1.　強化学習の概要

これから勉強していく「**強化学習**」は、「教師あり学習」や「教師なし学習」と並んで機械学習の問題設定を表す単語です。以下では、強化学習の問題設定とその数学的な定式化を説明していきましょう。

1.1　強化学習の問題設定はどんなものか？

1.1.1 登場人物 : 環境とエージェント

強化学習の問題設定を説明するために、2人の登場人物「**環境**（environment）」と「**エージェント**（agent）」を紹介します。一言でいえば、環境は制御される対象・エージェントは環境を制御するものです。

Example 1.1.1 : イメージを掴むために、プラント内の温度の最適化を考えます。プラント内の温度は放っておくと以下のグラフの様に上昇していきます。ところが、適温は125°Cから130°Cの間（グラフの帯の部分）だというわけです。そこで、冷却パイプを入れる / 入れないことで温度を制御しようと考えます。

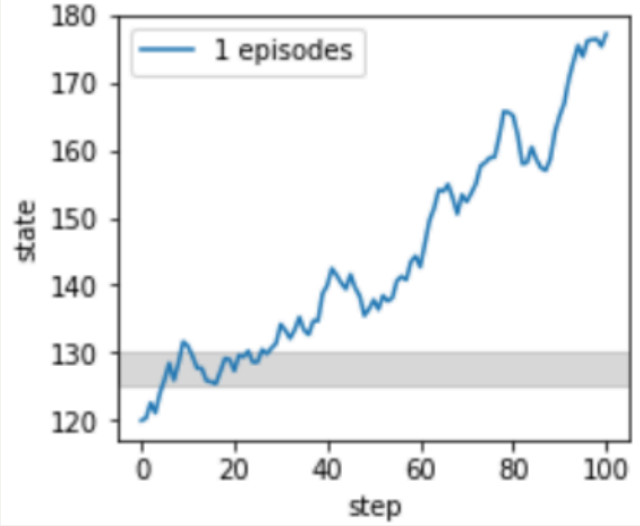

Fig. プラント内の温度最適化

この場合は、「プラント内の温度」が制御される対象で環境と呼ばれるものになります。また、「冷却パイプ」のコントローラーが環境を制御するものでエージェントと呼ばれるわけです。

1.1.2 強化学習の問題設定

環境に対してエージェントは何度も試行錯誤できる、またはすでに試行錯誤したデータがあるとします。この試行錯誤の結果から、なるべく分析者の目的に適ったエージェントの「行動指針」を求めたい。これが強化学習の問題設定です。強化学習では、このエージェントの「行動指針」を**方策**（policy）と呼びます。

強化学習の問題設定は、現在の方策に基づいてエージェントに試行錯誤をさせ、その結果に基づいて方策を改善していく**オンライン学習型の強化学習**と、すでに試行錯誤した結果のデータを持っていて、このデータから最適な方策を推定する**バッチ学習型の強化学習**の2種類に大別されています。以下では、オンライン型の強化学習に絞って説明していきます。

Example 1.1.2 : 1.1.1節の「プラント内の温度最適化」の例の場合、現在のプラント内の温度に応じて冷却パイプを入れるか・入れないかを決めることが、方策を決めることに対応します。既に人手でプラント内の温度最適化を試みたデータがあり、このデータから最適な方策を推定しようと考えた場合、これはバッチ学習型の強化学習を考えていることになります。

1.1.3 強化学習が活躍するとき

強化学習は、エージェントの「行動（action）」によって環境の「状態（state）」がどの様に遷移するか、要するに環境の仕組みは**未知**であるという仮定の下で活躍します。もしこの情報が既知であれば、最適な方策を求める上で試行錯誤するデータをわざわざ取らずとも、エージェントの行動によって得られる結果がわかるためです。

環境の仕組みが既知の場合は「**プランニング問題**」と呼ばれ、強化学習と区別されます。例えば、動的計画法と呼ばれる**数理最適化**の手法群によって最適な方策を求められる可能性があります。

1.2 強化学習を数学的に考える準備をしよう

1.2.1 環境とエージェントの関係をまとめてみよう

環境とエージェントの関係を復習しましょう。ある時刻 t において、まずエージェントが環境の状態 s_t を観測します。そして、これまでの環境の状態遷移 $s_{1:t} = (s_1, \cdots, s_t)$, 行動の履歴 $a_{1:t-1} = (a_1, \cdots, a_{t-1})$, 報酬の履歴 $r_{1:t-1} = (r_1, \cdots, r_{t-1})$ を参考に、現在の方策 $\pi^{(t)}$ に従って新たな行動 a_t を取ります。また、環境はエージェントの行動 a_t の働きかけに応じて、状態をs_t から s_{t+1} に遷移させます。

なお、その時刻 t において観測された環境の状態 s_t とこれに対してエージェントが取った行動 a_t に対しては、それがどれだけ分析者の理想に適っているか評価を行います。この評価値を r_t と書き、**報酬**（reward）と言います。