変分推論の理論

宮本　崇

2018/5/2

本資料は，ベイズの定理などの基本的な確率計算を既知として，確率モデリングされた問題のベイズ推論による解き方と，ベイズ推論の近似手法であり MAP 推定・最尤推定の一般化である変分推論の理論を参考文献 [1] に従って整理する．特に，ベイズ推論に関するキーワードを，確率モデリングや変分推論といった方法論に位置づけられる概念と，共役事前分布や平均場近似・勾配法による KL ダイバージェンスの最小化など，モデリングされた問題を実際に解析的/近似的に解くための具体的なテクニックに位置づけられる概念の区別に主眼をおき，ここでは主に変分推論に至るまでの方法論に関する部分を詳述する．

1　ベイズ推論

1.1　ベイズ推論の例題

ベイズ推論を説明するためによく用いられる例題として，次のようなものがある．

- 袋 A には赤球が 3 つと白球が 1 つ，袋 B には赤球が 1 つと白球が 3 つ入っている．今，A か B か分からない袋を 1 つ選び，選択した袋から球を 1 個取り出して袋に戻す操作を 4 回繰り返したところ，出現した球の色は順に { 赤，白，赤，赤 } となった．このとき，最初に選んだ袋は A と B のどちらと思われるか．

直感的には，赤の方が多く出たため袋は赤球が多く入っている袋 A が選ばれたように予想される．あるいは，もう少し確率的に

1. 袋 A を選んだときに { 赤，白，赤，赤 } となる確率：$\frac{3}{4}\times\frac{1}{4}\times\frac{3}{4}\times\frac{3}{4}=\frac{27}{256}$
2. 袋 B を選んだときに { 赤，白，赤，赤 } となる確率：$\frac{1}{4}\times\frac{3}{4}\times\frac{1}{4}\times\frac{1}{4}=\frac{3}{256}$

と計算し，「観測した結果が起こる確率は袋 A を選んだ時の方が高いので，選ばれた袋 A と思われる」と答える，最尤推定の考え方を用いることもできる．

これらに比較して，ベイズ推論では「袋 A（もしくは B）が選ばれた確率」を，観測データに基づいてベイズの定理に従って計算する．袋 A（もしくは B）が選ばれたことを $x=A$（もしくは $x=B$），k 回目に赤が出たことを $y_k=r$，白が選ばれたことを $y_k=w$ と表すこととすると，観測結果 $\boldsymbol{y}=\{y_1,y_2,y_3,y_4\}=\{r,w,r,r\}$ の下で各袋が選ばれたと思われる確率は

$$p(x=A|\boldsymbol{y})=\frac{p(\boldsymbol{y}|x=A)\cdot p(x=A)}{p(\boldsymbol{y})}=\frac{(\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4})\cdot\frac{1}{2}}{\sum_{E=A,B}p(\boldsymbol{y}|x=E)\cdot p(x=E)}=\frac{\frac{27}{256}\cdot\frac{1}{2}}{\frac{27}{256}\cdot\frac{1}{2}+\frac{3}{256}\cdot\frac{1}{2}}=\frac{9}{10}\quad(1)$$

$$p(x=B|\boldsymbol{y})=\frac{p(\boldsymbol{y}|x=B)\cdot p(x=B)}{p(\boldsymbol{y})}=\frac{(\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4})\cdot\frac{1}{2}}{\sum_{E=A,B}p(\boldsymbol{y}|x=E)\cdot p(x=E)}=\frac{\frac{3}{256}\cdot\frac{1}{2}}{\frac{27}{256}\cdot\frac{1}{2}+\frac{3}{256}\cdot\frac{1}{2}}=\frac{1}{10}\quad(2)$$

となる．ただし，上式の式変形においてはベイズの定理

$$p(\boldsymbol{x}|\boldsymbol{y})=\frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}=\frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int p(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}}=\frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}}\quad(3)$$

を用いており，最初にどちらの袋を選んだかはランダムだとして $p(x=A)=p(x=B)=\frac{1}{2}$ であると**設定している**.

したがって，事前分布 $p(x=A)=p(x=B)=\frac{1}{2}$ という設定と観測結果 $\boldsymbol{y}$ に基づくと，袋 A が選ばれていた確率の方が 9 倍高く，袋 A が選ばれたと結論付ける方が妥当であると考えられる.

1.2 確率モデリング

上記の例題は，ベイズの定理を用いた計算の典型的な例題であるが，ある観測データ $\boldsymbol{D}$（上記の例題では球を 4 回取り出した結果 $\boldsymbol{y}$）に基づいて，求めたい別のパラメタ $\boldsymbol{\theta}$（上記の例題では x）を決定的な値（$x=A$ or $x=B$）として求めるのでなくその確率分布 $p(\boldsymbol{\theta})$（$p(x=A)$ および $p(x=B)$）を求めようとする，問題設定そのものが特徴的であり，そのような問題のモデル化の考え方を**確率モデリング**と呼ぶ．確率モデリングされた問題において実際に変数の確率分布を求める際に，ベイズ推論では事前分布 $p(\boldsymbol{\theta})$ を設定したうえで，データ $\boldsymbol{D}$ が観測された後の事後分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ をベイズの定理に従って求めている.

