論 文 の 内 容 の 要 旨

論文題目　Unconscious Inference in Neural Networks:
Electrophysiology and Learning Theory
（神経回路網における無意識的推論：電気生理と学習理論）

氏　　名　磯村　拓哉

人はどのようにして外界を知覚しているのだろうか？19 世紀の物理／生理学者 von Helmholtz は、脳は無意識的に、常に感覚入力の背後にあるダイナミクスと原因を推論する装置であるとし、これを無意識的推論と呼んだ。この推論を実現するために、内部モデル仮説（人は脳内に感覚入力を生成する外界の力学システム（生成モデル）を再構築し、それに基づき常に無意識的に推論を行い、背後にある隠れた原因を推定し未来の入力を予測しているという仮説）が提案されている。内部モデル仮説に基づき、脳の認知機能は経験的・現象論的には機械学習的な数理モデルにより説明できると考えられている。その内のひとつが Helmholtz machine であり、脳内に生成モデルを模倣した内部モデルを形成し、外界のダイナミクスを脳内にコピーすることで推論を行う。近年、これらの機械学習則を統一的に説明する法則“自由エネルギー原理”が提唱されている。自由エネルギー原理は、脳を構成する神経細胞・シナプス結合・神経修飾物質は、学習を実行するため、入力信号の予測不可能性を意味する自由エネルギーを最小化させるように振る舞うとする仮説である。特に本論文では、無意識的推論の数理的基礎とされる、“内部モデル仮説＋自由エネルギー最小化による最適化の仕組み（学習則)”のことを自由エネルギー原理と呼ぶこととする。重要なことに、無意識的推論は様々な能力の複合であり、自由エネルギー原理によると、多様な認知機能を推論のモデルとして統一的に説明可能である。

そこで無意識的推論の構成要素である背後の原因の推論および未来の入力の予測に着目する。背後の原因の推論は、ブラインド信号源分離（blind source separation; BSS; 混ぜ合わさった複数の入力からその背後にある個々の信号源を取り出す手法）として知られ、カクテルパーティ効果（うるさい場所でも特定の話者の声を聞き分けられる能力）を脳が実行するためのメカニズムであると考えられている。未来の入力の予測に関しては、予測的符号化（predictive coding; PC）仮説が提案されており、大脳皮質は内部モデルに基づき現在の感覚入力から未来の入力を予測していると考えられている。

しかし、(i) 既存の研究は学習の結果形成された内部モデルを観察しているのみであり、その学習過程は観察されていないため、脳活動計測からの神経回路網レベルの学習則の特定は困難

である。そのため、実際の神経回路網が内部モデルを学習する過程を自由エネルギー原理により説明可能か否か、どのような構造・学習則で実装されているのかは未解明である。また、(ii) 内部モデル仮説は機能面を説明しているに過ぎず、自由エネルギー原理において主に用いられる Helmholtz machine 型モデルは、神経回路網には実装できない構造であるため生理学的に妥当ではない。さらに、(iii) 自由エネルギー原理は生成モデルが複雑な場合にも推論可能かの議論が不十分である。

そこで本研究では、自由エネルギー原理の妥当性・有用性を示すため、(i) *in vitro* 神経回路網が BSS・PC を通して入力情報の背後にある生成モデル（確率的力学システム）を学習する過程の観察、理論との整合性の確認、(ii) 実験結果と整合する生理学的に妥当な数理アルゴリズム（学習則）の開発、工学的有用性のデモンストレーション、および (iii) 数理的・心理学的観点から自由エネルギー原理の複雑な問題への適用範囲の拡張を行う。

2 章では、神経回路網における内部モデルの獲得過程を観察し学習則を同定するための生理学実験を行った。実験では、ラット胎児の大脳皮質から取得した細胞を、64 点の電極が配置された微小電極アレイ（microelectrode array; MEA）デバイス上で分散培養し、刺激・計測両用の微小電極から神経細胞に電気刺激を入力し、訓練刺激に対する神経細胞の誘発応答を継続的に計測した。訓練刺激として、隠れた構造とダイナミクスを持つ確率的力学システムから生成した信号を電気パルスに変換し約 14 時間に渡り入力した。BSS 課題では、入力は、2 つの隠れた信号源から独立に生成した時系列を、確率的に混合した信号を 32 の電極から印加した。混合された複数の電気刺激を十分な時間印加し、その訓練の間神経細胞の誘発応答スパイクを計測し続けることで、培養神経回路網の学習過程を観察した。PC 課題では、100 ms 間隔の 4 種類のランダムドットパターンや 7 種類の数字のパターンからなるシークエンスを MEA 上の培養神経回路網に対して入力し、回路網がシークエンス入力に適応する過程を観察した。

BSS 課題に関しては、計測した誘発応答を基に計算した応答特異性（片方の信号源のみに応答する性質）の増加量を学習達成の指標とした。継続的に訓練を行った結果、神経細胞が BSS を行い、応答特異性が増加して混合前の信号の一方を再現する様に変化することを発見した。つまりカクテルパーティ効果により片方の人の声を聞き分けられる様に、神経回路網のレベルにおいても片方の信号源の情報を抽出できることを示した。次に PC 課題に関しては、各パターンに対する応答の増加量を比較したところ、シークエンス内で近接する前後のパターンに対する応答が訓練により増加した。また変化が見られた神経細胞は、直前か直後の一方のパターンに対する応答が増加する傾向にあった。したがって、培養神経回路網が入力の順序を予測できることを示した。

