ベイズ法を用いたパラメータベースの方策探索

Bayesian Parameter-Based Exploration

星　野　　力

要　約　強化学習は，教師信号が明示的に与えられず，報酬だけが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，学習に頑健性がないことや，ハイパーパラメータの細かいチューニングが必要なため，新しいタスクへの適用は容易ではなく，強化学習の広い適用の障害となっている．頑健性がない理由のひとつとして，既存の手法は，報酬や方策パラメータの不確実性を明示的に考慮していないため，学習過程での確率的なゆらぎに強く影響を受けてしまうことが考えられる．本稿では，ロボットの制御などを含む連続的な状態空間と行動を持つ環境を対象に，パラメータベースの方策探索について，ベイズ法を用いて不確実性の問題点を解決する手法を提案する．はじめに，報酬と方策パラメータの不確実性を明示的に考慮に入れた目的関数を定義する．さらに，設定した目的関数を効率的に最適化するアルゴリズムを与える．数値実験を行い提案手法の有効性を調べたところ，提案手法は，方策パラメータの分布のエントロピーを通じて，報酬の探索と獲得の適切なバランスをとることにより，従来の手法より安定してタスクを成功させることがわかった．この結果は，提案手法が，安定した学習によりタスク達成までの試行回数を減らすことによって，強化学習の適用範囲を拡大する可能性があることを示唆している．

Abstract　Reinforcement learning is suitable for many tasks that cannot provide an explicit training sample and can only provide rewards. Despite some high-impact successes, reinforcement learning is not easily applicable to many new situations due to lack of robustness and the need for hard hyperparameter tuning. One reason for this problem is that the existing methods do not account for the uncertainties of rewards and policy parameters. In this paper, we use Bayesian methods in parameter-based policy exploration to define an objective function that explicitly accounts for uncertainty of reward and policy parameters. In addition, we provide an efficient algorithm that optimizes this objective function under continuous state and action spaces. Numerical results show that the proposed method is more robust than existing non-Bayesian parameter-based policy exploration by balancing reward acquisition and exploration.

1. は　じ　め　に

人間のみが行えると考えられていた知的な情報処理を計算機やロボットで代替する技術として，人工知能がある．高度な人工知能の構築とその応用は，産業革命の一つに値するほどの技術革新となるのではないかと，大きな注目を集めている．

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基盤技術としている．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．自然言語処理への応用（“BERT”[3]）では，教師なし学習による良いデータ表現の獲得により，テキスト分類や，文書要約など様々なタスクへ活用できることが示されている．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．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，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，統計的な学習を通した知的な情報処理の大きな枠組みとして，これら三つの学習の適切な組み合わせを構成するアプローチ（“World Models”[6][4]）が模索されている．

本稿では，この中で強化学習に焦点をあてる．強化学習は，ビデオゲーム[13]や碁[20]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．しかしながら，ロボット制御を含む実環境でのタスクを解くことは未だ困難な状況にある．本稿では，連続的な状態空間と行動を持つ場合の強化学習に関する三つの問題に焦点をあてる．

一点目の問題は，状態ベースで探索を促すための確率的な方策は，それぞれの時刻で確率的に行動を選択するため，状態の軌道がなめらかにならないことである．このことにより，“REINFORCE”[26]などの方策勾配を用いる手法では，勾配の推定において分散が大きくなってしまう．“Deep Deterministic Policy Gradient（DDPG）”[12]や“Twin Delayed Deep Deterministic policy gradient（TD3）”[5]などの手法は，状態ベースの方策探索において，勾配の分散が小さくなる決定論的方策が利用できるような改良を提案している．一方で，“Policy Gradients with parameter-based exploration”[19]のようなパラメータベースの方策探索は，方策のパラメータについて分布を導入する．この分布による階層構造の導入は，探索を行動空間からパラメータの空間に移行させることにより，決定論的な方策を可能にする．

二点目の問題は，学習の過程において，報酬の値は確率的に与えられるため，そこに不確実性があることに起因する．例えば，たまたま良い初期値から始まった場合や，たまたま報酬の値が大きかった場合は，その試行の中では，行動がうまくいったように見えるが，報酬の期待値を最大化する目的には十分でない．そのため，不確実性を考慮せず，一試行で大きな報酬が得られたことを過度に評価すると，その行動に固着して新しい行動を探索することが不足したり，うまくいった行動に飛びついて方策が不安定になってしまう．一方で，不確実性を過度に考慮すると，学習のスピードが停滞し，いつまでたっても最適な報酬に収束しないため，学習に必要な試行数が増加してしまう．この問題は，報酬の探索と獲得のトレードオフと呼ばれ，トレードオフの解消に精密なハイパーパラメータの設定が必要となり，強化学習の新しいタスクへの適用を困難にしている．

