自由エネルギーに基づく強化学習

Reinforcement Learning based on Free Energy

星　野　　力

要　約　大規模言語モデル（LLM）の進展により，モデルの大きさ，データ量，学習時の計算量からスケーリングされる形で，世界にある情報を確率分布としてどの程度精密に学習できるかが明らかにされるとともに，大規模な学習済みモデルを利用したアプリケーションからもその有用性が強く示されている．今後は，学習された基盤モデルをもとに外部とのインタラクションを行い具体的なタスクを実行するエージェントへの発展が期待されている．外部の環境との相互作用を通じた試行錯誤に基づく学習は，強化学習とよばれ，教師信号が明示的に与えられず，勝敗や成功が報酬として得られるタスクで利用され，多くの成功を収めている．しかしながら，強化学習は，学習に頑健性がないこと，ハイパーパラメータの細かいチューニングを要する新しいタスクへの適用は容易ではないこと，統一的な視点に基づいた設計法が確立されておらずそれぞれの問題に特有の手法が要ることなど，残された課題も多い．本稿では，学習システムの統一的な理解を目指して，確率分布の学習で基本的な量をなす自由エネルギーを指標とした強化学習の定式化を提案する．自由エネルギーに基づく定式化を通じて，方策パラメータのエントロピーによる頑健化と，報酬の分散による探索と活用の適切なバランスが導出される．さらに，設定した目的関数を正規分布で近似する場合の局所的な最適化アルゴリズムを与え，解析結果を他の手法と比較し，提案手法との関係性を明らかにする．得られた結果は，統一的な視点からの学習システムの設計に対し，有用な知見を与えるものであると考える．

Abstract　With the advancement of large-scale language models, scaled with model size, data volume, and computational resources during training it has become clear to what extent these models can learn the information available in the world as a probability distribution. The utility of applications utilizing large pre-trained models has also been strongly demonstrated. In the future, there is an expectation for the development of agents that can perform specific tasks through interactions with external environments based on these learned foundational models. Learning based on trial and error through interactions with external environments is referred to as reinforcement learning, which is employed in tasks where explicit teacher signals are not provided and where victories, defeats, or successes are received as rewards. This approach has achieved considerable success. However, reinforcement learning also faces several challenges, including a lack of robustness in learning, the need for fine-tuning hyperparameters making it difficult to apply to new tasks, and the absence of a unified design methodology causing problem-specific techniques. This paper proposes a formulation of reinforcement learning based on the concept of free energy, which serves as a fundamental quantity in learning probability distributions, aiming for a unified understanding of learning systems. By means of this formulation via free energy, we derive robustness through the entropy of policy parameters and a proper balance between exploration and exploitation through reward variance. Furthermore, we present a local optimization algorithm for approximating a specified objective

function using a normal distribution, and we compare the analytical results with other methods to clarify their relationships. The results obtained are expected to provide valuable insights for the design of learning systems from a unified perspective.

1. はじめに

生成 AI の発展により，人間のみが行えると考えられてきた知的作業を計算機上のシステムとして構築できる可能性が広がっており，OpenAI 社の“ChatGPT”[1]などのアプリケーションを通じて一般にもその未来が見えてきている．さらに研究開発においては，複数のチームが汎用人工知能 Artificial General Intelligence（AGI）が今後 10 年程度で達成されるとのビジョンを掲げ，活発な競争を繰り広げている．実際に AGI が達成されれば，現状人間が行っている少なからぬ作業が自動化し代替されるとともに，人とシステムが高いレベルで協調することにより，研究開発を含む科学的な知識やテクノロジーのフロンティアが加速度的に発展することなどが考えられ，社会に対して産業革命の名に値するインパクトを持つことになるであろう．

現状の人工知能は，機械学習，特に，統計的な学習システムの発展を主な技術基盤としている．生成 AI の成功要因は，Transformer[*1] を含む深層学習モデル，GPU クラスター[*2] による大規模並列計算，WWW[*3] をソースとする大量のデータ等の技術革新により，スケーラブルな統計的学習が可能となり，これまでは不可能であった言語や画像などの情報の詳細な確率分布を獲得できるようになった点にある．例えば，Meta 社により構築され，正式に情報が公開されている大規模言語モデル“Llama3”[2]では，4050 億のパラメータを持つモデルを 15.6 兆のデータ（トークン）で学習しており，これは 10 年前では想像していなかった規模である．大きなモデルを大量のデータを使って解像度高く学習する方法の利点として，データの背後にある様々な構造（ルール）をモデルが自動的に抽出し，それをシームレスな形で統合することによって，精度をスケールさせることが可能になった点があげられる．例えば言語処理では，これまで，単語の分割，文法の推定，意味の計算とモジュールを分けたパイプラインで情報が処理されていたが，それらが次の単語の分布を正確に予測するという単純な学習により，ある程度自動的にボトムアップで獲得できることが示唆されている．従来の方法では，トップダウンなモジュール分けにより特化した高度な処理を入れ込むことができるが，切り分けた後でモジュール間の複雑な相互作用を取り込むのは難しい場合が多く，システム設計上のボトルネックとなっていた．

