2H1-2

# e射影に基づく方策探索法

Direct Policy Search with e-projection

植野 剛
Tsuyoshi Ueno

科学技術振興機構 ERATO 湊離散構造処理系プロジェクト
Japan Science and Technology Agency ERATO Minato Discrete Structure Manipulation System Project

In this study, we propose a novel framework for direct policy search on first-exists finite-horizon Markov decision processes (MDPs). In this framework, following the parameter-based exploration scheme, we introduce an optimization problem for the probabilistic distribution of policy parameters which is a generalization of the original direct policy search problem. We then develop a novel iteration algorithm, KL divergence with parameter-based exploration (KLPE), which can discover the optimal distribution for the generalized problem according to the minimization of KL divergence. Furthermore, we show that the resulting distribution produced by KLPE converges to the global optimal parametric policy of the original DPS problem. Although KLPE has desirable theoretical aspects, it requires to compute the distribution with an intractable normalizing constant at each iteration. To cope with such difficulty, we derive an approximate algorithm based on the Gaussian approximation with e-projection.

## 1. はじめに

マルコフ決定過程における最適意思決定問題の解法である直接方策探索法は，方策をパラトリックモデルで特徴づけ，与えられたパラメータ空間から適切なパラメータを探索することにより，方策の最適化を実現する [Deisenroth 13]．よって，古典的な動的計画法規範の方法—価値関数の同定により方策の最適化を実現する方法と異なり，知りたい関数である方策に対する最適化問題を考えるため，理論的な見通しが良く，アルゴリズムも扱いやすい．したがって，ロボット制御，ゲーム AI の設計，オペレーションズ・リサーチなどさまざまな問題に応用され，従来法を凌ぐ性能を持つ方策の獲得に成功している．

直接方策探索法における重要な発見として，Kullback-Leibler (KL) ダイバージェンスの最小化に基づく方策探索法，KL 制御がある [Theodorou 10, Azar 12, Rawlik 12]．KL 制御は方策探索問題を確率分布の推論問題として再定式化し，確率推論を通して方策を最適化する．この定式化は，確率，統計分野で培われてきた洗練された近似推論アルゴリズム [Bishop 06] を方策探索に適用することが原理的には可能とするため，潜在変数を持つ大規模な意思決定問題の解法として期待される．しかし，KL 制御で必要とされる後ろ向きメッセージ伝搬の計算は煩雑であるため，現在の定式化ではそのような問題への応用は現実的ではない．

本研究では初期状態が固定された有限長マルコフ決定過程に注目し，簡潔でかつ理論的に美しい KL 制御法，KL control with parameter-based exploration (KLPE) を提案する．KLPE は parameter-based exploration [Sehnke 10, Rückstieß 10] の枠組みを応用し，方策パラメータの分布の最適化を反復することによって方策を更新する．興味深いことに，KLPE によって得られる方策パラメータの漸近分布は大域的に最適な方策を導く．しかし一方で，KLPE の各反復での解—方策パラメータの分布は解析計算不可能な積分を正規化項として持つ．このため，KL ダイバージェンスによる確率分布の近似法である e 射影 [Bishop 06] を用いて，各反復での KLPE の解を単峰なガウス分布で近似するアルゴリズム e-projection based KLPE(e-KLPE) を提案する．

連絡先: 植野 剛，科学技術振興機構，〒 0530-0012 大阪府大阪市北区芝田 1 丁目 4-14 芝田町ビル 5 階 C 号室，06-6147-2035，ueno@ar.sanken.osaka-u.ac.jp

## 2. 有限長マルコフ決定過程

本研究を通して次の 5 つの要素で定義される離散時間有限長マルコフ決定過程 $(X,U,p,c,\phi)$ について考える．$X \subseteq \mathbb{R}^m$，$U \subseteq \mathbb{R}^n$ はそれぞれ状態，行動空間であり，$X$，$U$ の要素である $x \in X$，$u \in U$ はそれぞれ状態，行動をあらわす．$p(x'|x,u)$ は状態 $x$, 行動 $u$ が与えられたとき，次状態 $x'$ への状態遷移をあらわす確率分布である．$c(x,u,t) \in \mathbb{R}^+$，$\phi(x,T) \in \mathbb{R}^+$ は時刻 $t$ における状態行動対 $(x,u)$ に与える即時コスト，終端時刻 $T$ における状態 $x$ に与える終端コストをあらわし，これらは常に正の値をとる．

