymbol{A}$ の各要素の二乗和の平方根を表し、フロベニウスノルム $||\boldsymbol{A}||_F$ ともよばれる

MAP 推定の例

- 従って、$||\boldsymbol{A}||_2^2$ は、行列 $\boldsymbol{A}$ の各要素の二乗和である
- 上記から

$$\begin{aligned} & \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W}) \\ = & \sum_i \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}) + \sum_i \ln p(\boldsymbol{z}_i) \\ = & -\frac{\beta}{2}||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 - \lambda||\boldsymbol{Z}||_1 + \text{Const.} \end{aligned} \tag{80}$$

  となるので、$\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ の最大化は、以下の関数 $J(\boldsymbol{Z}, \boldsymbol{W})$ の最小化である

$$J(\boldsymbol{Z}, \boldsymbol{W}) = ||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 + ||\boldsymbol{Z}||_1 \tag{81}$$

- スパース符号化の学習方法のまとめ
  - 関数 $J(\boldsymbol{Z}, \boldsymbol{W})$ を、$\boldsymbol{Z}$ と $\boldsymbol{W}$ について交互に最小化すればよい

MAP 推定の例

- $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ の最大化は、$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{W})p(\boldsymbol{Z})$ の最大化、即ち $J(\boldsymbol{Z}, \boldsymbol{W})$ の最小化と等価であった
- 従って、$J(\boldsymbol{Z}, \boldsymbol{W})$ の各パラメータ $\boldsymbol{Z}, \boldsymbol{W}$ についての最小化は、事後確率を大きくする方向に働く

- 関数 $J(\boldsymbol{Z}, \boldsymbol{W})$ をもう一度みてみよう

$$J(\boldsymbol{Z}, \boldsymbol{W}) = ||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 + ||\boldsymbol{Z}||_1 \tag{82}$$

- 第 1 項は、明らかに再構成誤差を表している
- 第 2 項は、データ $\boldsymbol{X}$ の表現 $\boldsymbol{Z}$ がスパースになるように付加した、正則化項である

- $\boldsymbol{Z}\boldsymbol{W}^T$ は、データ $\boldsymbol{X}$ の内部表現 $\boldsymbol{Z}$ に、重み $\boldsymbol{W}$ を掛けることで、データ $\boldsymbol{X}$ を復元したもの

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

変分推論

- これまでの話の流れ
  1. 変分推論から、MAP 推定と最尤推定が導出できることを確認した
  2. MAP 推定は、$q(\boldsymbol{Z})$ を無限に鋭い確率分布 (デルタ関数) として、$\boldsymbol{Z}$ が特定の値しか取らないと仮定した場合であった
  3. 最尤推定は、MAP 推定において、$\boldsymbol{Z}$ の事前確率分布を設けない場合であった
  4. MAP 推定の例として、スパース符号化を扱った
- これからの話の流れ
  - 話題を変えて、分解 $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ によって $p(\boldsymbol{Z}|\boldsymbol{X})$ を近似するときの、弊害を調べる

変分推論

- $q(\boldsymbol{Z})$ の分解による近似の性質
  - 変分推論では、真の事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ を、分解により近似する
  - 分解で近似することによって、どのような不正確さが生じるのか?

- ガウス分布の分解による近似
  - ガウス分布を、分解されたガウス分布で近似することを考えてみよう
  - 分解による近似で、どのような問題が起こるのか考えてみよう

  - 2 つの変数 $\boldsymbol{z} = (z_1, z_2)$ 間には、相関があるとする
  - $\boldsymbol{z}$ はガウス分布 $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ に従っているとする

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \tag{83}$$

  - 精度行列 $\boldsymbol{\Lambda}$ は対称行列であるから、$\Lambda_{12} = \Lambda_{21}$

変分推論

- この分布 $p(\boldsymbol{z})$ を、分解されたガウス分布 $q(\boldsymbol{z}) = q_1(z_1)q_2(z_2)$ で近似する
- 各因子 $q_1(z_1), q_2(z_2)$ の関数形については何の仮定も置いていないことに注意

- 最適な因子 $q_1(z_1), q_2(z_2)$ の計算
  - 最適な因子 $q_1^*(z_1)$ を、先程の結果を使って求める

$$\ln q_j(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{84}$$

  - 従って、$q_1^*(z_1)$ を計算する式は次のようになる

$$\ln q_1^*(z_1) = \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} \tag{85}$$

$$\mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] = \int \ln p(\boldsymbol{z}) q_2(z_2) dz_2 \tag{86}$$

  - 上式の右辺では、z_1 に依存する項だけを考えればよい

変分推論

- z_1 の関数を求めようとしているため
- z_1 に依存しない項は、全て定数項 (正規化定数) に含まれてしまうため

- 従って $q_1^*(z_1)$ は

$$\begin{aligned} \ln q_1^*(z_1) &= \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} && (87) \\ &= \mathbb{E}_{z_2}\left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\right] + \text{Const.} && (88) \end{aligned}$$

但し

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ = & \ln \left(\frac{1}{(2\pi)^{\frac{2}{2}}} \frac{1}{|\boldsymbol{\Lambda}^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \left(\boldsymbol{\Lambda}^{-1}\right)^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) \right) \right) \\ = & \ln \left(\frac{1}{2\pi} \frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{z}-\boldsymbol{\mu}) \right) \right) \\ = & -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{z}-\boldsymbol{\mu}) + \text{Const.} \qquad (89) \end{aligned}$$

z_1 に依存する項だけを取り出せば

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) \\
= \quad & -\frac{1}{2}\left[z_1-\mu_1, z_2-\mu_2\right]\begin{bmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{bmatrix}\begin{bmatrix} z_1-\mu_1 \\ z_2-\mu_2\end{bmatrix} \\
= \quad & -\frac{1}{2}\Big[\Big\{(\Lambda_{11}(z_1-\mu_1)+\Lambda_{21}(z_2-\mu_2)\Big\}(z_1-\mu_1)+ \\
& \qquad \Big\{\Lambda_{12}(z_1-\mu_1)+\Lambda_{22}(z_2-\mu_2)\Big\}(z_2-\mu_2)\Big] \\
= \quad & -\frac{1}{2}\left(\Lambda_{11}(z_1-\mu_1)^2+2\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\right. \\
& \qquad \left.\Lambda_{22}(z_2-\mu_2)^2\right) \quad (\because \Lambda_{21}=\Lambda_{12}) \\
= \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2-\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\text{Const.} \quad (90)
\end{aligned}
$$

これを代入して

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ = \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 - \Lambda_{12}(z_1-\mu_1)(z_2-\mu_2) + \text{Const.} \end{aligned} \tag{91}$$

従って

$$\begin{aligned} & \ln q_1^*(z_1) \\ = \quad & \mathbb{E}_{z_2}\left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\right] + \text{Const.} \\ = \quad & \mathbb{E}_{z_2}\left[-\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 - \Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)\right] + \text{Const.} \\ = \quad & -\frac{1}{2}\Lambda_{11}z_1^2 + \Lambda_{11}\mu_1 z_1 - \Lambda_{12}z_1\left(\mathbb{E}[z_2]-\mu_2\right) + \text{Const.} \end{aligned} \tag{92}$$

変分推論

- これより、$q_1^*(z_1)$ は次のように書ける

$$
\begin{aligned}
& q_1^*(z_1) \\
\propto \quad & \exp\left(-\frac{1}{2}\Lambda_{11}z_1^2 + \Lambda_{11}\mu_1 z_1 - \Lambda_{12}z_1\left(\mathbb{E}[z_2] - \mu_2\right)\right) \\
= \quad & \exp\left(-\frac{1}{2}\Lambda_{11}\left(z_1 - (\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2))\right)^2 + \right. \\
& \left. \frac{1}{2}\Lambda_{11}\left(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2)\right)^2\right) \\
\propto \quad & \exp\left(-\frac{1}{2\Lambda_{11}^{-1}}\left(z_1 - (\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2))\right)^2\right) \\
= \quad & \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) \qquad (93)
\end{aligned}
$$

但し m_1 は次のようにおいた

$$
m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) \qquad (94)
$$

変分推論

- 対称性から、$q_2^*(z_2)$ も次のように求められる

$$\begin{aligned} \ln q_2^*(z_2) &= \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} && (95) \\ &= \mathbb{E}_{z_2}[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})] + \text{Const.} && (96) \end{aligned}$$

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) \quad (97)$$

但し m_2 は次のようにおいた

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) \quad (98)$$

- これより $q_1^*(z_1), q_2^*(z_2)$ は

$$\begin{aligned} q_1^*(z_1) &= \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) && (99) \\ m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) && (100) \\ q_2^*(z_2) &= \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) && (101) \\ m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) && (102) \end{aligned}$$

変分推論

- $\mathbb{E}[z_2]$ は、z_2 の $q_2(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1})$ による平均であるから、$\mathbb{E}[z_2] = m_2$ である (同様に、$\mathbb{E}[z_1] = z_1$)

- これより m_1, m_2 を連立させれば

$$
\begin{aligned}
m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) && (103)\\
m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) && (104)
\end{aligned}
$$

  であるから、m_1 について解けば

$$
\begin{aligned}
m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) && (105)\\
&= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) - \mu_2\right) && (106)\\
&= \mu_1 + \Lambda_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) && (107)\\
&= \mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}(m_1 - \mu_1) && (108)\\
&= \left(1 - \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}m_1 && (109)
\end{aligned}
$$

従って

$$\left(1 - \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1 = \left(1 - \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)m_1 \tag{110}$$

精度行列 $\mathbf{\Lambda}$ が正則であれば

$$\begin{aligned}
& \Lambda_{11}\Lambda_{12} - \Lambda_{12}\Lambda_{21} \neq 0 && (111)\\
\Rightarrow \quad & \Lambda_{11}\Lambda_{12} - \Lambda_{12}^{2} \neq 0 && (112)\\
\Rightarrow \quad & \left(\Lambda_{11}\Lambda_{12}\right)\left(1 - \Lambda_{11}^{-1}\Lambda_{12}^{-1}\Lambda_{12}^{2}\right) \neq 0 && (113)\\
\Rightarrow \quad & \left(1 - \Lambda_{11}^{-1}\Lambda_{12}^{-1}\Lambda_{12}^{2}\right) \neq 0 && (114)
\end{aligned}$$

が成立するから、分布 $p(\boldsymbol{z})$ が非特異 (精度行列が正則) ならば

$$m_1 = \mu_1 \tag{115}$$

が唯一の解であるほか、対称性から、m_2 についても以下を得る

$$m_2 = \mu_2 \tag{116}$$

変分推論

- ゆえに、$q_1^*(z_1), q_2^*(z_2)$ は

$$q_1^*(z_1) = \mathcal{N}(z_1|\mu_1, \Lambda_{11}^{-1}) \quad (117)$$
$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, \Lambda_{22}^{-1}) \quad (118)$$

