予測符号化の哲学的含意と過学習の問題
Philosophical Implication of Predictive Coding and the Problem of Overfitting

柴田　翔平

Abstract

The aim of this paper is to show the philosophical relevance of predictive coding, a theory of perception in neuroscience. In philosophy of perception, not much attention has been paid to predictive coding. However, it can elucidate a philosophical problem which has not been discussed much, namely the relationship between perception and its generality. I start by introducing predictive coding with a statistical method necessary for understanding it and then argue that we see not only individual objects immediately present in our perception, but also general aspects of the objects.

1　研究テーマ

本論文では神経科学のみならず哲学においても影響力を増しつつあるリサーチプログラムである予測符号化を用いて知覚の哲学への含意を検討する。予測符号化に従えば知覚は神経系において事前知識を用い、感覚入力の原因を予測する過程である。予測符号化はより一般化され「自由エネルギー原理」として知覚に限らず行為や認知、価値などの多岐に渡る現象の説明に用いられている [1]。哲学的な関心も高まっており、Clark[2] や Hohwy[3] が知覚の哲学に応用している。予測符号化の特徴として統計・機械学習の手法を応用していることが挙げられるが、しかしこの点に注目した哲学的研究はまだ少ない。そこで予測符号化のこの点の応用の一例として知覚における過学習の問題を取り上げ、そこから知覚は個物にのみ関わっているのではなく、一般性の度合いを持ちうることを示唆する。この結論は知覚が目の前の個物に関するものであるという典型的な理解と対立することとなる。そうした典型的な理解として次の一節が代表的である [4]。

> 冬の最中に春の庭や、秋の葉っぱに覆われた庭を想像できるし、現在そうではないようなあらゆる仕方でその庭について考えることができる。これは知覚においては可能でない。なぜなら知覚は現在与えられているものにのみ直面しうるからである。

本論文における過学習からの議論が成功していれば、知覚の内容は現在与えられているものに限定されておらず、一般性の度合いを持つことが示唆される。予測符号化の統計的な側面を哲学的議論に持ち込むという新たな観点

から、知覚対象が個物なのか一般者なのかという知覚の哲学における議論 [5] に貢献することが本論文の目的である [1]。

本論文で示すのは以下のことである。(1) 知覚 (特に視覚) はベイズ推論によって捉えられること、(2) ベイズ推論は予測誤差最小化によって実現されていること、(3) 予測誤差最小化は統計的な手法であり過学習しうること、(4) したがって知覚でも過学習が起きうること、(5) 知覚における過学習は対象の持つ個別偶然的な性質に注意を過剰に向けることの帰結であることを示す。

知覚の哲学において予測符号化や統計的手法は頻繁に用いられないことを考慮し、数理的な説明は議論にとって必要最小限に留める。最小限の数理的な説明ののち、予測符号化の持つ哲学的な含意を検討する。

2 研究の背景・先行研究

予測符号化を理解するためには、知覚が一対多対応の関係を一対一対応にもたらす過程であることを認識する必要がある。つまり、私たちがなにかを見ているとき、見ているものは得られた感覚入力から一意的に得られるものではなく、知覚におけるなんらかの作用によって複数の候補の中から決定されているのである。というのも、感覚入力とその原因の間には一対多対応が成り立っているため、感覚入力だけからは可能な原因が複数存在してしまうからである。これは網膜が二次元であることを考えると分かりやすい。網膜上に映る像は二次元であるにも関わらず、私たちの見る対象はあたかも三次元上のものとして見られている。しかし、二次元上の対象を三次元に投射する仕方は一意的に決まらない。得られる感覚入力は二次元上のものであり、目の前の対象が自分からどれだけ離れているかという情報はそこにはないため、私たちが見ているものは単なる網膜上の像以上のものである。

