今年の生理研研究会は「認知神経科学の先端 脳の理論から身体・世界へ」と題して、自由エネルギー原理(Free-energy principle, FEP)をテーマに9/2に開催。これに先立つ8/31-9/1には「脳の自由エネルギー原理チュートリアル・ワークショップ」というタイトルでFEP入門のためのレクチャーとハンズオン。

これらに向けてFEP入門の資料を作りました。今日はその第1回。「Sec.1 変分自由エネルギー VFE の定義」です。これから数日かけて連続してポストして、最終版はPDFでアップロードします。これを使って予習してみてください。

参加募集開始は連休明けの予定。もう少々お待ちください。

Sec.1 変分自由エネルギー VFE の定義

このセクションでは、変分自由エネルギー VFE について、教科書的な定義と基礎についてまとめてみた。ちゃんとした説明のためには「パターン認識と機械学習(PRML)」の10章とかの機械学習の教科書を読んでみてほしい。

だいたい知っている人のためにまとめておくとこのとおり： 以降のセクションを読むために理解していほしいことはたった二つだ。(a) 変分推定を使うためには、 問題となっている状況の生成モデルがどういう構造になっているかを因果グラフ、因子グラフを用いてきっちり決めてやる必要がある。(b) そのうえで、変分自由エネルギー F を定義するためには、なにが観測データで、なにが潜在変数を理解して、あとはそれは式(5)に代入すれば一意に決まる。

[1-1. ベイズ推定とは]

そもそも変分自由エネルギー variational free energy VFE F とはなにかというと、変分推定を行うときに使われる値だ。変分推定というのはある種間接的な推定法なので、それの前にもっと直接的な推定(ベイズ推定)について説明する。ベイズ推定というのは確率論的な推定法だ。

まず問題設定はこうだ。観測データ $Y = \{y_1, y_2\}$ (y_1 :窓ガラスが濡れている or y_2 :窓ガラスが濡れていない)と 潜在変数 $Z(= \{z_1, z_2\})$ (z_1 :外は雨が降っている or z_2 :外は雨が降っていない) がある。いま窓ガラスが濡れている、を観測した($Y = y_1$)。このとき外が雨が降っている確率 $Prob(Z = z_1)$ を推定したい。これだけだと問題の解きようがないから、潜在変数 Z (雨が降っているかどうか)という原因が観測データ Y (窓ガラスが濡れているかどうか)という結果を引き起こすときの関係を、これまでの経験から、両者の同時確率 $Prob(Z, Y)$ として持っている。この同時確率のことを生成モデル $p(Z, Y)$ と表記する。

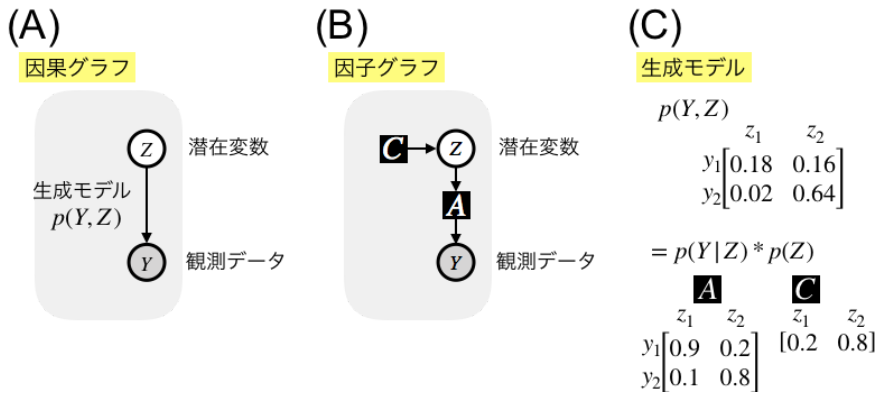

図1-1: 窓ガラスYから 降雨Zを推測する

このときの潜在変数 Z と観測データ Y の因果関係を有向グラフで表したのが図1-1Aの因果グラフ。両者をつなぐ関係が生成モデル $p(Z,Y)$ となっている。生成モデルは同時確率なので、具体的な例としては図1-1Cのような2x2の行列で表すことができる。同時確率なので、4つのセルの確率を全部足すと1になる。

しかし図1-1Aでは生成モデル $p(Z,Y)$ とZ および Y に対する関係がよくわからないので、それを明示したのが図1-1Bの因子グラフ。このような表現にすると、生成モデル $p(Z,Y)$ は Z の事前分布である $C=p(Z)$ 、それからZ から Y を生成する観測モデル $A=p(Y|Z)$ の二つに分解(因子化)できることができる。こうすると生成モデルの意味も理解しやすくなる(図1-1C)。事前分布 C からわかるように、そもそも雨が降る確率は低い($p(z_1)=0.2$)。また、観測モデル A からわかるように、雨が降ってなくても(z_2)、窓ガラスが濡れている(y_1)という 可能性はある($p(y_1|z_2)=0.2$)。