- 注意点 1
  - $q_1^*(z_1)$ は、$q_2^*(z_2)$ を使って計算される $p(\boldsymbol{z})$ の期待値 $\mathbb{E}[z_2]$ に依存 (逆も成り立つ)
  - $q_1^*(z_1)$ と、$q_2^*(z_2)$ は相互に依存しているため、2 つを同時に求めることはできない
  - その代わりに、次のように最適化すればよい
  - $q_1(z_1), q_2(z_2)$ を適当に初期化したあと、$q_1^*(z_1), q_2^*(z_2)$ の式を使って、交互に $q_1(z_1)$ と $q_2(z_2)$ を更新していく (収束するまでこれを繰り返す)
- 注意点 2

変分推論

- $q_1(z_1), q_2(z_2)$ の具体的な関数形については、何の仮定も置かなかった
- $q_i^*(z_i)$ がガウス分布だという仮定は置いていないが、$\mathrm{KL}(q||p)$ を最適化する変分推論によって、結果的にガウス分布が得られた

変分推論

- $\mathrm{KL}(q||p)$ の最適化と $\mathrm{KL}(p||q)$ の最適化の比較
  - 上記の結果は、$\mathrm{KL}(q||p)$ の最適化 (エビデンス下界 $\mathcal{L}(q)$ の最適化) によって得た
  - $\mathrm{KL}(q||p)$ ではなく、$\mathrm{KL}(p||q)$ を最適化したらどうなるか?
  - 変分推論ではない、もう一つの近似推論の方法である、EP 法で使われる考え方

  - $q(\boldsymbol{Z})$ を $p(\boldsymbol{Z}|\boldsymbol{X})$ に近づけたいのであれば、$\mathrm{KL}(q||p)$ と $\mathrm{KL}(p||q)$ のどちらを最小化してもよいはず
  - なぜなら、KL ダイバージェンスは、確率分布間の (擬似的な) 距離を表すため

- $\mathrm{KL}(p||q)$ の最適化
  - $q(\boldsymbol{Z})$ が平均場近似によって分解できるとき、$\mathrm{KL}(p||q)$ を最適化したい

変分推論

- KL ダイバージェンス $\mathrm{KL}(p||q)$ は、次のように書ける

$$
\begin{aligned}
\mathrm{KL}(p||q) &\equiv \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X})||q(\boldsymbol{Z})) && (119)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} && (120)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln q(\boldsymbol{Z}) - \ln p(\boldsymbol{Z}|\boldsymbol{X})\right) d\boldsymbol{Z} && (121)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} - \\
&\qquad -\underbrace{\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z}}_{q \text{ には依存しない定数項}} && (122)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \prod_i q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \mathrm{Const.} && (123)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \sum_i \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \mathrm{Const.} && (124)
\end{aligned}
$$

$$= \quad -\sum_i \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \text{Const.} \qquad (125)$$

定数項は、$p(\boldsymbol{Z}|\boldsymbol{X})$ のエントロピーであり、q には依存しない

- 各因子 $q_j(\boldsymbol{Z}_j)$ について $\mathrm{KL}(p||q)$ を最適化したい
- このとき、$i \neq j$ となる、全ての $q_i(\boldsymbol{Z}_i)$ は固定する

- $q_j(\boldsymbol{Z}_j)$ に依存する項を取り出せば、次のようになる

$$\sum_i \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}$$

$$= \quad \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z} \qquad (126)$$

$$= \quad \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \qquad (127)$$

$$= \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \left(\prod_{i \neq j} d\boldsymbol{Z}_i \right) \tag{128}$$

$$= \int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \tag{129}$$

- $q_j(\boldsymbol{Z}_j)$ は確率分布であるから、以下の条件を満たさなければならない

$$\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j = 1 \quad (\text{規格化条件}) \tag{130}$$

$$q_j(\boldsymbol{Z}_j) \geq 0 \tag{131}$$

- 従って $\mathrm{KL}(p||q)$ を $q_j(\boldsymbol{Z}_j)$ について最適化するとき、ラグランジュの未定乗数法を使って、規格化条件を組み込む必要がある

変分推論

- $q_j(\boldsymbol{Z}_j) \geq 0$ という条件は、$\ln q_i(\boldsymbol{Z}_i)$ という項が既にあるから、何もしなくても常に満たされる (ラグランジュ関数に、制約条件を改めて取り入れる必要がない)

- 結局、ラグランジュ汎関数 $\mathcal{L}[q_j]$ は、次のようになる

$$\begin{aligned} \mathcal{L}[q_j] &= -\int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \\ & \lambda \left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j - 1 \right) \end{aligned} \tag{132}$$

- 上記は $q_j(\boldsymbol{Z}_j)$ についての汎関数となっていることに注意

- 次の公式を使って、$\mathcal{L}[q_j]$ を変分最適化する

$$\frac{\delta}{\delta y(x)} \int G(y(x), x) dx = \frac{\partial}{\partial y} G(y(x), x) \tag{133}$$

- 従って

$$
\begin{aligned}
& \frac{\delta}{\delta q_j(\boldsymbol{Z}_j)} \mathcal{L}[q_j] \\
= & -\frac{\delta}{\delta q_j(\boldsymbol{Z}_j)} \int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \\
& \frac{\delta}{\delta q_j(\boldsymbol{Z}_j)} \lambda \left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j - 1 \right) \qquad (134) \\
= & -\frac{\partial}{\partial q_j} \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) + \\
& \lambda \frac{\partial}{\partial q_j} q_j(\boldsymbol{Z}_j) \qquad (135) \\
= & -\left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \frac{1}{q_j(\boldsymbol{Z}_j)} + \lambda = 0 \qquad (136)
\end{aligned}
$$

変分推論

- これより、未定乗数 λ は

$$\lambda q_j(\boldsymbol{Z}_j) = \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \frac{1}{q_j(\boldsymbol{Z}_j)} \tag{137}$$

$$\Rightarrow \quad \int \lambda q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j = \int \underbrace{\left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right)}_{=p(\boldsymbol{Z}_j|\boldsymbol{X})} d\boldsymbol{Z}_j \tag{138}$$

$$\Rightarrow \quad \lambda \underbrace{\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j}_{=1} = \underbrace{\int p(\boldsymbol{Z}_j|\boldsymbol{X}) d\boldsymbol{Z}_j}_{=1} \tag{139}$$

$$\Rightarrow \quad \lambda = 1 \tag{140}$$

変分推論

- 結局、最適解 $q_j^*(\boldsymbol{Z}_j)$ は次のようになる

$$-\left(\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i\right)\frac{1}{q_j(\boldsymbol{Z}_j)}+\lambda=0 \tag{141}$$

$$\Rightarrow \quad q_j^*(\boldsymbol{Z}_j)=\underbrace{\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i}_{=p(\boldsymbol{Z}_j|\boldsymbol{X})} \tag{142}$$

$$\Rightarrow \quad q_j^*(\boldsymbol{Z}_j)=p(\boldsymbol{Z}_j|\boldsymbol{X}) \tag{143}$$

- $q_j^(\boldsymbol{Z}_j)$ の最適解は、$p(\boldsymbol{Z}|\boldsymbol{X})$ を、$i\neq j$ である全ての $\boldsymbol{Z}_i$ について周辺化した分布
- これは閉じた解であり、繰り返しを必要としない

- 最適な因子 $q_1(z_1), q_2(z_2)$ の計算

変分推論

- 今回は $p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z})$ の場合を考えており、かつ $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ であった

- 従って、$q_1^*(z_1)$ は、$p(\boldsymbol{z})$ を z_2 について周辺化すればよいから

$$
\begin{aligned}
& q_1^*(z_1) \\
= & \int p(\boldsymbol{z}) dz_2 \\
= & \int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) dz_2 && (144) \\
= & \frac{1}{2\pi} \frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}} \int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right) dz_2 && (145)
\end{aligned}
$$

変分推論

- ここで、指数の内側を、積分変数 z_2 に依存する項と、そうでない項に分ける

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) \\
= \quad & -\frac{1}{2}\left(\Lambda_{11}(z_1-\mu_1)^2+\right. \\
& \qquad \left. 2\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\Lambda_{22}(z_2-\mu_2)^2\right) \qquad (146) \\
= \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2+\Lambda_{12}(z_1-\mu_1)\mu_2+ \\
& \qquad -\frac{1}{2}\left(2\Lambda_{12}(z_1-\mu_1)z_2+\Lambda_{22}(z_2-\mu_2)^2\right) \qquad (147)
\end{aligned}
$$

そして

$$
\begin{aligned}
& -\frac{1}{2}\left(2\Lambda_{12}(z_1-\mu_1)z_2+\Lambda_{22}(z_2-\mu_2)^2\right) \\
= \quad & -\frac{1}{2}\left(\Lambda_{22}z_2^2-2\Lambda_{22}\mu_2 z_2+2\Lambda_{12}(z_1-\mu_1)z_2+\Lambda_{22}\mu_2^2\right) \qquad (148)
\end{aligned}
$$

$$= -\frac{1}{2}\left(\Lambda_{22}z_2^2 - 2\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)z_2 + \Lambda_{22}\mu_2^2\right) \quad (149)$$

$$= -\frac{1}{2}\Big(\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)\right)^2 - \Lambda_{22}\left(\Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)\right)^2 + \Lambda_{22}\mu_2^2\Big) \quad (150)$$

$$= -\frac{1}{2}\left(\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2 - \Lambda_{22}^{-1}m^2 + \Lambda_{22}\mu_2^2\right) \quad (151)$$

ゆえ ($m = \Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)$ とおいた)

$$-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})$$

$$= -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 - \frac{1}{2}\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2 + \frac{1}{2}\Lambda_{22}^{-1}m^2 \quad (152)$$

- これより、積分変数 z_2 の依存項だけを取り出せたので

$$\begin{aligned}
&\int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)dz_2 \\
= \quad &\exp\left(-\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2+\Lambda_{12}(z_1-\mu_1)\mu_2-\right. \\
&\left.\frac{1}{2}\Lambda_{22}\mu_2^2+\frac{1}{2}\Lambda_{22}^{-1}m^2\right) \\
&\int \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2-\Lambda_{22}^{-1}m\right)^2\right)dz_2 \qquad (153)
\end{aligned}$$

であって、右側の積分は、中身が (正規化されていない) ガウス分布であるから

$$\int \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2-\Lambda_{22}^{-1}m\right)^2\right)dz_2$$

$$
\begin{aligned}
&= (2\pi\Lambda_{22}^{-1})^{\frac{1}{2}} \cdot \int \frac{1}{(2\pi\Lambda_{22}^{-1})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2\right) dz_2 \\
&= \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}} \qquad (154)
\end{aligned}
$$