本稿では，上で述べた生成 AI での成功例も考慮し，より高度な知的システムの構築には，統一的な手法によりスケーラブルなアーキテクチャーを構築することが重要なファクターになるとの仮説を置く．一般に，統計的な学習システムは，その機能をもとに分類すると主に三つの手法から成り立っている．一つ目は，例えば大量の犬と猫のラベル付きの画像を学習し，学習したモデルを使って新しい画像を判定する手法であり，教師あり学習と呼ばれる．二つ目は，ラベルがないデータから，データの背後にある構造（データを生成している過程）を推測し，それを活用してデータの良い表現を与える手法で，教師なし学習と呼ばれる．LLM は明示的なラベルは与えられないが，次の単語をマスクして予測することで教師あり学習と教師なし学習を接続し，データの精密な分布およびその背後にある構造を学習することに成功した．三つ目は，強化学習で，ラベル付きデータの代わりに，各状態について，その状態の良さを表す報酬と呼ばれる値が得られる設定である．対戦型ゲームであれば，勝ちが確定した状態に対して

正の報酬を与えたり，ロボットがあるタスクを達成した状態に対して正の報酬を与える．強化学習の目標は，得られる累積報酬を最大化するように，与えられた状態に対して行動を決定する“方策”を学習することにある．例えば，自転車に乗る動作の学習などは，正解データを与えるのが困難なため教師あり学習は使いにくいが，転ばずに進めた距離を報酬として強化学習は適用可能である．現状では，これら三つの学習を統合するアプローチとしてアーキテクチャーの側面からの，“World Models[*4]”[5][3]や原理的な側面からの自由エネルギー原理[*5][4]等が提案され活発に研究されている．これらの提案が，計算論的神経科学とよばれる，人間を含む生物の脳の機能の解明を，大まかなモジュールとして大脳，小脳，中脳を持ち，それらの相互作用を使って様々な課題を統一的に学習するシステムとして探求する分野から多くの示唆を得ていることも興味深い．

仮説の検証に近づくための具体的な課題として，強化学習を見通しの良い形で定式化し解析することに焦点をあてる．強化学習は，ビデオゲーム[6]や碁[12]など閉じたシミュレート可能な環境で大きな成功を収めているが，連続的な状態空間と行動を持つロボット制御を含む実環境でのタスクも含め，統一されたアーキテクチャーをもとに設計することは未だ困難な状況にある．また，LLM を含む基盤モデルを外部とのインタラクションを通じて実際にタスクを実行するエージェントとして利用する場合にも，これまでにはない複雑な環境との相互作用やそこでのプランニングを要し，その機能の獲得には，基盤モデルと強化学習のシームレスな接続が不可欠となる．

また，統一的な視点のための指標としては，“自由エネルギー”を採用する．自由エネルギーは，統計的な学習では，統計モデルにおいて観測データが得られる確率を示す指標であり，対数周辺尤度もしくはエビデンスとも呼ばれる．この指標は，ハイパーパラメータの最適化，複数の仮説の事後確率に基づくモデル選択，汎化誤差の理論計算などの指標としても広く利用され，自由エネルギー原理もこの量を中心とした定式化がなされている．

本稿では，連続的な状態空間と行動を持つ環境におけるパラメータベースの方策探索において，報酬と方策の不確実性を考慮に入れた指標を利用する強化学習法を提案する．まず 2 章で方策探索を定義し，3 章で方策パラメータと報酬の不確実性の導入について述べる．4 章で提案手法を説明し，5 章で今後の課題について述べる．本稿における貢献は以下の 3 点に集約される．

- 方策と報酬の不確実性を明示的に考慮し，方策の頑健性および，報酬の探索と活用のトレードオフを最適化する指標を定義したこと
- 定義した指標を使って，高次元の連続空間上でも効率的に動作する学習アルゴリズムを構築したこと
- アルゴリズムの解析の結果から，ハイパーパラメータを自動的に設定し，探索空間を削減する手法を提案したこと