本稿では方策として，$\ell$ 次元パラメータ $w \in W$，$W \subseteq \mathbb{R}^\ell$ によって特徴付けられた時間非依存の決定論的な関数 $u = \bar{u}(x,w)$ を考える．ここで，$W$ は方策パラメータの空間をである．本研究の目的は，初期状態 $x_1$ が与えられたとき，期待累積コストが最小となる最適な方策パラメータ $w^* \in W$ を見つけることである．

$$w^* = \operatorname*{argmin}_{w \in W} f(w) \tag{1}$$

ただし

$$f(w) := f(w,x_1) := \mathbb{E}\left[\phi(x_T,T) + \sum_{t=1}^{T-1} c(x_t,u_t,t) \middle| w\right]$$

である．また $\mathbb{E}[\cdot|w] := \mathbb{E}[\cdot|w,x_1]$ は，初期状態 $x_1$, 方策パラメータ $w$ が与えられたとき，分布 $\prod_{t=1}^{T} p(x_{t+1}|x_t,\bar{u}(x_t,w))$ による系列 $(x_t)_{t=2}^{T}$ に関する条件付き期待値である．本稿では初期状態 $x_1$ が固定されている場合についてのみ議論するため，以降，$x_1$ に関する依存性は省略する．

上述の有限長マルコフ決定過程における方策探索の問題設定は，ロボット制御問題，ゲーム AI 設計などで実際に用いられる汎用的な問題設定である [Bertsekas 07, Deisenroth 13]．この問題の標準的な解法では，方策をパラメトリックな確率分布であらわし，パラメータを調節したあと，$u$ に関して期待値をとることにより最適な方策を得る．現在の KL 制御 [Theodorou 10,

Azar 12, Rawlik 12] も基本的にこの枠組みに倣っている．これに対して本稿では，方策パラメータ $w$ の確率分布を考え，その確率分布の最適化を通して方策の最適化を実現する枠組み，parameter-based exploration [Sehnke 10, Rückstieß 10] を採用する．そして KL 制御の考えを応用し，方策パラメータの分布に関する新しい最適化アルゴリズムを提案する．

## 3. 提案法

以下の議論では，方策の最適化にのみ着目するため，$f(w)$ は既知であると仮定する．この仮定は現実にはあり得ないため，実際に応用する際は観測履歴から推定する必要がある．この話題については 5 節でふれる．

### 3.1 KLPE の導出

ここでは，方策パラメータ $w$ を確率変数であると考え，最適化問題 (1) を一般化した方策パラメータの確率分布に対する最適化問題を定義する．$\pi(w) \in \Pi$ を方策パラメータ $w$ に関する確率分布とし，$\Pi$ は $w$ に関する確率分布の集合とする．このとき，決定論的な方策 $\bar{u}(x,w)$ と確率分布 $\pi(w)$ によって構成される階層構造を持つ確率的な方策 $\delta_u(u-\bar{u}(x,w))\pi(w)$ を考えることができる．ただし，$\delta(\cdot)$ はディラックのデルタ関数である．

本稿では，式 (1) で与えられるパラメータ $w$ に関する最小化問題ではなく，確率分布 $\pi(w)$ に関する最適化問題について考える．

$$\min_{\pi\in\Pi} \eta(\pi) \tag{2}$$

$\eta(\pi)$ は $\pi(w)$ によって周辺化された期待累積コストである．

$$\eta(\pi) := \mathbb{E}_\pi[f(w)] := \int \pi(w) f(w) \mathrm{d}w$$

$\mathbb{E}_\pi[\cdot]$ は $\pi(w)$ による $w$ に関する期待値である．最適化問題 (2) は方策パラメータに関する最適化 (1) を方策パラメータの確率分布へ自然に拡張したものであり，parameter-based exploration [Sehnke 10, Rückstieß 10] による方策探索法の最適化問題の一般化に対応している．

本研究では，KL ダイバージェンス最小化による確率的な方策の最適化法，KL 制御の枠組み [Azar 12, Rawlik 12] を応用し，最適化問題 (2) を解く．まず分布 $\pi$ で特徴付けられる $w$ の関数 $g(w,\pi)$ を定義する．

$$g(w,\pi) := \frac{\pi(w)\exp[-f(w)]}{\mathbb{E}_\pi[\exp[-f(w)]]}$$

そして，$\pi(w)$ と $g(w,\pi')$ の間の KL ダイバージェンスを考える．

$$D_{\mathrm{KL}}[\pi\|g_{\pi'}] := \mathbb{E}_\pi\left[\ln\frac{\pi(w)}{g(w,\pi')}\right]$$

