1　一般の EM アルゴリズム

前節までに完全データ尤度関数の期待値を最大化することで、負担率および分布のパラメタを iterative に求めるアルゴリズムを取り扱った。しかしながら、各パラメタを求めるにあたって、完全データ尤度関数の期待値を最大化することの正当性については、これまで議論がなされてこなかった。

今回は、完全データ尤度関数の期待値を最大化することが、元々最大化すべき関数であった観測データの尤度関数（不完全データ尤度関数）を最大化することと等価であることを示す。これにより我々は、前節までに学んだフレームワークについて、負担率および分布のパラメタ最適化が正しく行えるものであるという保証を与えることができる。

本レジュメの流れは次の通りである：まずはじめにカルバックライブラーダイバージェンス[*1]の性質を利用して、不完全データ尤度関数の最大化と完全データ尤度関数の期待値最大化が等価であることを確かめる。そして、等価性を示す中で、EM アルゴリズム内の各ステップにおける演算の中身を整理する。最後に、これまで学んできた EM アルゴリズムが適用できないケースについて、その対応方法、つまり拡張された EM アルゴリズムを紹介する。。

●完全データ尤度関数の期待値最大化

ここでは、KL ダイバージェンスを導入し、最大化したい尤度関数の分解表現およびその解釈を考える。

まずは、最大化したい尤度関数を確認する。尤度関数は次の形で表される。

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \tag{1}$$

各変数やパラメタの設定は教科書を確認されたい。また、既に対数尤度関数で表現していることに注意せよ。

次に、KL ダイバージェンスを用いて、尤度関数 1 について新しい表現（式 9.70）を得ることを考える。これは尤度関数の「分解」と呼ばれている。以下では、分解した形（式 9.70-72）の右辺を変形し、尤度関数 1 に戻る計算過程を示す。

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) + KL(q \| p) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln\{\frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})}\} - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln\{\frac{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})}\} \ (\because (\text{式 } 9.71\text{-}72)) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \{\ln\{\frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})}\} - \ln\{\frac{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})}\}\} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \{\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) - (\ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}) - \ln q(\boldsymbol{Z}))\} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \{\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) - \ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})\} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \{\ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}) + \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) - \ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})\} \ (\because (\text{式 } 9.73)) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) \\
&= \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) \ (\because \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) = 1)
\end{aligned} \tag{2}$$

式 2 を得られたことから、不完全データ尤度関数について分解を用いた表現ができることが確かめられた。こ

[*1] 以下、KL ダイバージェンスと書く。

こで、分解された各項（式 9.71-72）の意味について考えよう。説明の都合上、まず（式 9.72）の意味から考える。（式 9.72）は $KL(q\|p)$ であり、モデル分布 $q(\boldsymbol{Z})$ と事後分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ との間の KL ダイバージェンスに他ならない。

ここで、振り返りとして KL ダイバージェンスの性質について簡単に復習する[*2]。強調しておきたいポイントとしては、KL ダイバージェンスは非負であるということである。この非負であるということから、式 2 のもうひとつの項（式 9.71）の意味についても考えることができる。いま明らかに以下の関係が成立する。

$$\begin{aligned} \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) &= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q\|p) \\ &\geq \mathcal{L}(q, \boldsymbol{\theta}) + 0 \ (\because \text{KL ダイバージェンスは非負}) \\ &= \mathcal{L}(q, \boldsymbol{\theta}) \end{aligned} \tag{3}$$

したがって、$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta})$ となり、（式 9.71）は不完全データ尤度関数の下界であると解釈することができる。なお、これは任意の $q(\boldsymbol{Z})$ について成立する。ここで、図 9.11 を見ながら分解における下界のイメージを確かめよう。

● EM アルゴリズム内の各ステップにおける演算

次に EM アルゴリズムの各ステップにおける演算について手を動かしながら理解していこう。結論として、E ステップが分解項（式 9.71）について $q(\boldsymbol{Z})$ を動かして最大化すること（$\boldsymbol{\theta}$ は固定）、M ステップが（式 9.71）について $\boldsymbol{\theta}$ を動かして最大化すること（$q(\boldsymbol{Z})$ は固定）にそれぞれ対応していることを押さえて欲しい。