2. 方策探索

強化学習は，行動を通じて環境と相互作用しながら，与えられる報酬を最大化するために，方策と呼ばれる状態から行動への最適な写像を求める問題である．以下，この問題の数学的定式化とエピソードベースの方策探索について述べる．

2.1 仮定

時刻 T ステップまでの状態 $x_{1:T+1} \equiv (x_1, x_2, \ldots, x_{T+1})$，行動 $u_{1:T}$，および報酬 $r_{1:T}$ の結合確率分布が以下のように定式化されるマルコフ決定過程を仮定する．

$$p(x_{1:T+1}, u_{1:T}, r_{1:T}) = p(x_1)\prod_{t=1}^{T} p(x_{t+1}\,|\,x_t, u_t)p(r_t\,|\,x_t, u_t)p(u_t\,|\,x_t).$$

ただし，$p(x_1)$ は初期状態の確率分布，$p(x_{t+1}|x_t, u_t)$ は状態遷移確率分布，$p(r_t|x_t, u_t)$ は報酬確率分布，$p(u_t|x_t)$ は方策の確率分布を表わす．

2.2 方策勾配法

方策勾配法[16]では，方策 π は現在の状態 x_t と方策パラメータ θ に対して確率的な行動 u_t を出力し，それが探索のドライブとなるように設定される．

$$u_t = \pi(x_t, \theta).$$

また，エピソードベースの方策探索の場合，パラメータ θ の一つの試行 h での評価は，1 エピソードにおける報酬の和 r_h で定義される．

$$h \equiv (x_1, u_1, \ldots, x_T, u_T, x_{T+1}), r_h \equiv \sum_{t=1}^{T} r_t.$$

これらの仮定のもと，方策探索の目的は方策パラメータ θ の下での報酬の和 r_h の期待値を最大にする θ^* を求めることである．

$$r(\theta) = \int r_h p(r_h\,|\,\theta) dr_h, \theta^* = \arg\max_{\theta} r(\theta).$$

ただし，$r(\theta)$ はパラメータ θ に対する報酬の和 r_h の平均である．一般に，θ^* を求めるには勾配法を用いる．勾配法ではパラメータの更新の大きさを与えるハイパーパラメータである学習率が用いられる．

3. 不確実性の導入

4 章で報酬と方策パラメータの不確実性をともに考慮した目的関数を定義するので，本章ではそのための準備を行う．また，目的関数を最適化するアルゴリズムの計算量を削減するためのガウス近似について説明する．

3.1 方策パラメータの不確実性

本稿では，方策 $\pi(x, \theta)$ はニューラルネットワークで表現する．既存の手法の多くは，パラメータ θ を点推定で行っている．特にニューラルネットワークを含む複雑な非線形モデルでは，パラメータの摂動に対して，関数が大きく変動し，そのことが学習過程の揺らぎに対して不安定に応答する原因になっている．学習を安定化する有力な方法の一つに方策パラメータの分布の最適化を行う方法がありパラメータベースの方策探索と呼ばれ[11]，クロスエントロピー法[10]と強い関係を持つ．パラメータベースの方策探索では，方策パラメータの分布を考え，それを $p(\theta|\rho)$ とおき，目的関数 J を期待報酬 r_h の方策パラメータについて期待値をとったもので定義する．

$$J(\rho) = \int r(\theta) p(\theta \,|\, \rho) d\theta = E_{p(\theta|\rho)}[r(\theta)].$$

ただし，ρ は分布のハイパーパラメータである．さらに方策 $\pi(x_t, \theta)$ は決定論的で，探索は θ の $p(\theta|\rho)$ からのサンプルにともなう揺らぎにより与えられる．最終的な行動 u_t^* は，最適化された方策のパラメータの分布 $p(\theta|\rho^*)$ による出力のアンサンブル平均で与えられる．

$$u_t^* = \int \pi(x_t, \theta) p(\theta \,|\, \rho^*) d\theta.$$

さらに $J(\rho)$ の発展として負の自由エネルギー（報酬とエネルギーで符号が逆なため，“負の”自由エネルギーとなる．）を目的関数にする方法が提案されている[9]．負の自由エネルギーは，逆温度 $\beta > 0$ を持つ以下の関数で定義され，平均 $J(\rho)$ より大きな値を持つ．

$$F(\rho) \equiv \frac{1}{\beta} \log \int \exp(\beta r(\theta)) p(\theta \,|\, \rho) d\theta \geq J(\rho). \tag{1}$$