三点目の問題は，方策パラメータの不確実性の問題である．一般にニューラルネットワークを含む非線形モデルは，学習できる関数の範囲が非常に広く，そのことが様々なタスクへの適用可能性を広げているが，逆にそれがあだともなり，パラメータの摂動に対して関数が大きく動いてしまう．そのため最尤推定を含む，パラメータの不確実性を考慮しない学習アルゴリズムは，データの有限性からくるばらつきによって方策が不安定化し，学習の頑健性が下がってしまう．

ベイズ法は，不確実性を分布として正面から扱う手法である．既存研究では，“K-Learning”[15][16]が報酬の不確実性をベイズ法の枠組みで導入し，探索と獲得のトレードオフを効率的に最適化できることを示している．しかしながら，“K-Learning”は離散的な状態空間と行動を仮定していて，そのアルゴリズムの連続的な状態空間と行動への適用は決して自明ではない．

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，報酬と方策パラメータの不確実性を考慮に入れた新たな強化学習法を提案する．本稿における貢献は以下の 3 点に集約される．

- ベイズ法を用いて，報酬と方策パラメータの不確実性を明示的に考慮した目的関数を定義したこと
- 定義した目的関数を効率的に最適化するアルゴリズムを構築したこと
- 数値実験を通じて，提案手法が，報酬の探索と獲得のトレードオフを最適化し，そのことにより従来の手法より安定して学習可能なことを示したこと

2. パラメータベースの方策探索

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パラメータベースの方策探索について述べる．

2.1 仮定

時刻 T ステップの状態 $x_{1:T+1}$，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．

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

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．

パラメータベースの方策探索では，方策 π は現在の状態 x_t と方策パラメータ θ に対して，決定論的な行動 u_t を出力する．

$$u_t = \pi\left(x_t, \theta\right)$$

また，パラメータ θ の一つの試行 h での評価は，一試行における報酬の和 r_h で定義される．

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

2.2 数学的定式化と方策探索

これらの仮定のもと，本稿で扱う問題は，報酬の和 r_h の分布 $p(r_h|\theta)$ および方策パラメータ $p(\theta)$ に関する期待値 $J(p)$ を最大化する方策パラメータの分布 $p(\theta)^*$ を求めることと定式化される．

$$r\left(\theta\right) = \int r_h p\left(r_h\middle|\theta\right)dr_h,\ J\left(p\right) = \int r\left(\theta\right)p\left(\theta\right)d\theta.$$

ただし，$r(\theta)$ はパラメータ θ に対する，報酬 r_h の平均である．さらに，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 $p(\theta)^*$ による出力のアンサンブル平均で与えられる．

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

3. 不確実性を考慮した目的関数

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が，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 報酬の不確実性の考慮

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 $r\left(\theta\right) = \int r_h p\left(r_h \middle| \theta\right) dr_h$ を何らかの方法で推定する必要がある．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 $\tilde{r}\left(\theta\right) = \frac{1}{K}\sum_{k=1}^{K} r_{h_k}$ 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”[15][16]を提案した．“K-learning”は単純平均の代わりに，ベイズ法を用いてキュムラントを利用することにより，不確実性を考慮している．キュムラントは，単純平均に比べてより探索的にふるまう推定量であり，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．

$$p\left(r \middle| \theta, w\right), \tag{1}$$

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 $D \equiv \left\{\left(r_1, \theta_1\right), \ldots, \left(r_K, \theta_K\right)\right\}$ 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．

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

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

$$K\left(\theta\right) \equiv \log E_{p\left(w|D\right)}\left[\exp\left(\beta \int r p\left(r \middle| \theta, w\right) dr\right)\right].$$

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．

$$C\left(\theta, \gamma\right) \equiv \log E_{p\left(w|D\right)}\left[\exp\left(\gamma\beta \int r p\left(r \middle| \theta, w\right) dr\right)\right].$$

このとき，$K(\theta)$ の二次近似は，以下で与えられる[24]．

$$K(\theta) = C(\theta, 1) \approx \left.\frac{d}{d\gamma}\right|_{\gamma=0} C(\theta, \gamma) + \frac{1}{2}\left.\frac{d^2}{d\gamma^2}\right|_{\gamma=0} C(\theta, \gamma).$$

一次と二次の微分は，それぞれ，平均と分散に対応する．

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

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

したがって，$K(\theta)$ の近似は，報酬の平均 $\tilde{r}(\theta)$ にその不確実さを表わす分散 $\tilde{v}(\theta)$ をボーナスとして加えた形で与えられる．ただし，分散 $\tilde{v}(\theta)$ 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が必要で，平均推定量の分散を使うことは，“Thompson sampling”[21] と強い関係性を持つ．