確率モデリングでは，ある観測データ $\boldsymbol{D}$ に基づいて，求めたい別のパラメタ $\boldsymbol{\theta}$ の確率分布 $p(\boldsymbol{\theta})$ を求める問題を設定する．問題を抽象化してパラメタの確率分布を設定することによって，例えば以下のような問題も確率モデリングによって定式化することができる.

- 確率分布のパラメータ推定：ある確率分布 $p(\boldsymbol{x}|\boldsymbol{\theta})$ に従って生起した n 個のサンプル $\boldsymbol{X}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$ から，元の分布のパラメータ $\boldsymbol{\theta}$ の値を推定する問題
- 線形回帰：入力値 $\boldsymbol{X}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$ と出力値 $\boldsymbol{Y}=\{\boldsymbol{Y_1},\cdots,\boldsymbol{Y_n}\}$ から，入出力関係を表す線形モデル $\boldsymbol{y}=\boldsymbol{w}^T\boldsymbol{x}$ の係数 $\boldsymbol{w}$ を求める問題
- ニューラルネットワーク：入力値 $\boldsymbol{X}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$ と出力値 $\boldsymbol{Y}=\{\boldsymbol{Y_1},\cdots,\boldsymbol{Y_n}\}$ から，入出力関係を再現するニューラルネットワーク $\boldsymbol{y}=f(\boldsymbol{x}|\boldsymbol{\theta})$ のモデルパラメタ $\boldsymbol{\theta}$ を求める問題

上記の問題はそれぞれ，最尤推定法や最小二乗法，勾配降下法などの決定論的なアルゴリズムを用いて解くことによって，求めるパラメタを確定値として得ることも可能である．こうした問題に対してあえて確率的な問題設定を行い，変数の確率分布を求める手順を経ることが確率モデリングのアプローチであり，問題の不確実性や解の信頼度を定量的に表現できるなどのメリットを有している.

1.3 ベイズ推論における求解の手順

確率モデリングされた問題をベイズ推論によって解く手順は，以下のように整理される.

1. 観測データ $\boldsymbol{D}$ と求めたいパラメタ $\boldsymbol{\theta}$ の同時確率分布 $p(\boldsymbol{D},\boldsymbol{\theta})$ を問題に応じてモデリングする
2. 事前情報などに応じて事前確率分布 $p(\boldsymbol{\theta})$ を設定する
3. ベイズの定理に基づいて事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ を求める.

$$p(\boldsymbol{\theta}|\boldsymbol{D})=\frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{D},\boldsymbol{\theta})d\boldsymbol{\theta}}=\frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{4}$$

4.（必要な場合）未観測のデータ $\boldsymbol{D}_{\mathrm{new}}$ に関する予測分布を求める.

$$p(\boldsymbol{D}_{\mathrm{new}})=\int p(\boldsymbol{D}_{\mathrm{new}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{D})d\boldsymbol{\theta} \tag{5}$$

以下に，それぞれの手順を詳細に解説する.

1.3.1 同時分布のモデリング

1. におけるデータとパラメタの同時分布のモデリングとは，パラメタとデータの関係性を具体的に定式化することを意味している．例として，あるデータ D が平均 μ，分散 σ の正規分布から生成される問題で，求めたいパラメタが μ であれば，同時分布は

$$p(D,\mu)=\mathcal{N}(D|\mu,\sigma)p(\mu) \tag{6}$$

としてモデル化される（ただし，この段階ではまだ事前分布 $p(\mu)$ は具体化されていない.）

別の例として，入力ベクトル $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ が線形変換によって出力スカラー $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ に変換される線形回帰の問題を考える．線形変換時には，平均 0，分散 σ の正規分布に従うわずかなノイズ ϵ が出力に乗るとし，求めたいパラメタは線形変換の係数 $\boldsymbol{w}$ であるとすると，同時分布は

$$p(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) = \Pi_{k=1}^{n} \mathcal{N}(y_k | \boldsymbol{w}\boldsymbol{x}_k, \sigma) p(\boldsymbol{w}) \tag{7}$$

とモデル化することができる（ここでも，同様にまだ事前分布 $p(\boldsymbol{w})$ は具体化されていない.）なお，ここでは分かりやすさのために $\boldsymbol{X}$ と $\boldsymbol{y}$ を分けて記述しているが，両者はともに観測されるデータなので $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ とまとめて書くこともできる．したがって，ベイズ推論で求めるべき事後確率分布は $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$ となる．

上の 2 つの例のように，同時分布のモデリングとは，物理現象を数式を用いてモデリングするように，観測データとパラメタの関係性を数式や確率の考え方を用いてモデリングすることを意味している．

1.3.2 事前確率分布の設定

2. における事前確率分布 $p(\boldsymbol{\theta})$ は，自動的に決まるものではなく設定に任意性を有している．どのような確率分布の形状を設定するか，その確率分布のパラメタの初期値をどのようにするか，自由に設定することができる．

