目次

EM アルゴリズムが困難な場合

- EM アルゴリズムで行う計算
  - E ステップでは、潜在変数の事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}})$ を計算
  - M ステップでは、完全データ対数尤度 $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ の期待値を計算

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \tag{1}$$

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) = \int_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \tag{2}$$

そして、$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}})$ を最大化するパラメータ $\boldsymbol{\theta}^{\mathrm{new}}$ を求める

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) \tag{3}$$

EMアルゴリズムが困難な場合

- EMアルゴリズムの困難さ
  - 実際に扱うモデルでは、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}})$ の計算や、事後分布に従った期待値 $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}})$ の計算が、不可能であることが多い
  - 隠れ変数の次元が高すぎる
  - 事後分布が複雑な形をしていて、期待値を解析的に計算できない
  - 連続変数であれば、積分が閉形式の解を持たないかもしれない
  - 空間の次元の問題や、被積分項の複雑さから、数値積分すら困難かもしれない
  - 離散変数であれば、期待値を計算するためには、潜在変数の可能な全ての組み合わせについての和を取る必要がある
  - 隠れ変数の次元が高くなると、組み合わせ数が指数的に増大する
  - 計算量が大きすぎて、期待値の厳密な計算がもはや不可能

EM アルゴリズムが困難な場合

- 近似法
  - EM アルゴリズムが困難であるとき、何らかの方法で近似しなければならない
  - 近似法は、確率的な近似と、決定的な近似の 2 つに分けられる
- 確率的な近似
  - マルコフ連鎖モンテカルロ法などの手法がある
  - 無限の計算資源があれば、厳密な結果が得られる
  - 実際には計算量が有限であるため、得られる解は近似解となる
- 決定的な近似
  - 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ を解析的に近似する
  - 事後分布に対して、何らかの仮定をおく
  - 例えば、単純な項の積として分解できる、あるいは、(ガウス分布などの特別な) パラメトリックな分布であるといった仮定

変分推論

- ここで扱う近似法
  - 変分推論法 (Variational inference) あるいは変分ベイズ法 (Variational Bayes) について扱う

- 変分推論 (Variational inference)
  - 18 世紀のオイラー、ラグランジュらによる変分法 (Calculus of variations) に起源をもつ
  - まずは、変分法について説明をしていく

変分法

- 関数と汎関数の違い
  - 通常の関数は、入力として値をとり、出力として関数の値を返す
  - 通常の関数は、値から値への写像である

  - 関数の導関数は、入力値を微小に変えたときに、出力の関数値がどの程度変わるかを表す

  - 汎関数 (Functional) とは、入力として関数をとり、出力として汎関数の値を返す
  - 汎関数は、関数から値への写像である

  - 汎関数微分 (Functional derivative) とは、入力関数が微小に変わったときに、出力の汎関数値がどの程度変わるかを表す

  - 汎関数の微分を、変分という

変分法

- 汎関数の例
  - エントロピー $H[p]$ は、確率分布 $p(x)$ を入力として、以下の量を返す汎関数

$$H[p] = -\int p(x) \ln p(x) dx \tag{4}$$

- 汎関数の最適化
  - 多くの問題は、汎関数の値を最適化する問題として定式化できる
  - 汎関数の最適化とは、可能な全ての入力関数の中から、汎関数の値を最大化、あるいは最小化するような関数を選び出すことである
  - 通常の最適化では、可能な全てのパラメータ (入力値) の中から、関数を最大化、あるいは最小化するような 1 つのパラメータを選び出す
  - 次は、いよいよ変分の計算について説明する

変分法

- 変分法
  - 通常の微分を使えば、ある関数 $y(x)$ を最大化 (最小化) するような x の値が求められる

  - 変分法を使えば、汎関数 $F[y]$ を最大化 (最小化) するような、関数 $y(x)$ が求められる
  - 従って、可能な全ての関数 $y(x)$ の中から、$F[y]$ を最大 (最小) にするような関数が得られる

- 変分法によって解ける問題の例
  - 2 点を結ぶ最短経路は? (答えは直線)
  - 最速降下曲線は? (答えはサイクロイド)
  - エントロピーが最大になるような確率分布は? (答えはガウス分布)

変分法

- 通常の微分の表現
  - 関数 $y(x+\epsilon)$ のテイラー展開は次のように記述できた

$$
\begin{aligned}
y(x+\epsilon) &= \sum_{n=0}^{\infty} \frac{y^{(n)}(x)}{n!}\epsilon^n &(5)\\
&= y(x) + \frac{dy}{dx}\epsilon + \frac{1}{2!}\frac{d^2y}{dx^2}\epsilon^2 + \frac{1}{3!}\frac{d^3y}{dx^3}\epsilon^3 + \cdots &(6)\\
&= y(x) + \frac{dy}{dx}\epsilon + O(\epsilon^2) &(7)
\end{aligned}
$$

  - これより微分 dy/dx は、次のように求められる
  - 変数 x に微小な変化 ϵ を加え、このときの関数値 $y(x+\epsilon)$ を ϵ の累乗形として表現する
  - 最後に $\epsilon \to 0$ の極限をとればよい

$$
\frac{dy}{dx} = \lim_{\epsilon\to 0} \frac{y(x+\epsilon) - y(x)}{\epsilon} \tag{8}
$$

変分法

- 多変数関数 $y(x_1, \ldots, x_D)$ の偏微分の表現
  - 多変数関数 $y(x_1, \ldots, x_D)$ のテイラー展開は次のように記述できた

$$\boldsymbol{D}^n = \left(\epsilon_1 \frac{\partial y}{\partial x_i} + \cdots + \epsilon_D \frac{\partial y}{\partial x_D} \right)^n \tag{9}$$

上記のような演算子 $\boldsymbol{D}$ を考えれば

$$\begin{aligned} & y(x_1 + \epsilon_1, \ldots, x_D + \epsilon_D) \\ = & \sum_{n=0}^{\infty} \frac{1}{n!} (\boldsymbol{D}^n y)(x_1, \ldots, x_D) \qquad (10) \\ = & y(x_1, \ldots, x_D) + \sum_{i=1}^{D} \frac{\partial y}{\partial x_i} \epsilon_i + \frac{1}{2!} \sum_{i=1}^{D} \sum_{j=1}^{D} \frac{\partial^2 y}{\partial x_i x_j} \epsilon_i \epsilon_j + \\ & \frac{1}{3!} \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} \frac{\partial^3 y}{\partial x_i x_j x_k} \epsilon_i \epsilon_j \epsilon_k + \cdots \qquad (11) \end{aligned}$$

変分法

であるから

$$
\begin{aligned}
& y(x_1+\epsilon_1,\ldots,x_D+\epsilon_D) \\
= \quad & y(x_1,\ldots,x_D)+\sum_{i=1}^{D}\frac{\partial y}{\partial x_i}\epsilon_i+O(\epsilon^2)
\end{aligned} \tag{12}
$$

- これより偏微分 $\partial y/\partial x_i$ は、次のように求められる

$$
\begin{aligned}
\frac{\partial y}{\partial x_i}=\lim_{\epsilon_i\to 0}\frac{1}{\epsilon_i}\,&(y(x_1,\ldots,x_{i-1},x_i+\epsilon_i,x_{i+1},\ldots,x_D)- \\
&y(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_D))
\end{aligned} \tag{13}
$$