3.2　方策パラメータの不確実性の考慮

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 を最尤推定などの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ベイズ推定がある．ベイズ推定は，エントロピーの効果を使って，学習を安定化するアンサンブルをパラメータの事後分布として得る手法である．

ベイズ推定導出の第一段階として，次の ρ に対する，積分 $I(\rho)$ を考える．

$$I(\rho) = \int w(\theta) p(\theta|\rho)\, d\theta, w(\theta) \geq 0.$$

ただし，$p(\theta|\rho)$ はハイパーパラメータ ρ を持つ方策のパラメータ θ の分布であり，$w(\theta)$ は θ に関する重みを表す関数である．この先，$I(\rho)$ に関する最大化は，重み付き尤度の推定に帰着されることを示す．この帰着を経由して，方策パラメータの不確実性を考慮するベイズ法を導出する．

はじめに，現在のハイパーパラメータ ρ および，最適化するハイパーパラメータ ρ' を用いて，$I(\rho')$ と $I(\rho)$ の比を評価する．イェンセンの不等式を用いると以下のように評価ができる．

$$\log \frac{I(\rho')}{I(\rho)} = \log \int \frac{w(\theta) p(\theta|\rho)}{I(p)} \frac{p(\theta|\rho')}{p(\theta|\rho)} d\theta \geq \int \frac{w(\theta) p(\theta|\rho)}{I(\rho)} \log \frac{p(\theta|\rho')}{p(\theta|\rho)} d\theta.$$

このとき，関数 $Q(\rho', \rho)$ を次のように定義すると，

$$Q(\rho', \rho) = \int w(\theta) p(\theta|\rho) \log p(\theta|\rho')\, d\theta. \tag{2}$$

以下の不等式が成り立つ．

$$\log I(\rho') \geq \log I(p) + \frac{Q(\rho', \rho) - Q(\rho, \rho)}{I(\rho)}$$

この不等式は，$Q(\rho', \rho)$ の ρ' についての最大化は，$I(\rho')$ の下限を与えることを示している[23]．さらに，(2)を $p(\theta|\rho)$ からのサンプルである J 個の θ_j とそれに対応する $w(\theta_j)$ で近似する．

$$D_p \equiv \left\{\left(\theta_1, w\left(\theta_1\right)\right), \ldots, \left(\theta_J, w\left(\theta_J\right)\right)\right\}, \theta_j \sim p\left(\theta\middle|\rho\right),$$

$$Q\left(\rho', \rho\right) \approx \frac{1}{J}\sum_{j=1}^{J} w\left(\theta_j\right) \log p\left(\theta_j\middle|\rho'\right).$$

そのとき，Q の ρ' について，一次の最適性条件は，

$$\sum_{j=1}^{J} w\left(\theta_j\right) \nabla_{\rho'} \log p\left(\theta_j\middle|\rho'\right) = 0, \tag{3}$$

と記述される．この式は，Q が以下の重み付き尤度と関連することを示していて，

$$\prod_{j=1}^{J} p\left(\theta_j\middle|\rho'\right)^{w\left(\theta_j\right)}, \tag{4}$$

論文[22][9]は，$J \to \infty$ の条件のもと，重み付き尤度(4)のハイパーパラメータ ρ' に対する最適な重みが，(3)の解に収束することを示した．

この重み付き尤度の表現を経由すると，方策パラメータのベイズ推定は，以下の予測分布 $p^*(\theta)$ として得られる．

$$p^*\left(\theta\right) = \int p\left(\theta\middle|\rho\right) p\left(\rho\middle|D_p\right) d\rho,$$

$$p\left(\rho\middle|D_p\right) = \frac{\prod_{j=1}^{J} p\left(\theta_j\middle|\rho\right)^{w\left(\theta_j\right)} p\left(\rho\right)}{\int \prod_{j=1}^{J} p\left(\theta_j\middle|\rho\right)^{w\left(\theta_j\right)} p\left(\rho\right) d\rho}, \theta_j \sim p^*_{old}\left(\theta\right).$$

ただし，$p(\rho)$ は，ハイパーパラメータ ρ の事前分布を表わす．

3. 3　計算量削減のためのガウス近似

報酬の分布 $p(r|\theta, w)$ と方策パラメータの分布 $p(\theta|\rho)$ を学習アルゴリズムの計算量削減のためガウス分布に制限する．この制限のもとでは，提案するアルゴリズムは，“EM-based policy hyperparameter exploration”[23]や“Covariance Matrix Adaptation Evolution Strategy”[8]と関連を持つ．これらアルゴリズムは，重み関数，共分散を推定するか否か，ベイズ推定か最尤推定か，に違いがある．