この事前確率分布の設定には，パラメタについて元から知っていたことを数理的に反映させることができるため，事前知識を明示的に表現できるという意味で事前確率分布の設定の任意性はメリットとなる．もし，パラメタの分布について何も情報がなければ無情報分布と呼ばれる確率分布を用いることができる．冒頭の例題に即すと，袋 A と袋 B を選ぶ事前確率にそれぞれ $\frac{1}{2}$ を設定することは，どちらが選ばれやすいかは完全に分からない無情報分布を設定することに相当する．あるいは，A の方が若干選ばれやすいことが例えば過去の経験から知られているようなケースであれば，$p(x = A) = \mu > \frac{1}{2}$ と設定することにより，そのような知識を反映させることができる．

モデリングされた問題を実際に解くときは，次のステップである事後確率分布の計算が解析的に行えるように都合のよい事前分布を設定したり，パラメタが取りうる値の範囲に応じて決まる自然な確率分布形状を設定したりすることによって，問題を扱いやすくする．そのような設定の下でも，十分な数の観測データの下では得られる事後確率分布の形状は似たものとなっていき，妥当な結論を得ることができる．

1.3.3 事後確率分布の計算

ベイズの定理に従って事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ を求めるプロセスは，冒頭の例のような簡単な例であれば直接計算することができるが，実際の問題では式内の積分が解析的に計算できなかったり，あるいは計算量が膨大で現実的でなかったりすることが多い．そのため，この計算を解析的/効率的に行うためのテクニックが様々に開発されており，**共役事前分布**の利用や**ギブスサンプリング**，**変分近似**による近似解の導出はそうしたテクニックに位置づけることができる．

1.3.4 予測分布の導出

予測分布の導出は，観測データ $\boldsymbol{D}$ によって更新された事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ とモデリングした同時分布 $p(\boldsymbol{D}, \boldsymbol{\theta})$ を用いて未知の観測データ $\boldsymbol{D}_{\mathrm{new}}$ に関する知見を確率分布 $p(\boldsymbol{D}_{\mathrm{new}})$ として得ることを意味している．内部パラメタに関する知見を得ることが目的となる場合はこのプロセスは行わないが，線形回帰やニューラルネットワークのように未知の入力データに対する出力をモデルから得たい場合は，出力が取りうる値の範囲や確率を確率分布として得ることができる．

2　変分近似と変分推論

2.1　変分推論による求解

パラメタと観測可能なデータの関係が複雑なモデルでは，ベイズの定理式 (4) が解析的に計算できず，求めたい事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ の解析解が求められない場合が多くある．そのような場合において解を近似的に求める手法の 1 つとして**変分近似**がある．

求めたいパラメタ $\boldsymbol{\theta}$ に関する確率分布 $q(\boldsymbol{\theta})$ を考え，この確率分布によって本来の事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ を近似することを考える．このとき，最良の近似関数 $q_{\mathrm{opt}}(\boldsymbol{\theta})$ は次の変分問題を解くことで得ることができる．

$$q_{\mathrm{opt}}(\boldsymbol{\theta}) = \arg\min_{q} \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{D})] \tag{8}$$

ただし，$\mathrm{KL}[q(\boldsymbol{x})||p(\boldsymbol{x})]$ は，次式で表される KL ダイバージェンスである．

$$\begin{aligned}\mathrm{KL}[q(\boldsymbol{x})||p(\boldsymbol{x})] &= \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} \\ &= < \ln q(\boldsymbol{x}) >_{q(\boldsymbol{x})} - < \ln p(\boldsymbol{x}) >_{q(\boldsymbol{x})}\end{aligned} \tag{9}$$

ここで，$< f(\boldsymbol{x}) >_{q(\boldsymbol{x})}$ は，関数 $f(\boldsymbol{x})$ に関する確率分布 $q(\boldsymbol{x})$ についての期待値を表している．KL ダイバージェンスは，確率分布間の相違度を定量化した量であり，確率分布間の一種の距離を表している．

式 (8) には本来求めたい $p(\boldsymbol{\theta}|\boldsymbol{D})$ が表れており，近似解を求める計算に本来の解が含まれているのは不思議に見えるが，後述する**平均場近似**の例のように近似確率分布 $q(\boldsymbol{\theta})$ をうまく置くことによって，事後確率分布を陽に用いることを回避できる．

式 (8) の変分問題の解 $q_{\mathrm{opt}}(\boldsymbol{\theta})$ によって本来の事後確率分布を近似することを**変分近似**と呼び，確率モデリングされた問題を変分近似によって解くことを**変分推論**と呼ぶ．

2.2　平均場近似による変分推論の例

変分近似の一例として，ここでは**平均場近似**と呼ばれる方法を説明する．

事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ の近似分布として，$\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_n\}$ の n 個のパラメタが互いに独立であることを仮定した $q(\boldsymbol{\theta}) = q_1(\theta_1) \cdot q_2(\theta_2) \cdots \cdot q_n(\theta_n)$ という分布をおく．この分布を変分問題 (8) を解くことによって求める．このような，パラメタ間の独立性を仮定した近似を**平均場近似**と呼ぶ．