いま知りたいのはある $Y=y_1$ のときの Z の確率分布 $p(Z|y_1)$ だから、ベイズの法則が使える。

$$p(Z|y_1)=\frac{p(Z,y_1)}{\sum_Z p(Z,y_1)} \tag{1}$$

このようにして、生成モデル $p(Z,Y)$ から事後分布(以下true posteriorと呼ぶ) $p(Z|y_1)$ を計算するのがベイズ推定だった。

(なお。以下のすべての説明で、総和の記号 $\sum$ のみで説明できる状況の話だけをする。積分記号 $\int$ は出てこない。つか積分記号嫌い。)

[1-2. 変分推定とは]

しかし、式(1)は分母で全ての可能な潜在変数 Z で足し算をするという作業が入っている。そんな事できないときもあるし、現実的な場面での応用では生成モデル $p(Z,Y)$ は変数が多くて計算量的に難しいという事情もある。そこで使われる近似的方法が変分推定だ。

True posterior $p(Z|y_1)$ を直接計算する代わりに、それを近似する確率分布として推測 $q(Z)$ というものを設定する。推測 $q(Z)$ の分布の形を変えて、true posterior $p(Z|y_1)$ に一致させることができれば、true posterior $p(Z|y_1)$ を計算できたのと同じことだ。

(この文書では q のことは 一貫して「推測」と呼ぶことにする。より正確な表現ではapproximate posteriorと呼ぶべきだろう。この呼び方だと、true posteriorと同じ変数が入っているということがわかりやすい。True posteriorの変数は潜在変数 Z だから、推測q = approximate posteriorの変数も潜在変数 Z だ。どちらも全ての Z で和を取れば1になる。)

(それでもここでq を「推測」と呼ぶのは、 今回の説明では近似法(平均場近似やラプラス近似)を使わないexactな計算だけに話を絞るので、q を近似によって得られたものと混同しないようにしたいという意図がある。機械学習での変分ベイズとは 近似推定であり、exactな計算ができるならこんな回り道をする必要はない。しかし自由エネルギー原理においての変分推定の本質とは、q という汎関数functionalを操作することにある、というのが私の理解。なのにFEP論文の説明のほとんどは平均場近似やラプラス近似をしたあとのテクニックに終始していて、FEPの正確な理解が阻まれている、というのが私の現状認識。)

[1-3. KL距離とは]

このために、推測 $q(Z)$ とtrue posterior $p(Z|y_1)$ というふたつの確率分布の近似度を計算する指標としてカルバック・ライブラー距離 D_{KL} (以下 KL距離, KLDと 略する)というのを使う。

$$D_{KL}[q(Z)\|p(Z|y_1)] = \sum_Z q(Z) \ln \frac{q(Z)}{p(Z|y_1)} \tag{2}$$

y_1 はすでに観察されて確定している値で、Z で総和をとっているから、KL距離は $q(Z)$ の分布の形だけによって決まる定数だ。もし推測 $q(Z)$ とtrue posterior $p(Z|y_1)$ が完全一致していたらKL距離は0になる。それ以外は>0となってる。

(距離とはいうけど逆向きは同じ距離ではない。つまり、$D_{KL}[q(Z)\|p(Z|y_1)] \neq D_{KL}[p(Z|y_1)\|q(Z)]$ 。ではここでなぜ前者を使うのかというと、KL距離と交差エントロピーとの関係から説明できるはずだが、後回しで。)

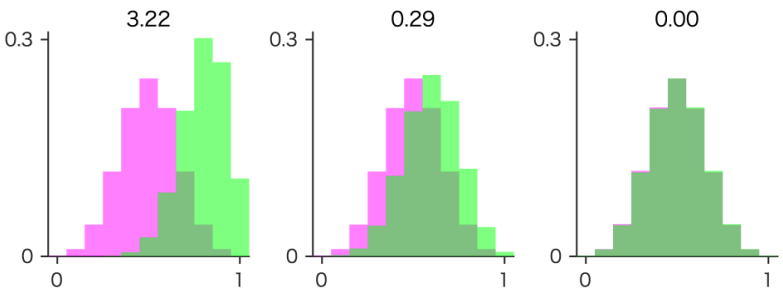

図1-2: KL距離の例

図1-2にKL距離の例として、二つの確率分布(二項分布から作成)のヒストグラムを書いて、そのKL距離(単位はbits)を表示した。離れた分布ほどKL距離は大きく(図1-2左)、二つの確率分布が完全に一致するときはKL距離=0なることがわかる(図1-2右)。

[1-4. KL距離の最小化]