共役事前分布のもとでのガウス分布のベイズ推定は以下の通り[14]．尤度関数はガウス分布を仮定する．

$$p\left(x\middle|\mu, \sigma^2\right) = \mathcal{N}\left(x\middle|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right).$$

さらに，事前分布は，それに共役な“Normal-Gamma”を用いる．

$$\mathcal{NG}\left(\mu, \lambda, \mu_0, \kappa_0, \alpha_0, \beta_0\right) = \mathcal{N}\left(\mu\middle|\mu_0, \left(\kappa_0\lambda\right)^{-1}\right) \mathcal{G}\left(\lambda\middle|\alpha_0, \beta_0\right) \tag{5}$$

$$= \frac{1}{Z_{NG}} \lambda^{\frac{1}{2}} \exp\left(-\frac{\kappa_0 \lambda}{2}\left(\mu - \mu_0\right)^2\right) \lambda^{\alpha_0 - 1} \exp\left(-\lambda \beta_0\right), \tag{6}$$

$$Z_{NG} = \frac{\Gamma\left(\alpha_0\right)}{\beta_0^{\alpha_0}} \left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}},$$

また，t 分布は以下のように定義する．

$$t_\nu\left(x \middle| \mu, \sigma^2\right) = c\left[1 + \frac{1}{\nu}\frac{\left(x-\mu\right)^2}{\sigma}\right]^{-\frac{\nu+1}{2}}, c = \frac{\Gamma\left(\nu/2 + 1/2\right)}{\Gamma\left(\nu/2\right)} \frac{1}{\sqrt{\nu\pi}\sigma}.$$

θ を固定したもとでの平均報酬の分布 $p_\theta(r|w)$ については，K 個のサンプル $D_r \equiv \left(r_1, \ldots, r_K\right)$ が与えられたもと，平均報酬の平均 $\tilde{r}(\theta)$ と，分散 $\tilde{v}(\theta)$ はそれぞれ，

$$\tilde{r}\left(\theta\right) = \mu_K, \tag{7}$$

$$\tilde{v}\left(\theta\right) = \frac{\beta_K}{\left(\alpha_K - 1\right)\kappa_K}, \tag{8}$$

で与えられる．ただし，

$$\overline{r} = \frac{\sum_{k=1}^K r_k}{K},$$

$$\kappa_K = k_0 + K, \mu_K = \frac{\kappa_0 \mu_0 + K\overline{r}}{\kappa_0 + K},$$

$$\alpha_K = \alpha_0 + \frac{K}{2}, \beta_K = \beta_0 + \frac{1}{2}\sum_{k=1}^K \left(r_k - \overline{r}\right)^2 + \frac{k_0 K\left(\overline{r} - \mu_0\right)^2}{2\left(\kappa_0 + K\right)}.$$

式(7)，(8)は，以下の報酬の平均の周辺事後分布から導出できる．

$$p_\theta\left(\mu \middle| D_r\right) = t_{2\alpha_K}\left(\mu \middle| \mu_K, \frac{\beta_K}{\alpha_K \kappa_K}\right).$$

また，方策パラメータの分布 $p(\theta|\rho)$ については，J 個の重みとサンプル $D_\rho \equiv \left\{\left(\theta_1, w\left(\theta_1\right)\right), \ldots, \left(\theta_J, w\left(\theta_J\right)\right)\right\}$ が与えられたもとで，方策パラメータの予測分布は以下で与えられる．

$$p\left(\theta \middle| D_\rho\right) = t_{2\alpha_J}\left(\theta \middle| \mu_J, \frac{\beta_J\left(\kappa_J + 1\right)}{\alpha_J \kappa_J}\right), \tag{9}$$

ただし，

$$\overline{J} = \sum_{j=1}^J w\left(\theta_j\right), \overline{\theta} = \frac{\sum_{j=1}^J w\left(\theta_j\right)\theta_j}{\overline{J}},$$

$$\kappa_J = k_0 + \overline{J}, \mu_J = \frac{\kappa_0 \mu_0 + \overline{J}\overline{\theta}}{\kappa_0 + \overline{J}},$$

$$\alpha_J = \alpha_0 + \frac{\overline{J}}{2},\ \beta_J = \beta_0 + \frac{1}{2}\sum_{j=1}^{J} w\left(\theta_j\right)\left(\theta_j - \overline{\theta}\right)^2 + \frac{k_0 \overline{J}\left(\overline{\theta} - \mu_0\right)^2}{2\left(\kappa_0 + \overline{J}\right)}.$$