ただし，$D_{\mathrm{KL}}[a\|b]$ は $a$ と $b$ の間の KL ダイバージェンス—$D_{\mathrm{KL}}[a\|b] = \int a(\bullet)[\ln a(\bullet) - \ln b(\bullet)]\mathrm{d}\bullet$ である．このとき，次の補題が成り立つ．

**補題 1** 任意の $\pi' \in \Pi$ に関して，次の 2 つの最適化問題は等価である．

$$\min_\pi D_{KL}[\pi\|g_{\pi'}] \iff \min_\pi \eta(\pi) + D_{KL}\left[\pi\|\pi'\right]$$

KL ダイバージェンスは常に 0 より大きい値をとるため，$D_{\mathrm{KL}}[\pi\|g_{\pi'}]$ は $\eta(\pi)$ の上限と等しい—$D_{\mathrm{KL}}[\pi\|g_{\pi'}]$ の最小化は直接的には $\eta(\pi)$ の最小化を導かない．しかし，$\eta(\pi)$ の単調な減少を実現することはできる．

**補題 2** 任意の $\pi' \in \Pi$ に関して，$D_{KL}[\pi\|g_{\pi'}] \le D_{KL}[\pi'\|g_{\pi'}]$ が成り立つならば, $\eta(\pi) \le \eta(\pi')$ が成り立つ．

補題 2 より，KL ダイバージェンス $D_{\mathrm{KL}}[\pi\|g_{\pi'}]$ の最小化を繰り返すことにより，$\eta(\pi)$ の単調減少を保証するアルゴリズムを構築できる．

$$\pi^{(k+1)}(w) = \operatorname*{argmin}_\pi D_{\mathrm{KL}}\left[\pi\,\middle\|\,g_{\pi^{(k)}}\right]$$

ここで，KL ダイバージェンスの性質より，$D[\pi\|g_{\pi^{(k)}}]$ の最小化解は次の式で得られる．

$$\pi^{(k+1)}(w) = g\left(w,\pi^{(k)}\right) = \frac{\pi^{(k)}(w)\exp[-f(w)]}{\mathbb{E}_{\pi^{(k)}}[\exp[-f(w)]]} \tag{3}$$

式 (3) による更新を繰り返すアルゴリズムを KLPE と呼ぶ．このアルゴリズムは，従来の KL 制御の方法 [Azar 12, Rawlik 12] と異なり，煩雑なメッセージ伝搬式を解く必要がない点は大きな強みである．

注意すべきこととして，KLPE の解 $\pi^{(k+1)}$ を計算するためには，1 ステップ前の解 $\pi^{(k)}$ による $\exp[-f(w)]$ に関する積分を評価しなければならない．この積分は一般に解析計算困難であるため，近似が必要となる．KLPE の近似解法は 4 節において議論する．

### 3.2 KLPE の理論的な性質

KLPE の優れた性質として，最適化問題 (2) の大域的な最適解への収束が保証されることである．

**定理 3** $\pi^* \in \Pi$ を $\eta(\pi)$ を最小にする確率分布とする．式 (3) の反復によって得られる系列 $(\pi^{(k)})_k$ は，$k\to\infty$ において $\pi^*$ に収束する．

証明は文献 [Rawlik 12] の定理 4 と同様の手順で確認できる．

続いて，KLPE の結果として得られる階層的な方策 $\delta_u(u-\bar{u}(x,w))\,\pi^*(w)$ と，最適な決定論的な方策 $\bar{u}(x,w^*)$ との関係性について考える．式 (3) は $\pi(w)$ に関する漸化式とみなせるため，初期条件 $\pi^{(0)} \in \Pi$ を用いて，$k+1$ ステップの KLPE 解，$\pi^{(k)}(w)$ は次のように書き直せる．

$$\pi^{(k)}(w) \propto \exp\left[\ln\pi^{(0)}(w) - kf(w)\right] \tag{4}$$

式 (4) は，$-f(w)$ をエネルギー関数，ステップ数 $k$ を温度パラメータとしたソフトマックス関数に初期分布 $\ln\pi^{(0)}(w)$ をバイアス項として付加したものと等価である．よって，異なる $f(w)$ の値を実現する $w$ と $w'$ の確率分布の出力の差分，$|\pi(w)-\pi(w')|$ は $k$ の増加に対して指数的なオーダーで増大する．したがって，$\pi^{(0)}(w)$ が任意の $w\in W$ において有限であり，かつ $f(w)$ を最小にする最適な方策パラメータが $C$ 個存在する—$\{(w_c^*)_{c=1}^C | w_c^* := \min_w f(w)\}$ ならば，KLPE 法によって得られる漸近分布 $\pi^*$ は，$w_c^*$ を中心とするデルタ分布の混合となる．この事実は，KLPE の結果として得られる階層的な方策は $f(w)$ を最小にする最適な方策に一致することを意味する．