となって、z_2 を積分により消去できる

- また指数の残りの部分から、z_1 に依存する項だけを取り出して

$$
\begin{aligned}
& -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 + \frac{1}{2}\Lambda_{22}^{-1}m^2 \\
= & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 + \\
& \quad \frac{1}{2}\Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)^2 \\
= & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 + \\
& \quad \frac{1}{2}\Lambda_{22}\mu_2^2 - \mu_2\Lambda_{12}(z_1-\mu_1) +
\end{aligned}
$$

$$\frac{1}{2}\Lambda_{22}^{-1}\Lambda_{12}^{2}(z_1-\mu_1)^2 \tag{155}$$

$$= -\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1-\mu_1)^2 \tag{156}$$

$$= -\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)z_1^2 + \frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1 z_1 + \text{Const.} \tag{157}$$

- これより結局、z_2 による積分は次のようになる

$$\int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)dz_2$$

$$= \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1-\mu_1)^2\right) \tag{158}$$

変分推論

- 従って、$q_1^*(z_1)$ は次のようになる

$$
\begin{aligned}
& q_1^*(z_1) \\
= & \frac{1}{2\pi} \frac{1}{|\mathbf{\Lambda}|^{-\frac{1}{2}}} \int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \mathbf{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right) dz_2 && (159) \\
= & \frac{1}{2\pi} \frac{1}{\left(\Lambda_{11}\Lambda_{22}-\Lambda_{12}^2\right)^{-\frac{1}{2}}} \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}} \\
& \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) && (160) \\
= & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)^{-\frac{1}{2}}} \\
& \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) && (161) \\
= & \mathcal{N}(z_1|\mu_1, (\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2)^{-1}) && (162)
\end{aligned}
$$

変分推論

- 共分散行列 $\boldsymbol{\Sigma}$ を、精度行列 $\boldsymbol{\Lambda}$ を使って次のように定めれば

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \quad (\Sigma_{12} = \Sigma_{21}) \tag{163}$$

次が成り立つから

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \boldsymbol{I} \tag{164}$$

各成分に注目すれば

$$\begin{cases} \Sigma_{11}\Lambda_{11} + \Sigma_{12}\Lambda_{21} = 1 \\ \Sigma_{11}\Lambda_{12} + \Sigma_{12}\Lambda_{22} = 0 \end{cases} \tag{165}$$

これを Σ_{11} について解けば

$$\Sigma_{11}\Lambda_{11} + \left(-\Lambda_{22}^{-1}\Lambda_{12}\Sigma_{11}\right)\Lambda_{21} = 1 \tag{166}$$

$$\Rightarrow \quad \Sigma_{11}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}\Lambda_{21}\right) = 1 \tag{167}$$

$$\Rightarrow \quad \Sigma_{11}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right) = 1 \tag{168}$$

$$\Rightarrow \quad \Sigma_{11} = \left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right)^{-1} \tag{169}$$

- これから、$q_1^*(z_1)$ は次のようにも書ける

$$\begin{aligned} & q_1^*(z_1) \\ = \quad & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2})^{-\frac{1}{2}}} \\ & \exp\left(-\frac{1}{2}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1 - \mu_1)^2\right) \qquad (170) \\ = \quad & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\Sigma_{11}^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Sigma_{11}^{-1}(z_1 - \mu_1)^2\right) \qquad (171) \\ = \quad & \mathcal{N}(z_1|\mu_1, \Sigma_{11}) \qquad (172) \end{aligned}$$

変分推論

- $q_2^*(z_2)$ は、対称性から次のようになる

$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, (\Lambda_{22} - \Lambda_{11}^{-1}\Lambda_{12}^2)^{-1}) = \mathcal{N}(z_2|\mu_2, \Sigma_{22}) \tag{173}$$

- 2 つの解の比較
  - $\mathrm{KL}(q||p)$ の最小化によって次の解を得た

$$q_1^*(z_1) = \mathcal{N}(z_1|\mu_1, \Lambda_{11}^{-1}) \tag{174}$$
$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, \Lambda_{22}^{-1}) \tag{175}$$

変分推論

- $\mathrm{KL}(p||q)$ の最小化では、次の解を得た

$$
\begin{aligned}
q_1^*(z_1) &= \mathcal{N}(z_1|\mu_1, \Sigma_{11}) && (176)\\
\Sigma_{11} &= \left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^2\right)^{-1} && (177)\\
q_2^*(z_2) &= \mathcal{N}(z_2|\mu_2, \Sigma_{22}) && (178)\\
\Sigma_{22} &= (\Lambda_{22} - \Lambda_{11}^{-1}\Lambda_{12}^2)^{-1} && (179)
\end{aligned}
$$

- $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の平均は $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$ であったので、いずれの場合も、平均は正しく捉えている
- しかし、両者の間では、分散が異なっている
- また、変数 z_1 と z_2 の間の相関は消えてなくなっている

- 両者の違いを、次の図 3 に示す

- 緑色の線が真の分布 $p(\boldsymbol{z})$ を表す
- 左側の赤線は、$\mathrm{KL}(q||p)$ の最小化によって得られた分布 $q(\boldsymbol{z})$
- 右側の赤線は、$\mathrm{KL}(p||q)$ の最小化によって得られた分布 $q(\boldsymbol{z})$

Figure 10.2 Comparison of the two alternative forms for the Kullback-Leibler divergence. The green contours corresponding to 1, 2, and 3 standard deviations for a correlated Gaussian distribution $p(\mathbf{z})$ over two variables z_1 and z_2, and the red contours represent the corresponding levels for an approximating distribution $q(\mathbf{z})$ over the same variables given by the product of two independent univariate Gaussian distributions whose parameters are obtained by minimization of (a) the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$, and (b) the reverse Kullback-Leibler divergence $\mathrm{KL}(p\|q)$.

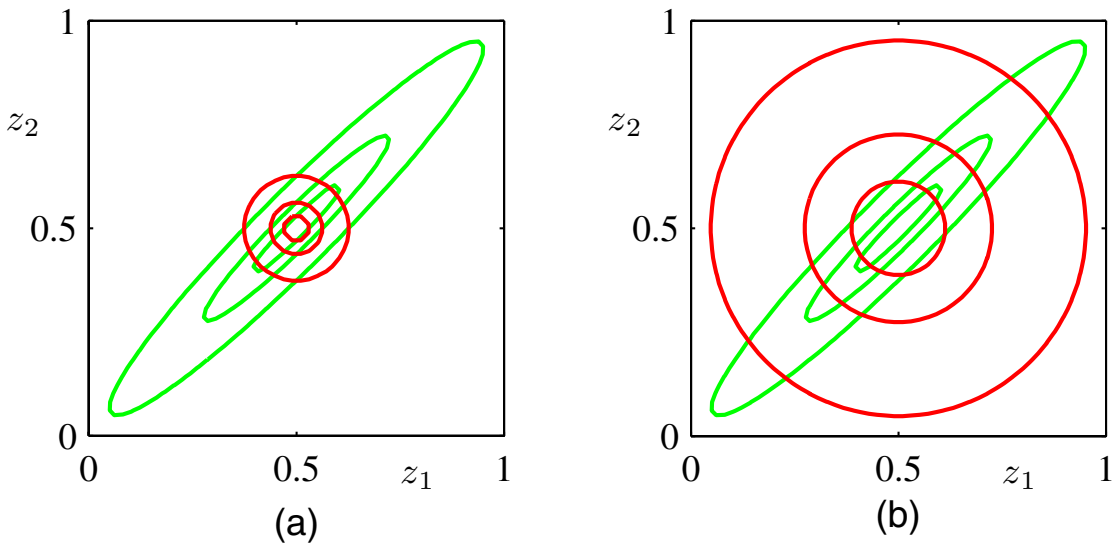

図 3: KL ダイバージェンスの 2 つの形の比較

変分推論

- 2 つの解の比較
  - $\mathrm{KL}(q||p)$ の最小化で得られる $q(\boldsymbol{z})$ は、分散が小さくなる方向に制御されていて、それと直交する方向の分散は、大きく過小評価されている
  - 分解による近似では、一般に事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ をコンパクトに近似しすぎる

  - $\mathrm{KL}(p||q)$ の最小化で得られる $q(\boldsymbol{z})$ は、分散が大きくなる方向に制御されていて、それと直交する方向の分散は、過大評価されている
  - 非常に低い確率しか持たないはずの領域にも、多くの確率質量が割り当てられている

  - 変分推論では、計算コストの観点から $\mathrm{KL}(q||p)$ の方を用いる
  - $\mathrm{KL}(q||p)$ の計算には、q に関する期待値の評価が含まれるので、q を単純な形に制限することで、必要な期待値を単純化できる
  - $\mathrm{KL}(p||q)$ の計算には、真の事後分布 p に関する期待値の計算が必要である

変分推論

- 違いが生じる理由
  - KL ダイバージェンス $\mathrm{KL}(q||p)$ は次のようであった

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{180}$$

  - $\mathrm{KL}(q||p)$ が大きくなる主要因は、$p(\boldsymbol{Z})$ がほとんど 0 で、$q(\boldsymbol{Z})$ はそうでない領域
  - 従って、$\mathrm{KL}(q||p)$ を最小化すると、$q(\boldsymbol{Z})$ は、$p(\boldsymbol{Z})$ が小さい領域を避けるようになる

  - また、$\mathrm{KL}(p||q)$ は次のようであった

$$\mathrm{KL}(p||q) = -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} \tag{181}$$

変分推論

- $\mathrm{KL}(p||q)$ が大きくなる主要因は、$q(\boldsymbol{Z})$ がほとんど 0 で、$p(\boldsymbol{Z})$ はそうではない領域
- 従って、$\mathrm{KL}(p||q)$ を最小化すると、$q(\boldsymbol{Z})$ は、$p(\boldsymbol{Z})$ が 0 でない領域にも、必ず確率を持たせるようになる

- 別の分布について、この両者の振舞いの違いを観察してみよう

変分推論

- 多峰性のある分布を、単峰の分布で近似する場合
  - $\mathrm{KL}(q||p)$ を最小化する変分近似では、多数ある峰のうちの 1 つを再現
  - $\mathrm{KL}(p||q)$ を最小化する変分近似では、全ての峰を平均したような分布が得られる
  - 多峰性のある分布を平均してしまうと、予測性能の悪化をもたらす
  - これらの比較を次の図 4 に示す