このために、知覚は場合に応じて可能な候補から一つを選び取る仕組みを持っていると考えるのは自然であるが、この仕組みが何かというのは自明ではない。この仕組みを明らかにするのが知覚研究における課題の 1 つである。Rao and Ballard[6] は網膜における感覚入力が視覚野などの高次の領域に伝達されるだけでなく、高次の領域から低次の領域へと感覚入力に関する予測が伝達されることを発見した。予測符号化においては感覚入力と予測に基づき、ベイズ推論によって最も蓋然性の高い感覚入力の原因を推定し、その推定結果が知覚内容になると主張される。こうした観点から知覚を説明した研究として Buckley et al.[7] を解説する。

Buckley et al. はまず、周囲の状態 ϑ は直接知覚できず、感覚入力 φ から推測されなければならないということを仮定する。ただし、$p(\vartheta)$ と $p(\varphi)$ は ϑ と φが取りうる値の確率分布であり、ϑ と φは開区間 $(0, 1)$ に実数値を取る

確率変数である。例えば物体の凹凸の度合いが ϑ、網膜における光量が φ で表されるとすると、それぞれの確率分布は最もよくある凹凸度合いや光量に高い確率を割り当てる。ここで感覚入力 φ から状態 ϑ の推測をしたい。それを条件付き確率によって表したのが、「ある感覚入力が与えられたときの状態に関する確率分布」である事後分布 $p(\vartheta|\varphi)$ である。そしてそれを計算すること、つまり感覚入力に基づいた状態の推測を可能にするのがベイズの定理である。ベイズの定理より以下の等式が成り立つ。

$$p(\vartheta|\varphi) = \frac{p(\vartheta)p(\varphi|\vartheta)}{p(\varphi)}$$

事後分布 $p(\vartheta|\varphi)$ はある感覚入力 φ のもとで、周囲の状態 ϑ が取りうる確率である。ここで注意すべきは、φ がすでに観測された一つの値であるのに対して、ϑ は実現しうるであろう値であって、$p(\vartheta|\varphi)$ はそれがどのように分布しているかを確率的に表現しているということである。事前分布 $p(\vartheta)$ は周囲の状態 ϑ がどのように分布しているかを確率的に表している。エージェントは直接周囲の状態を知ることはできないが、これまでの経験からそれに関する信念を持つことはできるだろう。例えば、究極的には知り得ないことであるが、ある対象がどのような凹凸を持っているかということは、これまでのその対象との接触経験から設定することができる。尤度 $p(\varphi|\vartheta)$ の解釈は注意が必要である。尤度も事後分布と同様に条件付き確率の形式をしているが、一般にここでも変数となるのは状態 ϑ である。そのため尤度はどの状態を所与とするとある感覚入力が尤もらしいかを表していると解釈できる。残りの周辺尤度 $p(\varphi)$ は感覚入力 φ に関する確率分布であるが以下の式で求められる。

$$p(\varphi) = \int d\vartheta p(\vartheta, \varphi) = \int d\vartheta p(\vartheta)p(\varphi|\vartheta)$$

そのため事前分布と尤度の情報さえあれば事後分布を求めることができる。従ってベイズの定理においては ϑ のみが変数であり感覚入力 φ は固定された値であるから、ベイズ推論の使用は与えられた感覚入力 φ から状態 ϑ を予測するという目的に合致している。

しかし、周辺尤度を求めるための積分は計算機でさえもしばしば困難であるため、これを脳が行なっているというのはとても強い仮定である。そこで、$q(\vartheta)$ という別の確率分布を考え、$q(\vartheta)$ による $p(\vartheta|\varphi)$ の積分を用いない近似を考えることとする。$q(\vartheta)$ は認識モデルと呼ばれる。Buckley et al. はここから $p(\vartheta|\varphi)$ と $q(\vartheta)$ のカルバック・ライブラー距離を求め、ベイズ自由エネルギーの最小化のためにラプラス近似を行う。途中の導出は省略するが、最終的に認識モデルによる事後分布の近似のために必要な最小化の対象は以下のようになる。

$$E(\mu,\varphi) = -lnp(\varphi|\mu) - lnp(\mu)$$
$$= \frac{1}{2\sigma_z}\varepsilon_z^2 + \frac{1}{2\sigma_w}\varepsilon_w^2 + \frac{1}{2}ln(\sigma_z\sigma_w)$$