変分法

- 変分の表現
  - 多少不正確だが、変分をどのように定義すればよいか考えてみる
  - ここで、各 x_i に対する関数の値 $z_i = y(x_i)$ を個別の変数とみなして、次の関数 $F(z_1, \ldots, z_D)$ について考えてみよう

$$\begin{aligned} & F(z_1 + \epsilon\eta(x_1), \ldots, z_D + \epsilon\eta(x_D)) \\ = \quad & F(z_1, \ldots, z_D) + \sum_{i=1}^{D} \frac{\partial F}{\partial z_i} \epsilon\eta(x_i) + O(\epsilon^2) \end{aligned} \tag{14}$$

$z_i = y(x_i)$ を代入してみると

$$\begin{aligned} & F(y(x_1) + \epsilon\eta(x_1), \ldots, y(x_D) + \epsilon\eta(x_D)) \\ = \quad & F(y(x_1), \ldots, y(x_D)) + \sum_{i=1}^{D} \frac{\partial F}{\partial y(x_i)} \epsilon\eta(x_i) + O(\epsilon^2) \end{aligned} \tag{15}$$

変分法

- ここで $D \to \infty$ の極限を取り、$x_1, \ldots, x_D$ が、ある連続した区間 $[a, b]$ に含まれる、全ての実数を表すことにする
- このとき $y(x_1), \ldots, y(x_D)$ は、実数の区間 $[a, b]$ で定義される連続関数 $y(x)$ として書けることが分かる
- 同様に $y(x_1) + \epsilon\eta(x_1), \ldots, y(x_D) + \epsilon\eta(x_D)$ は、実数の区間 $[a, b]$ で定義される連続関数 $y(x) + \epsilon\eta(x)$ として、まとめることができる
- 関数 $\eta(x)$ も、実数の区間 $[a, b]$ で定義される連続関数
- $\epsilon\eta(x)$ は、$y(x)$ に加わる摂動として、考えることができる

変分法

- 関数 F は、関数 $y(x)$ や $y(x)+\epsilon\eta(x)$ を入力として受け取る、汎関数 $F[y]$ として解釈できるから、次のように書ける

$$F(y(x_1)+\epsilon\eta(x_1),\ldots,y(x_D)+\epsilon\eta(x_D)) = F[y(x)+\epsilon\eta(x)] \quad (16)$$

$$F(y(x_1),\ldots,y(x_D)) = F[y(x)] \quad (17)$$

- 以下の項は、入力を $y(x)$ に摂動を加えて $y(x)+\epsilon\eta(x)$ へと微小に変化させたときの、汎関数の ($F[y(x)]$ から $F[y(x)+\epsilon\eta(x)]$ への) 変化量を表している

$$\sum_{i=1}^{D} \frac{\partial F}{\partial y(x_i)} \epsilon\eta(x_i) \quad (18)$$

- 点 x_i における汎関数 F の変化量を、$x_1,\ldots,x_D$ の範囲について、即ち、実数の区間 $[a,b]$ について足し合わせていると解釈する

変分法

- $D \to \infty$ のとき、$x_1, \ldots, x_D$ は区間 $[a, b]$ における全ての実数を表すから、総和を積分に置き換えられそうである
- 汎関数の微分 $\dfrac{\delta F}{\delta y(x)}$ を使えば、次のように書ける

$$\sum_{i=1}^{D} \frac{\partial F}{\partial y(x_i)} \epsilon \eta(x_i)$$
$$\Rightarrow \quad \int_a^b \frac{\delta F}{\delta y(x)} \epsilon \eta(x) dx = \epsilon \int_a^b \frac{\delta F}{\delta y(x)} \eta(x) dx \tag{19}$$

- 結局、変分 $\dfrac{\delta F}{\delta y(x)}$ は次のように定義できる

$$F[y(x) + \epsilon \eta(x)] = F[y(x)] + \epsilon \int_a^b \frac{\delta F}{\delta y(x)} \eta(x) dx + O(\epsilon^2) \tag{20}$$

変分法

- $F[y]$ は、区間 $[a,b]$ で定義される関数 y を受け取るとする
- 変分 $\delta F/\delta y$ は、入力関数 $y(x)$ に、任意の微小な変化 $\epsilon\eta(x)$ を加えたときの、汎関数 $F[y]$ の変化量として定義できる
- $\eta(x)$ は x についての任意の関数

変分法

Figure D.1 A functional derivative can be defined by considering how the value of a functional $F[y]$ changes when the function $y(x)$ is changed to $y(x) + \epsilon\eta(x)$ where $\eta(x)$ is an arbitrary function of x.

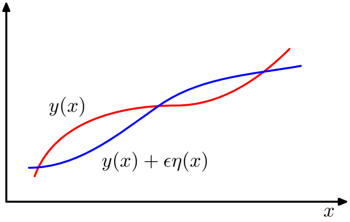

図 1: $y(x)$ と $y(x) + \epsilon\eta(x)$ の表現

変分法

- 変分法の例
  - 次の図 2 を使って、実際に変分を求めてみよう
  - 汎関数 $F[y]$ は、次のように定義されるとする

$$F[y] = \int_a^b y(x)dx \tag{21}$$

  - 汎関数の値 $F[y(x)], F[y(x) + \epsilon\eta(x)]$ は次のようになる

$$F[y(x)] = \int_a^b y(x)dx \tag{22}$$

$$F[y(x) + \epsilon\eta(x)] = \int_a^b (y(x) + \epsilon\eta(x))\,dx \tag{23}$$

変分法

- $F[y(x)+\epsilon\eta(x)]$ は次のように分解できる

$$F[y(x)+\epsilon\eta(x)] = \int_a^b y(x)dx + \epsilon\int_a^b \eta(x)dx \tag{24}$$

$$= F[y(x)] + \epsilon\int_a^b \eta(x)dx \tag{25}$$

- ここで、変分の定義式は

$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \epsilon\int_a^b \frac{\delta F}{\delta y(x)}\eta(x)dx + O(\epsilon^2) \tag{26}$$

であったので、上の 2 つの式を見比べれば、変分 $\delta F/\delta y$ は結局

$$\frac{\delta F}{\delta y(x)} = 1 \tag{27}$$

となることが分かる

変分法

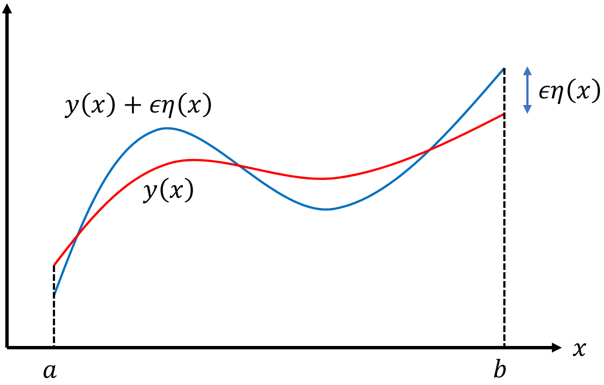

図 2: 区間 $[a, b]$ で定義された関数 $y(x)$ の表現

変分法

- 汎関数の最適化
  - 汎関数 $F[y]$ が最大 (最小) となるとき、関数 $y(x)$ の微小な変化に対して、汎関数は変化しないはず
  - 即ち、汎関数が最大 (最小) となるとき、$F[y(x) + \epsilon\eta(x)] = F[y(x)]$ が成り立つ
  - 従って、変分の定義式から、以下が成り立つ

$$\int_a^b \frac{\delta F}{\delta y(x)} \eta(x) dx = 0 \tag{28}$$

  - 上式は任意の $\eta(x)$ について成立しなければならない
  - 従って、変分 $\delta F / \delta y$ は、任意の x について 0 とならなければならない