ここで，$F(\rho)$ の性質を調べるため，積分を $p(\theta|\rho)$ からの n 個のサンプル $\theta_1, \ldots, \theta_n$ で近似する．そのとき，以下の式が成り立つ．

$$F(\rho) \simeq \frac{1}{\beta} \log \frac{1}{n} \sum_{i=1}^{n} \exp(\beta r(\theta_i)) = E_q[r(\theta)] + \frac{1}{\beta} H(q) - \frac{1}{\beta} \log n. \tag{2}$$

ただし，q は n 次元の離散分布であり，各 q_i は

$$q_i = \frac{\exp(\beta r(\theta_i))}{\sum_{j=1}^{n} \exp(\beta r(\theta_j))}$$

で与えられ，$E_q[r(\theta)] = \sum_{i=1}^{n} q_i r(\theta_i)$ は q での期待値，$H(q)$ は q のエントロピー（$H(q) = -\sum_{i=1}^{n} q_i \log q_i$）である．式(2)から，$F(\rho)$ の最大化は q でみたときの逆温度 β でバランスされた期待報酬とエントロピーの和の最大化になる．

さらに，局所的な近似を用いて $p(\theta|\rho)$ を正規分布と仮定する．

$$p(\theta \,|\, \rho) \equiv \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta - \rho)^2}{2\sigma_0^2} \right).$$

仮定のもとでの目的関数の最大化は，定義した自由エネルギー $F(\rho)$ の ρ についての極値条件 $\dfrac{dF(\rho)}{d\rho} = 0$ を求めることになる．“log derivative trick” $\dfrac{d}{d\rho} \int p(\theta \,|\, \rho) f(\theta) d\theta = \int p(\theta \,|\, \rho) \dfrac{d}{d\rho} \log p(\theta \,|\, \rho) f(\theta) d\theta$ を用いて $p(\theta|\rho)$ からの n 個のサンプル $\theta_1, \ldots, \theta_n$ でその条件を近似すると，

$$\frac{1}{n} \sum_{i=1}^{n} \frac{d \log p(\theta_i \,|\, \rho)}{d\rho} \exp(\beta(r(\theta_i))) = 0$$

となる．この条件は，$p(\theta|\rho)$ が正規分布の場合は対数微分が

$$\frac{d}{d\rho} \log p(\theta \,|\, \rho) = \frac{\theta - \rho}{\sigma_0^2}$$

となることから

$$\rho = \frac{\sum_{i=1}^{n} \exp(\beta r(\theta_i))\theta_i}{\sum_{i=1}^{n} \exp(\beta r(\theta_i))} \tag{3}$$

としてρについて明示的に解くことができる．そのため，勾配法で用いられる学習率の調整は不要である．

3. 2　報酬の不確実性

パラメータベースの方策探索において，パラメータθが与えられたもとでの真の報酬の分布$p(r_h|\theta)$を知ることはできない．したがって，報酬の平均$r(\theta) = \int r_h p(r_h|\theta) dr_h$を何らかの方法で推定しなければならない．多くの既存研究では，固定されたθのもと，単純なサンプル平均$\tilde{r}(\theta) = \frac{1}{K}\sum_{k=1}^{K} r_{h_k}$で推定を行っている．しかしながら，この推定値は，推定量の不確実性を考慮に入れていない．不確実性の考慮がない場合，学習の確率的なゆらぎによっては，報酬を過大に推定し，探索不足になることが知られている．この問題に対し，O'Donoghue らは“K-learning”[7][8][13]を提案した．“K-learning”は単純平均の代わりに，キュムラントを利用することにより，不確実性を考慮したものになっている．提案されている統計量は，単純平均に比べてより探索的にふるまい，確率変数の独立性に対して加法性を持つなど数学的にも良い性質を持つ．しかしながら，彼らの提案は，状態空間および行動が離散値を持つ場合であり，それらを連続値を持つ場合へ展開することは，連続値の分布ではエントロピーが負の値を持ちえることなどもあり，決して自明ではない．