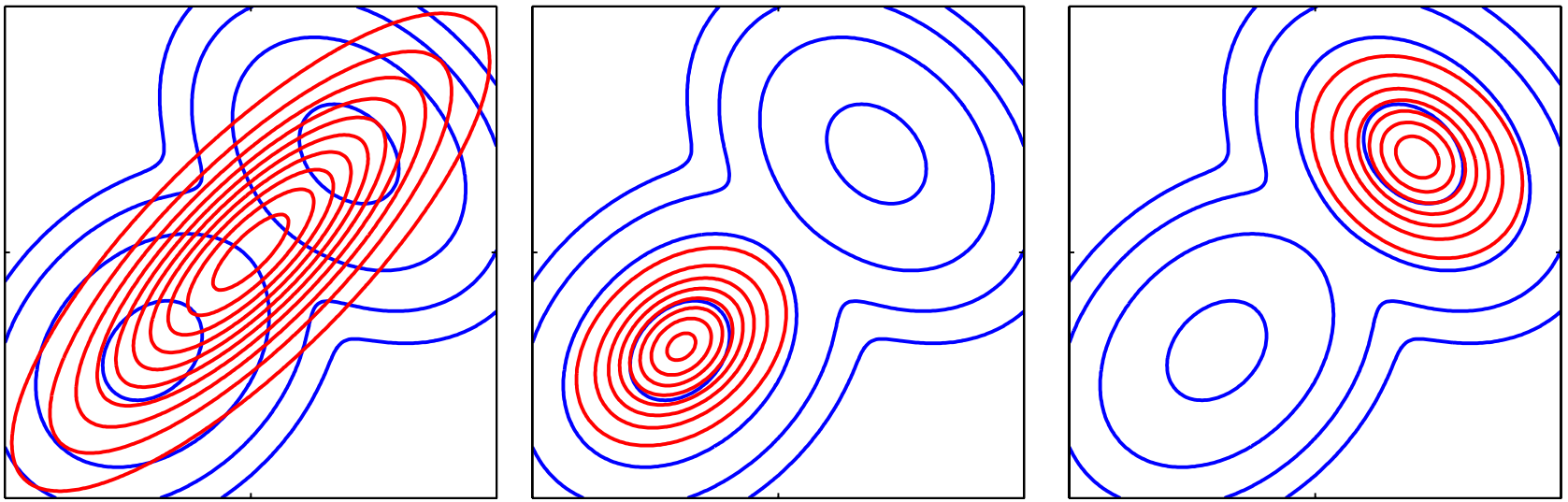

Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

図 4: KL ダイバージェンスの 2 つの形の別の比較

変分推論

- ここまでの話の流れ
  1. 分解 $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ を使ったエビデンス下界 $\mathcal{L}(q)$ の最適化は、$\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}))$ の最小化と等価である
  2. $\mathrm{KL}(q||p)$ と、$\mathrm{KL}(p||q)$ を最小化する変分近似を、2 変数のガウス分布を例として試した
  3. $\mathrm{KL}(q||p)$ の最小化を使って求めた $q(\boldsymbol{Z})$ は、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ をコンパクトに近似する傾向にあった
  4. $\mathrm{KL}(p||q)$ の最小化によって求めた $q(\boldsymbol{Z})$ は、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を大きく捉えて近似する傾向にあった
- これからの話の流れ
  - 変分推論の具体的な例について更に見ていく

変分推論

- 離散潜在変数のモデル (二値スパース符号化モデル) に変分推論を適用する
- 連続潜在変数の場合は、簡単な確率モデルを使って、変分推論を試してみる

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

離散潜在変数の変分推論

- 離散潜在変数をもつ変分推論の概要
  - ここでは単純な場合を扱う
  - データ $\boldsymbol{x}_i$ に対する潜在変数 $\boldsymbol{z}_i$ は、各要素が二値であるとする
  - データ $\boldsymbol{x}$ は D 次元、潜在変数 $\boldsymbol{z}$ は K 次元とする

  - 分布 $q(\boldsymbol{z}_i|\boldsymbol{x}_i)$ は、平均場近似によって、次のように分解できるとする

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_k q(z_{ik}|\boldsymbol{x}_i) \tag{182}$$

  - 潜在変数は二値であるから、$q(z_{ik} = 1|\boldsymbol{x}_i) = \widehat{z_{ik}}$ と書くことにする
  - このとき、$q(z_{ik}|\boldsymbol{x}_i)$ は次のようになる

$$q(z_{ik}|\boldsymbol{x}_i) = \widehat{z_{ik}}^{z_{ik}} \left(1 - \widehat{z_{ik}}\right)^{(1-z_{ik})} \tag{183}$$

  - 各パラメータ $\widehat{z_{ik}}$ について、下界 $\mathcal{L}(q)$ を順番に最適化することを、$\mathcal{L}(q)$ が収束するまで繰り返し行う

離散潜在変数の変分推論

- 即ち、以下の不動点方程式を、各 $\widehat{z_{ik}}$ について繰り返し解く

$$\frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) = 0 \tag{184}$$

- 離散潜在変数の場合は、単なる標準的な最適化問題を解くことになる

- 二値スパース符号化モデル
  - 二値スパース符号化モデルでのデータ生成過程は、次のようになる
  - 各データ $\boldsymbol{x}_i$ には、潜在変数 $\boldsymbol{z}_i \in \{0,1\}^K$ が対応する
  - $\boldsymbol{z}_i$ に対し、重み $\boldsymbol{W}$ を用いて線形変換を施し、更にガウスノイズを足し合わせることで、データ $\boldsymbol{x}_i$ が生成される
  - 潜在変数 $\boldsymbol{z}_i$ は、K 次元ベクトルであり、その各要素は 0 または 1 である

離散潜在変数の変分推論

- 従って、確率分布は次のようになる

$$p(z_{ik}=1) = \sigma(b_{ik}) \tag{185}$$
$$p(\boldsymbol{x}_i|\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \tag{186}$$

- $\sigma(\cdot)$ はシグモイド関数、$\boldsymbol{b}_i = [b_{i1}, \ldots, b_{iK}]^T$ は学習可能なバイアス、$\boldsymbol{W}$ は学習可能な重み行列、$\boldsymbol{\beta}$ は学習可能な対角精度行列である
- $\boldsymbol{\beta} = \mathrm{diag}\left(\beta_1, \beta_2, \ldots, \beta_D\right)$ と書ける

- $p(\boldsymbol{z}_i)$ は次のように計算できる

$$p(\boldsymbol{z}_i) = \prod_k p(z_{ik}) \tag{187}$$
$$= \prod_k \left(\sigma(b_{ik})\right)^{z_{ik}} \left(1-\sigma(b_{ik})\right)^{1-z_{ik}} \tag{188}$$
$$= \prod_k \left(\sigma(b_{ik})\right)^{z_{ik}} \left(\sigma(-b_{ik})\right)^{1-z_{ik}} \tag{189}$$

離散潜在変数の変分推論

- 事後分布 $p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ は複雑であり、表現することも、計算することもできない
- 従って、最尤推定の手法 (EM アルゴリズム) により学習することはできない

- 例えばバイアス b_{ik} に関する微分を考えると

$$\begin{aligned} & \frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{z}_i|\boldsymbol{x}_i) \\ = & \frac{\partial}{\partial b_{ik}} \ln \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i)}{p(\boldsymbol{x}_i)} \qquad (190) \\ = & \frac{\partial}{\partial b_{ik}} \left(\ln p(\boldsymbol{x}_i, \boldsymbol{z}_i) - \ln p(\boldsymbol{x}_i)\right) \qquad (191) \end{aligned}$$

であって、第 1 項の $p(\boldsymbol{x}_i, \boldsymbol{z}_i)$ は容易に計算できるが、第 2 項について考えると

$$\frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{x}_i)$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \frac{\partial}{\partial b_{ik}} p(\boldsymbol{x}_i) \tag{192}$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \frac{\partial}{\partial b_{ik}} \sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i, \boldsymbol{z}_i) \tag{193}$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \frac{\partial}{\partial b_{ik}} \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i) p(\boldsymbol{x}_i|\boldsymbol{z}_i) \tag{194}$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i|\boldsymbol{z}_i) \frac{\partial}{\partial b_{ik}} p(\boldsymbol{z}_i) \tag{195}$$

$$= \sum_{\boldsymbol{z}_i} \frac{1}{p(\boldsymbol{x}_i)} \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i)}{p(\boldsymbol{z}_i)} \frac{\partial}{\partial b_{ik}} p(\boldsymbol{z}_i) \tag{196}$$

$$= \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i|\boldsymbol{x}_i) \frac{1}{p(\boldsymbol{z}_i)} \frac{\partial}{\partial b_{ik}} p(\boldsymbol{z}_i) \tag{197}$$

$$= \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i|\boldsymbol{x}_i) \frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{z}_i) \tag{198}$$

離散潜在変数の変分推論

$$= \quad \mathbb{E}_{\boldsymbol{z}_i \sim p(\boldsymbol{z}_i|\boldsymbol{x}_i)} \left[\frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{z}_i) \right] \tag{199}$$

何とか $p(\boldsymbol{z}_i)$ と $p(\boldsymbol{x}_i|\boldsymbol{z}_i)$ を使って計算できないかと試行錯誤したが、結局、$p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ に関する期待値の計算が必要になった

- $p(\boldsymbol{x}_i, \boldsymbol{z}_i)$ と、$p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ のグラフ構造を次の図 5 に示す
- 図では、$\boldsymbol{h} = \boldsymbol{z}$、$\boldsymbol{v} = \boldsymbol{x}$ のように読み替える必要がある

離散潜在変数の変分推論

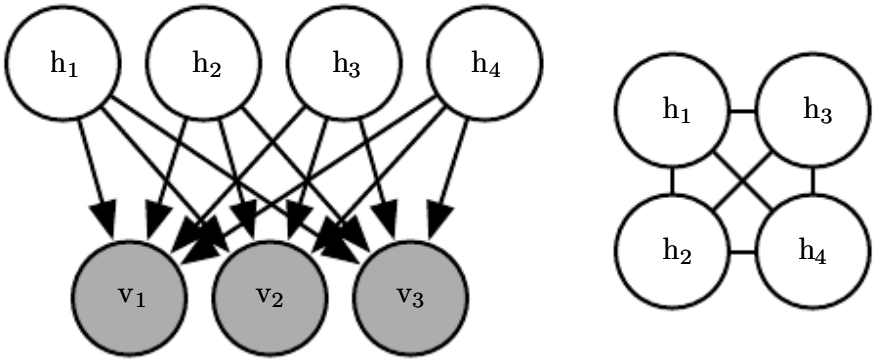

Figure 19.2: The graph structure of a binary sparse coding model with four hidden units. *(Left)*The graph structure of $p(\boldsymbol{h}, \boldsymbol{v})$. Note that the edges are directed, and that every two hidden units are co-parents of every visible unit. *(Right)*The graph structure of $p(\boldsymbol{h} \mid \boldsymbol{v})$. In order to account for the active paths between co-parents, the posterior distribution needs an edge between all of the hidden units.