式 (8) を解くために, 今, $q_i(\theta_i)$ 以外の確率分布が所与で固定されているとし, $q_i(\theta_i)$ のみを最適化する場合を考える．$\boldsymbol{\theta}$ から変数 θ_i のみを除いた集合を $\boldsymbol{\theta}_{\setminus i}$ とおき，$q_1(\theta_1) \cdots q_{i-1}(\theta_{i-1}) \cdot q_{i+1}(\theta_{i+1}) \cdots q_n(\theta_n) = q(\boldsymbol{\theta}_{\setminus i})$ と表すこととすると，変分問題は次のように式変形できる．

$$
\begin{aligned}
q_(\theta_i) &= \arg\min_{q_i} \mathrm{KL}[q_i(\theta_i)\cdot q(\boldsymbol{\theta}_{\setminus i})||p(\boldsymbol{\theta}|\boldsymbol{D})] \\
&= \arg\min_{q_i} \left\langle \ln \frac{q_i(\theta_i)\cdot q(\boldsymbol{\theta}_{\setminus i})}{p(\boldsymbol{\theta}|\boldsymbol{D})} \right\rangle_{q(\boldsymbol{\theta})} \\
&= \arg\min_{q_i} \left\langle \left\langle \ln \frac{q_i(\theta_i)\cdot q(\boldsymbol{\theta}_{\setminus i})}{p(\boldsymbol{\theta}|\boldsymbol{D})} \right\rangle_{q(\boldsymbol{\theta}_{\setminus i})} \right\rangle_{q_i(\theta_i))} \\
&= \arg\min_{q_i} \left\langle \langle \ln q_i(\theta_i)\rangle_{q(\boldsymbol{\theta}_{\setminus i})} + \left\langle \ln q(\boldsymbol{\theta}_{\setminus i})\right\rangle_{q(\boldsymbol{\theta}_{\setminus i})} - \langle \ln p(\boldsymbol{\theta}|\boldsymbol{D})\rangle_{q(\boldsymbol{\theta}_{\setminus i})} \right\rangle_{q_i(\theta_i)} \\
&= \arg\min_{q_i} \left\langle \ln q_i(\theta_i) - \langle \ln p(\boldsymbol{\theta}|\boldsymbol{D})\rangle_{q(\boldsymbol{\theta}_{\setminus i})} \right\rangle_{q_i(\theta_i)} + \text{const.} \\
&= \arg\min_{q_i} \left\langle \ln \frac{q_i(\theta_i)}{\frac{\exp\left\{<\ln p(\boldsymbol{\theta}|\boldsymbol{D})>_{q(\boldsymbol{\theta}_{\setminus i})}\right\}}{C}} - \ln C \right\rangle_{q_i(\theta_i)} + \text{const.} \\
&= \arg\min_{q_i} \mathrm{KL}\left[q_i(\theta_i) || \frac{\exp\left\{< \ln p(\boldsymbol{\theta}|\boldsymbol{D}) >_{q(\boldsymbol{\theta}_{\setminus i})}\right\}}{C} \right] \qquad (10)
\end{aligned}
$$

ただし, C は $\frac{\exp\left\{<\ln p(\boldsymbol{\theta}|\boldsymbol{D})>_{q(\boldsymbol{\theta}_{\setminus i})}\right\}}{C}$ が確率分布となるような正規化係数であり, $C = \int \exp\left\{< \ln p(\boldsymbol{\theta}|\boldsymbol{D}) >_{q(\boldsymbol{\theta}_{\setminus i})}\right\} d\theta_i$ である.

したがって，式 (10) を最小化する $q_i(\theta_i)$ は以下のように得られる.

$$
\begin{aligned}
q_i(\theta_i) &= \frac{\exp\left\{< \ln p(\boldsymbol{\theta}|\boldsymbol{D}) >_{q(\boldsymbol{\theta}_{\setminus i})}\right\}}{C} \qquad (11) \\
\therefore \ln q_i(\theta_i) &= < \ln p(\boldsymbol{\theta}|\boldsymbol{D}) >_{q(\boldsymbol{\theta}_{\setminus i})} + \text{const.} \\
&= \left\langle \ln \frac{p(\boldsymbol{D},\boldsymbol{\theta})}{p(\boldsymbol{D})} \right\rangle_{q(\boldsymbol{\theta}_{\setminus i})} + \text{const.} \\
&= \langle \ln p(\boldsymbol{D},\boldsymbol{\theta})\rangle_{q(\boldsymbol{\theta}_{\setminus i})} - \langle p(\boldsymbol{D})\rangle_{q(\boldsymbol{\theta}_{\setminus i})} + \text{const.} \\
&= \langle \ln p(\boldsymbol{D},\boldsymbol{\theta})\rangle_{q(\boldsymbol{\theta}_{\setminus i})} + \text{const.} \qquad (12)
\end{aligned}
$$

式 (12) における定数部分は確率分布を正規化する項であり，1 項目の期待値計算を解析的に行うことによって $q_i(\theta_i)$ の具体的な分布形状が得られる．$q_i(\theta_i)$ を式 (12) によって更新する手続きを，i を変更しながら KL ダイバージェンスが収束するまで十分な回数繰り返すことによって，平均場近似による変分推論の解を得ることができる.