この問題に対し，一般化して考察する．パラメータベースの方策探索において，目的変数を報酬r，説明変数を方策パラメータθとするベイズ法による回帰問題を考える．

$$p(r|\theta, w).$$

ただし，wは回帰のパラメータであり，その事前分布を$p(w)$とする．K個のサンプル$D \equiv \{(r_1, \theta_1), \ldots, (r_K, \theta_K)\}$が与えられたもとで，パラメータ$w$の事後分布は以下のように記述される．

$$p(w|D) = \frac{\prod_{k=1}^{K} p(r_k|\theta_k, w)p(w)}{\int \prod_{k=1}^{K} p(r_k|\theta_k, w)p(w)dw}.$$

はじめに，平均報酬$\int rp(r|\theta, w)\, dr$と$w$の事後分布$p(w|D)$を使って$K(\theta)$を定義する．

$$K(\theta) \equiv \frac{1}{\beta}\log \int \exp(\beta \int rp(r|\theta, w)dr)p(w|D)dw. \tag{4}$$

次に，パラメータγを用いて，キュムラント母関数を定義する．

$$C(\theta, \gamma) \equiv \log \int \exp(\gamma\beta \int rp(r|\theta, w)dr)p(w|D)dw.$$

このとき，$K(\theta)$の二次近似は，以下で与えられる[15]．

$$K(\theta) \approx \frac{1}{\beta}\frac{d}{d\gamma}\bigg|_{\gamma=0} C(\theta,\gamma) + \frac{1}{2\beta}\frac{d^2}{d\gamma^2}\bigg|_{\gamma=0} C(\theta,\gamma)$$
$$= R(\theta) + \frac{\beta}{2}V(\theta).$$

ただし，一次と二次の微分は，それぞれ，平均推定量の平均 $R(\theta)$ と分散 $V(\theta)$ に対応する．

$$\frac{d}{d\gamma}\bigg|_{\gamma=0} C(\theta,\gamma) = \int p(w|D)\beta \int rp(r|\theta,w)drdw \equiv \beta R(\theta),$$

$$\frac{d^2}{d\gamma^2}\bigg|_{\gamma=0} C(\theta,\gamma) = \int p(w|D)\left(\beta\int rp(r|\theta,w)dr\right)^2 dw - \left(\int p(w|D)\beta\int rp(r|\theta,w)drdw\right)^2 \equiv \beta^2 V(\theta).$$

したがって，$K(\theta)$ の近似は，報酬の平均 $R(\theta)$ にその不確実さを表わす分散 $V(\theta)$ をボーナスとして加えた形で与えられる．ただし，分散 $V(\theta)$ は報酬の平均推定量の分散であり，r の予測分布の分散ではないことには注意を要し，平均推定量の分散を使うことは，“Thompson sampling”[14] と強い関係性を持つ．さらに，本稿では，$R(\theta)$ および $V(\theta)$ を θ を固定したもとでの m 回の試行で得られる報酬 $r_{1\theta}, \ldots, r_{m\theta}$ を用いて，

$$R(\theta) \equiv \frac{1}{m}\sum_{i=1}^{m} r_{m\theta}, \sigma^2(\theta) \equiv \frac{1}{m-1}\sum_{i=1}^{m}(r_{i\theta} - R(\theta))^2, V(\theta) \equiv \frac{\sigma^2(\theta)}{m} \tag{5}$$

として，報酬に正規分布を仮定した推定値を用いる．

4. 提案手法

本章では，本稿の提案手法である，3 章で考察した方策パラメータと報酬の不確実性を同時に考慮する指標を定式化し，その指標を用いた方策パラメータの分布の最適化アルゴリズムを述べる．

4.1 指標の設定

一般的な設定として，$p(\theta|\rho)$ を方策パラーメータ θ の分布（ρ は分布のハイパーパラメータ），$p(r|\theta, w)$ を θ が与えられたもとでの報酬 r の回帰モデル（w は回帰モデルのパラメータ），$p(w|\theta^n, r^n)$ を n 回の試行で得られたサンプルを使った回帰パラメータの事後分布とする．その時，イェンセンの不等式を使って下記が成り立つ．

$$\int\left(\int\left(\int rp(r|\theta,w)dr\right)p(w|\theta^n,r^n)dw\right)p(\theta|\rho)d\theta \tag{6}$$

$$\leq \frac{1}{\beta}\log\int\exp\left(\beta\int\left(\left(\int rp(r|\theta,w)dr\right)p(w|\theta^n,r^n)dw\right)\right)p(\theta|\rho)d\theta \tag{7}$$

$$\leq \frac{1}{\beta}\log\int\left(\int\exp\left(\beta\left(\int rp(r|\theta,w)dr\right)\right)p(w|\theta^n,r^n)dw\right)p(\theta|\rho)d\theta \tag{8}$$

$$\simeq \frac{1}{\beta}\log\int\exp\left(\beta R(\theta) + \frac{\beta^2}{2}V(\theta)\right)p(\theta|\rho)d\theta \equiv F_0(\rho). \tag{9}$$