4. 提案手法

この章では，本稿の提案手法である，目的関数の定式化とその最適化アルゴリズムについて述べる．

4.1 目的関数の定式化

$q(\theta)$ を最適化する方策パラメータの分布，$p(\theta)$ を現状の行動出力に使われる方策パラメータの分布，$\phi(\theta)$ を方策パラメータの事前分布とする．初めに，変分法を用いて，以下のカルバック距離を最小化することを考える．

$$KL\left(q\left(\theta\right)\middle\|p\left(\theta\right)\right),$$

ただし，

$$p\left(\theta\right) = \frac{\exp\left(E\left(\theta\right)\right)}{Z},\ Z = \int \exp\left(E\left(\theta\right)\right) d\theta,$$

かつ，

$$E\left(\theta\right) \equiv \beta\tilde{r}\left(\theta\right) + \frac{\beta^2}{2}\tilde{v}\left(\theta\right) + \alpha \log p\left(\theta\right) + \left(1-\alpha\right)\log\phi\left(\theta\right).$$

この定義では，$0 \le \alpha \le 1$ は，エントロピーによる正則化の強さを表し，$\beta > 0$ は逆温度である．最適化アルゴリズムでは，α は利用者が設定する定数とし，β は最適化する変数であるとする．

カルバック距離の正値性を用いると，本稿の鍵となる不等式が導出される．

$$\begin{aligned}
\tilde{J}\left(q\right) &= \int \tilde{r}\left(\theta\right) q\left(\theta\right) d\theta \\
&\le \int \left(\tilde{r}\left(\theta\right) + \frac{\beta}{2}\tilde{v}\left(\theta\right)\right) q\left(\theta\right) d\theta \qquad (10) \\
&\le \frac{1}{\beta}\log Z + \frac{\alpha}{\beta}\left(KL\left(q\left(\theta\right)\middle\|p\left(\theta\right)\right)\right) + \frac{1-\alpha}{\beta}\left(KL\left(q\left(\theta\right)\middle\|\phi\left(\theta\right)\right)\right) \qquad (11)
\end{aligned}$$

よって，目的関数 $F(\beta)$ を

$$F\left(\beta\right) \equiv \frac{1}{\beta}\left(\log Z + \alpha\left(KL\left(q\left(\theta\right)\middle\|p\left(\theta\right)\right)\right) + \left(1-\alpha\right)\left(KL\left(q\left(\theta\right)\middle\|\phi\left(\theta\right)\right)\right)\right), \qquad (12)$$

で定義すると，$F(\beta)$ の β に対する最小化は，$\tilde{J}(q)$ の上限を与える．また，目的関数の主要項 $\log Z$ は平均の期待値 $\int rp\left(r\middle|\theta, w\right)dr$ の平均報酬の回帰パラメータ $p(w|D_r)$ と方策パラメータの予測分布 $p(\theta|D_p)$ に関するキュムラントの近似になっている．

$$\log Z \approx \log E_{p(\theta|D_p)}\left[\left(\frac{\phi(\theta)}{p(\theta|D_p)}\right)^{1-\alpha} E_{p(w|D_r)}\left[\exp\left(\beta \int r p(r|\theta, w)\, dr\right)\right]\right].$$

定義した目的関数(11)は，明確な解釈と既存手法との関連性を持つ[11]．第一項は，探索不足を回避するために，報酬の平均の推定値として，単純平均ではなく，そのキュムラントを用いているが，それは“K-learning”[15][16]と関連を持つ．第二項は，現在の探索を行う方策パラメータの分布から，更新する新しい方策パラメータへのカルバック距離 $KL(q(\theta)\|p(\theta))$ であり，分布が大きく更新されることを防ぐ効果がある．この項を適切に調整することにより，方策の不安定さを防ぐ手法は，“Trust Region Policy Optimization”[18]や“Relative Entropy Policy Search”[17]で用いられている．第三項は，現在の探索を行う方策パラメータの分布から，方策パラメータの事前分布へのカルバック距離 $KL(q(\theta)\|p(\theta))$ である．仮に，$p(\phi) \equiv 1$ と置くと（後の数値実験はこの設定で行った.），この項は方策パラメータの分布 $q(\theta)$ のエントロピーとなり，“Soft Actor Critic”[7]で用いられている．

4. 2　最適化アルゴリズム

目的関数 $F(\beta)$ は以下の重点サンプリングで推定する．

$$F(\beta) \approx \frac{1}{\beta}\left(\log \frac{1}{J}\sum_{j=1}^{J} \exp\left(H(\theta_j)\right) + \alpha KL\left(q(\theta)\middle\|p(\theta)\right) + (1-\alpha) KL\left(q(\theta)\middle\|\phi(\theta)\right)\right), \tag{13}$$