  - 汎関数 $F[y]$ が最大 (最小) となるとき、$\delta F / \delta y = 0$ が成立することが分かった (通常の微分と同じ)

変分法

- 変分法の例
  - 様々な汎関数について、変分を導出してみよう
  - また、その汎関数が最大 (最小) となるときに成り立つ条件を、導出してみよう

変分法

- 汎関数の例 (1)
  - $y(x)$ とその微分 $y'(x) = dy/dx$、そして x によって決まる関数 $G(y(x), y'(x), x)$ があるとする
  - 汎関数 $F[y]$ を、$G(y(x), y'(x), x)$ を区間 $[a, b]$ にわたって積分した結果を出力する関数として、次のように定める

$$F[y] = \int_a^b G(y(x), y'(x), x)dx \tag{29}$$

  - 積分区間は無限であってもよいとする ($a = -\infty, b = \infty$ でもよい)

変分法

- $y(x)$ に摂動 $\epsilon\eta(x)$ を加えたときの、汎関数の値 $F[y(x)+\epsilon\eta(x)]$ を使って、変分 $\delta F/\delta y$ を調べてみる

$$F[y(x)+\epsilon\eta(x)] = \int_a^b G(y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x), x)dx \tag{30}$$

ここで、被積分項のテーラー展開を考えれば

$$\begin{aligned} & G(y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x), x) \\ = \ & G(y(x), y'(x), x) + \frac{\partial G}{\partial y}\epsilon\eta(x) + \\ & \quad \frac{\partial G}{\partial y'}\epsilon\eta'(x) + \frac{\partial G}{\partial x}\cdot 0 + O(\epsilon^2) & (31) \\ = \ & G(y(x), y'(x), x) + \epsilon\left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) + O(\epsilon^2) & (32) \end{aligned}$$

であるから

$$
\begin{aligned}
& F[y(x)+\epsilon\eta(x)] \\
= \; & \int_a^b G(y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x), x)dx \\
= \; & \int_a^b \Bigg(G(y(x), y'(x), x) + \\
& \quad \epsilon\left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) + O(\epsilon^2)\Bigg) dx \qquad (33) \\
= \; & \int_a^b G(y(x), y'(x), x)dx + \\
& \quad \epsilon\int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) dx + O(\epsilon^2) \qquad (34) \\
= \; & F[y(x)] + \epsilon\int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) dx + O(\epsilon^2) \qquad (35)
\end{aligned}
$$

変分法

ここで

$$
\begin{aligned}
&\int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x) \right) dx \\
= \; & \int_a^b \left(\frac{\partial G}{\partial y}\eta(x) \right) dx + \int_a^b \left(\frac{\partial G}{\partial y'}\eta'(x) \right) dx \qquad (36) \\
= \; & \int_a^b \left(\frac{\partial G}{\partial y}\eta(x) \right) dx + \\
& \qquad \left[\frac{\partial G}{\partial y'}\eta(x) \right]_a^b - \int_a^b \frac{d}{dx}\left(\frac{\partial G}{\partial y'} \right) \eta(x) dx \qquad (37) \\
= \; & \left[\frac{\partial G}{\partial y'}\eta(x) \right]_a^b + \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'} \right) \right) \eta(x) dx \qquad (38)
\end{aligned}
$$

である

変分法

- 途中の式変形では、部分積分を使っていることに注意

- いま、積分区間の両端において、$y(x)$ の値は固定されているとする
- これを固定端条件という (図 3)

- このとき、$\eta(a) = \eta(b) = 0$ であるから、上式の最初の項が消えて

$$\int_a^b \left(\frac{\partial G}{\partial y} \eta(x) + \frac{\partial G}{\partial y'} \eta'(x) \right) dx$$

$$= \left[\frac{\partial G}{\partial y'} \eta(x) \right]_a^b + \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx} \left(\frac{\partial G}{\partial y'} \right) \right) \eta(x) dx \tag{39}$$

$$= \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx} \left(\frac{\partial G}{\partial y'} \right) \right) \eta(x) dx \tag{40}$$

のようになる

変分法

- 従って、摂動を加えたときの汎関数の値 $F[y(x)+\epsilon\eta(x)]$ は

$$\begin{aligned} & F[y(x)+\epsilon\eta(x)] \\ = \; & F[y(x)]+\epsilon\int_a^b \left(\frac{\partial G}{\partial y}-\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right)\eta(x)dx+O(\epsilon^2) \end{aligned} \tag{41}$$

  となる

- 上式を、変分の定義式と比べれば

$$F[y(x)+\epsilon\eta(x)] = F[y(x)]+\epsilon\int_a^b \frac{\delta F}{\delta y(x)}\eta(x)dx+O(\epsilon^2) \tag{42}$$

  変分 $\delta F/\delta y$ は次のように書ける

$$\frac{\delta F}{\delta y(x)} = \frac{\partial G}{\partial y}-\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) \tag{43}$$

変分法

- 汎関数 $F[y]$ が最大 (最小) になるとき、変分 $\delta F/\delta y$ が 0 になる
- 従って、汎関数が最大 (最小) になるとき、以下の方程式が成り立つ

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{44}$$

- これをオイラー-ラグランジュ方程式という

- オイラー-ラグランジュ方程式は、次のような考え方で導出することもできる

- $F[y]$ が最大 (最小) であれば、摂動 $\epsilon\eta(x)$ によって $y(x)$ が少し変化しても、$F[y]$ の値は変化しないはず
- 従って、$F[y]$ が最大 (最小) であるとき、$F[y]$ の ϵ による微分は 0 になるはず

変分法

- これを数式で表現すると、次のようになる

$$\left.\frac{\partial F[y]}{\partial \epsilon}\right|_{\epsilon=0} = 0 \tag{45}$$