**Corollary 4** KLPE は $f(w)$ を最小にする方策を導く．

## 4. KLPE の近似

前述の通り，KLPE を実行するためには，正規化項として $\pi^{(k)}(w)$ による $\exp[-f(w)]$ に関する積分を計算する必要がある．この積分は解析計算が困難であることから，各ステップでの KLPE 解を正規化項の計算が容易な分布で近似しなければならない．本研究では，各ステップの KLPE 解を単峰なガウス分布で置き換えることで，KLPE を近似的に実行する方法を提案する．

$s(w,\theta)$ はパラメータ $\theta \in \Theta, \Theta \subseteq \mathbb{R}^{\ell(\ell+3)/2}$ で特徴付けられる単峰なガウス分布とする．確率分布の近似法として幅広く用いられているのは，KL ダイバージェンスの最小化に基づくものである．KL ダイバージェンスによる近似は，その非対称性により 2 種類存在する，すなわち e 射影と m 射影である．e 射影は KL ダイバージェンス $D_{\mathrm{KL}}[s_\theta \| \pi^{(k+1)}]$ を最小化することで $\theta$ を求めるが，m 射影は e 射影の反対—$D_{\mathrm{KL}}[\pi^{(k+1)} \| s_\theta]$ を最小化する．

m 射影と e 射影はともに KL ダイバージェンスの最小化であるが，異なる性質の近似分布を導くことが知られている．m 射影によるガウス近似は 2 次のモーメントマッチングと等価であり [Bishop 06]，被近似分布の平均と共分散を保存するようにパラメータ $\theta$ が調節される．ゆえに，被近似分布が歪みのない単峰であれば良い近似を実現するが，多峰な場合は近似精度が大きく損なわれることがある．一方，e 射影によるガウス近似は被近似分布の峰を捉えるようにパラメータ $\theta$ が調節される．したがって，被近似分布が多峰な場合でも近似精度の著しい悪化を招かないことが報告されている [*1]．方策探索問題においては，方策 $\bar{u}(x,w)$ として非線形関数を設定すると $f(w)$ は多峰な関数となることが多いため，KLPE 解 (3) は多峰な分布となることが想定される．したがって，本研究では e 射影に基づく近似を考える．

初期分布 $\pi^{(0)}(w)$ はガウス分布 $s(w,\theta^{(0)})$ であるとする．このとき, 最初の KLPE の解，$\pi^{(1)}(w)$ は次の式で得られる．

$$\pi^{(1)}(w) = \frac{s(w,\theta^{(0)})\exp[-f(w)]}{\mathbb{E}_{\theta^{(0)}}[\exp[-f(w)]]}$$

ただし，$\mathbb{E}_\theta[\cdot]$ は $s(w,\theta)$ による $w$ に関する期待値である．
e 射影によるガウス近似 $s(w,\theta^{(1)})$ は $D_{\mathrm{KL}}[s_\theta \| \pi^{(1)}]$ の最小化によって得られる．$D_{\mathrm{KL}}[s_\theta \| \pi^{(1)}]$ は以下のように展開できる．

$$D_{\mathrm{KL}}\left[s_\theta \middle\| \pi^{(1)}\right] = \mathbb{E}_\theta[\ln s(w,\theta)] - \mathbb{E}_\theta\left[\ln s\left(w,\theta^{(0)}\right)\right] + \mathbb{E}_\theta[f(w)] + \mathrm{const.} \quad (5)$$

ここで，右辺の第 1 項，第 2 項は解析的に計算できるが，第 3 項 $\mathbb{E}_\theta[f(w)]$ は，$f(w)$ が 2 次形式，ガウスカーネルの和の場合など特殊な場合を除き，解析的に計算できない．また $D_{\mathrm{KL}}[s_\theta \| \pi^{(1)}]$ は非凸最適化問題であるため，数値最適化が必要となる．本研究では，ガウス分布による数値積分と勾配法ベースの数値最適化を組み合わせこの問題に対処する．

式 (5) の右辺の第 1 項, 第 2 項の微分は解析的に計算できるため，勾配法ベースな数値最適化を実行するためには，第 3 項 $\mathbb{E}_\theta[f(w)]$ の微分が評価できればよい．$\mathbb{E}_\theta[f(w)]$ のパラメータ微分は次の式で書ける．