ただし，θ_j は $p(\theta)$ からのサンプル $\theta_j \sim p(\theta)$ であり，$H(\theta)$ は，

$$H(\theta) \equiv \beta \tilde{r}(\theta) + \frac{\beta^2}{2}\tilde{v}(\theta) + (1-\alpha)\left(\log \phi(\theta) - \log p(\theta)\right), \tag{14}$$

で定義される．この場合には，それぞれのパラメータ θ_j の重み係数 $\tilde{w}(\theta_j)$ は以下のように得られる．

$$\tilde{w}(\theta_j) = \frac{\exp\left(H(\theta_j)\right)}{\sum_{j=1}^{J} \exp\left(H(\theta_j)\right)}. \tag{15}$$

$KL(q(\theta)\|p(\theta))$ および $KL(q(\theta)\|\phi(\theta))$ については下記の手法で推定する．まず，$KL(q(\theta)\|\phi(\theta))$ は，計算の容易さのため，更新されたあとの方策パラメータの分布でなく，データを生成する方策パラメータの分布 $p(\theta)$ を用いることとする．

$$KL\left(q(\theta)\middle\|\phi(\theta)\right) \approx KL\left(p(\theta)\middle\|\phi(\theta)\right) \approx \frac{1}{J}\sum_{j=1}^{J} \log \frac{p(\theta_j)}{\phi(\theta_j)}, \theta_j \sim p(\theta). \tag{16}$$

$KL(q(\theta)\|p(\theta))$ については，許容量 δ_1 を設定し，$KL(q(\theta)\|p(\theta)) < \delta_1$ を満たすよう定める．$KL(q(\theta)\|p(\theta))$ は，学習速度と学習の頑健性のトレードオフを表現していて，勾配法による学習で用いられる学習係数に相当する．さらに，アルゴリズムの頑健性は，重点サンプリング(13)の精度にも依存している．重点サンプリングの精度は，以下に定義される“Effective Sample Size（ESS）”[10]で計測する．

$$ESS = \frac{1}{\sum_{j=1}^{J} \tilde{w}\left(\theta_j\right)^2}.$$

これらの考察から，定数$KL\left(q\left(\theta\right)\|p\left(\theta\right)\right)$を次の二つの拘束条件を満たすなかでの最大値とする．

$$KL\left(q\left(\theta\right)\middle\|p\left(\theta\right)\right) \leq \delta_1, ESS \leq \delta_2. \tag{17}$$

$KL\left(q\left(\theta\right)\|p\left(\theta\right)\right)$，$\beta$の探索には線形探索を用いる．探索は，$KL\left(q\left(\theta\right)\|p\left(\theta\right)\right) = 0$から始め，(17)が満たされている間，それぞれの$KL\left(q\left(\theta\right)\|p\left(\theta\right)\right)$で$\beta$を最小化しながら，$KL\left(q\left(\theta\right)\|p\left(\theta\right)\right)$を増加させる．

目的関数の最適化アルゴリズムを Algorithm 1 に記述する．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　目的関数の最適化アルゴリズム

入力：方策パラメータの初期分布$p\left(\theta_0\right)$，エントロピーの係数α，カルバック距離の閾値δ_1，ESS の閾値δ_2,
出力：方策パラメータの分布$p\left(\theta\right)$
for $i = 1$ to I（反復数）**do**
　for $j = 1$ to J（集団のサイズ）**do**
　　$\theta_j \propto p\left(\theta\right)$でサンプリング
　　$E\left[j\right] \leftarrow \log p\left(\theta_j\right)$
　　for $k = 1$ to K（エピソードの長さ）**do**
　　　方策θ_jでエピソードを実行，結果を$R\left[k\right] \leftarrow r_h$に代入
　　end for
　　$R\left[.\right]$と(7)，(8)を使って$\tilde{r}\left(\theta_j\right)$，$\tilde{v}\left(\theta_j\right)$を計算
　end for
　$E\left[.\right]$，(16)を使って，$KL\left(p\left(\theta\right)\|\phi\left(\theta\right)\right)$を計算
　(17)の制約下で，(13)をβ，$KL\left(q\left(\theta\right)\|p\left(\theta\right)\right)$について最適化
　(15)を使って$\tilde{w}\left(\theta.\right)$を計算
　$\theta.$, $w\left(\theta.\right) = J \times \tilde{w}\left(\theta.\right)$，(9)を使って$p\left(\theta\right)$を更新
end for
return $p\left(\theta\right)$

