PILCO におけるカーネル関数の変更による予測精度の向上

加藤 鳳人　　指導教員：小林 邦和

1　はじめに

近年の強化学習・深層強化学習の研究が進んでいる一方で, 強化学習・深層強化学習には膨大なデータが必要であり, データを収集するためのコストが問題となっている. そのコストを削減するために環境から遷移モデルを学習し, 遷移モデルからデータを得るモデルベース強化学習が存在する. しかし, 遷移モデルの学習においても多数のデータが必要である問題が存在する. その解決方法としてガウス過程 [1] を用いて少ないデータで遷移モデルを近似する手法である PILCO(probabilistic inference for learning control)[2] が提案されている. しかし, PILCO はガウス過程回帰の出力の期待値を求める必要がある. そのため, カーネル関数を変更するたびに期待値を解析的に求めなければならず, カーネル関数の変更によって予測精度の向上を図ることが困難である. 本研究では, 上記の問題を解決し, PILCO のガウス過程のカーネル関数を容易に変更することで予測精度の向上を図ることを目的とする.

2　PILCO

PILCO は, M.P.Deisenroth と C.E.Rasmussen によって提案された確率的なモデルを用いたモデルベース強化学習法である. 最初に環境からデータを生成し, 遷移モデルを学習する. その後, 図 1 に従って強化学習を行う. なお, 状態は (多変量) 正規分布に従うと仮定し, ガウス過程にはガウスカーネル [1] が使用されている.

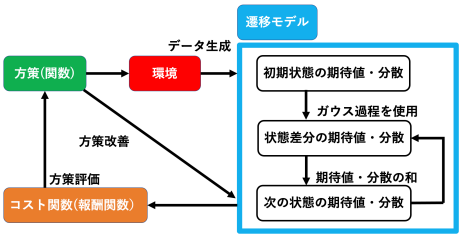

図 1　PILCO の実行手順

3　提案手法

本研究で提案する手法は, PILCO における状態差分の期待値・分散をモンテカルロ法によって推定する手法である. 状態差分の期待値・分散を推定することでカーネル関数を変更することが容易となる. 具体的には, マルコフ連鎖モンテカルロ法 [3] を利用してガウス過程の入力が従う (多変量) 正規分布から十分大きなサンプルを得ることで出力である状態差分から期待値・分散を推定する. 入力が従う (多変量) 正規分布のサンプルを事前に標準化された (多変量) 正規分布からマルコフ連鎖モンテカルロ法を用いてサンプルを得ておき, 入力が従う (多変量) 正規分布の平均ベクトル・分散共分散行列を用いて標準化と逆の計算をすることで得るアプローチである.

4　計算機シミュレーション

本研究では,「OpenAI Gym」を使用して提案手法の性能検証を行う. 具体的には「OpenAI Gym」の「cartpole」の行動 (制御信号) を連続値に変更した「continuous cartpole」[4] を使用する. ポールを倒さないようにカートを前後に移動させることが目的のタスクである. 従来手法 [2] と提案手法についてカーネルを変更した結果を表 1, 表 2 に示す. 表 1 からカーネルにより遷移モデルの予測精度が異なることがわかる. 例えば, RationalQuadratic カーネル [5] は, 他のカーネルに比べてカートの位置の予測精度が低くなり, ポールの角速度の予測精度が高くなっている. また, 表 2 から提案手法は学習に時間がかかる欠点があることがわかる. 最後にカーネルを変更して学習した方策による 100 回の試行のポールの角度の平均値を図 2 に示す. 図 2 からカーネルを変更することで方策が変化することがわかる. また, ガウスカーネルの場合, 従来手法と提案手法で誤差が生じている. これは期待値の推定誤差によるものだと考えられる.

表 1　100 回の試行の平均値と遷移モデルの期待値の平均平方二乗誤差

状態の要素	従来手法 (ガウスカーネル)	提案手法 (ガウスカーネル)
カートの位置 [m]	1.37876×10^{-2}	1.53714×10^{-2}
カートの速度 [m/s]	1.03938×10^{-2}	1.07902×10^{-2}
ポールの角度 [rad]	6.48482×10^{-3}	8.70724×10^{-3}
ポールの角速度 [rad/s]	1.10095×10^{-2}	1.13271×10^{-2}
状態の要素	提案手法 (Matérn5 カーネル [1])	提案手法 (RationalQuadratic カーネル)
カートの位置 [m]	2.57775×10^{-2}	6.15311×10^{-2}
カートの速度 [m/s]	1.45332×10^{-2}	2.05330×10^{-2}
ポールの角度 [rad]	6.04185×10^{-3}	9.37516×10^{-3}
ポールの角速度 [rad/s]	1.63083×10^{-2}	6.10205×10^{-3}

表 2　従来手法と提案手法の学習時間とデータ収集時間

手法	学習時間 [s]	データ収集時間 [s]
従来手法 (ガウスカーネル)	2.34228×10^{2}	2.20076×10^{1}
提案手法 (ガウスカーネル)	4.22184×10^{3}	2.01243×10^{1}
提案手法 (Matérn5 カーネル)	4.72848×10^{3}	2.00920×10^{1}
提案手法 (RationalQuadratic カーネル)	5.00656×10^{3}	2.04909×10^{1}

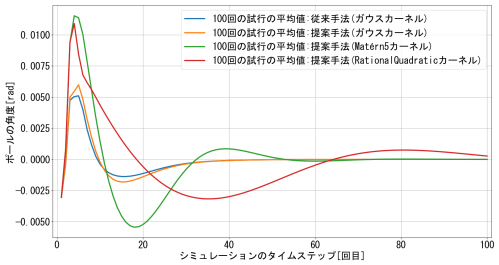

図 2　学習した方策関数を使用した結果

5　おわりに

本研究では, モンテカルロ法を用いることで PILCO におけるカーネル関数を変更し, 予測精度の向上を図った. その結果, PILCO のカーネル関数の変更を容易にすることに成功した. 今後の課題としては, 状態差分の予測において最適なカーネル関数の検討や学習時間の短縮などが挙げられる.

参考文献

[1] 持橋 大地, 大羽 成征:「ガウス過程と機械学習」, 講談社, (2019)
[2] M.P.Deisenroth and C.E.Rasmussen: “PILCO: A Model-Based and Data-Efficient Approach to Policy Search”, Proceedings of ICML 2011, pp.465–472 (2011)
[3] 須山 敦志:「ベイズ深層学習」, 講談社, (2019)
[4] Ian Danforth: continuous_cartpole, https://gist.github.com/ iandanforth/e3ffb67cf3623153e968f2afdfb01dc8, 参照日 (2022 /12/03)
[5] D. Duvenaud: “The Kernel Cookbook: Advice on Covariance functions””, (2014), https://www.cs.toronto.edu/˜duvenaud/cookbook/, 参照日 (2023/01/08)