ISS-P-016

動的報酬多腕バンディット問題における UCBQ の短期的調整法

蜂谷　直寛†　　中野　秀洋†
† 東京都市大学大学院総合理工学研究科情報専攻

1. はじめに

本稿では，報酬確率分布が時間経過により変化する動的報酬多腕バンディット問題[1]において，UCBQ 手法[2]の探索傾向を動的に調整する手法を提案する．

2. 提案手法

2.1 UCBQ の探索傾向の動的調整

Q 学習と UCB 手法を組み合わせた UCBQ 手法[2]において，探索の傾向を決めるパラメータを動的に変化させる．2.2 節にて後述する方法で環境変化を察知したとき，式 (1) を用いて UCB 値を計算することで選択回数が少ないアームの探索を強める．

$$UCB_i = Q_i + C \times \frac{T}{r+K} \times \sqrt{\frac{\ln N}{n_i}} \quad (1)$$

ここで，iはアームの番号，Q_iはアームiの Q 値，Cは定数，Nは総選択回数，n_iはアームiの選択回数である．TとKはそれぞれ探索の強さと期間を制御するパラメータ，rは環境変化察知からの経過ステップ数である．Tの大きさに応じて探索を強めるだけでなく，rとKによって探索の効果を一時的なものとすることにより，環境変化への迅速かつ柔軟な対応が可能ある．

2.2 環境変化の察知

損失率を一定ステップs毎に前と後で比較し，損失率が増加していた際，環境変化を察知する．損失率を式(2)に示す．

$$1 - \frac{total\_reward}{total\_reward^*} \quad (2)$$

ここで，$total_reward$は実際に得られた累計報酬，$total_reward^*$は得られる最大の累計報酬である．

表 1 実験で用いたパラメータ

環境	試行回数	10000
	アーム数	5
	報酬	1
	環境変化ステップ数	5000
	終了ステップ数	10000
Q 学習	Q 値の初期値	0
	学習率	0.1
	割引率	0.9
	ランダム選択確率	0.1
提案手法	C	1
	T	10000
	K	5000
	s	50

3. 結果及び考察

提案手法の性能を評価するため，通常の UCBQ 手法と比較実験を行う．実験環境を表 1 に示す．表 2 に本実験で使用したアーム 5 本の報酬確率を示す．図 1,2 および表 3 にそれぞれ平均報酬と損失率の結果を示す．

表 2 各アームの報酬確率

	アーム 1	アーム 2	アーム 3	アーム 4	アーム 5
開始時	0.5	0.3	0.7	0.1	0.4
変化後	0.3	0.1	0.4	0.7	0.5

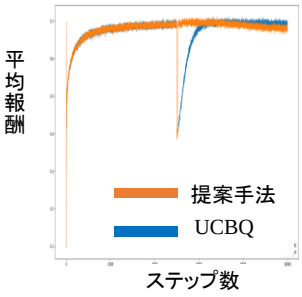

図 1 平均報酬

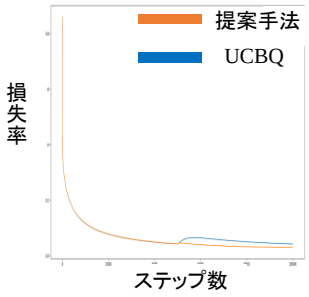

図 2 損失率

表 3 各手法の平均報酬，最終的な損失率

	平均報酬	損失率
UCBQ 手法	0.692	0.042
提案手法	0.681	0.031

図 1 より，提案手法は環境変化後の収束速度が優れていた．図 2 より，環境変化後，提案手法の損失率は素早く減少した．表 3 より，最終的な平均報酬は，若干低くなったが，最終的な損失率は，1%以上の差があるため提案手法が有効であったと考えられる．一方，探索を強めた結果，最終的な平均報酬に差が生じたと考えられる．このため，探索を強めた後，パラメータKの調整により，探索を通常時に戻す速度を早める必要があると考えられる．

4. むすび

本稿では，UCBQ 手法の探索の傾向を決めるパラメータを動的に変更する手法を提案し，比較実験を行った．今後の課題として，損失率だけでなく平均報酬の向上も見込める手法を検討する必要がある．

参考文献

[1] A. Garivier, et.al., Algorithmic learning theory (ALT' 11), pp.174-188, 2011.
[2] K.Saito, et.al, Proc. IEEE-ICMLA, DOI: 10.1109/ICMLA.2015.59, 2015.