表 1　実験“Pendulum-v1”の報酬

	報酬（平均）	報酬（標準偏差）
提案手法	−159.330	14.269
EPHE-RW	−306.554	193.666

表 2　“Pendulum-V1”方策パラメータの分布のエントロピー

	エントロピー（平均）	エントロピー（標準偏差）
提案手法	−0.620	0.333
EPHE-RW	−5.344	0.006

5. 数値実験

提案手法の有効性を確認するための数値実験を行った．結果の評価は二つの観点から行った．一つ目は，提案手法の頑健性の確認であり，二つ目は，提案手法が報酬の探索と獲得のトレードオフを，獲得報酬と方策パラメータの分布のエントロピーのバランスとして実現しているかの確認である．この目的のため，比較の対象として，“EM-based policy hyperparameter exploration using the REPS weighting scheme（EPHE-RW）”[23]を選んだ．EPHE-RW は，報酬の平均と，方策パラメータの分布(4)を，提案手法で用いるベイズ法ではなく，最尤法で推定した場合に相当する．さらに，報酬の平均を最尤推定する場合は，平均の推定値の分散は 0 となる $(\tilde{v}(\theta)=0)$．したがって，これら二つの手法を比較することで，ベイズ法により報酬と方策パラメータの不確実性を考慮した効果を確認できる．

タスクとして，OpenAI Gym[2]の“Pendulum-v1”を選択した．このタスクは，状態空間と行動が連続値を持つ非線形の制御タスクである．方策には，40 個の隠れ素子を持ち，“tanh”の活性化関数を持つ，3 層のニューラルネットワークを用いた．集団のサイズを $J=26$，エピソードの長さを $K=26$ として，方策パラメータの反復数を $I=1000$ とした．エントロピーの係数 α は，$|\theta|$ を方策のニューラルネットワークのパラメータ数として，$\alpha=1-\dfrac{1}{|\theta|}$ に設定した．カルバック距離に関するパラメータは $\delta_2=0.5$，ESS に関するパラメータは $\delta_1=0.5\times J$ に設定した．また，方策パラメータの初期分布 $p_0(\theta)$ は，標準正規分布 $\mathcal{N}(\theta|0, 1.0)$ とした．

Normal-Gamma の事前分布(6)に関しては，$\mu_0=0$，$\alpha_0=-0.5$，$\beta_0=0.5\mathrm{e}-7$，$\kappa_0=1.0e-6$ と設定した．この設定では，α_0 に関しては，リファレンス事前分布[1][25]，残りは無情報に近い事前分布となる．実験は，10 回の異なる初期値で行い，1000 ステップ中の最後の 100 ステップの報酬と方策パラメータの分布のエントロピーを方策パラメータの数で割った値で評価した．

数値実験の結果は，表 1 のとおりである．この結果は，提案手法が，“EPHE-RW”に比べて，平均獲得報酬が高く，獲得報酬の標準偏差も小さい．

図 1 は，より詳細な比較のため，学習ステップごとの軌跡を示している．この図から，学習の初期段階では，提案手法より“EPHE-RW”の方が獲得報酬の増加の速度が速い場合があることが確認される．しかしながら，“EPHE-RW”は，しばしば局所解にはまってそこから抜け出せないことがあるのに対して，提案手法は，最終的にはほとんど確実に，最適な獲得報酬