ただし、$\varepsilon_z \equiv \varphi - g(\mu;\psi)$、$\varepsilon_w \equiv \mu - \bar{\mu}$ とする。ここで μ は ϑ の確率分布を表現するための脳状態を表す確率変数であり、$\bar{\mu}$ はその平均、そして $g(\mu;\psi)$ は μ を変数、ψ をパラメータとして取る関数であり、φ の予測値である。すると ε_z は実際の感覚入力 φ と、その予測値である $g(\mu;\psi)$ の誤差を表しており、ε_w は実際の脳状態 μ とその平均の $\bar{\mu}$ の間の誤差を表していると解釈できる。そして、それぞれ φ の分散 σ_z と μ の分散 σ_w の逆数によって重みづけられている。最後の項は変数 φ と μ を含まない定数であるから最小化に当たって無視できる。すると、分散の逆数によって重みづけられた感覚入力の予測誤差と脳状態に関する予測誤差の最小化問題に帰結する。そのため感覚入力から状態の推測は予測誤差最小化によって達成される。予測誤差最小化については過学習の問題を論じる際に再び触れることとする。

ここまでが Buckley et al. によるベイズ推論の近似に関する説明である。しかしなぜ他の近似法ではなくラプラス近似を使うのかに関する説明がなかった。ベイズ推論は様々な近似法がある。しかし、予測符号化理論の発端である Rao and Ballard の発見は感覚入力と予測が神経系における高次の領域と低次の領域の間で伝達されることだけでなく、2 つの間の誤差が最小化されるということも含んでいる。この研究成果と整合的である理論を作ることが知覚研究における課題である。そのため知覚をベイズ推論によって特徴づけるためには神経系において予測誤差最小化が起きているという経験的制約と整合的であることが求められる。ラプラス近似の使用はこの観点から正当化できるだろう。

このようにして、予測誤差最小化が導かれた。ここから (1) 感覚入力の予測値と実際の値、そして (2) 脳状態の予測値と実際の値の誤差を 0 に近づけることが要請される。そして、感覚入力と脳状態のそれぞれの誤差は分散の逆数によって重みづけられる。例えば、感覚入力の誤差が小さく、脳状態の誤差が大きかったとしても、感覚入力誤差の分散が小さければ、結果として誤差は大きくなり、脳状態の誤差を上回ることもあり得る。その場合、脳状態の誤差よりも感覚入力の誤差を小さくするように動機づけられるだろう。この点はのちに取り上げる過学習の問題の際に重要になる。なお、Rao and Ballard に従えば予測誤差最小化は神経系のそれぞれの領域で行われるため、Buckley et al. はこの後予測誤差最小化を階層構造に拡張している。しかし、ここでの

目的には一つの層の中での予測誤差最小化で十分であるためこれ以上は踏み込まない。

3　筆者の主張

これまでこうした統計的な側面に注意を払った予測符号化の哲学における応用はされてこなかったが、こうした側面にも哲学的問題がある。その一例として過学習の問題を考える。過学習とは統計的学習理論の用語であり、統計モデルがすでに持っているデータに対して過度に適合してしまうために予測精度が低くなってしまうことを指す [8]。特にデータの分散が大きく不確実性が大きい時に、強引にパターンを見出そうとするとそのデータにとって偶然的な性質までもパターンとしてみなされてしまう。なおここでは不確実であるということを、分散が大きいことと交換可能なものとして扱う。つまり、何かが不確実であると言う時の主語は一つではなく、複数の感覚入力や脳状態などのデータである。データの分散はデータ全体の平均からの離れ度合いとして定義されるから、分散が大きいというのは予測が難しく不確実な状態ということである。逆にデータの分散が小さく平均の周囲に密集していれば予測は簡単だろう。例えば人間の身長と体重に関するデータを得た場合、それは普通 1 次関数に従う過程によって得られるが、測定誤差を含んだデータに適合するパターンを作ろうとすればデータは直線から少しずれるため 1 次関数よりも複雑なパターンになる。しかしそうしたパターンは今後得られるデータを適切に予測できなくなるだろう。そのため誤差を無視する必要があるが、過度にデータに適合してしまうとこうしたことが起き、これが過学習として知られる現象である。