図 5: 二値スパース符号化モデルのグラフ構造 (4 つの潜在変数をもつ場合)

離散潜在変数の変分推論

- 変分推論による二値スパース符号化モデルの学習
  - 最尤推定の手法を諦める代わりに、変分推論を使うことで、困難を解決できる
  - 分布 $p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ を $q(\boldsymbol{z}_i)$ で表現し、更に平均場近似を行う
  - 但し $\boldsymbol{z}_i = [z_{i1}, z_{i2}, \ldots, z_{iK}]^T$ とする

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_k q(z_{ik}|\boldsymbol{x}_i) \tag{200}$$

  - 潜在変数の各要素は二値であるから、各 $q(z_{ik}|\boldsymbol{x}_i)$ はベルヌーイ分布とすればよい
  - 即ち、$q(z_{ik} = 1|\boldsymbol{x}_i) = \widehat{z_{ik}}$ とする
  - $\widehat{z_{ik}} \neq 0, 1$ という制約を課すことで、$\ln \widehat{z_{ik}}$ を計算できる

  - このようにすれば、潜在変数 $z_{ik}, z_{il}(k \neq l)$ 間の相関を断ち切ることができる
  - 先程の図 5 の右側から、潜在変数間の辺を、全て消し去ることになる

離散潜在変数の変分推論

- これで、平均場近似により、因数分解可能な q を表現することができる

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_k q(z_{ik}|\boldsymbol{x}_i) \tag{201}$$

$$= \prod_k \widehat{z_{ik}}^{z_{ik}} \left(1 - \widehat{z_{ik}}\right)^{1-z_{ik}} \tag{202}$$

$$q(\boldsymbol{Z}|\boldsymbol{X}) = \prod_i q(\boldsymbol{z}_i|\boldsymbol{x}_i) \tag{203}$$

$$= \prod_i \prod_k \widehat{z_{ik}}^{z_{ik}} \left(1 - \widehat{z_{ik}}\right)^{1-z_{ik}} \tag{204}$$

- ソフトウェア上では、丸め誤差などによって $\widehat{z_{ik}}$ が 0 や 1 になり、計算を続行できなくなるかもしれない
- これを回避するためには、パラメータ $\widetilde{z}_i$ を使って二値スパース符号化モデルを学習させる

離散潜在変数の変分推論

- そして、$\widehat{\boldsymbol{z}}_i = \sigma(\widetilde{\boldsymbol{z}}_i)$ の関係によって、$\widehat{\boldsymbol{z}}_i = [\widehat{z_{i1}}, \ldots, \widehat{z_{iK}}]^T$ を得るようにする
- $\ln \widehat{z_{ik}} = \ln \sigma(\widetilde{z_{ik}}) = -\zeta(-\widetilde{z_{ik}})$ によって、コンピュータ上で安全に $\ln \widehat{z_{ik}}$ を計算できる ($\zeta(\cdot)$ はソフトプラス関数)

- 変分推論のために、まずはエビデンス下界 $\mathcal{L}(q)$ を計算する

$$
\begin{aligned}
& \mathcal{L}(q) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z}|\boldsymbol{X})} && (205) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{p(\boldsymbol{Z})p(\boldsymbol{X}|\boldsymbol{Z})}{q(\boldsymbol{Z}|\boldsymbol{X})} && (206) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) + \ln p(\boldsymbol{X}|\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) && (207) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) + \\
\end{aligned}
$$

離散潜在変数の変分推論

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \tag{208}$$

- $q(\boldsymbol{Z}|\boldsymbol{X})$ についての期待値を取っていることに注意する
- 全ての $\boldsymbol{Z}$ についての和を取ればよいので、第 1 項は次のようになる

$$\begin{aligned}
& \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) \\
= & \sum_{i} \sum_{k} \sum_{z_{ik}} q(z_{ik}|\boldsymbol{x}_i) \left(\ln p(z_{ik}) - \ln q(z_{ik}|\boldsymbol{x}_i)\right) && (209) \\
= & \sum_{i} \sum_{k} q(z_{ik}=1|\boldsymbol{x}_i) \left(\ln p(z_{ik}=1) - \ln q(z_{ik}=1|\boldsymbol{x}_i)\right) + \\
& \quad q(z_{ik}=0|\boldsymbol{x}_i) \left(\ln p(z_{ik}=0) - \ln q(z_{ik}=0|\boldsymbol{x}_i)\right) && (210) \\
= & \sum_{i} \sum_{k} \{\widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \\
& \quad (1-\widehat{z_{ik}}) \left(\ln\left(1-\sigma(b_{ik})\right) - \ln\left(1-\widehat{z_{ik}}\right)\right)\} && (211)
\end{aligned}$$

離散潜在変数の変分推論

$$= \sum_i \sum_k \{\widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln \left(1 - \widehat{z_{ik}}\right)\right)\} \tag{212}$$

- また第 2 項は、次のようになる

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z})$$

$$= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i|\boldsymbol{x}_i) \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i) \tag{213}$$

$$= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i|\boldsymbol{x}_i) \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \tag{214}$$

離散潜在変数の変分推論

- 但し、$\ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1})$ は次のように分解できる

$$
\begin{aligned}
& \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \\
= \quad & \ln \left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\beta}^{-1}|^{\frac{1}{2}}} \right. \\
& \qquad \left. \exp\left(-\frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right)^T \boldsymbol{\beta} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right) \right) \right) \qquad (215) \\
= \quad & -\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\beta}^{-1}| - \\
& \qquad \frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right)^T \boldsymbol{\beta} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right) \qquad (216)
\end{aligned}
$$

離散潜在変数の変分推論

- ここで、$\boldsymbol{\beta} = \mathrm{diag}\left(\beta_1, \beta_2, \ldots, \beta_D\right)$ であるので

$$\begin{aligned} \ln\left|\boldsymbol{\beta}^{-1}\right| &= \ln\left|\boldsymbol{\beta}\right|^{-1} = -\ln\left|\boldsymbol{\beta}\right| \\ &= -\ln\prod_{j=1}^{D}\beta_j = -\sum_{j=1}^{D}\ln\beta_j \end{aligned} \tag{217}$$

となるほか

$$\left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right)^T \boldsymbol{\beta}\left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right) = \sum_{j=1}^{D}\beta_j\left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2 \tag{218}$$

であるから

$$\begin{aligned} &\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \\ = &-\frac{1}{2}\sum_j \ln 2\pi + \frac{1}{2}\sum_j \ln\beta_j - \frac{1}{2}\sum_j \beta_j\left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2 \end{aligned} \tag{219}$$

離散潜在変数の変分推論

$$= \frac{1}{2}\sum_j \left(-\ln 2\pi + \ln \beta_j - \beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{220}$$

$$= \frac{1}{2}\sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{221}$$

$$= \frac{1}{2}\sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \sum_k W_{jk} z_{ik}\right)^2\right) \tag{222}$$

- 上式の対数を外すと、確率分布の積になっていることが分かる

$$\mathcal{N}(\boldsymbol{x}_i | \boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1})$$

$$= \exp\left(\sum_j \left(\frac{1}{2}\ln \frac{\beta_j}{2\pi} - \frac{1}{2}\beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right)\right) \tag{223}$$

$$= \prod_j \exp\left(\frac{1}{2}\ln \frac{\beta_j}{2\pi} - \frac{1}{2}\beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{224}$$

離散潜在変数の変分推論

$$= \prod_j \exp\left(\frac{1}{2}\ln\frac{\beta_j}{2\pi}\right)\exp\left(-\frac{1}{2}\beta_j\left(x_{ij}-\boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \quad (225)$$

$$= \prod_j \left(\frac{\beta_j}{2\pi}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\beta_j\left(x_{ij}-\boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \quad (226)$$

$$= \prod_j \left(\frac{1}{(2\pi\beta_j^{-1})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\beta_j\left(x_{ij}-\boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right)\right) \quad (227)$$

$$= \prod_j \mathcal{N}(x_{ij}|\boldsymbol{W}_{j:}\boldsymbol{z}_i,\beta_j^{-1}) \quad (228)$$

- 精度行列 $\boldsymbol{\beta}$ は対角行列であるから、$\boldsymbol{x}_i$ の各成分は無相関である
- 従って、$\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i,\boldsymbol{\beta}^{-1})$ は、各成分 x_{ij} についての確率 $\mathcal{N}(x_{ij}|\boldsymbol{W}_{j:}\boldsymbol{z}_i,\beta_j^{-1})$ の積として、記述できる

離散潜在変数の変分推論

- さて、第 2 項は

$$
\begin{aligned}
& \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \\
= & \sum_{i} \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i|\boldsymbol{x}_i) \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) && (229) \\
= & \sum_{i} \mathbb{E}_{\boldsymbol{z}_i \sim q(\boldsymbol{z}_i|\boldsymbol{x}_i)} \left[\ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \right] && (230) \\
= & \sum_{i} \mathbb{E}_{\boldsymbol{z}_i} \left[\frac{1}{2} \sum_{j} \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \sum_{k} W_{jk} z_{ik} \right)^2 \right) \right] && (231) \\
= & \frac{1}{2} \sum_{i} \sum_{j} \mathbb{E}_{\boldsymbol{z}_i} \left[\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \sum_{k} W_{jk} z_{ik} \right)^2 \right] && (232) \\
= & \frac{1}{2} \sum_{i} \sum_{j} \left(\ln \frac{\beta_j}{2\pi} - \beta_j \mathbb{E}_{\boldsymbol{z}_i} \left[\left(x_{ij} - \sum_{k} W_{jk} z_{ik} \right)^2 \right] \right) && (233)
\end{aligned}
$$

離散潜在変数の変分推論

ここで

$$\mathbb{E}_{\boldsymbol{z}_i}\left[\left(x_{ij}-\sum_k W_{jk}z_{ik}\right)^2\right]$$

$$= \mathbb{E}_{\boldsymbol{z}_i}\left[x_{ij}^2 - 2x_{ij}\sum_k W_{jk}z_{ik} + \left(\sum_k W_{jk}z_{ik}\right)^2\right] \quad (234)$$

$$= x_{ij}^2 - 2x_{ij}\mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k W_{jk}z_{ik}\right] + \mathbb{E}_{\boldsymbol{z}_i}\left[\left(\sum_k W_{jk}z_{ik}\right)^2\right] \quad (235)$$

各項を順番に計算すると

$$\mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k W_{jk}z_{ik}\right]$$

$$= \sum_k W_{jk} \mathbb{E}_{\boldsymbol{z}_i} [z_{ik}] \tag{236}$$

$$= \sum_k W_{jk} \mathbb{E}_{z_{ik}} [z_{ik}] \tag{237}$$

$$= \sum_k W_{jk} \left(q(z_{ik} = 1 | \boldsymbol{x}_i) \cdot 1 + q(z_{ik} = 0 | \boldsymbol{x}_i) \cdot 0 \right) \tag{238}$$

$$= \sum_k W_{jk} \widehat{z_{ik}} \tag{239}$$

また

$$\mathbb{E}_{\boldsymbol{z}_i} \left[\left(\sum_k W_{jk} z_{ik} \right)^2 \right]$$

$$= \mathbb{E}_{\boldsymbol{z}_i} \left[\sum_k \sum_l W_{jk} W_{jl} z_{ik} z_{il} \right] \tag{240}$$

