仮想研究課題レポート
大脳基底核の強化学習はどこまで正しいのだろうか

雨森 賢一・玉井 信也

はじめに

大脳基底核では強化学習が行われているのではないかという仮説がある。この「強化学習＝基底核」の適用範囲について議論したい。 Barto らの仮説によると基底核の学習方式は、 TD 学習であるとされている。本仮想研究では、この TD 学習の適用範囲と、基底核における学習の適応範囲を比較することにする。

基底核 - 視床 - 大脳皮質回路

まず簡単に、基底核回路の構造について説明する。基底核は、大脳皮質の様々な領野から興奮性の投射を、まず線条体 (striatum) で受ける。線条体はストリオゾームとマトリックスと呼ばれるコンパートメントに分けられており、ストリオゾームは黒質緻密部のドーパミンニューロンに、マトリックスは淡蒼球外節と黒質網様部に投射している、マトリックスからの２つの経路は、淡蒼球外節、視床下核を経て黒質網様部へ至る間接経路と、直接、黒質網様部へ至る直接経路があり、直接経路に投射する線条体ニューロンの興奮は対応する視床 - 皮質の回路を興奮させ、間接経路へ投射する線条体ニューロンの興奮は、視床 - 皮質の回路を抑制することになる。

黒質緻密部

黒質緻密部には、線条体のストリオゾームと、扁桃体からの投射がある。この扁桃体からの投射により、黒質緻密部のドーパミン細胞は、ジュースなどの報酬刺激により強く反応する。ところがサルに、何らかのタスクをさせたのちに報酬を与える場合、このドーパミン細胞は報酬を予測させる刺激に反応し、逆に報酬自体には反応しなくなるという変化を見せる。これは、ドーパミン細胞の反応が、線条体からの投射により、報酬自体ではなく、報酬を予期させる感覚入力に反応するように変換されていることを示唆する。このことから、 Barto らはドーパミンニューロンの反応は、強化学習における TD 誤差に対応するのではないかと主張した。将来の報酬の累積を $\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}$ とし、時刻 t の報酬の予測値を P_t とすると、隣り合った時刻の予測値に対して $P_{t-1} = r^t + \gamma P_t$ が成り立つ。時刻 t に実際に得られる報酬が r_t であったとすると、この予測誤差（TD 誤差) $\hat{r}_t = r_t - \gamma P_t - P_{t-1}$ を小さくするように学習が進行すれば良い。この TD 誤差は、報酬そのものではなく、報酬の予測の誤差に相当し、ドーパミン細胞の活動は、まさにこの TD 誤差と同様の活動度を示すものと言える。いま、扁桃体からの入力は r_t であると言えるから、線条体からの入力は $\gamma P_t - P_{t-1}$ である。ドーパミン細胞の興奮は線条体の入力部分に投射し、予測よりもより多く報酬を得られた直接の空間状況を強化し、 P_t を増大させる。

基底核の学習が TD 学習だったとすると？

強化学習の特徴は、報酬を得る直前の行動に強い価値を割り当てるというものである。この時、学習はそれぞれの時刻における報酬の予測誤差を小さくするという形で進行する。 TD 学習の特徴は隣り合った時刻の報酬の予測値 P_{t-1} のみを用いて、報酬の累積を推定するマルコフ性が成り立っているという点である。すなわち、問題空間がやや複雑であるとか、問題空間が変化してしまうと言った場合、この学習法は不向きであるということができる。そこで、強化学習、特に TD 学習に不向きなタスクを、報酬に基づいて学習させた場合、どのような振る舞いをするのかという

問題を立てることができる。つまり、基底核の学習を調べるタスクとして、次の２つのタスクを用意する。一つは、強化学習で行えるであろうマルコフタスクで、もう一つは強化学習には難しいと思われる非マルコフタスクである。この２つのタスクを報酬に基づいて学習させ、振る舞いの違いを見ることにする。

強化学習に向いているタスク

強化学習は、問題空間が固定されている場合に向いていて、報酬を予測する状態へと報酬が伝播することによって学習する。４つのボタンからなるボタン押しタスクを考えよう。たとえば、 A、 B、 C、 D の順にボタンを押すと、 D を押したときに報酬が得られるとする。この時学習によって、 D、 C、 B、 A の順に価値が割り当てられ、刺激 A でドーパミンニューロンが反応するようになるものと思われる。

強化学習では難しいタスク

強化学習では難しく、推論によって比較的簡単に解けるであろうタスクを考案する。強化学習は、単一の問題空間に対して試行を繰り返すことに諺チて学習するため、問題空間が毎回変更されるような問題は向いていないと考えられる。そのため、問題空間が毎回隠れた規則で更新されるタスクを考案した。それは次のようなものである。やはり、A、 B、 C、 D の４つのボタンからなるものとする。今あるセッションで、ボタン A を押し、報酬を得たものとする。この時、強化学習に基づく考え方では、基底核に入力される空間情報は、ボタン A だから、ボタン A と報酬を結び付けるように、強化されると思われる。そこで、次のセッションでは必ず、前回のセッションで報酬がでた場所と違う場所に報酬を割り当てる、という隠れた規則を導入しよう。すると、強化学習のように報酬の直前の状態に価値を割り当てる限り、いつまでも学習が収束しないと考えられる。

強化学習に不向きなタスクができるか?

もし、強化学習に不向きなタスクが遂行された場合、次の可能性が考えられる。一つは強化学習によって、タスク間の相互の関係を含む長期予測が可能になっているということである。しかし、これは、タスク間のインターバルを十分に長くとるなどして、排除することが可能であろう。もう一つの可能性は、タスク間の関係に関して何らかの推論を行っているという可能性である。この時の脳活動はどのようになっているのかを調べたい。

実験課題

基底核で行われている学習は、問題空間の性質に応じて限定できるのではなかろうか? この仮説に基づき、次のようなタスクを考案する。まず、サルに強化学習で可能な課題を遂行させ、その課題が達成されたとしよう。この時、強化学習では難しい課題が遂行できるのかどうかを調べる。強化学習では難しい問題が急に解けなくなる場合、やはり、強化学習が行われていることが示唆される。反対に、比較的簡単に遂行された場合、次の２つの可能性が考えられる。一つは、強化学習が行われていると仮定するのは疑わしいのではないかということ。もう一つは、報酬に基づくタスクにも関わらず、基底核ループによる学習ではない、という可能性である。この場合、基底核の入力が大脳皮質であることから、大脳皮質に部分的にムシモールなどを投与することで、強化学習に不向きな推論を行う部位を特定できる可能性があるかもしれない。

さいごに

TD 学習のマルコフ性に基づき、基底核で行われている学習が強化学習であるかを推定するタスクを考案した。上記の議論から、大脳皮質からの投射を切除、あるいは麻痺させることによって、推論を行う経路が特定できるのではないかと考えられる。その場合、マルコフ問題は基底核で解かれ、非マルコフ問題は大脳皮質で解かれているなどの、問題空間によって、活動する場所の変化が見られるのではないかという期待ができるかもしれない。

補足

本仮想研究は、短時間で作成する必要があったため、非常に未成熟なもので、欠陥も多い。現実的には、マルコフ性、非マルコフ性は、 TD 学習の枠組みのみで議論すべきではないと思っている。この研究課題は基底核の学習に不馴れな筆者らの勉強の過程ととらえていただきたい。