なお，$q(\boldsymbol{\theta}) = q_1(\theta_1)\cdot q_2(\theta_2)\cdot\cdots\cdot q_n(\theta_n)$ の $q_i(\theta_i)$ 以外の関数を固定し，$q_i(\theta_i)$ を最適化する手続きを繰り返すことによって目的の関数 (すなわち，KL ダイバージェンス $\mathrm{KL}[q||p]$) を最適化する手続きは**座標降下法**と呼ばれている他，モデルにおける潜在変数とパラメタに独立性を仮定して同様の定式化を行うことで得られるアルゴリズムは**拡張 EM アルゴリズム**や**変分 EM アルゴリズム**と呼ばれる.

2.3 尤度下限最大化としての変分推論

上記のように，変分推論とは事後確率分布 $p(\boldsymbol{\theta}|\boldsymbol{D})$ を KL ダイバージェンスの意味で最良近似する確率分布 $q(\boldsymbol{\theta})$ を求める問題として定式化・解釈される．一方で，変分推論を周辺尤度の下限値を最大化する確率分布を求める問題として定式化されることもあるため，本節ではそのような定式化の手続きとその解釈を説明する.

2.3.1 変分問題の定式化

ある確率モデル $p(\boldsymbol{D},\boldsymbol{\theta})$ と観測データ $\boldsymbol{D}$ が与えられた時，この観測データの周辺尤度について次のような式が成り立つ.

$$
\begin{aligned}
\ln p(\boldsymbol{D}) &= \ln \int p(\boldsymbol{D}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \ln \int q(\boldsymbol{\theta}) \frac{p(\boldsymbol{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\geq \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \mathcal{L}[q(\boldsymbol{\theta})]
\end{aligned} \tag{13}
$$

ここで 3 行目の不等号は，上に凸な関数 $f(\cdot)$ と確率分布 $p(x)$ に関する以下のイェンセンの不等式から成立している．

$$
f(\int y(x)p(x)dx) \geq \int f(y(x))p(x)dx \tag{14}
$$

式 (13) より，$\mathcal{L}(q(\boldsymbol{\theta}))$ は周辺尤度（エビデンスとも呼ばれる）の対数値の下限となっているため，$q(\boldsymbol{\theta})$ に対する **ELBO(evidence lower bound)** や**変分下限 (variational lower bound, VLB)** と呼ばれている．

次に，周辺尤度の対数と ELBO の差分を以下のように計算する．

$$
\begin{aligned}
\ln p(\boldsymbol{D}) - \mathcal{L}[q(\boldsymbol{\theta})] &= \ln p(\boldsymbol{D}) - \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \ln p(\boldsymbol{D}) - \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{D})p(\boldsymbol{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \ln p(\boldsymbol{D}) - \int q(\boldsymbol{\theta}) \left\{ \ln p(\boldsymbol{\theta}|\boldsymbol{D}) + \ln p(\boldsymbol{D}) - \ln q(\boldsymbol{\theta}) \right\} d\boldsymbol{\theta} \\
&= \ln p(\boldsymbol{D}) - \int \ln p(\boldsymbol{D}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \left\{ \ln q(\boldsymbol{\theta}) - \ln p(\boldsymbol{\theta}|\boldsymbol{D}) \right\} d\boldsymbol{\theta} \\
&= \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{D})]
\end{aligned} \tag{15}
$$

上式より，周辺尤度の対数と ELBO の差分は $q(\boldsymbol{\theta})$ と $p(\boldsymbol{\theta}|\boldsymbol{D})$ の間の KL ダイバージェンスとなっていることが分かる．また，対数尤度 $\ln p(\boldsymbol{D})$ は $q(\boldsymbol{\theta})$ **に依存しない一定値を取る**ため，上式の KL ダイバージェンスを最小化する $q(\boldsymbol{\theta})$ は同時に ELBO を最大化する．

$$
\begin{aligned}
q_{\mathrm{opt}}(\boldsymbol{\theta}) &= \arg\max_q \mathcal{L}[q(\boldsymbol{\theta})] \\
&= \arg\max_q \left\{ \ln p(\boldsymbol{D}) - \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{D})] \right\} \\
&= \arg\min_q \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{D})]
\end{aligned} \tag{16}
$$

式 (16) は式 (8) と同じ形であり，以上のように変分推論を周辺尤度の下限を最大化する変分問題として定式化することができる．また，解析力学などにおける定理の再定式化と同様に，ベイズ推論を汎関数の最小化問題として定式化するという発想の下で

$$
-\mathcal{L}[q(\boldsymbol{\theta})] = \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{D}, \boldsymbol{\theta})} d\boldsymbol{\theta} \tag{17}
$$

を**変分自由エネルギー**と呼び，これを最小化する $q(\boldsymbol{\theta})$ を求める問題として変分推論が説明されることもある．

2.3.2 変分問題の解釈

ELBO を最大化する $q(\boldsymbol{\theta})$ を用いることによって，周辺尤度を ELBO で近似することができる．周辺尤度の近似値として計算された ELBO の大小は確率モデルが複数設定されうるときに，モデルの良し悪しを選択する定量的な基準として用いることができる．