離散潜在変数の変分推論

$$= \mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k \left(W_{jk}^2 z_{ik}^2 + \sum_{l\neq k}(W_{jk}W_{jl}z_{ik}z_{il})\right)\right] \tag{241}$$

$$= \sum_k \left(W_{jk}^2 \mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}^2\right] + \sum_{l\neq k} W_{jk}W_{jl}\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}z_{il}\right]\right) \tag{242}$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}^2\right] \\
=\ &\mathbb{E}_{z_{ik}}\left[z_{ik}^2\right] && (243)\\
=\ &q(z_{ik}=1|\boldsymbol{x}_i)\cdot 1^2 + q(z_{ik}=0|\boldsymbol{x}_i)\cdot 0^2 && (244)\\
=\ &\widehat{z_{ik}} && (245)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}z_{il}\right] \\
=\ &\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}\right]\mathbb{E}_{\boldsymbol{z}_i}\left[z_{il}\right] \quad (\because z_{ik} \text{ と } z_{il} \text{ は独立であるため}) && (246)
\end{aligned}$$

$$= \mathbb{E}_{z_{ik}}[z_{ik}]\mathbb{E}_{z_{il}}[z_{il}] \quad (247)$$

$$= \widehat{z_{ik}}\widehat{z_{il}} \quad (248)$$

従って

$$\mathbb{E}_{\boldsymbol{z}_i}\left[\left(x_{ij} - \sum_k W_{jk}z_{ik}\right)^2\right]$$

$$= x_{ij}^2 - 2x_{ij}\mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k W_{jk}z_{ik}\right] + \mathbb{E}_{\boldsymbol{z}_i}\left[\left(\sum_k W_{jk}z_{ik}\right)^2\right] \quad (249)$$

$$= x_{ij}^2 - 2x_{ij}\sum_k W_{jk}\widehat{z_{ik}} +$$

$$\sum_k\left(W_{jk}^2\widehat{z_{ik}} + \sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{ik}}\widehat{z_{il}}\right) \quad (250)$$

離散潜在変数の変分推論

これより

$$\begin{aligned}
& \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \\
= & \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \mathbb{E}_{\boldsymbol{z}_i} \left[\left(x_{ij} - \sum_k W_{jk} z_{ik} \right)^2 \right] \right) \quad (251) \\
= & \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2 x_{ij} \sum_k W_{jk} \widehat{z_{ik}} + \right. \right. \\
& \left. \left. \sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \quad (252)
\end{aligned}$$

離散潜在変数の変分推論

- よって、エビデンス下界 $\mathcal{L}(q)$ は

$$
\begin{aligned}
& \mathcal{L}(q) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) + \\
& \quad \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \qquad (253) \\
= & \sum_{i} \sum_{k} \{\widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \\
& \quad (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}})\right)\} + \\
& \quad \frac{1}{2} \sum_{i} \sum_{j} \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2 x_{ij} \sum_{k} W_{jk} \widehat{z_{ik}} + \right.\right. \\
& \quad \left.\left. \sum_{k} \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \qquad (254)
\end{aligned}
$$

離散潜在変数の変分推論

- この式にはあまり美的魅力がない (Somewhat unappealing aesthetically) が、事後分布よりは計算しやすい
- 事後分布 $\ln p(\boldsymbol{Z}|\boldsymbol{X})$ を最大化する代わりに、上記の下界 $\mathcal{L}(q)$ を q について最大化することができる

- データ $\boldsymbol{x}_i$ に対するパラメータ $\{\widehat{z_{i1}}, \ldots, \widehat{z_{iK}}\}$ をまとめて、ベクトル $\widehat{\boldsymbol{z}_i}$ として記述する
- パラメータ $\widehat{\boldsymbol{z}_i} = [\widehat{z_{i1}}, \ldots, \widehat{z_{iK}}]^T$ は、データ $\boldsymbol{x}_i$ に対応する潜在変数 $\boldsymbol{z}$ の各要素が、1 となる確率を集めたベクトルである

- $\widehat{z_{ik}} = q(z_{ik} = 1|\boldsymbol{x}_i)$ と定義されていることに注意

- よってベクトル $\widehat{\boldsymbol{z}_i}$ は、データ $\boldsymbol{x}_i$ に対応する二値のスパース符号である

- 勾配上昇法を利用しない理由
  - データ $\boldsymbol{X}$ と、潜在変数 $\boldsymbol{Z}$ についての勾配上昇法を用いれば、学習することが可能

離散潜在変数の変分推論

- 但し、その方法では、各 x_i について平均場パラメータ $\widehat{z_i}$ を保管する必要がある
- 各事例について、動的に更新されるベクトルが必要であるため、そのアルゴリズムを、数十億もの事例に対して適用することは困難である

- また、収束するまで繰り返し計算を行うため、データ x から、パラメータ $\widehat{z_i}$ を素早く抽出することができない
- 現実にデプロイされるときは、$\widehat{z_i}$ をリアルタイムで計算できなければならない

- 不動点方程式による平均場パラメータ $\widehat{z_i}$ の推定
  - 勾配上昇法の代わりに、不動点方程式を使ってパラメータ $\widehat{z_{ik}}$ を素早く推定できる

  - $\nabla_{\widehat{z_i}} \mathcal{L}(q) = 0$ をみたす、$\widehat{z_i}$ の極大値を見つけ出す

離散潜在変数の変分推論

- $\widehat{z}_i$ の全ての成分について同時に解くことはできないので、各成分 $\widehat{z_{ik}}$ について繰り返し解く
- 即ち、各パラメータ $\widehat{z_{ik}}$ について、下界 $\mathcal{L}(q)$ を順番に最適化する手続きを、$\mathcal{L}(q)$ の収束基準を満たすまで繰り返す

$$\frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) = 0 \tag{255}$$

- 平均場不動点方程式を導くためには、$\mathcal{L}(q)$ を $\widehat{z_{ik}}$ で微分する必要がある

$$\begin{aligned} & \frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) \\ = \quad & \frac{\partial}{\partial \widehat{z_{ik}}} \sum_i \sum_k \{ \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) + \\ & \qquad (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}}) \right) \} + \end{aligned}$$

離散潜在変数の変分推論

$$\frac{\partial}{\partial \widehat{z_{ik}}} \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} + \sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \tag{256}$$

- 前半部分は

$$\frac{\partial}{\partial \widehat{z_{ik}}} \sum_i \sum_k \{ \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) + (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}}) \right) \} \tag{257}$$

$$= \frac{\partial}{\partial \widehat{z_{ik}}} \{ \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) + (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}}) \right) \} \tag{258}$$

$$= \frac{\partial}{\partial \widehat{z_{ik}}} \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) +$$

$$\frac{\partial}{\partial \widehat{z_{ik}}} \left(1 - \widehat{z_{ik}}\right) \left(\ln \sigma(-b_{ik}) - \ln\left(1 - \widehat{z_{ik}}\right)\right) \tag{259}$$

$$\begin{aligned} = \quad & \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \widehat{z_{ik}} \left(-\frac{1}{\widehat{z_{ik}}}\right) + \\ & \left(-\left(\ln \sigma(-b_{ik}) - \ln\left(1 - \widehat{z_{ik}}\right)\right)\right) + \\ & \left(1 - \widehat{z_{ik}}\right) \frac{1}{1 - \widehat{z_{ik}}} \end{aligned} \tag{260}$$

$$= \quad \ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} - 1 - \ln \sigma(-b_{ik}) + \ln(1 - \widehat{z_{ik}}) + 1 \tag{261}$$

$$= \quad \ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} - \ln \sigma(-b_{ik}) + \ln(1 - \widehat{z_{ik}}) \tag{262}$$

$$= \quad -\zeta(-b_{ik}) - \ln \widehat{z_{ik}} + \zeta(b_{ik}) + \ln(1 - \widehat{z_{ik}}) \tag{263}$$

$$= \quad (\zeta(b_{ik}) - \zeta(-b_{ik})) - \ln \widehat{z_{ik}} + \ln(1 - \widehat{z_{ik}}) \tag{264}$$

$$= \quad b_{ik} - \ln \widehat{z_{ik}} + \ln(1 - \widehat{z_{ik}}) \tag{265}$$

離散潜在変数の変分推論

- ここで、シグモイド関数 $\sigma(\cdot)$ と、ソフトプラス関数 $\zeta(\cdot)$ に関する、以下の公式を用いた

$$\ln \sigma(x) = -\zeta(-x) \tag{266}$$

$$\zeta(x) - \zeta(-x) = x \tag{267}$$

- 後半部分は

$$\frac{\partial}{\partial \widehat{z_{ik}}} \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} + \sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \tag{268}$$

$$= \frac{1}{2} \sum_j \beta_j \frac{\partial}{\partial \widehat{z_{ik}}} \left(2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} -$$

$$\sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \Bigg) \tag{269}$$

$$= \frac{1}{2} \sum_j \beta_j \frac{\partial}{\partial \widehat{z_{ik}}} \Bigg(2 x_{ij} \sum_k W_{jk} \widehat{z_{ik}} -$$

$$\left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) -$$

$$\sum_{m \neq k} \left(W_{jm}^2 \widehat{z_{im}} + \sum_{l \neq m} W_{jm} W_{jl} \widehat{z_{im}} \widehat{z_{il}} \right) \Bigg) \tag{270}$$

$$= \frac{1}{2} \sum_j \beta_j \left(2 x_{ij} W_{jk} - W_{jk}^2 - \right.$$

$$\left. \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{il}} - \sum_{m \neq k} W_{jm} W_{jk} \widehat{z_{im}} \right) \tag{271}$$

$$= \frac{1}{2}\sum_j \beta_j \left(2x_{ij}W_{jk} - W_{jk}^2 - 2\sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{il}} \right) \tag{272}$$

$$= \sum_j \beta_j \left(x_{ij}W_{jk} - \frac{1}{2}W_{jk}^2 - \sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{il}} \right) \tag{273}$$

$$= \sum_j x_{ij}\beta_j W_{jk} - \frac{1}{2}\sum_j W_{jk}\beta_j W_{jk} - \sum_j \beta_j \sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{il}} \tag{274}$$

$$= \sum_j x_{ij}\beta_j W_{jk} - \frac{1}{2}\sum_j W_{jk}\beta_j W_{jk} - \sum_{l\neq k} \left(\sum_j W_{jl}\beta_j W_{jk} \right) \widehat{z_{il}} \tag{275}$$

$$= \quad \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}} \tag{276}$$

- これより、$\mathcal{L}(q)$ の $\widehat{z_{ik}}$ による微分は次のようになる

$$\begin{aligned} & \frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) \\ = \quad & b_{ik} - \ln \widehat{z_{ik}} + \ln(1 - \widehat{z_{ik}}) + \\ & \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}} \end{aligned} \tag{277}$$