これらの式は，それぞれ既存のアルゴリズムと対応を持つ．まず式(6)は Parameter based

policy Gradient[11]となる．式(6)の方策の分布$p(\theta|\rho)$を指数関数の外に出すと，式(7)のRelative Entropy Policy Search[9]となる．最後の式(8)が提案する指標であり，報酬の事後分布$p(w|\theta^n, r^n)$も指数関数の外に出したものとなる．この変換により，キュムラントを通じて平均だけでない高次の統計量が指標に導入される．

4. 2　アルゴリズムの導出

任意の試行分布$q(\theta)$から$p(\theta|\rho)$に$F_0(\rho)$と対応するよう重みづけされた分布へのカルバック・ライブラ情報量を考える．

$$\frac{1}{\beta} KL\left(q(\theta) \middle\| \frac{\exp\left(\beta R(\theta) + \frac{\beta^2}{2} V(\theta)\right) p(\theta|\rho)}{\int \exp\left(\beta R(\theta) + \frac{\beta^2}{2} V(\theta)\right) p(\theta|\rho) d\theta} \right)$$

$$= \frac{1}{\beta} KL(q(\theta) \| p(\theta|\rho)) - E_{q(\theta)}[R(\theta)] - \frac{\beta}{2} E_{q(\theta)}[V(\theta)] + \frac{1}{\beta} \log \int \exp\left(\beta R(\theta) + \frac{\beta^2}{2} V(\theta)\right) p(\theta|\rho) d\theta \geq 0.$$

最後の項の積分を$p(\theta|\rho)$からのn点のサンプルθ_iで近似し，さらに$R(\theta)$および$V(\theta)$を固定したθでのm回の試行により式(5)を利用して推定した値を代入して整理すると，

$$\frac{1}{\beta} \log \sum_{i=1}^{n} \exp\left(\beta R(\theta_i) + \frac{\beta^2}{2} \frac{\sigma^2(\theta_i)}{m}\right) + \frac{1}{\beta} KL(q(\theta) \| p(\theta|\rho)) \tag{10}$$

$$\geq E_{q(\theta)}[R(\theta)] + \frac{\beta}{2} E_{q(\theta)}\left[\frac{\sigma^2(\theta)}{m}\right] + \frac{1}{\beta} \log n \tag{11}$$

$$\geq E_{q(\theta)}[R(\theta)] + \sqrt{E_{q(\theta)}[\sigma^2(\theta)]} \sqrt{\frac{2 \log n}{m}} \tag{12}$$

が任意のβに対して成り立つ．ただし，最後の式は，式(11)をβについて最小化した値$\beta^* = \sqrt{\frac{2m \log n}{E_{q(\theta)}[\sigma^2(\theta)]}}$を代入し整理した．式(12)は，$q(\theta)$から$n$点サンプルをとり，それぞれの$\theta_i$で$m$回の試行を行って得た$r_{i\theta}$を使って求めた$E_{q(\theta)}[R(\theta)]$の近似的な上界を与えるため[13]，式(10)は，その上界を上から抑えていることがわかる．これらの事実を使うと，式(10)による逆温度βの最適化と，求めたβを用いた式(3)による方策のハイパーパラメータρの更新を繰り返して，式(10)を最適化するアルゴリズムを以下のようにして構成できる．

まず，$G(\beta, \rho, c)$を以下に定義する．

$$G(\beta, \rho, c) \equiv \frac{1}{\beta} \log \sum_{i=1}^{n} \exp\left(\beta R(\theta_i) + \frac{\beta^2}{2} \frac{\sigma^2(\theta_i)}{m}\right) + \frac{1}{\beta} c. \tag{13}$$

ただし，$c \geq 0$を満たす定数である．これは，式(2)を用いて以下のように書き換えられる．

$$G(\beta,\rho,c) \equiv E_{q'}[R(\theta)] + \frac{\beta}{2} E_{q'}\left[\frac{\sigma^2(\theta)}{m}\right] + \frac{1}{\beta} H(q') + \frac{1}{\beta} c.$$