左辺は通常の偏微分であり、これを計算すると

$$\begin{aligned}
&\frac{\partial F[y]}{\partial \epsilon} \\
&= \frac{\partial}{\partial \epsilon}\int_a^b G(y, y', x)dx \qquad (46)\\
&= \int_a^b \frac{\partial}{\partial \epsilon} G(y, y', x)dx \qquad (47)\\
&= \int_a^b \left(\frac{\partial G}{\partial y}\frac{\partial y}{\partial \epsilon} + \frac{\partial G}{\partial y'}\frac{\partial y'}{\partial \epsilon} + \frac{\partial G}{\partial x}\frac{\partial x}{\partial \epsilon}\right) dx \qquad (48)\\
&= \int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) dx \qquad (49)
\end{aligned}$$

$$= \left[\frac{\partial G}{\partial y'}\eta(x)\right]_a^b + \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right)\eta(x)dx \tag{50}$$

$$(\because \eta(a) = \eta(b) = 0)$$

$$= \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right)\eta(x)dx \tag{51}$$

$$= 0$$

- 上の式変形では、$y = y(x) + \epsilon\eta(x)$ であるから

$$\frac{\partial y}{\partial \epsilon} = \eta(x) \tag{52}$$

$$\frac{\partial y'}{\partial \epsilon} = \frac{\partial}{\partial \epsilon}\left(\frac{\partial y}{\partial x}\right) = \frac{\partial}{\partial \epsilon}\left(y'(x) + \epsilon\eta'(x)\right) = \eta'(x) \tag{53}$$

$$\frac{\partial x}{\partial \epsilon} = 0 \tag{54}$$

が成立することを利用している

変分法

- 任意の $\eta(x)$ について、上式が恒等的に成り立つためには

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{55}$$

でなければならないことが分かり、先程と同様に、オイラー-ラグランジュ方程式を得る

変分法

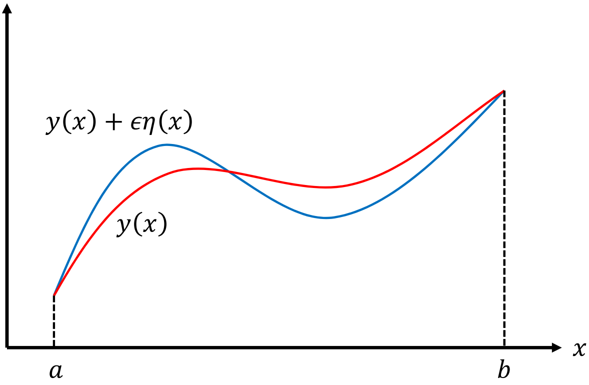

図 3: 制約条件を含んでいる場合の表現

変分法

- 汎関数の例 (2)
  - 上では $G(y(x), y'(x), x)$ について考えて、変分を導出した
  - $y(x)$ と x のみによって決まり、$y'(x)$ には依存しない関数 $G(y(x), x)$ を考えよう

  - 汎関数 $F[y]$ は、先程と同様に以下で表されるとする

$$F[y] = \int_a^b G(y(x), x)dx \tag{56}$$

  - このとき変分 $\delta F/\delta y$ を求めるのは、非常に簡単である
  - 先程の式に、$\partial G/\partial y' = 0$ を代入すれば直ちに得られる

$$\frac{\delta F}{\delta y(x)} = \frac{\partial G}{\partial y} \tag{57}$$

変分法

- あるいは以下のように書ける

$$\frac{\delta}{\delta y(x)}\int_a^b G(y(x),x)dx = \frac{\partial}{\partial y}G(y(x),x) \tag{58}$$

- $F[y]$ が最大 (最小) であるとき、以下のオイラー-ラグランジュ方程式が成り立つ

$$\frac{\partial G}{\partial y} = 0 \tag{59}$$

変分法

- 汎関数の例 (3)
  - 今度は、$y'(x)$ と x のみによって決まり、$y(x)$ には依存しない関数 $G(y'(x), x)$ を考えよう
  - この場合も変分 $\delta F/\delta y$ を求めるのは簡単である
  - $G(y(x), y'(x), x)$ の変分の式に、$\partial G/\partial y = 0$ を代入すればよい

$$\frac{\delta F}{\delta y(x)} = -\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) \tag{60}$$

  - オイラー-ラグランジュ方程式は次のようになる

$$-\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{61}$$

変分法

- 汎関数の例 (4)
  - $y(x)$ と $y'(x)$ によって決まる関数 $G(y(x), y'(x))$ を考えよう
  - このときのオイラー-ラグランジュ方程式を導出してみよう

  - $G(y(x), y'(x))$ を x で微分すれば

$$
\begin{aligned}
& \frac{d}{dx} G(y, y') \\
= \quad & \frac{\partial}{\partial y} G(y, y') \frac{dy}{dx} + \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} && (62) \\
= \quad & y' \frac{\partial}{\partial y} G(y, y') + \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} && (63)
\end{aligned}
$$

となるから

$$
y' \frac{\partial}{\partial y} G(y, y') = \frac{d}{dx} G(y, y') - \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} \qquad (64)
$$

変分法

また、オイラー-ラグランジュ方程式の両辺に y' を掛けたものは

$$y'\frac{\partial}{\partial y}G(y,y') - y'\frac{d}{dx}\left(\frac{\partial}{\partial y'}G(y,y')\right) = 0 \tag{65}$$

これらを連立させて

$$y'\frac{d}{dx}\left(\frac{\partial}{\partial y'}G(y,y')\right) = \frac{d}{dx}G(y,y') - \frac{\partial}{\partial y'}G(y,y')\frac{dy'}{dx} \tag{66}$$

$$y'\frac{d}{dx}\left(\frac{\partial}{\partial y'}G(y,y')\right) + \frac{\partial}{\partial y'}G(y,y')\frac{dy'}{dx} = \frac{d}{dx}G(y,y') \tag{67}$$

$$\frac{d}{dx}\left(y'\cdot\frac{\partial}{\partial y'}G(y,y')\right) = \frac{d}{dx}G(y,y') \tag{68}$$

$$\begin{aligned}\int\left(\frac{d}{dx}\left(y'\cdot\frac{\partial}{\partial y'}G(y,y')\right)\right)dx = \\ \int\left(\frac{d}{dx}G(y,y')\right)dx + C\end{aligned} \tag{69}$$

変分法

$$G(y, y') = y' \cdot \frac{\partial}{\partial y'} G(y, y') + C \tag{70}$$

となるので、結局オイラー-ラグランジュ方程式は

$$G - y' \frac{\partial G}{\partial y'} = \mathrm{Const.} \tag{71}$$

と書ける

変分法で解ける問題の例

- 変分法で解ける問題の例 (1)
  - 2 点 $P(0,0)$、$Q(a,b)$ を結ぶ最短経路は?
  - 2 点を結ぶ経路 $y=f(x)(0 \leq x \leq a)$ の長さ l は、次のようになる

$$l=\int_0^a \sqrt{1+\left(\frac{dy}{dx}\right)^2}dx=\int_0^a \sqrt{1+y'^2}dx \tag{72}$$