ただし，ここで変分問題が最大化しているものは周辺尤度 $p(\boldsymbol{D})$ の下限値であり，周辺尤度そのものを最大化しようとしていないことに注意する必要がある．この点は，最尤推定の概念，特に EM アルゴリズムとの類推から変分推論の意味を理解しようとしたときに混乱を招く部分である（この変分問題の本質的な意味を，2.4.2 節で詳述する．）

これまでにも述べたように，周辺尤度は現象のモデル化としての $p(\boldsymbol{D}, \boldsymbol{\theta})$ と事前分布 $p(\boldsymbol{\theta})$ を設定すると一意に決まるものであり，そもそも $q(\boldsymbol{\theta})$ には依存していない．したがって，ELBO を最大化する $q(\boldsymbol{\theta})$ を求めても，モデルによる観測データの周辺尤度そのものは変化していない．

変分推論による $q(\boldsymbol{\theta})$ を用いて求めた周辺尤度の近似値 $p(\boldsymbol{D}) \approx \mathcal{L}[q(\boldsymbol{\theta})]$ は，近似分布 $q(\boldsymbol{\theta})$ の良し悪しの尺度ではなく，予め設定した確率モデル $p(\boldsymbol{D}, \boldsymbol{\theta})$ と事前分布 $p(\boldsymbol{\theta})$ の組がどれくらい観測データにフィットしているかという尺度となる．また，近似事後分布を構成した後に，新しいデータに対する尤度を近似的に求める場合にも ELBO が計算される．

2.3.3 例題を通した変分問題の意味の確認

上記のような，ELBO の最大化としての変分問題の意味を，具体例を通して確認してみる．再び冒頭の例に戻り，赤球 3 つと白球 1 つの入った袋 A と，赤球 1 つと白球 1 つの入った袋 B のどちらが選ばれたか分からない状況で観測結果 $\boldsymbol{D} = \{r, w, r, r\}$ が観測されたとき，袋 A と袋 B のどちらが選ばれたか，その確率を推定する問題を考える．これは，事前確率を冒頭と同様に $p(\theta = A) = p(\theta = B) = \frac{1}{2}$ と設定し，求めたい確率分布を

$$q(\theta) = \begin{cases} \mu \ (\theta = A) \\ 1 - \mu \ (\theta = B) \end{cases} \tag{18}$$

として，μ を $0 \leq \mu \leq 1$ の範囲で ELBO を最大化（すなわち KL ダイバージェンスを最小化）する変分問題と定式化することができる．確率分布 $q(\theta)$ は変数 μ にのみ依存しているため，この変分問題は実際には μ に関する最適化問題となる．

はじめに，計算に必要な各種の確率を以下のように整理しておく．

$$\begin{aligned} p(\theta) : p(\theta = A) &= \frac{1}{2} \\ p(\theta = B) &= \frac{1}{2} \end{aligned} \tag{19}$$

$$p(\boldsymbol{D}|\theta) : p(\boldsymbol{D}|\theta = A) = \frac{3}{4} \times \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{256}$$

$$p(\boldsymbol{D}|\theta = B) = \frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{3}{256} \tag{20}$$

$$p(\boldsymbol{D}) : p(\boldsymbol{D}) = \sum_{\theta} p(\boldsymbol{D}|\theta) p(\theta) = \frac{27}{256} \cdot \frac{1}{2} + \frac{3}{256} \cdot \frac{1}{2} = \frac{15}{256} \tag{21}$$

$$p(\theta|\boldsymbol{D}) : p(\theta = A|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta = A) \cdot p(\theta = A)}{p(\boldsymbol{D})} = \frac{9}{10}$$

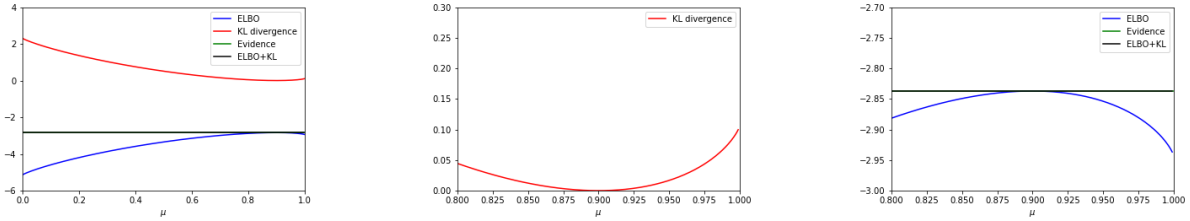
$$p(\theta|\boldsymbol{D}) : p(\theta = B|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta = B) \cdot p(\theta = B)}{p(\boldsymbol{D})} = \frac{1}{10} \tag{22}$$