ただし，q' は各 i が

$$q'_i = \frac{\exp\left(\beta R(\theta_i) + \frac{\beta^2}{2}\frac{\sigma^2(\theta_i)}{m}\right)}{\sum_{j=1}^{n} \exp\left(\beta R(\theta_j) + \frac{\beta^2}{2}\frac{\sigma^2(\theta_j)}{m}\right)}$$

を満たす n 次元の離散分布である．

$G(\beta,\rho,c)$ は β に対して凸なので，c 固定のもと，G を最小にする β^* は

$$\beta^* = \sqrt{\frac{2m(H(q')+c)}{E_{q'}[\sigma^2(\theta)]}}$$

を満たす．さらに，c を 0 から大きくし，各 c で最適化した β^* を用いて式(3)を用いて ρ^* を更新した方策を $p(\theta|\rho^*)$ とし，更新前後のカルバック・ライブラ情報量が $KL(p(\theta|\rho^*)||p(\theta|\rho)) = c$ を満たすような c^* を探索する．見つけた c^* を式(13)へ代入したものは式(10)と等しい．$p(\theta|\rho^*)$ と $p(\theta|\rho)$ の分散が等しい場合は，$KL(p(\theta|\rho^*)||p(\theta|\rho)) \geq 0$ は有限なので必ずこのような c^* を見つけることができる．

この (β^*,ρ^*,c^*) の組を G に代入して整理すると，

$$G(\beta^*,\rho^*,c^*) = E_{q'}[R(\theta)] + \sqrt{E_{q'}[\sigma^2(\theta)]}\sqrt{\frac{2(H(q') + KL(p(\theta\,|\,\rho^*)\,\|\,p(\theta\,|\,\rho)))}{m}}$$

となる．$H(q')$ は $0 \leq H(q') \leq \log n$ を満たし，$KL(p(\theta|\rho^*)||p(\theta|\rho))$ の平均はサンプリングによる誤差により $KL(p(\theta|\rho^*)||p(\theta|\rho)) \geq \frac{d}{2n}$（ただし，$d$ は方策ネットワークのパラメータ数）を満たす．n, m を適応的に調整して $H(q') + KL(p(\theta|\rho^*)||p(\theta|\rho)) \simeq \log n$ を満たすようにすれば，$\beta^* \simeq \sqrt{\frac{2m\log n}{E_{q'}[\sigma^2(\theta)]}}$ となる．

これらの設定のもと，$F(\rho)$（式(1)）の被積分関数である指数関数の中身は $\sigma_0^2 = \frac{1}{\sqrt{2m\log n}}$ とおけば

$$\sqrt{2m\log n}\left(\frac{R(\theta)}{\sqrt{E_{q'}[\sigma^2(\theta)]}} - \frac{1}{2}(\theta-\rho)^2\right)$$

以下となり，アルゴリズムの収束は，報酬の平均 $R(\theta)$ の θ 方向の連続性と各 θ での報酬の分布の裾の太さと分散 $\sigma_0^2 = \frac{1}{\sqrt{2m\log n}}$ の大きさの関係により決まることがわかる．

以上を整理すると，提案する方策パラメータの分布の最適化アルゴリズムは Algorithm 1 のように記述される．このアルゴリズムの計算量は，既存のパラメータベースの方策探索と同じオーダーである．

Algorithm 1　方策パラメータの分布の最適化アルゴリズム

入力：方策パラメータの初期分布 $p(\theta|\rho_0)$，反復数 L，集団のサイズ n，エピソードの繰り返し数 m
出力：目的関数 $F(\rho)$（式(1)）を最大化する方策パラメータの分布 $p(\theta|\rho)$

for $l = 1$ to L **do**
　$\sigma_0^2 \leftarrow \dfrac{1}{\sqrt{2m\log n}}$
　for $i = 1$ to n **do**
　　$\theta_i \sim p(\theta|\rho)$ でサンプリング
　　for $j = 1$ to m **do**
　　　方策 θ_i でエピソードを実行，結果を $R[j] \leftarrow r_h$ に代入
　　end for
　　$R[.]$ を使って $R(\theta_i) = \dfrac{1}{m}\sum_{j=1}^{m} R[j], \sigma^2(\theta_i) = \dfrac{1}{(m-1)}\sum_{j=1}^{m}(R[j] - R(\theta_i))^2$ を計算
　end for
　$c = 0$
　while true **do**
　　固定した c で(13)を最小化する β^* を探索し，β^* を使って式(3)で ρ を ρ^* へ更新
　　if $KL(p(\theta|\rho^*)||p(\theta|\rho)) \le c$ **then**
　　　break
　　end if
　　$c \leftarrow c + step(>0)$
　end while
　得られた (c^*, β^*) を使って式(3)で ρ を ρ_{new} へ更新
　if $H(q') + KL(p(\theta|\rho_{new})||p(\theta|\rho)) > \log n$ **then**
　　$n \leftarrow n + 1, m \leftarrow m - 1, m \leftarrow \max(2, m)$
　else
　　$m \leftarrow m + 1$
　end if
end for
return $p(\theta|\rho)$

