# 1　EP 法

今回は変分推論とは異なるもうひとつの近似推論法を取り扱う。これまでに学んだ変分推論においては、周辺分布の対数をその下界と KL ダイバージェンスとに分解することを考えた[*1]。ここで現れる KL ダイバージェンスは reverse の KL ダイバージェンスであり、式（10.4）のように定義される。そして KL ダイバージェンスを最小化するという手続きによって、モデル分布 $q(\boldsymbol{Z})$ を求めることができた。

新しく学ぶ EP 法[*2]においては、KL ダイバージェンスが forward の KL ダイバージェンスとなる。異なる KL ダイバージェンスを用いて、変分推論との違いについて数学的操作や導出結果の観点から理解したい。

以下では $p(\boldsymbol{z})$ を固定された確率分布（以下、真の分布）としたときに、$KL(p\|q)$ を $q(\boldsymbol{z})$（以下、モデル分布）について最小化する問題を考える。ここで $q(\boldsymbol{z})$ は指数型分布族であると仮定する[*3]。いま $q(\boldsymbol{z})$ が式（10.184）で表されるとすると、$KL(p\|q)$ は次のようになる。なお、以下では $KL(p\|q)$ をパラメタ $\boldsymbol{\eta}$ の関数として捉えるものとする。

$$
\begin{aligned}
KL(p\|q) &= -\int p(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int (p(\boldsymbol{z}) \ln p(\boldsymbol{z}) - p(\boldsymbol{z}) \ln q(\boldsymbol{z})) d\boldsymbol{z} \\
&= -\int p(\boldsymbol{z}) \ln p(\boldsymbol{z}) d\boldsymbol{z} + const \\
&= -\int p(\boldsymbol{z}) \ln [h(\boldsymbol{z}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{z})\}] d\boldsymbol{z} + const \\
&= -\int p(\boldsymbol{z}) \ln h(\boldsymbol{z}) d\boldsymbol{z} - \int p(\boldsymbol{z}) \ln g(\boldsymbol{\eta}) d\boldsymbol{z} - \int p(\boldsymbol{z}) \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{z}) d\boldsymbol{z} + const \\
&= -\ln g(\boldsymbol{\eta}) \int p(\boldsymbol{z}) d\boldsymbol{z} - \boldsymbol{\eta}^T \int p(\boldsymbol{z}) \boldsymbol{u}(\boldsymbol{z}) d\boldsymbol{z} + const \\
&= \ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T E_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] + const
\end{aligned}
\tag{1}
$$

したがって $KL(p\|q)$ の最小化をおこなうためには、式 1 の最右辺について勾配をとり 0 としてやればよい。

$$
\begin{aligned}
\nabla_{\boldsymbol{\eta}} (\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T E_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] + const) &= 0 \\
-\nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) - E_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] &= 0 \\
-\nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) &= E_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})]
\end{aligned}
\tag{2}
$$

ここで式（2.226）[*4]より、この等式は次のようになる。

$$
E_{q(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] = E_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] \tag{3}
$$

以上より最適解が得られるのは、モデル分布と真の分布とで、十分統計量 $\boldsymbol{u}(\boldsymbol{z})$ についての期待値が一致するときである。これらの期待値を一致させることを moment matching と呼ぶ。

---

[*1] p.177 の式（10.2）を見よ。

[*2] Expectation propagation の頭文字をとったもの。

[*3] 逆に指数型分布族でなければ以降の議論は成立しないのだろうか。確認はできていない。

[*4] 上巻 p.113 最下部を見よ。今回は $q(\boldsymbol{z}) = h(\boldsymbol{z}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{z})\}$ を式（2.225）の真ん中の辺に当てはめればよい。

ここで moment matching の例について、図 10.3(a) を参照する。ここでは十分統計量として、平均と分散の一致を考える。平均の一致は理解できるが、分散の一致が理解できなかった。真の分布が多峰製の分布で、モデル分布が単峰性の分布なのに分散を一致させるのは原理的に不可能ではと思ってしまった。

以下では簡単なケースとして、真の分布が式（10.188）のような因子の積で書ける場合について、モデル分布を求める。そしてそれをアルゴリズムに落とし込んでいくところまで辿り着いてみせる。

なお、各因子は 1 データ点についての尤度 $f_n(\boldsymbol{\theta}) = p(\boldsymbol{x}_n \mid \boldsymbol{\theta})$ である。ただし $n = 0$ のときは $f_0(\boldsymbol{\theta}) = p(\theta)$ として事前分布の形で与えられる。

グラフィカルモデルの言語による表現のところは読み合わせ。特に無向グラフによる表現方法を確認させてほしい。

さて、我々にとって興味があるのは、いつも通りパラメタについての事後分布[*5]であり

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathcal{D}) &= \frac{p(\mathcal{D}, \boldsymbol{\theta})}{p(\mathcal{D})} \\ &= \frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta}) \end{aligned} \tag{4}$$

で表される。ここで $p(\mathcal{D})$ はモデルエビデンスであり、これはパラメタ $\boldsymbol{\theta}$ についての周辺化によって

$$p(\mathcal{D}) = \int \prod_i f_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{5}$$