離散潜在変数の変分推論

- これを 0 と等置して、$\widehat{z_{ik}}$ について解くと次のようになる

$$\ln \widehat{z_{ik}} - \ln(1 - \widehat{z_{ik}}) = b_{ik} + \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}} \tag{278}$$

右辺を A とおけば

$$\ln \widehat{z_{ik}} - \ln(1 - \widehat{z_{ik}}) = A \tag{279}$$

$$\Rightarrow \quad \ln \frac{\widehat{z_{ik}}}{1 - \widehat{z_{ik}}} = A$$

$$\Rightarrow \quad \frac{\widehat{z_{ik}}}{1 - \widehat{z_{ik}}} = \exp A$$

$$\Rightarrow \quad \frac{1}{1 - \widehat{z_{ik}}} - 1 = \exp A$$

$$\Rightarrow \quad 1 - \widehat{z_{ik}} = \frac{1}{1 + \exp A}$$

離散潜在変数の変分推論

$$\Rightarrow \quad \widehat{z_{ik}} = 1 - \frac{1}{1 + \exp A}$$

$$\Rightarrow \quad \widehat{z_{ik}} = \frac{\exp A}{1 + \exp A}$$

$$\Rightarrow \quad \widehat{z_{ik}} = \frac{1}{1 + \exp(-A)} \tag{280}$$

$$\Rightarrow \quad \widehat{z_{ik}} = \sigma(A) \tag{281}$$

従って、不動点方程式は

$$\widehat{z_{ik}} = \sigma\left(b_{ik} + \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}}\right) \tag{282}$$

- 不動点方程式の観察

離散潜在変数の変分推論

- 不動点方程式は次で表された

$$\widehat{z_{ik}} = \sigma\left(b_{ik} + \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2}\boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}}\right) \quad (283)$$

- 第 2 項 $\boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k}$ は、潜在変数のユニット k に対する入力
- 第 3 項 $-\frac{1}{2}\boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k}$ は、隠れユニット k から自身への入力
- 第 4 項 $-\sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}}$ は、他の隠れユニット $l \neq k$ から隠れユニット k への入力

- これより、平均場不動点方程式は、回帰結合型ニューラルネットワーク (RNN) との関係があることが分かる

- 隠れユニット k と l は、それらの重みベクトル $\boldsymbol{W}_{:l}$ と $\boldsymbol{W}_{:k}$ が互いに同調するとき (似たような重みを持っているとき) に、互いに抑制し合う

離散潜在変数の変分推論

- 即ち、2 つの隠れユニット k, l が、共に入力を説明するとき (入力から同じような表現を抽出するとき)、入力を最もよく説明するユニットのみがアクティブになる (強く活性化される)
- これはユニット間の競合の一形態である

- 従って、実際には多峰性の事後分布かもしれないが、そのうちの 1 つのみが選択される (図 4 の (b) と (c) 参照)

- 不動点方程式を更に以下のように変形する

$$\widehat{z_{ik}} = \sigma\left(b_{ik} + \left(\boldsymbol{x}_i - \sum_{l \neq k} \boldsymbol{W}_{:l}\widehat{z_{il}}\right)^T \boldsymbol{\beta}\boldsymbol{W}_{:k} - \frac{1}{2}\boldsymbol{W}_{:k}^T\boldsymbol{\beta}\boldsymbol{W}_{:k}\right) \quad (284)$$

- これより、ユニット k への入力は、$\boldsymbol{x}_i$ ではなく $\boldsymbol{x}_i - \sum_{l \neq k} \boldsymbol{W}_{:l}\widehat{z_{il}}$ であるとみなせる

離散潜在変数の変分推論

- ユニット k への入力は、他の全てのユニットによる $\boldsymbol{x}_i$ の再構成と、実際の入力 $\boldsymbol{x}_i$ との誤差である
- ユニット k は、この残差誤差を符号化していると分かるので、スパース符号化は、反復自己符号化器とみなせる

- スパース符号化では、入力 $\boldsymbol{x}_i$ の符号化 ($\widehat{z_{ik}}$ の計算) と、復号 ($\sum_{l \neq k} \boldsymbol{W}_{:l} \widehat{z_{il}}$ の計算) を繰り返す
- この反復のたびに、再構成の誤差を修正していく

- ダンピング
  - 1 つのユニットの更新則を (不動点方程式として) 導出した
  - 複数のユニットを同時に更新することは、二値スパース符号化モデルでは通常できない
  - 但し、ダンピングという発見的手法を使えば可能になる

  - 各要素 $\widehat{z_{ik}}$ についての最適値を計算し、その値の変化の方向に、他の要素 $\widehat{z_{il}}$ を小さいステップで動かす

離散潜在変数の変分推論

- 下界 $\mathcal{L}(q)$ が増加することはもはや保証されないが、多くの場合はうまくいく

離散潜在変数の変分推論のまとめ

- ここまでの話の流れ
  1. 離散潜在変数における変分推論の具体例として、二値スパース符号化モデルをみた
  2. 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ が複雑になるので、最尤推定 (EM アルゴリズム) が使えない
  3. 代わりに、別の分布 $q(\boldsymbol{Z})$ を使って事後分布を近似することにした
  4. エビデンス下界 $\mathcal{L}(q)$ を苦労して求めた
  5. 更に、下界 $\mathcal{L}(q)$ を (各パラメータについて) 最大化するための、不動点方程式を導出した
  6. 不動点方程式を観察し、回帰結合型ニューラルネットワークや、自己符号化器との関係を考えた
- これからの話の流れ

離散潜在変数の変分推論のまとめ

- 連続潜在変数に対する変分推論を、簡単な確率モデルを使って、試してみよう

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

連続潜在変数の変分推論

- 連続潜在変数をもつ変分推論の概要
  - 平均場近似を行う場合は、以下の式によって最適な因子 $q_j^*(\boldsymbol{Z}_j|\boldsymbol{X})$ が得られる

$$\ln q_j^*(\boldsymbol{Z}_j|\boldsymbol{X}) = \mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right] + \text{Const.} \tag{285}$$

$$\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right] = \int \ln p(\boldsymbol{X},\boldsymbol{Z}) \prod_{i\neq j} q_i(\boldsymbol{Z}_i|\boldsymbol{X}) d\boldsymbol{Z}_i \tag{286}$$

$$q_j^*(\boldsymbol{Z}_j|\boldsymbol{X}) = \frac{\exp\left(\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right]\right) d\boldsymbol{Z}_j} \tag{287}$$

  - 上式は、下界 $\mathcal{L}(q)$ を最大化する q であり、従って連続潜在変数の場合の不動点方程式とみなせる
  - 各因子 $q_j(\boldsymbol{Z}_j|\boldsymbol{X})$ を、上式を用いて順番に更新していく ($i \neq j$ である全ての q_i を固定した状態で、各 q_j について下界を最適化する)

連続潜在変数の変分推論

- このステップを、下界 $\mathcal{L}(q)$ が収束するまで繰り返し行う (座標降下法)
- 不動点方程式は、下界が最適な値に収束するかどうかには関係なく、q_j の最適解が取る関数形を提供してくれる

- 扱う確率モデルの表現
  - ここでは次のような確率モデルを対象として、変分推論を扱う

$$p(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{z}_i|0, \boldsymbol{I}) \tag{288}$$
$$p(x_i|\boldsymbol{z}_i) = \mathcal{N}(x_i|\boldsymbol{w}^T\boldsymbol{z}_i, 1) \tag{289}$$

  - 1 次元のデータ $x \in \mathbb{R}$ に対して、2 次元の潜在変数 $\boldsymbol{z} \in \mathbb{R}^2$ が存在する
  - 同時分布 $p(x_i, \boldsymbol{z}_i) = p(x_i|\boldsymbol{z}_i)p(\boldsymbol{z}_i)$ を $\boldsymbol{z}_i$ で積分消去すれば、x_i についての単なるガウス分布となる

- 真の事後分布 $p(\boldsymbol{z}_i|x_i)$ の計算

連続潜在変数の変分推論

- 正規化定数を無視して次のように計算できる

$$
\begin{aligned}
& p(\boldsymbol{z}_i|x_i) \\
\propto\ & p(\boldsymbol{z}_i|x_i)p(x_i) \\
=\ & p(x_i, \boldsymbol{z}_i) \\
=\ & p(\boldsymbol{z}_i)p(x_i|\boldsymbol{z}_i) \\
=\ & \mathcal{N}(\boldsymbol{z}_i|0, \boldsymbol{I})\mathcal{N}(x_i|\boldsymbol{w}^T\boldsymbol{z}_i, 1) \\
\propto\ & \exp\left(-\frac{1}{2}\boldsymbol{z}_i^T\boldsymbol{z}_i\right)\exp\left(-\frac{1}{2}\left(x_i-\boldsymbol{w}^T\boldsymbol{z}_i\right)^T\left(x_i-\boldsymbol{w}^T\boldsymbol{z}_i\right)\right) \\
=\ & \exp\left(-\frac{1}{2}\left(z_{i1}^2+z_{i2}^2\right)\right) \\
& \qquad \exp\left(-\frac{1}{2}\left(x_i-w_1z_{i1}-w_2z_{i2}\right)^T\left(x_i-w_1z_{i1}-w_2z_{i2}\right)\right) \\
=\ & \exp\left(-\frac{1}{2}\left(z_{i1}^2+z_{i2}^2+x_i^2+w_1^2z_{i1}^2+w_2^2z_{i2}^2-\right.\right.
\end{aligned}
$$

連続潜在変数の変分推論

$$2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2}) \Bigg) \tag{290}$$

- 正規化項を C とすれば次のように書ける

$$\begin{aligned} & p(\boldsymbol{z}_i | x_i) \\ = \quad & C \exp \Bigg(-\frac{1}{2} \left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 - \right. \\ & \left. 2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2} \right) \Bigg) \end{aligned} \tag{291}$$