しかしこのKL距離が直接計算できるならいいのだが、それができるなら直接true posteriorを計算すればいいだけだ。違いはどこに現れてくるかというと、KL距離を式変形すると、式(3)のようになる。(式変形で $p(Y,Z) = p(Z|Y)p(Y)$ を使用。)

$$\underbrace{\ln p(y_1)}_{\text{Log evindece}} = \underbrace{-\sum_Z q(Z)\ln\frac{q(Z)}{p(Z,y_1)}}_{\text{ELBO}=-\text{F}} + \underbrace{D_{KL}[q(Z)\|p(Z|y_1)]}_{\text{KLD}} \tag{3}$$

左辺は観測データ Y が生成モデルp のもとでどのくらいの確率であり得るかというevidence $p(Y)=\sum_Z p(Y,Z)$ のlogをとったものなので、log evidence (周辺対数尤度)と呼ぶ。

(なお、$p(Y)$ はただの 観測データ Y の出現確率 $Prob(Y)$ ではないことに注意。$p(Y)$ はいまある生成モデルに基づいたうえでデータの出現確率だから、尤度likelihoodなのだ。このことは後述の「暗い部屋問題」に関わってくる。)

(このことを明示するためにこの 生成モデルに m という名前をつけて、$p(Y|m)$ もしくは $p_m(Y)$ と表示することもある。ここではその表示は採用しない。そのかわりに、生成モデル $p()$ とただの確率 $Prob()$ を厳密に分けて表示している。)

でもって、このlog evidneceというやつはlogの中身が確率だから、かならず負の値を取る。(実際に観測されたデータがあるのだから、この値は決して $-\infty$ にはならない。) そして右辺の第2項はKLD ≥ 0 だから、第1項はlog evidenceよりも 必ず小さい。

$$\underbrace{\ln p(y_1)}_{\text{Log evindece}} \leq \underbrace{-\sum_Z q(Z)\ln\frac{q(Z)}{p(Z,y_1)}}_{\text{ELBO}=-\text{F}} \tag{4}$$

よって、この第 1項はlog evidneceの 下限になっている。そこでこの第 1項はEvidence lower bound, ELBOと呼ばれる。そして変分自由エネルギー F とは $F=-\text{ELBO}$ と単に符号逆転させただけのものだ。(ELBOは log evidneceよりも小さいのだから、必ず負の値だ。ということは変分自由エネルギー F はかならず正の値。)

[1-5. ちょっと脇道]

ちょっと脇道。といいつつFEPの理解に関しては、ここからが面白い。以上のことからわかるのは、以下の(1)-(7)が全部同じことだという点だ。

- (1) ベイズ推定をする
- (2) true posteriorを計算する
- (3) 推測q をtrue posteriorに一致させる
- (4) KL距離を0にする
- (5) ELBOを最大化する
- (6) 変分自由エネルギー F を最小化する
- (7) 周辺対数尤度 $\ln p(Y)$ を最大化する

ただし、 生物が扱うような複雑な状況においては、これらすべてを完全に実現することはできない。よって、どこかで近似が入ってくることになる。

自由エネルギー原理についての議論で、どれが目的でどれが結果か、どれがなにを近似しているのか、といった問題を理解するためには、この関係について考えればいい。

たとえば暗い部屋問題というのがある。自由エネルギー原理では (6) F を最小化することが生物の目的だというが、それだったら、(7') 一番ありうる 観測データをサンプルすればいいのだから、暗い部屋にじっとしていればいいじゃん、という議論。でもこれは(7)の解釈が間違っている。(7)は生成モデルの元での尤度なのだから、 暗い部屋という

観測データの出現確率が高い生成モデルをあらかじめ獲得していない限り、これは成り立たない。(だから、冬眠する動物は暗い部屋を選ぶ。我々だって夜は暗い部屋を選ぶ。)

それよりは、ネット言論がタコツボ化することについて当てはめる方がまだ尤もらしいだろう。つまり、我々は自分の持論(=世界についての生成モデル)を持っていて、それをより補強するデータばかり観測しようとする。だからネット言論ではみな自分が聞きたい意見ばかり集めるようになって、分断はより強調されていく。これは自由エネルギー原理から説明できる。なんだってー！

あと、ベイズ脳と自由エネルギー原理の関係についてもどちらが原因でどちらが結果かは自明でない。

- 議論A:「(1) われわれ生物はベイズ的に情報処理をしているんだ、というのが先にあって、そのための近似法として (6) 自由エネルギー最小化 をするように脳と身体を進化せさせてきた」
- 議論B:「なんらかの生物学的拘束条件から、われわれ生物は (6) 自由エネルギー最小化をするようになっていて、その結果として (1) 行動や知覚でベイズ推定をしているように見えている」

どちらも 議論としてはありえる。Friston自身は(6)を自由エネルギー「原理」と呼んだうえで、 (1) をベイズ脳「仮説」と呼んでいるので、議論Bに基づいていると考えるのが筋が通っていると私は思うのだが、Fristonの発言自体はそのつど言ってることがブレているように思う。