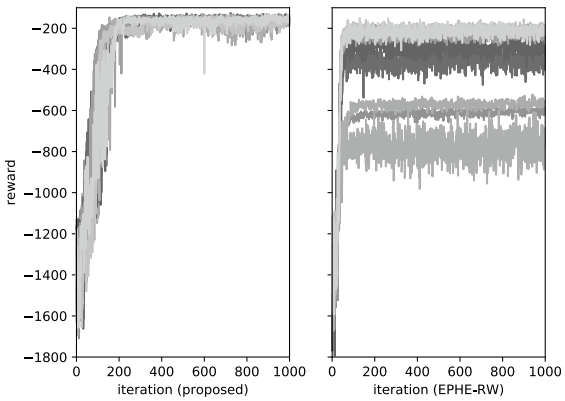

図 1　10 回の試行における獲得報酬の軌跡（左：提案手法，右：“EPHE-RW”）

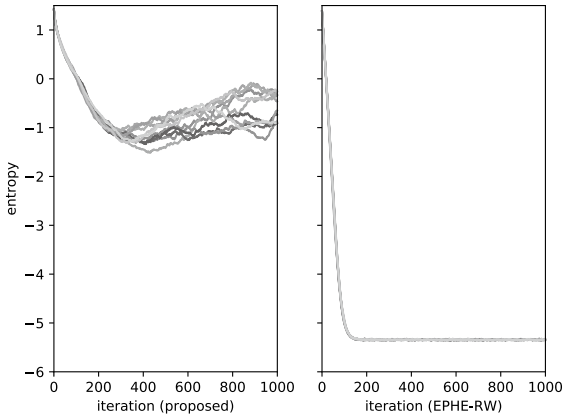

図 2　10 回の試行における方策パラメータの分布のエントロピーの軌跡（左：提案手法，右：“EPHE-RW”）

を達成していることがわかる．このことは，提案手法が，初期値などの試行間にある確率的なゆらぎに対して，強い頑健性を持つことを示している．

表 2 と図 2 は提案手法と，“EPHE-RW”の間に方策パラメータの分布のエントロピーの明確な違いがあることを示している．図 2 から，提案手法の方策パラメータのエントロピーは，学習の初期では，減り続け，アルゴリズムがほぼ最適な解を発見する 400 ステップくらいから徐々に増加することが確認できる．このふるまいは，提案手法が，探索を制御する方策パラメータの分布のエントロピーを通じて，報酬の探索と獲得のあいだでバランスをとり，探索不足を避けていることを示している．一方で，“EPHE-RW”においては，方策パラメータの分布のエントロピーは単調に減少し，報酬の探索と獲得のトレードオフを行っている様子を見ることができない．

6. お わ り に

本稿では，強化学習における学習の頑健性と探索不足の問題に対し，ベイズ法を利用して，報酬と方策パラメータの不確実性を考慮に入れた新しい目的関数の定式化を行った．さらに，状態区間と行動が連続値を持つ場合に，目的関数を近似的に最適化する既存のパラメータベースの方策探索と同程度の計算量を持つアルゴリズムを開発した．数値実験の結果，提案手法は，報酬の獲得と報酬の探索を促す方策のパラメータのエントロピーとのトレードオフを最適化し，有限のサンプル数による確率的なゆらぎに対して従来法より頑健であることがわかった．

学習が頑健になることによって，タスク達成までの試行回数を削減できることや，学習が進まず試行錯誤する場合に，ハイパーパラメータの設定や乱数のために学習ができないのか，そもそもタスクの難易度が高いことによるのかを判断することが容易となる利点がある．強化学習は，教師信号を作成するのが難しいタスクや，人による教師信号以外の方法を探索したい場合に適した手法であり，三つの学習機能のうち一つを占める重要なものであるが，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．

参考文献
[1] José M Bernardo and Adrian FM Smith. *Bayesian theory*, Vol. 405. John Wiley & Sons, 2009.
[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[4] 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.
[5] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 1582-1591. PMLR, 2018.
[6] David Ha and Jürgen 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.
[7] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 1856-1865. PMLR, 2018.
[8] Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Toshio Fukuda and Takeshi Furuhashi, editors, *Proceedings of 1996 IEEE International Conference on Evolutionary Computation, Nayoya University, Japan, May 20-22, 1996*, pp. 312-317. IEEE, 1996.
[9] Masaaki Imaizumi and Ryohei Fujimaki. Factorized asymptotic Bayesian policy search for pomdps. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 4346-4352. ijcai.org, 2017.
[10] Augustine Kong. A note on importance sampling using standardized weights. Technical report, University of Chicago, 1992.
[11] Tadashi Kozuno, Eiji Uchibe, and Kenji Doya. Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, Vol. 89 of *Proceedings of Machine Learning Research*, pp. 2995-3003. PMLR, 2019.
[12] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
[13] 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.
[14] Kevin P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, University of British Columbia, 2007.
[15] Brendan O'Donoghue. Variational Bayesian reinforcement learning with regret bounds. *arXiv preprint arXiv:1807.09647*, 2018.
[16] Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning*

Representations, 2020.
[17] Jan Peters, Katharina Mülling, and Yasemin Altün. 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.
[18] John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, Vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 1889-1897. JMLR.org, 2015.
[19] Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Policy gradients with parameter-based exploration for control. In Vera Kurková, Roman Neruda, and Jan Koutník, 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.
[20] 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.
[21] 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.
[22] Tsuyoshi Ueno, Kohei Hayashi, Takashi Washio, and Yoshinobu Kawahara. Weighted likelihood policy search with model selection. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 2366-2374, 2012.
[23] Jiexin Wang, Eiji Uchibe, and Kenji Doya. Adaptive baseline enhances em-based policy search: Validation in a view-based positioning task of a smartphone balancer. *Front. Neurorobot.*, Vol. 2017, , 2017.
[24] Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, Vol. 11, pp. 3571-3594, 2010.
[25] SumioWatanabe. Bayesian cross validation and waic for predictive prior design in regular asymptotic theory. *arXiv preprint arXiv:1503.07970*, 2015.
[26] 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.