まずは E ステップについて考える。不完全データ尤度関数を考えるかわりに、下界 $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ を考えてやるとうまくいくらしいが、発想がよくわかっていない。下界 $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ が最大になるのは、KL ダイバージェンスが 0 になるとき、つまりモデル分布 $q(\boldsymbol{Z})$ が事後分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old})$ と等しいときである[*3]。このとき下界は対数尤度に一致し、それらの関係性は図 9.12 のようになる。

それでは、モデル分布 $q(\boldsymbol{Z})$ が事後分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old})$ と等しいとき、E ステップの後の下界がどのような表式になるか考えよう。

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln\{\frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old})}\} \ (\because \ q(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old})) \\ &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + \text{const.} \ (\because \text{（式 9.30）}) \end{aligned} \tag{4}$$

ここで出てくる $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ は、以前に見たような完全データ対数尤度を潜在変数 $\boldsymbol{Z}$ で期待値をとったものになっている（式 9.30）。第二項は q のエントロピーであり、$\boldsymbol{\theta}$ とは独立なのでパラメタの最大化を考えるにあたっては定数項となる。この時点で、完全データ尤度関数の期待値を最大化することの正当性が言えそうなものだが、まだ腹落ちできていない。やはり E ステップの始まりで下界の最大化を考えれば良いという思想が理解できていないからだろう。以上から、完全データ尤度関数の期待値を最大化することが、元々最大化すべき関数であった観測データの尤度関数を最大化することと等価であることが示された。

続いて M ステップについて考える。ここでは式 4 の最大化をおこなう。そのために新しい $\boldsymbol{\theta}$ を探索することになる。なお、このステップにおいて、新しい $\boldsymbol{\theta}$ を持ってくると、それに対応する新しい尤度関数は、古い尤度関数より必ず増加する。その理由は次の通りである。先で固定していた古いパラメタ $\boldsymbol{\theta}^{new}$ とは異なる新

[*2] See 【201206 輪講.pdf】

[*3] KL ダイバージェンスが 0 になるときの等号成立のことを言っている。

しいパラメタ $\boldsymbol{\theta}^{new}$ を持ってくる。いま M ステップではモデル分布 q を固定しており、この q は古いパラメタで構成されている。したがって、q は新しい $\boldsymbol{\theta}^{new}$ を持つ事後分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{new})$ と一致することはない。これは KL ダイバージェンスが 0 にならないということを意味している。そして、KL ダイバージェンスが非負であることを踏まえると、我々が確かめたかった次のことが言える：**M ステップにおいて新しい $\boldsymbol{\theta}^{new}$ を持ってくると尤度関数を必ず増加させる**[*4]。あとは最大値にたどり着くまで、E ステップと M ステップを繰り返せばよい。

ここで、パラメタ空間における下界と最大化すべき不完全データ対数尤度との関係を見ながら、パラメタが更新されていく様子を視覚的に捉える（図 9.14）。そのために、EM アルゴリズムのステップをいま一度整理する。

・古いパラメタ $\boldsymbol{\theta}^{old}$ のもとで、下界 $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ を計算する（E ステップ）[*5]

・上で求めた下界について $\boldsymbol{\theta}$ を動かしながら、最大値を与える $\boldsymbol{\theta}^{new}$ を計算する（M ステップ）

以下では、図 9.14 を見ながら、各ステップについて詳細および注意点を確認していく。まず、E ステップにおける下界と不完全データ対数尤度の曲線の関係について、次の 2 つのポイントを押さえる。