[1-6. 変分自由エネルギーとは]

そういうわけで、変分自由エネルギー variational free energy VFE F は式(5)のとおりに定義される。

$$
\begin{aligned}
F &= \mathbb{E}_{q(Z)}[\ln q(Z) - \ln p(Y,Z)] \\
&= \underbrace{D_{KL}[q(Z)||p(Z|Y)]}_{\text{KLD}} + \underbrace{-\ln p(Y)}_{\text{Surprisal}}
\end{aligned} \tag{5}
$$

(ここまでは観測データには具体例での y_1 を使ってきた。しかしもちろんこの式は y_2 でも成り立つわけで、ここからは観測データ は $Y(=\{y_1, y_2\})$ を使って表示する。しかしこれまでのことからわかるように、Y はすでに確定した一つのデータであり、一方で Z はとりうる値が全部並んだ変数になっている。同じ変数みたいに見えてべつものなのだ。慣れた人にとっては当たり前だろうけど、私はこれが馴染むまで相当時間がかかった。この違いを明示するために $p(Z|y)$ や $p(Z|Y=y)$ のように表示するのが正確だと思うが、式が長くグロくなるので、短く書ける場所以外ではその表記はしてない。)

(ここでは期待値の記号として $\mathbb{E}$ を使っている。下付き文字の期待値で重み付けして、その期待値の変数 Z で和をとる。たとえば $\mathbb{E}_{q(Z)}\ p(Z) = \sum_Z q(Z)p(Z)$ というふうに。)

ここで第2項は(sensory) surprisalと呼んでいる。 この値はlog evidenceの符号逆転で、必ず正の値になっている。 得られた観測データが現在の生成モデル に基づいて予想外であれば、$p(Y)$ はより小さくなるので、surprisalは大きくなる。 Surprisalはシャノンのself informationとも呼ばれる。これをsurpriseと呼ぶこともあるが、 次に出てくるBayesian surpriseとの区別のために、この文書ではsurprisalという言葉で統一させてもらう。

[1-7. Bayesian surpriseとは]

F は式変形によって式(6)の形でも表現できる。($p(Y,Z) = p(Y|Z)p(Z)$ を使用。)

$$F = \underbrace{D_{KL}[q(Z)\|p(Z)]}_{\text{Bayesian surprise}} + \underbrace{-\mathbb{E}_{q(Z)} \ln p(Y|Z)}_{\text{Uncertainty}} \tag{6}$$

第1項のBayesian surpriseは、現在の推測q と生成モデルにおける事前分布 p との間の距離なので、推測q が観測データによってアップデートされると、より大きくなる。つまり、情報獲得の大きさと捉えることができる。

第2項の $Uncertainty$ は、Fristonの表記では $-Accuracy$ となるが、$Accuracy$ は必ず負なので、ここではわかりやすさ優先で、必ず正となる $Uncertainty$ という表記を採用している。この項については後回しで。(なお、ここでのuncertaintyは推定 q の分散とは別ものであることに注意。現在の例では推定 q は点推定をしているので、 分散ではないことはわかると思う。今回の文書では推定の分散があるような例は扱わない。)

第1項の意味は、次の近似(いまこの文書で初めて近似を使った！)を考えるとさらに具体的になる。もし推測$q(Z)$ をtrue posterior $p(Z|Y)$ に完全に一致させることができた場合には、式(7)の2行目が成り立つ。

$$\begin{aligned} F &= \underbrace{D_{KL}[q(Z)\|p(Z)]}_{\text{Bayesian surprise}} + \underbrace{-\mathbb{E}_{q(Z)} \ln p(Y|Z)}_{\text{Uncertainty}} \\ &\approx \underbrace{D_{KL}[p(Z|Y)\|p(Z)]}_{\text{BS}'} + \underbrace{-\mathbb{E}_{q(Z)} \ln p(Y|Z)}_{\text{Uncertainty}} \end{aligned} \tag{7}$$

ここの第1項BS' は、式(8)にあるように観測データの期待値 $p(Y)$ をかけてやると、観測データ Y と潜在変数 Z の間の相互情報量MIになっている。このことからも、Bayesian surpriseが情報獲得に関わっているということがわかるかと思う。

$$\begin{aligned} \mathbb{E}_{p(Y)} \text{BS}' &= \mathbb{E}_{p(Y)} D_{KL}[p(Z|Y)\|p(Z)] \\ &= MI(p(Z|Y), p(Z)) \end{aligned} \tag{8}$$

なお式(8)は後ほど出てくる期待自由エネルギーEFE G でのepistemic value と同じ形になっている(ただしepistemic value では p ではなくて q だが)。これについては再訪する。