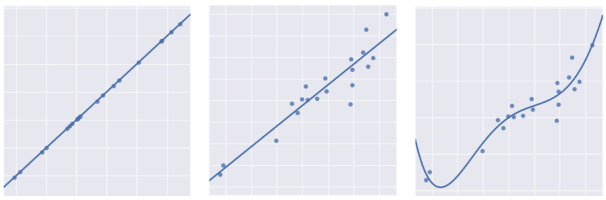

図1　図2　図3

以上は (図 1) 分散の小さいデータでの予測誤差最小化と、(図 2) 分散の大きいデータでの 1 次関数による予測、(図 3) 分散の大きいデータでの 5 次関数による複雑な予測である。図 1 の場合には適切な予測を立てることは容易だが、図 2 と図 3 の場合には容易ではない。ここでは図 2 と図 3 は図 1 のデータにノイズを加えて発生させたものである。そのため適切な予測は図 1 と同

じ、つまり図 2 である。他方で図 3 はここのデータにおける偶然的なノイズに適合し過学習しているため、予測がうまくいかなくなるだろう。図 3 よりも適合の度合いが強い予測も考えることができるが、そうした予測が現在のデータに偶然的な性質に適合してしまっているのは明らかだろう。

知覚の話に戻すと、知覚は予測誤差最小化を行なっているのであったが、それは以下の式で特徴付けられるのであった。

$$E(\mu,\varphi)=\frac{1}{2\sigma_z}\varepsilon_z^2+\frac{1}{2\sigma_w}\varepsilon_w^2+\frac{1}{2}ln(\sigma_z\sigma_w)$$

これはそれぞれの誤差に関する項にデータ（感覚入力、脳状態）の分散の逆数が乗じられている。データの分散はデータが平均から離れているほど大きくなる。その場合には分散の逆数は小さくなるため、誤差を小さくしなくても最小化したい値 $E(\mu,\varphi)$ は小さくなる。逆に予測誤差が小さくても分散が小さければ、$E(\mu,\varphi)$ が大きくなってしまうため、予測誤差をさらに小さくする必要がある。データの分散が小さく不確実性が小さい時には細かいところまで予測誤差を小さくする必要がある。そのため予測と大きく異なる感覚入力を得た場合、それと適合するように予測を改訂する必要がある。一方データの不確実性が大きい時に同様の仕方で予測を改訂すると、図 3 で見たようにデータへの適合が過剰になり過学習が生じてしまう。

感覚入力や脳状態が不確実であればあるほど分散が大きくなり弱く重みづけられるようになるため、予測誤差を小さくする必要性が少なくなる。分散が大きい分布から得られたデータに対して過剰に誤差を小さくしようとすると、そのデータの持つ偶然的な特徴にばかり適合してしまい、同じ確率分布から得られるであろう未知のデータに対する適合度が低くなってしまう。不確実性の中に無理やりパターンを見出そうとすれば、今後得られるデータにはないであろう特徴を見出してしまう。そして二次元上の像を三次元に投影する際に予測を使っているのであるから、予測が失敗すると正しい投影ができなくなり、視覚にも問題が出てしまう。

どんな知覚も以前と同じ状況ということはありえないから、知覚において予測をするためには目の前の対象を一般性において捉える必要がある。例えば目の前の机を見る時、その机の見え方は今まで見てきた机と完全に同一ということはないだろう。しかし、机が絵の中にある二次元上の対象ではないものとして、そして適切な距離感にあるものとして知覚するためにはこれまで机がどのように見えてきたかという情報を用いる必要がある。そのためにはこれまで見てきた机がどれだけ多様であったか（つまり不確実であるか）に応じて、予測誤差を小さくする必要がある。より不確実なときには目の前の机が持つ個別的な性質とこれまでの経験に基づく予測の間の誤差を小さくしす