  - 被積分項が y' のみの関数となっていることが分かる

変分法で解ける問題の例

- $G(y'(x), x)$ の場合の公式を使えば、l の $y = f(x)$ による変分が求まる

$$\frac{\delta l}{\delta f(x)} = -\frac{d}{dx}\left(\frac{\partial}{\partial y'}\sqrt{1+y'^2}\right) \tag{73}$$

$$= -\frac{d}{dx}\left(\frac{1}{2}\frac{1}{\sqrt{1+y'^2}}\frac{\partial}{\partial y'}\left(1+y'^2\right)\right) \tag{74}$$

$$= -\frac{d}{dx}\frac{y'}{\sqrt{1+y'^2}} \tag{75}$$

- l を最小化するような $y = f(x)$ は、上式の変分を 0 と等置すれば

$$-\frac{d}{dx}\frac{y'}{\sqrt{1+y'^2}} = 0 \tag{76}$$

$$\therefore \quad \frac{y'}{\sqrt{1+y'^2}} = \mathrm{Const.} \tag{77}$$

変分法で解ける問題の例

- これは、y' が定数であることを意味している
- 従って、$y = C_0x + C_1$ と書ける

- 以上より、2 点間を結ぶ最短経路は直線である

- $y = f(x)$ の形について、具体的な仮定は特に置いていないことに注意
- 変分法では、関数そのものを最適化する
- 従って、関数の具体的な形については、特に仮定する必要がない

変分法のまとめ

- 変分のまとめ
  - これまでの計算で、次の変分が明らかとなった

$$
\frac{\delta}{\delta y(x)}\int G(y(x),y'(x),x)dx = \\
\frac{\partial}{\partial y}G(y(x),y'(x),x) - \frac{d}{dx}\left(\frac{\partial}{\partial y'}G(y(x),y'(x),x)\right) \tag{78}
$$

$$
\frac{\delta}{\delta y(x)}\int G(y(x),x)dx = \frac{\partial}{\partial y}G(y(x),x) \tag{79}
$$

$$
\frac{\delta}{\delta y(x)}\int G(y'(x),x)dx = -\frac{d}{dx}\left(\frac{\partial}{\partial y'}G(y'(x),x)\right) \tag{80}
$$

変分法のまとめ

- ここまでの話の流れ
  1. 変分の定義や、変分の計算法について調査した
  2. 変分 (汎関数の微分) とは、入力関数が微小に変化したときの、出力値の変化量として定義される
  3. 汎関数が特定の形で表せるとき、変分がどのようになるか計算した
  4. 汎関数が最大 (最小) になるとき、オイラー-ラグランジュ方程式が成立した
  5. 変分法を用いて、2 点間を結ぶ最短経路が直線になることを確認した
- これからの話の流れ
  - 変分最適化を、どのように推論問題に適用するのかについて調べていく

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 4 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

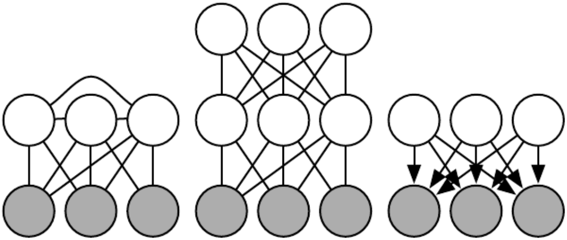

図 4: 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の計算が困難な場合

変分推論

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 5: 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

$$\begin{aligned} & p(z_1, z_2, z_3|\boldsymbol{X}) \\ = \; & p(z_1|\boldsymbol{X}, z_2, z_3)p(z_2|\boldsymbol{X}, z_3)p(z_3|\boldsymbol{X}) \end{aligned} \tag{81}$$

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

$$
\begin{aligned}
& p(\boldsymbol{Z}_1, \boldsymbol{Z}_2|\boldsymbol{X}) \\
= \quad & p(\boldsymbol{Z}_2|\boldsymbol{X})p(\boldsymbol{Z}_1|\boldsymbol{Z}_2) && (82) \\
= \quad & p(z_{21}, z_{22}, z_{23}|\boldsymbol{X})p(z_{11}, z_{12}, z_{13}|\boldsymbol{Z}_2) \\
= \quad & p(z_{21}|\boldsymbol{X})p(z_{22}|\boldsymbol{X})p(z_{23}|\boldsymbol{X}) \\
& \quad p(z_{11}|\boldsymbol{Z}_2)p(z_{12}|\boldsymbol{Z}_2)p(z_{13}|\boldsymbol{Z}_2) && (83)
\end{aligned}
$$

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まとめて、$\boldsymbol{X}, \boldsymbol{Z}$ 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

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{84}$$

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

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \tag{85}$$

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \tag{86}$$

- 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
  - $\mathcal{L}(q)$ を q について最大化し、従って $\mathrm{KL}(q||p)$ を最小化できれば、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

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

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i) \tag{87}$$

  - 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

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{88}$$

$$= \int \left(\prod_i q_i(\boldsymbol{Z}_i) \right) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{89}$$

$$= \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \tag{90}$$

$$
= \int \prod_i q_i(\boldsymbol{Z}_i)\,(\ln p(\boldsymbol{X}, \boldsymbol{Z}))\,d\boldsymbol{Z} - \int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \tag{91}
$$

ここで第 1 項は

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

$$
= \int q_j(\boldsymbol{Z}_j) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i) \right) (\ln p(\boldsymbol{X}, \boldsymbol{Z}))\,d\boldsymbol{Z} \tag{92}
$$

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

$$
= \int q_j(\boldsymbol{Z}_j) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i) \right) (\ln p(\boldsymbol{X}, \boldsymbol{Z}))\,d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \tag{93}
$$

$$= \int q_j(\boldsymbol{Z}_j)\,(\ln p(\boldsymbol{X},\boldsymbol{Z}))\left(\prod_{i\neq j} q_i(\boldsymbol{Z}_i)\right)\left(\prod_{i\neq j} d\boldsymbol{Z}_i\right) d\boldsymbol{Z}_j \quad (94)$$

$$= \int q_j(\boldsymbol{Z}_j)\left(\int (\ln p(\boldsymbol{X},\boldsymbol{Z}))\prod_{i\neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i\right) d\boldsymbol{Z}_j \quad (95)$$

$$= \int q_j(\boldsymbol{Z}_j)\ln\widetilde{p}(\boldsymbol{X},\boldsymbol{Z}_j) d\boldsymbol{Z}_j \quad (96)$$

- 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)

$$\ln\widetilde{p}(\boldsymbol{X},\boldsymbol{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\boldsymbol{X},\boldsymbol{Z})] + \mathrm{Const.} \quad (97)$$

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

$$\mathbb{E}_{i\neq j}[\ln p(\boldsymbol{X},\boldsymbol{Z})] = \int (\ln p(\boldsymbol{X},\boldsymbol{Z})) \prod_{i\neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{98}$$

- 第 2 項は

$$\begin{aligned}
&\int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \\
= & \sum_i \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} && (99) \\
= & \sum_i \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M && (100) \\
= & \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_j(\boldsymbol{Z}_j) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M +
\end{aligned}$$

変分推論

$$\sum_{i \neq j} \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \quad (101)$$

但し

$$\begin{aligned}
&\int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_j(\boldsymbol{Z}_j) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M && (102)\\
=& \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \left(\prod_{k \neq j} q_k(\boldsymbol{Z}_k) \right) \left(\prod_{k \neq j} d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_j && (103)\\
=& \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \left(\int \prod_{k \neq j} q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_j && (104)\\
=& \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \prod_{k \neq j} \underbrace{\left(\int q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)}_{=1} d\boldsymbol{Z}_j && (105)
\end{aligned}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \tag{106}$$