以上より、複数の信号が混合した入力の背後にある隠れ信号源を表現する神経細胞や、未来や過去の入力の期待値を表現する神経細胞が現れることを観察した。またこの学習過程は、シナプス可塑性の阻害薬（APV）によって抑制され、GABA ニューロンからの入力の増減の影響を受けたため、活動と GABA による修飾に依存したシナプス可塑性により行われていることが

強く示唆された。さらにこれらの学習が、自由エネルギー原理の予言通りに、自由エネルギーを最小化するようにシナプス結合が変化することで起きること、状態依存型 Hebb 型可塑性に従うことを発見した。これらの結果により、自由エネルギー原理が神経回路網の学習過程と整合することを初めて示した。

3 章では、2 章で観察した学習過程を説明できる生理学的に妥当な BSS 学習則“error-gated Hebbian rule (EGHR)”を開発した。EGHR は、Hebb 型可塑性を第 3 の要素が修飾する形の数式により、確率的力学システムの推論ができる点に新規性がある。既存モデルと異なり、実際の神経細胞がシナプス結合を介して知り得るローカルな情報のみを用いて学習が可能であり、かつ既存モデルよりもシンプルな実装が可能である。数理解析により解の安定性と信頼性を確認し、EGHR が BSS と PC を同時に実行し、内部モデルを学習できることを示した。手法間の比較により、EGHR が既存手法よりも正しい解を得られる信頼性が高く、工学的にも有用であることを示した。また複数の入力の同時確率を推定し、入力同士を連想することで、入力の補完や想起も行えることが分かった。

次に、自然画像（手書き数字や顔画像）を入力した場合には、入力情報の背後にある階層構造を抽出し、ラベルなしクラスタリングを行うことができた。楽曲（オーケストラ）を入力とした場合には、背後にある因果関係を学習し、楽曲を記憶・想起することができた。これらの結果により、EGHR は自然な視聴覚入力に対しても柔軟な推論能力を有することを示した。

さらに、EGHR は信号源より神経細胞の数が多い場合に BSS を実行できる唯一のローカル BSS 学習則であり、神経細胞数が多いとき過去に経験した環境への再適応（学習）速度が上昇すること、線形に近い非線形生成モデルから生成した入力からの信号源抽出（非線形 BSS）を実行できることを示した。最後に EGHR と自由エネルギー原理の関係を議論し、EGHR の応用により、従来の Helmholtz machine 型モデルとは異なるアプローチから、生理学的知見と自由エネルギー原理の双方と整合性のある無意識的推論の数理モデルを構築できることを示した。

4 章では、数理的および心理学的観点から、自由エネルギー原理の適用範囲の拡張を行った。前半では、コスト関数が定義される最適化問題における大域解探索の計算コストの期待値を求めた。自由エネルギー原理における最適化は勾配学習（ローカルサーチ）であるため、生成モデルが非線形の場合は局所解に陥る可能性がある（局所解問題）。そこで大域解を得るのに掛かる試行回数の期待値を解析計算し、必要な計算コストが現在知られている上限値よりも十分小さいことを示した。また、複数の内部モデルが脳内に同時に存在すると考え、遺伝的アルゴリズムとして知られる“選択と交叉”を用いて最適化を行うと、ローカルサーチのみの場合よりも効率的に大域解を探索できることを示した。

後半では、複数の内部モデルを持つことにより複数の推論を並列化できることをモデル研究により示した。他者を含む外界を推論の対象として捉え、他者の脳内状態（内部モデル）をコピーし自分の脳内に再構築することで、他者の思考の推論が可能である。そこで、外界の生成

モデルと他者の内部モデルを同時に推論するモデルを考案した。例として鳥の歌の生成モデルを用いて、提案モデルが複数の鳥の歌を別々に推論できることをシミュレーションにより示した。また提案モデルが、既存モデルとは異なり、自己と他者の思考を区別し推論できるかを問う課題（Sally-Anne test; 自閉症の診断方法の一つ）を解決できることを示した。

まとめると本研究では、(i) 実際の神経回路網が、自由エネルギー原理に従い BSS・PC を実行できることを初めて示し、(ii) 実験結果との類推から新たなローカル学習則 EGHR を発見し、無意識的推論の数理モデルとしての応用可能性を示し、(iii) 複数の内部モデルを用いると探索効率が上昇すること、自己と他者の思考を区別して推論できることを示した。本研究は、実際の神経回路網を用いて内部モデルの学習過程を観察し、未解明の仮説を検証した点において前例がなく独創的である。また提案した EGHR は従来のモデルよりも生理学的妥当性および解の安定性・信頼性の面で優れており、学術的新規性・工学的有用性が高い。本研究の結果は、培養神経回路網や人工ニューラルネットワークのような初期状態がランダムな回路網でも、Hebb 型可塑性を通じて内部モデルを構築し、確率的力学システムの無意識的推論を実装可能であることを強く支持している。また、生体由来の神経回路網における学習が、機械学習の理論によって説明可能であることを示唆しており、生物とコンピュータという違いに依らない共通の知能の存在を示唆している。本成果は、人工知能に入力情報の原因や他者の思考を柔軟に推論する能力を付与し、脳高次機能の神経メカニズムを統一的に説明するために利用できると期待される。