・下界の値と不完全データ対数尤度が $\boldsymbol{\theta}^{old}$ において一致すること

・両曲線の値が一致する点は接点であること、すなわち勾配が一致すること

前者については、KL ダイバージェンスが 0 になることから自明である。後者について、両曲線の $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$ における勾配を実際に計算することで確かめよう（演習 9.25）。まず、（式 9.70）の両辺について $\boldsymbol{\theta}$ で微分をとると次のようになる。

$$\left.\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{X} \mid \boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{old}} = \left.\frac{\partial \mathcal{L}(q, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{old}} + \left.\frac{\partial}{\partial \boldsymbol{\theta}} KL(q \| p)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{old}} \tag{5}$$

なお、微分計算のち $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$ を代入することに注意せよ。この等式において、右辺第二項、すなわち KL ダイバージェンスの微分項が 0 であることは自明である。なぜなら、モデル分布 q と潜在変数の事後分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ が等しいとき、KL ダイバージェンスが最小値となる、つまり停留条件を満たすからである。したがって、下界と不完全データ対数尤度が $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$ において接することが示された。

次に M ステップにおける下界と不完全データ対数尤度の曲線の関係については読み合わせとする。

i.i.d. の話は位置付けが謎。特に最後の「混合ガウスモデルの場合には、〜言っているに過ぎない」が何を目的として主張しているのかわからない。ひとまず負担率計算（式 9.75）の式変形は追いかけたので、それを

[*4] ただし既に最大化されている場合はその限りではない。

[*5] もちろん初期値を決めてかかるステップは存在するが、説明の都合上今回は割愛。

以下に記す。

$$
\begin{aligned}
p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}) &= \frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{p(\boldsymbol{X} \mid \boldsymbol{\theta})} \ (\because \text{条件付き確率の定義}) \\
&= \frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})} \\
&= \frac{\prod_n p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})}{\sum_{\boldsymbol{Z}} \prod_n p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})} \ (\because \text{i.i.d. 性}) \\
&= \frac{\prod_n p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})}{\prod_n p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} \ (\because \ \boldsymbol{Z} \text{ について周辺化}) \\
&= \prod_n \frac{p(\boldsymbol{z}_n, \boldsymbol{x}_n \mid \boldsymbol{\theta})}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} \\
&= \prod_n \frac{p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} \ (\because \text{条件付き確率の定義}) \\
&= \prod_n p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) \qquad (6)
\end{aligned}
$$

実は EM アルゴリズムはパラメタについての事前分布 $p(\boldsymbol{\theta})$ を導入した事後分布 $p(\boldsymbol{\theta} \mid \boldsymbol{X})$ についての最大化を行うこともできる。この事後分布は（式 9.70-72）の分解を用いた形で表現される。以下に、その分解計算を示す。

$$
\begin{aligned}
\ln p(\boldsymbol{\theta} \mid \boldsymbol{X}) &= \ln \frac{p(\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{X})} \ (\because \text{条件付き確率の定義}) \\
&= \ln p(\boldsymbol{\theta}, \boldsymbol{X}) - \ln p(\boldsymbol{X}) \\
&= \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) \\
&= \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \| p) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) \ (\because \text{（式 9.70）}) \\
&\geq \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) \ (\because \text{KL ダイバージェンスは非負}) \qquad (7)
\end{aligned}
$$

そして、これまでの EM アルゴリズムと同様に $q, \boldsymbol{\theta}$ を交互に最適化可能である。これができることは次の理由による：q については、分解項 $\mathcal{L}(q, \boldsymbol{\theta})$ の部分にしか現れず、従来アルゴリズムと等価であること。また、$\boldsymbol{\theta}$ についても、今までのアルゴリズムと同様に最適化可能であることが知られていることによる[*6]。

● EM アルゴリズムの拡張

最後に EM アルゴリズムの拡張について 2 つの側面で議論をする。まず、アルゴリズムにおいて各 E ステップ、M ステップの計算が実行できない場合についての対応方法を学ぶ。次に負担率の計算においてバッチ更新ではなくオンライン更新による手法を確認する。こっからはまじで無理だった。ごめん。

[*6] 実際、本当に問題ないのかはわからない。教科書に説明はない。