ぎるべきではないが、これは知覚内容が必ずしも目の前の個別的対象ではなく、一般者にも関わりうることを含意する。予測誤差を小さくせずそのままにするとき、知覚内容は目の前の机だけでなく、机一般にも関するものとなっている。仮に目の前の机の天板がこれまで見てきた机と微妙に異なる見た目であったとしても、予測誤差を小さくしようとして予測を改訂すれば、過学習の問題が示唆するようにその後の予測がうまくいかなくなるだろう。知覚が現在与えられているものに直ちに反応しなければならないとする知覚の典型的な理解はこの点を見落としている。得られる感覚入力が不確実な時には予測誤差を減らそうとしすぎれば予測誤差最小化によるベイズ推論の近似が成り立たなくなり、感覚入力から状態の推測がうまくいかなくなってしまう。

こうしたことが実際に知覚において起きうることの論拠として、自閉症スペクトラム患者の特徴的な知覚を過学習の結果とする研究が存在する [9]。そうした研究によれば、例えば自閉症スペクトラムの症状の一つである定型発達者ならば気づかないような対象の特徴に注意が向く傾向は、分散の逆数によって規定される度合いを超えた予測誤差の最小化によって過学習が起きると解釈される [10]。そうした例としてあげられるのがコミュニケーション時の顔の知覚である。人の顔は多様であり不確実であるため、目の位置や鼻の形、肌の質感などの予測誤差を下げようとしすぎると過学習してしまう。過学習すると感覚入力と知覚内容の一対多関係の解決に必要な予測がうまくいかなくなってしまう。顔を顔として認識できなければ相手の気持ちを顔から読み取ることも難しくなるためコミュニケーションが困難になるというのは自然だろう。そのためこれまでの経験に基づいた予測と異なる顔を見たとしても、今見ている顔に過剰に適合するような予測をするべきでないこと、そして個別的な顔を通して一般者としての顔を知覚する必要があることが以上の議論から理解できるだろう。

4　今後の展望

今後の展望としては、(1) ここで提示した知覚観を用いることで自閉症スペクトラムの特徴的な知覚をより詳細に分析すること、(2) 予測符号化をより一般化し、知覚のみならず行為や意識、価値などを説明することを試みる自由エネルギー原理においてここで検討した一般性の概念がどう関連しているかということ、(3) 予測符号化とは異なる観点から知覚の対象に一般性を帰する立場 [5] との関連性を検討すること、(4) 過学習の問題以外の予測符号化が用いている統計的手法が持つ哲学的含意を検討すること、これら 4 つの方向での研究を進めることが必要となるだろう。自由エネルギー原理において行為もまた予測誤差最小化を行なっているとされており [1]、現在は特に行為にお

ける過学習の研究を進めている。

注

[1] しかしながら既存の議論は予測符号化の研究を用いてこなかったため、ここでは予測符号化がそうした議論と関連しすることを示すに留め、既存の立場とどのような関係にあるかということは今後の研究の課題とする。

文献

[1] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127, 2010.

[2] Andy Clark. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015.

[3] Jakob Hohwy. *The predictive mind*. Oxford University Press, 2013.

[4] Tim Crane and Craig French. The problem of perception. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[5] Hannah Ginsborg. Perception, generality, and reasons. In Andrew Reisner and Asbjørn Steglich-Petersen, editors, *Reasons for Belief*, pages 131–157. Cambridge University Press, 2011.

[6] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

[7] Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 2017.

[8] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[9] Sander Van de Cruys, Lee de Wit, Kris Evers, Bart Boets, and Johan Wagemans. Weak priors versus overfitting of predictions in autism: Reply to pellicano and burr (tics, 2012). *i-Perception*, 4(2):95–97, 2013.

[10] Sander Van de Cruys, Kris Evers, Ruth Van der Hallen, Lien Van Eylen, Bart Boets, Lee de Wit, and Johan Wagemans. Precise minds in uncertain worlds: Predictive coding in autism. *Psychological review*, 121(4):649, 2014.

（北海道大学）