以上のもとで，このモデルにおける変分近似 $q(\theta)$ に対する周辺尤度下限 $\mathcal{L}[q(\theta)]$，KL ダイバージェンス $\mathrm{KL}[q(\theta)||p(\theta|\boldsymbol{D})]$ を定義に従って計算すると以下のようになる．

$$\begin{aligned} \mathcal{L}[q(\theta)] &= \sum_{\theta} q(\theta) \cdot \ln \frac{p(\boldsymbol{D}, \theta)}{q(\theta)} \\ &= q(\theta = A) \cdot \ln \frac{p(\boldsymbol{D}, \theta = A)}{q(\theta = A)} + q(\theta = B) \cdot \ln \frac{p(\boldsymbol{D}, \theta = B)}{q(\theta = B)} \\ &= \mu \cdot \ln \frac{\frac{27}{256}}{\mu} + (1 - \mu) \cdot \ln \frac{\frac{3}{256}}{1 - \mu} \end{aligned} \tag{23}$$

$$\begin{aligned} \mathrm{KL}[q(\theta)||p(\theta|\boldsymbol{D})] &= \sum_{\theta} q(\theta) \cdot \ln \frac{q(\theta)}{p(\theta|\boldsymbol{D})} \\ &= q(\theta = A) \cdot \ln \frac{q(\theta = A)}{p(\theta = A|\boldsymbol{D})} + q(\theta = B) \cdot \ln \frac{q(\theta = B)}{p(\theta = B|\boldsymbol{D})} \\ &= \mu \cdot \ln \frac{\mu}{\frac{9}{10}} + (1 - \mu) \cdot \ln \frac{1 - \mu}{\frac{1}{10}} \end{aligned} \tag{24}$$

このような $\mathcal{L}[q(\theta)]$，$\mathrm{KL}[q(\theta)||p(\theta|\boldsymbol{D})]$ を，$0 \leq \mu \leq 1$ の範囲で実際に計算したものを図 1 に示す．同図には，$\ln p(\boldsymbol{D}) = \mathcal{L}[q(\theta)] + \mathrm{KL}[q(\theta)||p(\theta|\boldsymbol{D})]$ として計算される周辺尤度と，式 (21) のように直接求めた周辺尤度の値 $\ln p(\boldsymbol{D}) = \ln \frac{15}{256}$ を同時にプロットしている．図からは，$\mu = 0.9$，すなわち変分近似した $q(\theta)$ が事後確率分布に一致するときに KL ダイバージェンスが最小値の 0 となり，このときに ELBO が最大化され周辺尤度に一致していることが分かる．また，ELBO と KL ダイバージェンスの和は常に一定値で，モデルの周辺尤度と等しいことが確認できる．



(a) 全体図　(b)$\mu = 0.9$ 付近：KL ダイバージェンス　(c)$\mu = 0.9$ 付近：ELBO

2.4 MAP 推定・最尤推定と変分推論

ベイズ推論以外に確率モデルのパラメタを求める方法として，最尤推定と MAP 推定が知られている．ここでは，これらの手法が変分推論の特殊なケースとして定式化されることを確認する．

2.4.1 KL ダイバージェンスからの導出

今，変分近似に用いる確率分布として，クロネッカーのデルタ関数 $q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ という確率分布を考える．これは，特定の値 $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ で無限に鋭い確率分布を考えていることに相当する．このとき，近似分布と事後分布の間の KL ダイバージェンスは

$$\begin{aligned}
\mathrm{KL}[\delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})||p(\boldsymbol{\theta}|\boldsymbol{D})] &= \int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln \frac{\delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}|\boldsymbol{D})} d\boldsymbol{\theta} \\
&= \int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) d\boldsymbol{\theta} - \int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln p(\boldsymbol{\theta}|\boldsymbol{D}) d\boldsymbol{\theta} \\
&= \ln \delta(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - \ln p(\hat{\boldsymbol{\theta}}|\boldsymbol{D}) \\
&= \ln \delta(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - \ln \frac{p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) p(\hat{\boldsymbol{\theta}})}{p(\boldsymbol{D})} \\
&= -(\ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) + \ln p(\hat{\boldsymbol{\theta}})) + \mathrm{const.}
\end{aligned} \tag{25}$$

と変形できる．ただし，$q(\boldsymbol{\theta})$ によらない部分は定数項 const. と表現している．$q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ はパラメタ $\hat{\boldsymbol{\theta}}$ にのみ依存する関数であり，したがって上記の KL ダイバージェンスを最小化する $\hat{\boldsymbol{\theta}}$ を求めることが，変分近似した確率分布を求めることに等しい．

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\hat{\boldsymbol{\theta}}} -(\ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) + \ln p(\hat{\boldsymbol{\theta}})) + \mathrm{const.} \\
&= \arg\max_{\hat{\boldsymbol{\theta}}} \ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) + \ln p(\hat{\boldsymbol{\theta}}) \\
&= \arg\max_{\hat{\boldsymbol{\theta}}} \ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) p(\hat{\boldsymbol{\theta}}) \\
&= \arg\max_{\hat{\boldsymbol{\theta}}} p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) p(\hat{\boldsymbol{\theta}})
\end{aligned} \tag{26}$$

と変形され，最後の式は MAP 推定の式と等しくなる．

また，式 (26) において更に事前情報分布を $p(\hat{\boldsymbol{\theta}}) = c$ と仮定する．これは，$\hat{\boldsymbol{\theta}}$ が取りうる範囲で確率分布が