として与えられる。

ベイズモデルのフレームワークに従えば、事後分布を求めたあとで周辺化によって予測分布を求め、そして $\boldsymbol{\theta}$ に関する周辺化によってモデルエビデンスを求めることになる。しかしながら、これらの周辺化は解析的に実行することはできないとしよう。そこで、以下では EP 法と呼ばれる近似手法を導入し、モデル分布の事後分布とモデルエビデンスを表現する方法を考えていく。

まずはモデル分布の事後分布について議論する。EP 法ではモデル分布の事後分布が次の近似によって与えられるとする。

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta}) \tag{6}$$

各変数の詳細は読み合わせとする。ここで気になったのは、事後分布なら引数は $\boldsymbol{\theta} \mid \mathcal{D}$ なのではということ。なお、$\tilde{f}_i(\boldsymbol{\theta})$ は指数型分布族であると仮定する。これにより $q(\boldsymbol{\theta})$ も指数型分布族となる。指数型分布族で表される関数の積は指数型分布族になる、というのは自明だよね？

以上の設定のもとで、最小化すべき $KL(p\|q)$ は次のようになる。

$$KL(p\|q) = KL(\frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta}) \| \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta})) \tag{7}$$

この最適化計算は解析的に実行するのが難しい。どこがどう難しいのか一緒に確認したい。

この計算を実行するために、まずはナイーブな手法を紹介する[*6]。そのやり方は次のようなものである。

・各因子 $\tilde{f}_i(\boldsymbol{\theta})$ ごと別個に $KL(f_i(\boldsymbol{\theta}) \| \tilde{f}_i(\boldsymbol{\theta}))$ を最小化する

・全 $i$ について計算を実行したのち、積をとることで $q(\boldsymbol{\theta})$ を構成する

しかしながら、この手法では精度がよくないということが知られているため、これは採用しない。

---

[*5] 予測分布を求めるために必要。もはやベイズモデルのフレームワークとも言えよう。

[*6] これは教科書にどうしても載せる必要があったのか？

そこで EP 法では、各因子を完全に独立に考えるのではなく、他の因子も計算式に織り込みながら最適化を進めていく。詳細については、以下で EP 法のステップを辿りながら説明をおこなう。ステップは大きく 4 つ：各因子の初期化／特定の 1 因子の更新（=改良）／改良された因子の出力／モデルエビデンスの計算となっている。それぞれ詳しく追っていくことにする。

1. 各因子 $\tilde{f}_i(\boldsymbol{\theta})$ を適当に初期化する
2. 特定の 1 因子 $\tilde{f}_j(\boldsymbol{\theta})$ を更新（=改良）する

2-1. $\tilde{f}_j(\boldsymbol{\theta})$ を総積 $\prod_i f_i(\boldsymbol{\theta})$ から除去する（$=\prod_{i\neq j} f_i(\boldsymbol{\theta})$ を得る）

2-2. 新しく得られる事後分布を $q^{new}(\boldsymbol{\theta})$ としたとき

$$q^{new}(\boldsymbol{\theta}) \propto \tilde{f}_j(\boldsymbol{\theta}) \prod_{i\neq j} f_i(\boldsymbol{\theta}) \tag{8}$$

と

$$f_j(\boldsymbol{\theta}) \prod_{i\neq j} f_i(\boldsymbol{\theta}) \tag{9}$$

とが近づくように KL ダイバージェンスを最小化する。式 8 の右辺に改良したい因子以外の因子も織り込むところがナイーブな手法との違いである。気になったのは、式 9 に近づけることの正当性について。真の分布はチルダがついていないもののはずであり、それと近しいモデル分布を構成しなくても大丈夫なのかなと思った。

2-2′. 実際は 2-2. で定義した式 8 を使うのではなく $q(\boldsymbol{\theta})$ を $\tilde{f}_j(\boldsymbol{\theta})$ で除した

$$q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\theta)}{\tilde{f}_j(\boldsymbol{\theta})} \tag{10}$$

を使う。「除算を行う方が簡単である」と教科書にはあるけど、簡単の意味がわからなかった。表式が簡潔ということなのか、計算コストが低いということなのか。そして真の分布の因子と規格化定数をかけてやると

$$\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \tag{11}$$

が得られる。これが真の分布（もどき？）となる。なお $Z_j$ は式（10.197）で与えられる。

2-3. 得られた真の分布（もどき？）とモデル分布とで約束される以下の KL ダイバージェンス

$$KL(\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \| q^{new}(\theta)) \tag{12}$$

を最小化する。

3. 改良された因子 $\tilde{f}_j(\boldsymbol{\theta})$ の保存。正直ここは追いきれなかったので一緒に確認したい。
4. モデルエビデンスの近似は

$$p(\mathcal{D}) \simeq \int \prod_i \tilde{f}_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{13}$$

で与えられ、これを使ってモデル比較を行えばよい。なお式 13 は式 5 において $f_i(\boldsymbol{\theta})$ を $\tilde{f}_i(\boldsymbol{\theta})$ に置き換えたものである。

最後に変分推論（=ベイズ）法と EP 法との違いを下表 1 のようにまとめる。なお $p$ が真の分布、$q$ がモデル分布を表す。

表 1　変分推論法と EP 法の比較

| | 変分推論法 | EP 法 |
|---|---|---|
| 最小化する KL ダイバージェンス | $KL(q\|\|p)$:reverse | $KL(p\|\|q)$:forward |
| 望ましい p の形状 | ー | ロジスティック分布 |
| 望ましくない p の形状 | ー | 多峰性分布 |