5. 今後の課題

今後は，提案したアルゴリズムを実際の問題に広範囲に適用し，評価する予定である．また，本稿では，計算の容易さのため，報酬の分布を固定したパラメータのもとでの試行で近似する手法を選択したが，データ効率やより広い範囲への適用可能性を考えると近似なしのベイズ法による回帰で考えるのが望ましく，その方向でも有効なアルゴリズムを構築したい．

6. おわりに

強化学習は，環境との相互作用による探索的な試行錯誤に基づく手法であり，学習システムの柱の一つをなしているが，ハイパーパラメータを含む各種設定の難しさから，一部の専門家以外は活用が難しいと考えられていた．本稿の結果は，この問題を解決し強化学習の広範囲な問題への適用と普及を推進するための一助となると考える．また，本稿で与えた一般的な設定での強化学習の定式化は，他の学習システムと組み合わせ，より高度な知的処理を実現する際の設計の指針の一つとなり，知的なシステムの今後の発展に寄与するものであると考える．

＊1　入力シーケンスを効果的に処理し，各位置のコンテキストに基づいて出力シーケンスを生成する，注意機構を用いたニューラルネットワークモデル
＊2　複数の GPU を連携させて大量のデータ処理や計算を高速化するためのシステム
＊3　インターネット上で情報を閲覧するためのハイパーテキストシステム
＊4　エージェントが環境の動的な振る舞いを学習し，将来の予測や意思決定を行うための内部的な表現を持つモデル
＊5　脳やシステムが予測誤差を最小化することで環境に適応し，効率的に行動する仕組みを説明する理論

参考文献

［1］ ChatGPT. https://chat.openai.com/.
［2］ Llama3. https://github.com/meta-llama/llama3.
［3］ Kenji Doya. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks,* Vol. 12, No. 7-8, pp. 961–974, 1999.
［4］ Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience,* Vol. 11, No. 2, pp. 127–138, 2010.
［5］ David Ha and Jurgen Schmidhuber. Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31,* pp. 2450–2462. Curran Associates, Inc., 2018.
［6］ Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature,* Vol. 518, No. 7540, pp. 529–533, 2015.
［7］ Brendan O'Donoghue. Variational Bayesian reinforcement learning with regret bounds. *arXiv preprint arXiv:1807.09647,* 2018.
［8］ Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations,* 2020.
［9］ Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010.* AAAI Press, 2010.
［10］ Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research,* Vol. 99, No. 1, pp. 89–112, 1997.
［11］ Frank Sehnke, Christian Osendorfer, Thomas Rucksties, Alex Graves, Jan Peters, and Jurgen Schmidhuber. Policy gradients with parameter-based exploration for control. In Vera Kurkova, Roman Neruda, and Jan Koutnik, editors, *Artificial Neural Networks -ICANN 2008, 18th International Conference, Prague, Czech Republic, September 3-6, 2008, Proceedings, Part I,* Vol. 5163 of *Lecture Notes in Computer Science,* pp. 387–396. Springer, 2008.
［12］ David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature,* Vol. 529, No. 7587, pp. 484–489, 2016.
［13］ Jean Tarbouriech, Tor Lattimore, and Brendan O'Donoghue. Probabilistic inference in reinforcement learning done right. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems,* Vol. 36, pp. 33687–33725. Curran Associates, Inc., 2023.
［14］ William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika,* Vol. 25, No. 3/4, pp. 285–294, 1933.
［15］ Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely ap-

plicable information criterion in singular learning theory. *Journal of Machine Learning Research,* Vol. 11, pp. 3571–3594, 2010.
［16］ Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning,* Vol. 8, pp. 229–256, 1992.

執筆者紹介 星　野　　力（Chikara Hoshino）
2000年　日本ユニシス株式会社入社.
2007年　東京工業大学にて博士（工学）を取得.
2000年－現在　統計的なシステムの設計およびその基盤理論の研究に従事.