であるほか

$$\begin{aligned}
&\sum_{i \neq j} \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i)\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \\
&= \sum_{i \neq j} \int q_i(\boldsymbol{Z}_i) q_j(\boldsymbol{Z}_j) \left(\prod_{k \neq i,j} \ln q_k(\boldsymbol{Z}_k) \right) \\
&\qquad \left(\ln q_i(\boldsymbol{Z}_i)\right) \left(\prod_{k \neq i,j} d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_i d\boldsymbol{Z}_j \qquad (107) \\
&= \sum_{i \neq j} \underbrace{\left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \right)}_{=1} \left(\int \prod_{k \neq i,j} \ln q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)
\end{aligned}$$

変分推論

$$
\begin{aligned}
& \int \ln q_i(\boldsymbol{Z}_i) q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i && (108)\\
= \; & \sum_{i \neq j} \prod_{k \neq i,j} \underbrace{\left(\int \ln q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)}_{=\text{Const.}} \underbrace{\int \ln q_i(\boldsymbol{Z}_i) q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i}_{=\text{Const.}} && (109)\\
= \; & \sum_{i \neq j} \text{Const.} = \text{Const.} && (110)
\end{aligned}
$$

となるから結局

$$
\begin{aligned}
& \int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \\
= \; & \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \text{Const.} && (111)
\end{aligned}
$$

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

$$\begin{aligned}\mathcal{L}(q) = \int q_j(\boldsymbol{Z}_j) \ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j - \\ \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \text{Const.}\end{aligned} \tag{112}$$

但し

$$\ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{113}$$

- $\mathcal{L}(q)$ 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

$$\mathcal{L}(q) = \int q_j(\boldsymbol{Z}_j) \ln \frac{\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)}{q_j(\boldsymbol{Z}_j)} d\boldsymbol{Z}_j + \mathrm{Const.} \tag{114}$$

$$= -\mathrm{KL}\left(q_j(\boldsymbol{Z}_j) || \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)\right) + \mathrm{Const.} \tag{115}$$

- 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

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \mathrm{Const.} \tag{116}$$

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

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{117}$$

  - 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

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]) d\boldsymbol{Z}_j} \tag{118}$$

  - 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

- 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

$$\ln q_j(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{119}$$

$$\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{120}$$

     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

変分推論

- MAP 推定の変分推論からの導出
    -

変分推論

- これまでの話の流れ
  1.
- これからの話の流れ
  - 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

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \tag{121}$$

  - 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

$$\ln q_j(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \mathrm{Const.} \tag{122}$$

  - 従って、$q_1^*(z_1)$ を計算する式は次のようになる

$$\ln q_1^*(z_1) = \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \mathrm{Const.} \tag{123}$$

$$\mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] = \int \ln p(\boldsymbol{z}) q_2(z_2) dz_2 \tag{124}$$

  - 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

$$
\begin{aligned}
\ln q_1^*(z_1) &= \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} && (125)\\
&= \mathbb{E}_{z_2}\left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\right] + \text{Const.} && (126)
\end{aligned}
$$

但し

$$
\begin{aligned}
&\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\\
&= \ln\left(\frac{1}{(2\pi)^{\frac{2}{2}}}\frac{1}{|\boldsymbol{\Lambda}^{-1}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\left(\boldsymbol{\Lambda}^{-1}\right)^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)\right)\\
&= \ln\left(\frac{1}{2\pi}\frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)\right)\\
&= -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) + \text{Const.} && (127)
\end{aligned}
$$

変分推論

z_1 に依存する項だけを取り出せば

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) \\
= \quad & -\frac{1}{2}\left[z_1-\mu_1, z_2-\mu_2\right]\left[\begin{array}{cc} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{array}\right]\left[\begin{array}{c} z_1-\mu_1 \\ z_2-\mu_2 \end{array}\right] \\
= \quad & -\frac{1}{2}\Bigg[\Big\{(\Lambda_{11}(z_1-\mu_1)+\Lambda_{21}(z_2-\mu_2)\Big\}(z_1-\mu_1)+ \\
& \quad \Big\{\Lambda_{12}(z_1-\mu_1)+\Lambda_{22}(z_2-\mu_2)\Big\}(z_2-\mu_2)\Bigg] \\
= \quad & -\frac{1}{2}\Big(\Lambda_{11}(z_1-\mu_1)^2+2\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+ \\
& \quad \Lambda_{22}(z_2-\mu_2)^2\Big) \quad (\because \Lambda_{21}=\Lambda_{12}) \\
= \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2-\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\mathrm{Const.}
\end{aligned} \tag{128}
$$

これを代入して

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ = \quad & -\frac{1}{2}\Lambda_{11}(z_1 - \mu_1)^2 - \Lambda_{12}(z_1 - \mu_1)(z_2 - \mu_2) + \text{Const.} \end{aligned} \tag{129}$$

従って

$$\begin{aligned} & \ln q_1^*(z_1) \\ = \quad & \mathbb{E}_{z_2}\left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\right] + \text{Const.} \\ = \quad & \mathbb{E}_{z_2}\left[-\frac{1}{2}\Lambda_{11}(z_1 - \mu_1)^2 - \Lambda_{12}(z_1 - \mu_1)(z_2 - \mu_2)\right] + \text{Const.} \\ = \quad & -\frac{1}{2}\Lambda_{11}z_1^2 + \Lambda_{11}\mu_1 z_1 - \Lambda_{12}z_1\left(\mathbb{E}[z_2] - \mu_2\right) + \text{Const.} \end{aligned} \tag{130}$$

変分推論

- これより、$q_1^*(z_1)$ は次のように書ける

$$
\begin{aligned}
& q_1^*(z_1) \\
\propto \quad & \exp\left(-\frac{1}{2}\Lambda_{11}z_1^2 + \Lambda_{11}\mu_1 z_1 - \Lambda_{12}z_1\left(\mathbb{E}[z_2]-\mu_2\right)\right) \\
= \quad & \exp\left(-\frac{1}{2}\Lambda_{11}\left(z_1 - (\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2]-\mu_2))\right)^2 + \right. \\
& \left. \frac{1}{2}\Lambda_{11}\left(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2]-\mu_2)\right)^2\right) \\
\propto \quad & \exp\left(-\frac{1}{2\Lambda_{11}^{-1}}\left(z_1 - (\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2]-\mu_2))\right)^2\right) \\
= \quad & \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1})
\end{aligned}
\tag{131}
$$

但し m_1 は次のようにおいた

$$
m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2]-\mu_2) \tag{132}
$$

変分推論

- 対称性から、$q_2^*(z_2)$ も次のように求められる

$$\begin{aligned} \ln q_2^*(z_2) &= \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} && (133) \\ &= \mathbb{E}_{z_2}[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})] + \text{Const.} && (134) \end{aligned}$$

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) \quad (135)$$

但し m_2 は次のようにおいた

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) \quad (136)$$

- これより $q_1^*(z_1), q_2^*(z_2)$ は

$$\begin{aligned} q_1^*(z_1) &= \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) && (137) \\ m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) && (138) \\ q_2^*(z_2) &= \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) && (139) \\ m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) && (140) \end{aligned}$$