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[f(w)] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln s(w,\theta) f(w)\right]$$

で得られる．ここで $(\partial/\partial\theta)s(w,\theta) = s(w,\theta)(\partial/\partial\theta)\ln s(w,\theta)$ の関係を利用した．$s(w,\theta)$ に関する期待値は以前として解析的には計算できないが，ガウス分布はサンプリングが容易であるため，数値積分により近似的に評価することができる．したがって，この近似した微分を勾配ベースの数値最適化に組み込み，KLPE 解の e 射影によるガウス近似を計算することができる．以降，全てのステップでの KLPE 解をこの近似したガウス分布で置き換えることで，KLPE を近似的に実行することができる．この反復アルゴリズムを e-KLPE と呼ぶ．

e-KLPE は，各ステップでの KLPE の解を単峰なガウス分布で近似しているため，前節で示した KLPE が持つ大域的な収束性はもはや保持していない．しかしながら，数値積分による微分の近似精度が十分であり，数値最適化として目的関数の単調減少性を保証するアルゴリズム (例えば BFGS，共役勾配法など) を用いれば，任意のステップにおいて，$D[s_{\theta^{(k+1)}} \| g_{\theta^{(k)}}] \leq D[s_{\theta^{(k)}} \| g_{\theta^{(k)}}]$ が成り立つ．これは，補題 2 より，e-KLPE の反復は $\eta(\theta) := \eta(\pi = s_\theta)$ を単調に減少させ，局所最小解へ収束することを示している．

## 5. まとめ

本研究では, 有限長マルコフ決定過程における新しい方策探索法, KLPE を提案し，KLPE は単調性，大域的な収束性を持つなど理論的に望ましい性質を持つことを示した．また，KLPE が持つ計算困難性を克服するため，e 射影に基づく近似法を提案した．

本研究を通して，期待累積コスト $f(w)$ は既知であると仮定したが，実問題においては $f(w)$ は未知である場合が多く，$f(w)$ は得られた観測履歴から推定する必要がある．$f(w)$ の有力な推定法として，再生核ヒルベルト空間上における確率推論法の利用が考えられる [Fukumizu 11]．この方法は，確率分布を直接モデル化することなく分布による期待値を推定でき，また確率分布の基本演算である和，鎖，そしてベイズ則を再生核ヒルベルト空間上で定義される線形演算で計算できる．したがって，$f(w)$ のように確率過程の系列の関数の期待値において，しばしば近似が必要となるメッセージ伝搬を，状態遷移分布 $p(x'|x,u)$ を同定することなく，かつ線形演算で評価できる．よって今後，e-KLPE にこの方法による $f(w)$ の推定を組み込み，より汎用的な枠組みに拡張することが期待できる．

## 参考文献


[Azar 12] Azar, M., Gómez, V., and Kappen, H.: Dynamic policy programming, *Journal of Machine Learning Research*, Vol. 13, No. 1, pp. 3207–3245 (2012)

[Bertsekas 07] Bertsekas, D.: *Dynamic Programming and Optimal Control*, Athena Scientific (2007)

[Bishop 06] Bishop, C. M.: *Pattern recognition and machine learning*, Springer (2006)

[Deisenroth 13] Deisenroth, M., Neumann, G., and Peters, J.: *Survey on Policy Search for Robotics* (2013)


[*1] これらの議論は，確率分布の近似推論法である変分ベイズ法と期待値伝搬法の違いとしてよく知られている．詳細は文献 [Bishop 06] の 10 節に見ることができる．

[Fukumizu 11] Fukumizu, K., Song, L., and Gretton, A.: Kernel Bayes' Rule, in *Advances in Neural Information Processing Systems*, pp. 1737–1745 (2011)

[Rawlik 12] Rawlik, K., Toussaint, M., and Vijayakumar, S.: On stochastic optimal control and reinforcement learning by approximate inference, in *Robotics: Science and Systems* (2012)

[Rückstieß 10] Rückstieß, T., Sehnke, F., Schaul, T., Wierstra, D., Sun, Y., and Schmidhuber, J.: Exploring parameter space in reinforcement learning, *Paladyn*, Vol. 1, No. 1, pp. 14–24 (2010)

[Sehnke 10] Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J.: Parameter-exploring policy gradients, *Neural Networks*, Vol. 23, No. 4, pp. 551–559 (2010)

[Theodorou 10] Theodorou, E. A., Buchli, J., and Schaal, S.: A generalized path integral control approach to reinforcement learning, *Journal of Machine Learning Research*, Vol. 11, pp. 3137–3181 (2010)