- z_{i1} と z_{i2} を乗算する項が存在する
- 従って、真の事後分布は、z_{i1} と z_{i2} のみの因子には分解できないことが分かる

- 平均場近似の計算

連続潜在変数の変分推論

- 平均場近似を次のように表現する

$$q(\boldsymbol{z}_i|x_i) = q_1(z_{i1}|x_i)q_2(z_{i2}|x_i) \tag{292}$$

- q_2 を固定した状態で、最適な $q_1^*(z_{i1}|x_i)$ を求める

$$\begin{aligned}
& \ln q_1^*(z_{i1}|x_i) \\
= \; & \mathbb{E}_{z_{i2}\sim q_2(z_{i2}|x_i)}\left[\ln p(\boldsymbol{z}_i, x_i)\right] + \text{Const.} \qquad (293) \\
= \; & \mathbb{E}_{z_{i2}}\left[\ln\left(C\exp\left(-\frac{1}{2}\left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 - \right.\right.\right.\right. \\
& \left.\left.\left.\left. 2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2}\right)\right)\right)\right] + \text{Const.} \\
= \; & \mathbb{E}_{z_{i2}}\left[-\frac{1}{2}\left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 - \right.\right. \\
& \left.\left. 2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2}\right)\right] + \text{Const.} \qquad (294)
\end{aligned}$$

連続潜在変数の変分推論

- $z_{i2} \sim q_2(z_{i2}|x_i)$ による期待値を取っている
- $q_2(z_{i2}|x_i)$ から得る必要があるのは、結局 $\mathbb{E}_{z_{i2}}[z_{i2}]$ と、$\mathbb{E}_{z_{i2}}[z_{i2}^2]$ の 2 つだけである
- $\langle z_{i2} \rangle = \mathbb{E}_{z_{i2}}[z_{i2}]$、$\langle z_{i2}^2 \rangle = \mathbb{E}_{z_{i2}}[z_{i2}^2]$ と書くことにする

- このとき次式が得られる

$$
\begin{aligned}
& \ln q_1^*(z_{i1}|x_i) \\
= \quad & -\frac{1}{2}\left(z_{i1}^2 + \langle z_{i2}^2 \rangle + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 \langle z_{i2}^2 \rangle - \right. \\
& \left. 2x_i w_1 z_{i1} - 2x_i w_2 \langle z_{i2} \rangle + 2w_1 w_2 z_{i1} \langle z_{i2} \rangle\right) + \\
& \text{Const.}
\end{aligned} \tag{295}
$$

連続潜在変数の変分推論

- これより、最適な $q_1^*(z_{i1}|x_i)$ はガウス分布の形であると分かる

$$
\begin{aligned}
& q_1^*(z_{i1}|x_i) \\
= \quad & C \exp\left(-\frac{1}{2} \left(z_{i1}^2 + \left\langle z_{i2}^2 \right\rangle + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 \left\langle z_{i2}^2 \right\rangle - \right. \right. \\
& \left. \left. 2x_i w_1 z_{i1} - 2x_i w_2 \left\langle z_{i2} \right\rangle + 2w_1 w_2 z_{i1} \left\langle z_{i2} \right\rangle \right) \right) \\
\propto \quad & \exp\left(-\frac{1}{2} \left(z_{i1}^2 + \left\langle z_{i2}^2 \right\rangle + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 \left\langle z_{i2}^2 \right\rangle - \right. \right. \\
& \left. \left. 2x_i w_1 z_{i1} - 2x_i w_2 \left\langle z_{i2} \right\rangle + 2w_1 w_2 z_{i1} \left\langle z_{i2} \right\rangle \right) \right) \qquad (296)
\end{aligned}
$$

- 対称性から、最適な $q_2^*(z_{i2}|x_i)$ もガウス分布であることが分かる
- ガウス分布同士の積もガウス分布になるので、結局 $q(\boldsymbol{z}_i|x_i) = q_1(z_{i1}|x_i)q_2(z_{i2}|x_i)$ はガウス分布である

連続潜在変数の変分推論

- $q(\boldsymbol{z}_i|x_i)$ が 2 つの因子に分解できるとは仮定したが、各因子の関数形については全く仮定していないことに注意
- ガウス分布は、下界 $\mathcal{L}(q)$ を q について変分最適化する過程で、自然に出現した

連続潜在変数の変分推論のまとめ

- ここまでの話の流れ
  1. 連続潜在変数における変分推論の例として、簡単な確率モデルを扱った
  2. 分布 $q(\boldsymbol{Z}|\boldsymbol{X})$ を、$\prod_i q_i(\boldsymbol{Z}_i|\boldsymbol{X})$ のように因数分解できるという仮定 (平均場近似) のみを置いた
  3. 各因子 $q_i(\boldsymbol{Z}_i|\boldsymbol{X})$ の関数形については全く仮定を置かなかった
  4. 変分推論によって、最適解となる q_i の関数形が自然に導出できた

- これからの話の流れ
  - 教科書で触れられている雑多な話題を扱う
  - 近似推論が、推論アルゴリズムの精度に影響することについて考える
  - 学習による近似推論の手法を考える

目次

1. 近似推論法
   - 変分推論
   - 座標降下法
   - MAP 推定および最尤推定と変分推論
   - 変分近似による弊害
   - 離散潜在変数の変分推論
   - 連続潜在変数の変分推論
   - 様々な話題

学習と推論の相互作用

- 近似推論が精度に与える影響
  - 近似推論では、以下の下界 $\mathcal{L}(q)$ を q について最適化する

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{297}$$

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{298}$$

  - 下界 $\mathcal{L}(q)$ を次のように分解する

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} && (299)\\
&= \int q(\boldsymbol{Z}) \left(\ln p(\boldsymbol{Z}|\boldsymbol{X}) + \ln p(\boldsymbol{X}) \right) d\boldsymbol{Z} + H[q] && (300)\\
&= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z} + \ln p(\boldsymbol{X}) + H[q] && (301)\\
&= \mathbb{E}_{\boldsymbol{Z} \sim q(\boldsymbol{Z})} \left[\ln p(\boldsymbol{Z}|\boldsymbol{X}) \right] + \ln p(\boldsymbol{X}) + H[q] && (302)
\end{aligned}$$

学習と推論の相互作用

- 分布に含まれるパラメータを明示的に記述する
- $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ と、$p(\boldsymbol{Z}|\boldsymbol{\theta})$ をモデリングする
- このとき、$p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta})$ と書ける
- 従って、下界 $\mathcal{L}(q)$ は次のようになる

$$\mathcal{L}(q,\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\right] + \ln p(\boldsymbol{X}|\boldsymbol{\theta}) + H[q] \tag{303}$$

- 下界 $\mathcal{L}(q,\boldsymbol{\theta})$ をパラメータ $\boldsymbol{\theta}$ について増加させるとする
- これは、$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\right]$ を、q を固定しつつ、$\boldsymbol{\theta}$ について増加させることに繋がる

- このとき、$q(\boldsymbol{Z})$ が高い確率となる $\boldsymbol{Z}$ について $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ が増大し、$q(\boldsymbol{Z})$ が低い確率となる $\boldsymbol{Z}$ について $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ が減少する

- これより、近似仮定が自己充足的予言となることが分かる
- 自己充足的予言とは、ある事象や状況に関する判断や思い込みが原因となり、その結果として、その判断や思い込みが現実化することである

学習と推論の相互作用

- 即ち、予め近似事後分布 $q(\boldsymbol{Z})$ に何らかの仮定を置いて訓練することで、結果としてその仮定に沿うような分布が得られてしまう

- 単峰性の近似事後分布 $q(\boldsymbol{Z})$ を使って訓練することを考える
- 得られる事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ は、厳密な推論によって得られた $p(\boldsymbol{Z}|\boldsymbol{X})$ と比べて、単峰性がはるかに強くなっている
- 事後分布が多峰性であるという仮定が、推定結果にも現れてしまう

- 変分推論によってモデルに課される損害
  - モデルを訓練した後に、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を推定し、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ との差が十分に小さい $(\mathcal{L}(q, \boldsymbol{\theta}) \simeq \ln p(\boldsymbol{X}|\boldsymbol{\theta}))$ ことを確認する
  - このとき、ある特定のパラメータ $\boldsymbol{\theta}$ について、変分近似が正確であるといえる

  - 但し、一般的に変分近似が正確であるとはいえない

学習と推論の相互作用

- また、変分推論が、学習過程に殆ど害を及ぼさないとも結論付けてはいけない

- 変分近似による真の損害は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を最大にする $\boldsymbol{\theta}^*$ が分からなければ、測定することができない

- ある $\boldsymbol{\theta}$ について $\mathcal{L}(q, \boldsymbol{\theta}) \simeq \ln p(\boldsymbol{X}|\boldsymbol{\theta})$ が成立しても、$\mathcal{L}(q, \boldsymbol{\theta}) \simeq \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \ll \ln p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ であるかもしれない
- あるパラメータ $\boldsymbol{\theta}$ については変分近似は正確だが、実際には近似が全く上手く行っていないことになる

- また $\max_q \mathcal{L}(q, \boldsymbol{\theta}^*) \ll \ln p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ であれば、$\boldsymbol{\theta}^*$ のときの事後分布は、制限された q にとって複雑すぎるため、訓練では $\boldsymbol{\theta}^*$ を見つけることができない
- この問題は、変分推論とは別の方法で $\boldsymbol{\theta}^*$ が求まる場合にしか、検知することができない

学習による近似推論

- 近似推論の実行の学習
  - 不動点方程式や勾配に基づく最適化を、繰り返し行う近似推論は、計算コストが非常に高い
  - 近似推論による最適化処理は、入力 $\boldsymbol{X}$ から、近似事後分布 $q^*(\boldsymbol{Z}) = \arg\max_q \mathcal{L}(\boldsymbol{X}, q)$ へと写像する関数 f とみなせる
  - これにより、反復的な最適化処理を単なる関数とみなして、関数 $\widehat{f}(\boldsymbol{X}, \boldsymbol{\theta})$ を、ニューラルネットワークで近似することができる
- Wake-Sleep アルゴリズム
  - データ $\boldsymbol{X}$ から潜在変数 $\boldsymbol{Z}$ を推論するときの困難は、正しい潜在変数 $\boldsymbol{Z}$ が分からず、従って教師あり訓練集合が得られないことである
  - データ $\boldsymbol{X}$ から潜在変数 $\boldsymbol{Z}$ への写像は、モデル $(p(\boldsymbol{X}|\boldsymbol{Z}|\boldsymbol{\theta}), p(\boldsymbol{Z}|\boldsymbol{\theta})$ の選択に依存し、更に $\boldsymbol{\theta}$ の変化に伴って変わり続ける

学習による近似推論

- Wake-Sleep アルゴリズムでは、$\boldsymbol{X}$ と $\boldsymbol{Z}$ の両方のサンプルを、モデル分布から抽出する
- モデルの分布が高い確率を示すような $\boldsymbol{X}$ に対してしか、推論ネットワークを学習できない

- モデルの分布が、データの分布に似ていないとき、推論ネットワークはデータに似たサンプルを学習できない

- 推論学習の他の形式
  - 変分自己符号化器は、生成モデリングで主要なアプローチとなった

  - 推論ネットワークは、単に下界 $\mathcal{L}(q)$ を定義するために使用される
  - ネットワークのパラメータは、下界 $\mathcal{L}(q)$ が増加するように適用される