[1-8. 実例で変分自由エネルギーの最小化を見てみる]

それでは、図1-1に出した例を使って、推測q を変化させながら変分自由エネルギーF の最小化をシミュレーションしてみよう。まず、先に正解をカンニングしてしまうと、図1-1の生成モデル p からtrue posteriorを直接計算して、

$$p(Z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} | Y = \begin{bmatrix} y_1 & y_2 \end{bmatrix}) = \begin{bmatrix} 0.529 & 0.030 \\ 0.470 & 0.969 \end{bmatrix} \tag{9}$$

となる。よって、$Y = y_1$ (窓ガラスが濡れている)のとき、$Z = z_1$ (雨が降っている)の確率は $p(Z = z_1|Y = y_1) = 0.529$ とわかった。事前分布 $p(Z = z_1) = 0.2$ であったことを考えると、 窓ガラスが濡れている、という観察はinformativeであったことがわかる。

では $q(Z)$ を変化させてみよう。ここで$q(Z)$ の形を変えると言ったが、実のところ、$q(Z=z_1)$ が決まれば、$q(Z=z_2)=1-q(Z=z_1)$ も決まる。そこでパラメーター $\phi_Z=q(Z=z_1)$ を$0-1$ の範囲で動かして$VFE, KLD, Surprisal$ を計算してプロットしてやる。すると図1-3Aのとおりになった。

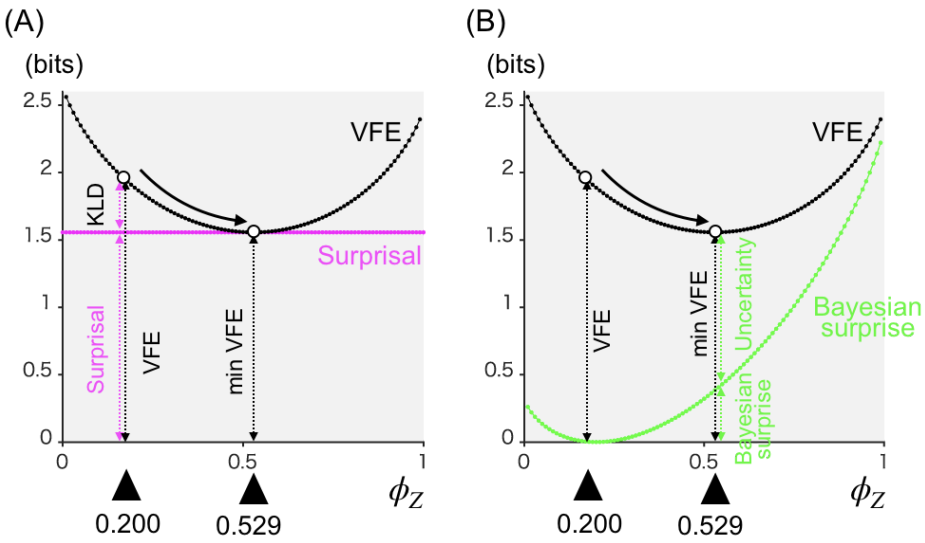

図1-3: 変分自由エネルギーVFEの最小化

$q(Z)$ の初期状態は事前分布 $p(Z)=\begin{bmatrix}0.2\\0.8\end{bmatrix}$ と等しいと考えてやれば、$\phi_Z=0.2$ からスタートするのが妥当だろう。このとき、

$$VFE=1.8879$$
$$KLD=0.3315$$
$$Surprisal=1.5564\ (bits)$$

となっている。ここからVFE が小さくなる方に ϕ_Z を動かしていくと、$\phi_Z=0.529$ で

$$VFE=1.5564$$
$$KLD=0$$
$$Surprisal=1.5564\ (bits)$$

となる。$KLD=0$ で VFE が最小化された状態では、$VFE=Surprisal$ となっている。この $\phi_Z=0.529$ は直接true posteriorを計算した結果の $p(Z=z_1|Y=y_1)=0.529$ と同じになっていることが確認できた。

そういうわけで、変分自由エネルギーを最小化することで、true posteriorを近似できるということが確認できた。自由エネルギー原理では、脳はこのようなやり方で観察データ(感覚入力)から潜在変数(外界の状態)を推定していると考える。以降のセクションでは、より知覚や行動に似せた状況での生成モデルを構築して、どのように VFE が計算されるかを見てゆく。

なお、図1-3から想像がつくように、もしVFE の曲線がlocal minimumを持っていたら、そこで停まってしまって、正しいtrue posteriorが計算できないであろうことも想像がつく。そし

て、もっと複雑な生成モデルではそのようなことはいかにも起こりそうだ。

[1-9. 変分自由エネルギーの最小化をBayesian surpriseから見てみる]