- $\mathbb{E}[z_2]$ 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

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) \quad (141)$$

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) \quad (142)$$

であるから、m_1 について解けば

$$\begin{aligned} m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) && (143)\\ &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) - \mu_2\right) && (144)\\ &= \mu_1 + \Lambda_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) && (145)\\ &= \mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}(m_1 - \mu_1) && (146)\\ &= \left(1 - \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}m_1 && (147) \end{aligned}$$

従って

$$\left(1-\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1=\left(1-\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)m_1 \tag{148}$$

精度行列 $\mathbf{\Lambda}$ が正則であれば

$$\begin{aligned}
& \Lambda_{11}\Lambda_{12}-\Lambda_{12}\Lambda_{21}\neq 0 && (149)\\
\Rightarrow\quad & \Lambda_{11}\Lambda_{12}-\Lambda_{12}^{2}\neq 0 && (150)\\
\Rightarrow\quad & (\Lambda_{11}\Lambda_{12})\left(1-\Lambda_{11}^{-1}\Lambda_{12}^{-1}\Lambda_{12}^{2}\right)\neq 0 && (151)\\
\Rightarrow\quad & \left(1-\Lambda_{11}^{-1}\Lambda_{12}^{-1}\Lambda_{12}^{2}\right)\neq 0 && (152)
\end{aligned}$$

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

$$m_1=\mu_1 \tag{153}$$

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

$$m_2=\mu_2 \tag{154}$$

変分推論

- ゆえに、$q_1^*(z_1), q_2^*(z_2)$ は

$$q_1^*(z_1) = \mathcal{N}(z_1|\mu_1, \Lambda_{11}^{-1}) \quad (155)$$
$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, \Lambda_{22}^{-1}) \quad (156)$$

- 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

$$
\begin{aligned}
\mathrm{KL}(p||q) &\equiv \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X})||q(\boldsymbol{Z})) && (157)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} && (158)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln q(\boldsymbol{Z}) - \ln p(\boldsymbol{Z}|\boldsymbol{X})\right) d\boldsymbol{Z} && (159)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} - \\
&\quad - \underbrace{\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z}}_{q\ \text{には依存しない定数項}} && (160)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \prod_i q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \mathrm{Const.} && (161)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \sum_i \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \mathrm{Const.} && (162)
\end{aligned}
$$

変分推論

$$= -\sum_i \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \text{Const.} \tag{163}$$

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

$$\begin{aligned} & \sum_i \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} \\ = & \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z} && (164) \\ = & \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M && (165) \end{aligned}$$

変分推論

$$= \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \left(\prod_{i \neq j} d\boldsymbol{Z}_i \right) \quad (166)$$

$$= \int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \quad (167)$$

- $q_j(\boldsymbol{Z}_j)$ 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

$$\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j = 1 \quad (\text{規格化条件}) \quad (168)$$

$$q_j(\boldsymbol{Z}_j) \geq 0 \quad (169)$$

- 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

$$\begin{aligned} \mathcal{L}[q_j] &= -\int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \\ & \qquad \lambda \left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j - 1 \right) \end{aligned} \tag{170}$$

- 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

$$\frac{\delta}{\delta y(x)} \int G(y(x), x) dx = \frac{\partial}{\partial y} G(y(x), x) \tag{171}$$

- 従って

$$
\begin{aligned}
&\frac{\delta}{\delta q_j(\boldsymbol{Z}_j)}\mathcal{L}[q_j] \\
=\quad & -\frac{\delta}{\delta q_j(\boldsymbol{Z}_j)}\int\left(\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i\right)\ln q_j(\boldsymbol{Z}_j)d\boldsymbol{Z}_j + \\
& \quad \frac{\delta}{\delta q_j(\boldsymbol{Z}_j)}\lambda\left(\int q_j(\boldsymbol{Z}_j)d\boldsymbol{Z}_j - 1\right) \qquad (172) \\
=\quad & -\frac{\partial}{\partial q_j}\left(\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i\right)\ln q_j(\boldsymbol{Z}_j) + \\
& \quad \lambda\frac{\partial}{\partial q_j}q_j(\boldsymbol{Z}_j) \qquad (173) \\
=\quad & -\left(\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i\right)\frac{1}{q_j(\boldsymbol{Z}_j)} + \lambda = 0 \qquad (174)
\end{aligned}
$$

変分推論

- これより、未定乗数 λ は

$$\lambda q_j(\boldsymbol{Z}_j) = \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i\right) \frac{1}{q_j(\boldsymbol{Z}_j)} \tag{175}$$

$$\Rightarrow \quad \int \lambda q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j = \int \underbrace{\left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i\right)}_{=p(\boldsymbol{Z}_j|\boldsymbol{X})} d\boldsymbol{Z}_j \tag{176}$$

$$\Rightarrow \quad \lambda \underbrace{\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j}_{=1} = \underbrace{\int p(\boldsymbol{Z}_j|\boldsymbol{X}) d\boldsymbol{Z}_j}_{=1} \tag{177}$$

$$\Rightarrow \quad \lambda = 1 \tag{178}$$

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

$$-\left(\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i\right)\frac{1}{q_j(\boldsymbol{Z}_j)} + \lambda = 0 \tag{179}$$

$$\Rightarrow \quad q_j^*(\boldsymbol{Z}_j) = \underbrace{\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i}_{=p(\boldsymbol{Z}_j|\boldsymbol{X})} \tag{180}$$

$$\Rightarrow \quad q_j^*(\boldsymbol{Z}_j) = p(\boldsymbol{Z}_j|\boldsymbol{X}) \tag{181}$$

- $q_j^(\boldsymbol{Z}_j)$ 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

$$\begin{aligned} & q_1^*(z_1) \\ = & \int p(\boldsymbol{z}) dz_2 \\ = & \int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) dz_2 && (182) \\ = & \frac{1}{2\pi} \frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}} \int \exp\left(-\frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{z} - \boldsymbol{\mu}) \right) dz_2 && (183) \end{aligned}$$

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

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) \\
= & -\frac{1}{2}\left(\Lambda_{11}(z_1-\mu_1)^2 + \right. \\
& \left. 2\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2) + \Lambda_{22}(z_2-\mu_2)^2\right) \qquad (184) \\
= & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 + \\
& -\frac{1}{2}\left(2\Lambda_{12}(z_1-\mu_1)z_2 + \Lambda_{22}(z_2-\mu_2)^2\right) \qquad (185)
\end{aligned}
$$

そして

$$
\begin{aligned}
& -\frac{1}{2}\left(2\Lambda_{12}(z_1-\mu_1)z_2 + \Lambda_{22}(z_2-\mu_2)^2\right) \\
= & -\frac{1}{2}\left(\Lambda_{22}z_2^2 - 2\Lambda_{22}\mu_2 z_2 + 2\Lambda_{12}(z_1-\mu_1)z_2 + \Lambda_{22}\mu_2^2\right) \qquad (186)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\left(\Lambda_{22} z_2^2 - 2\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right) z_2 + \Lambda_{22}\mu_2^2\right) && (187)\\
&= -\frac{1}{2}\Big(\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)\right)^2 - \\
&\qquad \Lambda_{22}\left(\Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)\right)^2 + \Lambda_{22}\mu_2^2\Big) && (188)\\
&= -\frac{1}{2}\left(\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1} m\right)^2 - \Lambda_{22}^{-1} m^2 + \Lambda_{22}\mu_2^2\right) && (189)
\end{aligned}
$$