一様であり，$\hat{\boldsymbol{\theta}}$ について事前情報がない無情報分布を用いることに相当する．すると，上式は

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\max_{\hat{\boldsymbol{\theta}}} p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) \cdot c \\ &= \arg\max_{\hat{\boldsymbol{\theta}}} p(\boldsymbol{D}|\hat{\boldsymbol{\theta}})\end{aligned} \tag{27}$$

と変形され，これは最尤推定の式と等しくなる．

以上のことから，MAP 推定は変分推論において無限に鋭い確率分布を仮定したケース，すなわち $\boldsymbol{\theta}$ が確定値のみをとるケースを定式化したものに相当し，最尤推定は更に MAP 推定において事前確率分布を設けないケースに相当していることが分かる．

2.4.2 変分下限からの導出と変分推論の本質

また，変分推論において最大化の対象となる変分下限 $\mathcal{L}[q(\boldsymbol{\theta})]$ は，以下のように分解される．

$$\begin{aligned}\mathcal{L}[q(\boldsymbol{\theta})] &= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \left(\ln p(\boldsymbol{D}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln q(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \ln p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \ln p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta})]\end{aligned} \tag{28}$$

変分推論では変分下限 $\mathcal{L}[q(\boldsymbol{\theta})]$ を最大化する $q(\boldsymbol{\theta})$ を求めるが，そのためには式 (28) の最下段における第 1 項はなるべく大きく，かつ第二項はなるべく小さくなる必要がある．このことは，変分推論の本質に関わる非常に大きな示唆を含んでいる．

第 1 項は，観測データ $\boldsymbol{D}$ の $\boldsymbol{\theta}$ に関する条件付き分布の，近似分布 $q(\boldsymbol{\theta})$ による期待値となっており，これは**近似分布 $q(\boldsymbol{\theta})$ による観測データの尤度の対数値**を表している．これを最大化するような確率分布 $q(\boldsymbol{\theta})$ を求めようとすることは，尤度を最大化するパラメタを求める最尤推定の発想と一致しており，この項のみを最大化する近似分布は $\ln p(\boldsymbol{D}|\boldsymbol{\theta})$ が最大となる $\hat{\boldsymbol{\theta}}$ で値が立つデルタ関数を考えれば良い．

これに対し，第 2 項は事前分布と近似分布の KL ダイバージェンスを表している．この項を小さくしようとすることは，近似分布 $q(\boldsymbol{\theta})$ が事前分布 $p(\boldsymbol{\theta})$ から大きく離れることを防ぐ．そのため，第 1 項で近似分布が無限に尖ろうとすることに対して**第二項は設定された事前分布を通してペナルティを与える正則化の効果を与えている**．これは MAP 推定において事前分布を考慮することと同じ効果であり，その結果，変分推論の解は第 1 項と第 2 項のバランスの取れたポイントとして得られることとなる．

また，上式における第 1 項が最尤推定の発想を，第 2 項が MAP 推定の発想を定式化したものであることを以下のように実際に確認する．先ほどと同様に，近似分布に $q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ という仮定をおく．すると，式 (28) は

$$\begin{aligned}\mathcal{L}[\delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})] &= \int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \mathrm{KL}[\delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})||p(\boldsymbol{\theta})] \\ &= \int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \left(\int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) d\boldsymbol{\theta} - \int \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ln p(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \\ &= \ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) - \ln \delta(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + \ln p(\hat{\boldsymbol{\theta}}) \\ &= \ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) + \ln p(\hat{\boldsymbol{\theta}}) + \text{const.} \qquad (29) \\ &= \ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) p(\hat{\boldsymbol{\theta}}) + \text{const.} \qquad (30)\end{aligned}$$

と変形できる．式 (30) の最大化は MAP 推定にほかならない．また，式 (28) の第 1 項に由来し最尤推定に等しい式 (29) の第 1 項のみで最適値が決定されないように，式 (28) の第 2 項に由来する式 (29) の第 2 項が正

則化の役割を与えている．ここから更に事前分布に無情報分布の仮定をおくと，式 (29) における第 2 項が定数項に吸収され，変分問題は最尤推定に一致する．

以上のことから，**尤度を最大化するパラメタを確定的に求める最尤推定に対して，事前分布を通した自然な正則化項の追加によって極端にデータにフィットする過剰適合を防いだのが MAP 推定であり，更にパラメタの確率分布を考慮することによってデータに基づく推論の不確実性を定量化した手法が変分推論（ひいてはベイズ推論）である**と解釈することができる．このように，変分推論が最尤推定や MAP 推定の自然な拡張理論となっているという式 (28) の意味するところからは，変分推論が 1 つの近似手法にとどまらず重要な意義を持った方法論であることが示唆される．

参考文献

[1] 須山敦志：ベイズ推論による機械学習入門，講談社，2017.

[2] 中島伸一, 杉山将 : 変分ベイズ学習理論の最新動向, 日本応用数理学会論文誌, Vol.23, No.3, pp.453-483, 2013.

[3] C.M. ビショップ：パターン認識と機械学習（上・下），丸善出版，2012.

[4] 中島伸一：変分ベイズ学習，講談社，2016.