図1-3Aでは式(5) $F = KLD + Surprisal$ に基づいて、ϕ_Z を0.2から0.529に動かして、そのときに $KLD = 0$ となり、F が最小化されるのを見た。では同じものを式(6)
$F = Bayesian\ surprise + Uncertainty$ に基づいて表示してみよう(図1-3B)。

Bayesian surpriseは $q(Z)$ の初期状態 $\phi_Z = 0.2$ では 0になっていることがわかる。これは $q(Z)$ の初期状態 $= Z$ の 事前分布 $q(Z) = p(Z)$ を採用したのが理由だ。まだこの段階では観察データ y_1 から獲得した情報がゼロということ。

$$\begin{aligned}\mathrm{VFE} &= 1.8879\\ \text{Bayesian surprise} &= 0\\ \mathrm{Uncertainty} &= 1.8879(bits)\end{aligned}$$

ϕ_Z は正負どちらに動かそうが、Bayesian surpriseは増える。現状よりもちがうというだけで情報獲得なので。ということで、Bayesian surpriseだけ見てても、VFEを下げることはできないということがわかる。

そして、ϕ_Z を0.2から0.529に動かしてVFE を最小化したとき、Bayesian surpriseが増えていることが確認できる。

$$\begin{aligned}\mathrm{VFE} &= 1.5564\\ \text{Bayesian surprise} &= 0.3845\\ \mathrm{Uncertainty} &= 1.1719(bits)\end{aligned}$$

図1-3Bのプロットを見ると、$\phi_Z = 0.529$ はBayesian surpriseを最大化する場所ではないこともわかる。

[1-10. 変分自由エネルギーを微分して解析的に解いてみる]

最後に変分自由エネルギーVFE F を ϕ_Z で微分して、自由エネルギーの最小化について解析的に解いてみることにしよう。ここで必要な数学的知識は高校数学IIIまででいける。$f'(x)$ は xxの関数 $f(x)$ をx で微分したものを意味するとしたうえで、
$(f(x) * g(x))' = f'(x) * g(x) + f(x) * g'(x)$ と $(\ln x)' = \frac{1}{x}$ の二つだけ。

式(5)に $Y = y_1, q(Z = z_1) = \phi_z, q(Z = z_2) = 1 - \phi_z$ を代入する。生成モデルから
$p(Y = y_1, Z = z_1) = 0.18, p(Y = y_1, Z = z_2) = 0.16$ も入れておく。

$$
\begin{aligned}
F &= \mathbb{E}_{q(Z)}[\ln q(Z) - \ln p(Y,Z)] \\
&= \sum_Z q(Z) * [\ln q(Z) - \ln p(Y,Z)] \\
&= q(Z=z_1) * [\ln q(Z=z_1) - \ln p(Y=y_1, Z=z_1)] \\
&\quad + q(Z=z_2) * [\ln q(Z=z_2) - \ln p(Y=y_1, Z=z_2)] \\
&= \phi_z * [\ln \phi_z - \ln 0.18] + (1-\phi_z) * [\ln(1-\phi_z) - \ln 0.16] \\
&= \phi_z \ln \phi_z - \phi_z \ln 0.18 + \ln(1-\phi_z) - \phi_z \ln(1-\phi_z) - \ln 0.16 + \phi_z \ln 0.16 \\
&= \phi_z \ln \phi_z + \ln(1-\phi_z) - \phi_z \ln(1-\phi_z) + \phi_z \ln \frac{0.16}{0.18} - \ln 0.16 \\
&= \phi_z (\ln \phi_z + \ln \frac{0.16}{0.18}) + (1-\phi_z) \ln(1-\phi_z) - \ln 0.16
\end{aligned} \tag{10}
$$

これを ϕ_Z で微分する。

$$
\begin{aligned}
F' &= [\phi_z (\ln \phi_z + \ln \frac{0.16}{0.18}) + (1-\phi_z) \ln(1-\phi_z) - \ln 0.16]' \\
&= \phi_z' (\ln \phi_z + \ln \frac{0.16}{0.18}) + \phi_z (\ln \phi_z + \ln \frac{0.16}{0.18})' \\
&\quad + (1-\phi_z)' \ln(1-\phi_z) + (1-\phi_z) \ln(1-\phi_z)' \\
&= \ln \phi_z + \ln \frac{0.16}{0.18} + \phi_z (\frac{1}{\phi_z}) - \ln(1-\phi_z) - (1-\phi_z) \frac{1}{1-\phi_z} \\
&= \ln \phi_z + \ln \frac{0.16}{0.18} - \ln(1-\phi_z) \\
&= \ln \frac{0.16}{0.18} - \ln \frac{1-\phi_z}{\phi_z}
\end{aligned} \tag{11}
$$