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

$$
\begin{aligned}
&-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\\
&= -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 - \\
&\qquad \frac{1}{2}\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1} m\right)^2 + \frac{1}{2}\Lambda_{22}^{-1} m^2 && (190)
\end{aligned}
$$

- これより、積分変数 z_2 の依存項だけを取り出せたので

$$
\begin{aligned}
&\int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)dz_2 \\
= \quad & \exp\left(-\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2+\Lambda_{12}(z_1-\mu_1)\mu_2-\right. \\
&\left.\frac{1}{2}\Lambda_{22}\mu_2^2+\frac{1}{2}\Lambda_{22}^{-1}m^2\right) \\
&\int \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2-\Lambda_{22}^{-1}m\right)^2\right)dz_2 \qquad (191)
\end{aligned}
$$

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

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

変分推論

$$
\begin{aligned}
&= (2\pi\Lambda_{22}^{-1})^{\frac{1}{2}} \cdot \int \frac{1}{(2\pi\Lambda_{22}^{-1})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2\right) dz_2 \\
&= \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}} \qquad (192)
\end{aligned}
$$

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

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

変分推論

$$
\begin{aligned}
& \frac{1}{2}\Lambda_{22}^{-1}\Lambda_{12}^{2}(z_1-\mu_1)^2 && (193)\\
= & -\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1-\mu_1)^2 && (194)\\
= & -\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)z_1^2+ \\
& \frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1 z_1+\mathrm{Const.} && (195)
\end{aligned}
$$

- これより結局、z_2 による積分は次のようになる

$$
\begin{aligned}
& \int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)dz_2 \\
= & \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1-\mu_1)^2\right) && (196)
\end{aligned}
$$

変分推論

- 従って、$q_1^*(z_1)$ は次のようになる

$$
\begin{aligned}
& q_1^*(z_1) \\
= \ & \frac{1}{2\pi} \frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}} \int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right) dz_2 \qquad (197) \\
= \ & \frac{1}{2\pi} \frac{1}{(\Lambda_{11}\Lambda_{22}-\Lambda_{12}^2)^{-\frac{1}{2}}} \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}} \\
& \quad \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) \qquad (198) \\
= \ & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2)^{-\frac{1}{2}}} \\
& \quad \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) \qquad (199) \\
= \ & \mathcal{N}(z_1|\mu_1, (\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2)^{-1}) \qquad (200)
\end{aligned}
$$

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

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \quad (\Sigma_{12} = \Sigma_{21}) \tag{201}$$

次が成り立つから

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \boldsymbol{I} \tag{202}$$

各成分に注目すれば

$$\begin{cases} \Sigma_{11}\Lambda_{11} + \Sigma_{12}\Lambda_{21} = 1 \\ \Sigma_{11}\Lambda_{12} + \Sigma_{12}\Lambda_{22} = 0 \end{cases} \tag{203}$$

これを Σ_{11} について解けば

$$\Sigma_{11}\Lambda_{11} + \left(-\Lambda_{22}^{-1}\Lambda_{12}\Sigma_{11}\right)\Lambda_{21} = 1 \tag{204}$$

$$\Rightarrow \quad \Sigma_{11}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}\Lambda_{21}\right) = 1 \tag{205}$$

変分推論

$$\Rightarrow \quad \Sigma_{11}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right) = 1 \tag{206}$$

$$\Rightarrow \quad \Sigma_{11} = \left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right)^{-1} \tag{207}$$

- これから、$q_1^*(z_1)$ は次のようにも書ける

$$\begin{aligned} & q_1^*(z_1) \\ = \; & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right)^{-\frac{1}{2}}} \\ & \qquad \exp\left(-\frac{1}{2}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1 - \mu_1)^2\right) && (208) \\ = \; & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\Sigma_{11}^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Sigma_{11}^{-1}(z_1 - \mu_1)^2\right) && (209) \\ = \; & \mathcal{N}(z_1 | \mu_1, \Sigma_{11}) && (210) \end{aligned}$$

変分推論

- $q_2^*(z_2)$ は、対称性から次のようになる

$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, (\Lambda_{22} - \Lambda_{11}^{-1}\Lambda_{12}^2)^{-1}) = \mathcal{N}(z_2|\mu_2, \Sigma_{22}) \tag{211}$$

- 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

$$q_1^*(z_1) = \mathcal{N}(z_1|\mu_1, \Lambda_{11}^{-1}) \tag{212}$$
$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, \Lambda_{22}^{-1}) \tag{213}$$

変分推論

- $\mathrm{KL}(p||q)$ の最小化では、次の解を得た

$$\begin{aligned} q_1^*(z_1) &= \mathcal{N}(z_1|\mu_1, \Sigma_{11}) && (214) \\ \Sigma_{11} &= \left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^2\right)^{-1} && (215) \\ q_2^*(z_2) &= \mathcal{N}(z_2|\mu_2, \Sigma_{22}) && (216) \\ \Sigma_{22} &= (\Lambda_{22} - \Lambda_{11}^{-1}\Lambda_{12}^2)^{-1} && (217) \end{aligned}$$

- $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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**??**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 $\mathrm{KL}(p\|q)$.

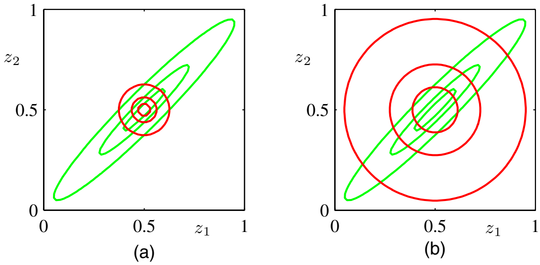

図 6: KL ダイバージェンスの 2 つの形の比較

変分推論

- 2 つの解の比較
  - $\mathrm{KL}(q||p)$ の最小化で得られる $q(z)$ 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 $q(z)$ 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

- 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

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{218}$$

変分推論

- $\mathrm{KL}(q||p)$ 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

$$\mathrm{KL}(p||q) = -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} \tag{219}$$

- $\mathrm{KL}(p||q)$ 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**??**に示す

変分推論

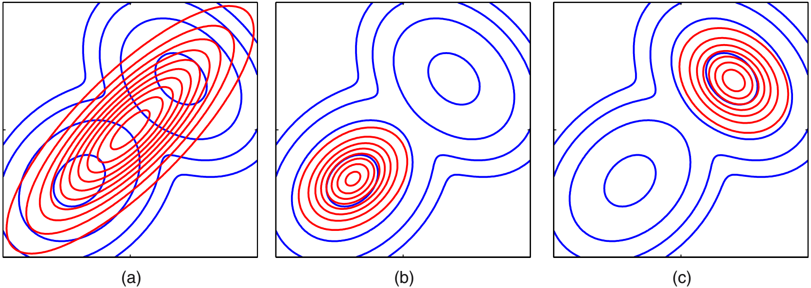

Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ 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 7: 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
  -