$F' = 0$ と置くと ϕ_z が 求まった。

$$
\begin{aligned}
0 &= \ln \frac{0.16}{0.18} - \ln \frac{1-\phi_z}{\phi_z} \\
\ln \frac{1-\phi_z}{\phi_z} &= \ln \frac{0.16}{0.18} \\
\frac{1-\phi_z}{\phi_z} &= \frac{0.16}{0.18} \\
\frac{1}{\phi_z} &= 1 + \frac{0.16}{0.18} \\
\phi_z &= \frac{0.18}{0.18 + 0.16} \\
\phi_z &= 0.529
\end{aligned} \tag{12}
$$

たしかに、$\phi_z = 0.529$ で F が最小化するのを確認できた。

これでなにがわかったかというと、われわれ生物が F を下げてゆくためには、

$$
F \Leftarrow F + \alpha * F' \tag{13}
$$

という更新ルール(α は更新のスピードを決めるパラメーター)と F' を持っていればいい。そうすると F' が計算可能でないといけないわけだけど、確認のため、式(11)に元に記号を入れて戻してやると、

$$
\begin{aligned}
F' &= \ln \frac{0.16}{0.18} - \ln \frac{1-\phi_z}{\phi_z} \\
&= \ln \frac{p(Y=y_1, Z=z_2)}{p(Y=y_1, Z=z_1)} - \ln \frac{q(Z=z_2)}{q(Z=z_1)} \\
&= \ln \frac{p(Y=y_1|Z=z_2)p(Z=z_2)}{p(Y=y_1|Z=z_1)p(Z=z_1)} - \ln \frac{q(Z=z_2)}{q(Z=z_1)}
\end{aligned} \tag{13}
$$

となるので、第1項は生成モデルの A と C (図1-1C参照)さえあれば計算できる。True posteriorを計算する必要はない。そしてこの第1項は、二つの対立仮説($Z=z_1$ vs. $Z=z_2$)での観察データ y_1 についての対数尤度比になっている。これと第2項の現在の推測 q による対数尤度比との差が F' の正体だった。

われわれ生物は(すくなくとも)このような単純な環境では、尤度比検定をすることによって認識・推論の過程を最適化している、というのが自由エネルギー原理の世界観だということになる。

[1-11. このセクションで言いたかったこと]

長々と書いてきたが、 以降のセクションを読むために理解していほしいことはたった二つだ。(a) 変分推定を使うためには、 問題となっている状況の生成モデルがどういう構造になっているかを因果グラフ、因子グラフを用いてきっちり決めてやる必要がある。(b) そのうえで、変分自由エネルギー F を定義するためには、なにが観測データで、なにが潜在変数を理解して、あとはそれは式(5)に代入すれば一意に決まる。

では このセクションのおさらいを兼ねて、新しい生成モデルでこの二つの作業をやってみよう。図1-3Aでは さきほどの図1-1の状況にひとつだけ要素が加わっている。図1-1 Aでは、観測データY 「窓ガラスが濡れているか」、潜在変数 Z 「雨が降っているか」の二つの要素だけがあった。図1-3Aではこれに加えて、 潜在変数 X として「スプリンクラーが作動しているかどうか」がある。この図ではスプリンクラーの動作は雨が降るかどうかとは無関係にしてある。(もし、雨が降ったときはスプリンクラーは動かさない、という因果関係を設定した場合には、Z からX への矢印も必要になる。)

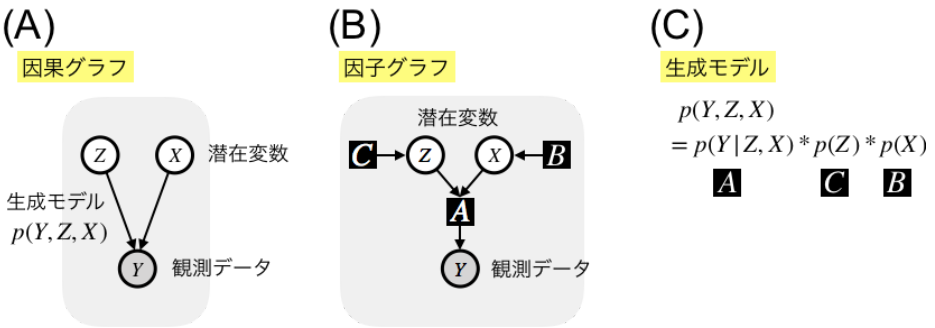

図1-4: 窓ガラスYから 降雨Zとスプリンクラーの動作Yを推測する

このような関係を因子グラフにすると図1-3Bになる。観察モデルA は二つの入力を受けて一つの出力を出すように変わった。ゆえに、雨が降っているかいないか $\{z_1, z_2\}$ とスプリンクラーが動作しているかいないか $\{x_1, x_2\}$ の4通りの組み合わせ $\{(z_1, x_1), (z_1, x_2), (z_2, x_1), (z_2, x_2)\}$ について、窓ガラスが濡れているかどうか $\{y_1, y_2\}$ の確率を知っている、これが観察モデルA だ(図1-3C)。

するとこのときのVFE F を計算するためには、観察データY と潜在変数 Z, X を式(5)の中に入れてやればいい。

$$
\begin{aligned}
F &= \mathbb{E}_{q(Z)}[\ln q(Z,X) - \ln p(Y,Z,X)] \\
&= \underbrace{D_{KL}[q(Z,X)\|p(Z,X|Y)]}_{\text{KLD}} + \underbrace{-\ln p(Y)}_{\text{Surprisal}} \\
&= \underbrace{D_{KL}[q(Z,X)\|p(Z,X)]}_{\text{Bayesian surprise}} + \underbrace{-\mathbb{E}_{q(Z,X)} \ln p(Y|Z,X)}_{\text{Uncertainty}}
\end{aligned} \tag{10}
$$

その結果が式(10)だ。つまり、式(5)でZ だったところが、Z, X になっているだけ。つまり、複数の変数は同時確率として扱われるので、カンマつけて並べておけばいい。

以上。以下に図1-3を作ったmatlabコードを貼っておく。

```
clearvars;
% 生成モデルを作る グラフィカルモデルから作成
% p(Z,Y)=p(Y|Z)p(Z)
% それぞれconditional probabilityで、縦方向に足すと1になる
gZ=[0.2 0.8]';
gY_Z = [[0.9 0.1];[0.2 0.8]]';

[Z,Y] = ndgrid(1:2, 1:2);
labels = sortrows([Z(:),Y(:)]);
gm = table(labels(:,1),labels(:,2), 'VariableNames', {'Z', 'Y'});

gmc = zeros(size(labels));
for ii=1:size(gmc,1)
    gmc(ii,1) = gZ(gm.Z(ii));
    gmc(ii,2) = gY_Z(gm.Y(ii),gm.Z(ii));
end
p_gm = prod(gmc,2); % sum=1になるのを確認した

% generative model
gm = [gm table(p_gm)];

% posteriorの計算
% Y (observed)でループ回す
unq = unique(gm.Y);
posterior = table([],[],[],'VariableNames', {'Z','Y', 'post'});
for jj = 1:length(unq)
    sel = find(gm.Y == unq(jj));
    for ii=1:length(sel)
        newrow = size(posterior,1) + 1;
        posterior(end+1,:) = ...
        table(gm.Z(sel(ii)),...
```

```
            gm.Y(sel(ii)),...
            p_gm(sel(ii)) / sum(p_gm(sel)));
    end
end
disp(gm)
disp(posterior)

phi_Zs = 0:0.01:1;
% Yごとに計算
Y = 1;
for ii = 1:length(phi_Zs)
    phi_Z = phi_Zs(ii);
    ret = calcVFEnew(gm, posterior, Y, phi_Z);
    VFE(1,ii) = ret.VFE;
    KLD(1,ii) = ret.KLD;
    Surprisal(1,ii) = ret.Surprisal;
    BayesSurp(1,ii) = ret.BayesSurp;
    Uncertainty(1,ii) = ret.Uncertainty;
end

% plot
fig1 = figure();
fig1.PaperOrientation = 'landscape';
fig1.PaperPosition = [4.8500 7.5000 20 8];

sp1 = subplot(1,2,1);
hold on
plot(phi_Zs, Surprisal, 'm.-');
plot(phi_Zs, VFE, 'k.-');
axis([0 1 0 2.6])
sp1.Box = 'off';
sp1.TickDir = 'out';

sp2 = subplot(1,2,2);
hold on
plot(phi_Zs, BayesSurp, 'g.-');
plot(phi_Zs, VFE, 'k.-');
axis([0 1 0 2.6])
sp2.Box = 'off';
sp2.TickDir = 'out';

function  ret = calcVFEnew(gm, posterior, Y, phi_Z)

    qZ = [phi_Z; 1-phi_Z];

    post = posterior.post(posterior.Y == Y,:);
    p_gm = gm.p_gm(gm.Y == Y,:);
    Surprisal = -log2(sum(p_gm));

    VFE = sum(qZ .* log2(qZ ./ p_gm));
    KLD = sum(qZ .* log2(qZ ./ post));

    % Zでループ回す
    unqx = unique(gm.Z);
    pZ = [];
    for ii = 1:length(unqx)
        % p(X) = sum_S(g(X,S))
        pZ(ii,1) = sum(gm.p_gm(gm.Z == unqx(ii)));
    end
    BayesSurp = sum(qZ .* log2(qZ ./ pZ));
    Uncertainty = VFE - BayesSurp;

    ret = [];
    ret.VFE = VFE;
    ret.KLD = KLD;
    ret.Surprisal = Surprisal;
    ret.BayesSurp = BayesSurp;
```

```
    ret.Uncertainty = Uncertainty;
end
```