情報処理学会論文誌 Vol.59 No.11 2063–2073 (Nov. 2018)

強化学習法によるデジタルカーリングの初歩的な行動知識の獲得

松井 亮平[1,a)] 保木 邦仁[1]

受付日 2018年2月16日, 採録日 2018年9月7日

概要：不確定性を持ち，ストーン配置やショットが浮動小数点数で表されるデジタルカーリングを題材に，一般化方策反復に基づく強化学習の一手法を検討した．強化学習はおおよそカーリングの予備知識を用いない行動集合とランダム方策から開始した．行動価値は重みの総数 1,000 万ほどの畳込みニューラルネットワークを用いて，挙動方策が生成した総数 6 億ほどの行動から推定した．行動集合が巨大であるため，グリーディ方策はモンテカルロ法により近似的に求めた．この実験によりグリーディ方策がサンプルプログラムに比する程度の強さを持ち，初歩的なショット知識に基づいた行動をとるようになる過程を明らかにした．

キーワード：ゲームにおける学習，デジタルカーリング，不確定ゲーム，強化学習，ニューラルネットワーク，方策オフ型モンテカルロ法

Reinforcement Learning of Elementary Action Knowledge of Digital Curling

Matsui Ryohei[1,a)] Hoki Kunihito[1]

Received: February 16, 2018, Accepted: September 7, 2018

Abstract: We examined one of the reinforcement learning methods on the basis of generalized policy iteration by means of digital curling in which stone coordinates and shots are represented by floating numbers. The reinforcement learning method was carried out by using an action set which was formed by little preliminary domain knowledge of the game, and by using random initial policy functions. Action values were inferred from about six hundred million actions generated by behavior policies by using convolutional neural networks with about ten million weights. Greedy policies were approximately computed by the Monte Carlo method because the action set is huge. Our experiment showed that the performance of the greedy policy is nearly comparable to a sample AI program and disclosed the process in which the policy starts to make actions on the basis of elementary knowledge of shots.

Keywords: learning in games, digital curling, stochastic game, reinforcement learning, neural network, off-policy Monte-Carlo method

1. はじめに

カーリングは 2 つのチームが競い合うウインタースポーツであり，氷上のチェスともいわれ，勝つためには戦略の高度な分析が求められる [1]．カーリングは現実世界で行われるスポーツであることから，チームを構成する選手の技量や自然環境の変化などあらゆる可能性を知り考慮することはおおよそ無理であり，戦略の分析は非常に困難な課題となる．

デジタルカーリングは [2]，カーリングにおいて抽象的な戦略の議論を行うために様々な可能性を排除した，二人零和有限不確定完全情報ゲームである[*1]．ここで，ゲーム

[1] 電気通信大学
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[a)] matsui.ryohei@uec.ac.jp

[*1] 二人零和有限不確定完全情報ゲームについては，たとえば，書籍 [3] 参照．本論文では，不確定ゲームとは，偶然手番によりプレイが確率的に分岐する展開型ゲームとする．なお，デジタルカーリングはゲーム木が有限ではあるが，ゲーム状況が浮動小数点数により記述され，このゲームのプレイヤには選択肢数が無限に近いような感覚を与える．

の抽象化は選手の技量差やスウィーピングの排除，氷の性質の均一化，各チームを 1 プレイヤと見なす 2 人ゲーム化などによって達成される．このゲームの特色はストーンの配置やショットが複数の浮動小数点数で表現される点にあり，ゲーム木の探索空間は膨大なものとなる．

ゲーム木の探索において，膨大な探索空間でも有効となりうる方法としてヒューリスティック探索が知られている [4]．デジタルカーリングにおいて現在主流となっている思考アルゴリズムも，ゲームプレイを何度も試行してゲーム木の末端節点を評価するモンテカルロ木探索（MCTS）や，静的に末端節点を評価して木探索する Expectimax に基づくものが多い [5], [6]．これらヒューリスティック探索の性能は，ゲームプレイの試行で用いられる行動確率や，木探索の末端節点の静的評価に大きく依存するため，デジタルカーリングにおいては強い既存プログラムのプレイ記録からこれらの値を機械学習する方法が提案されている [6], [7]．

近年，ゲーム木のヒューリスティック探索において，人間プレイヤや既存プログラムのプレイ記録を必要としない強化学習により行動確率や静的評価値を得る研究が顕著な成果をあげている．このような強化学習法を種々のゲームに適用する事例研究は，現在の人工知能分野において主要な興味の対象となっている [8], [9], [10]．

本研究ではデジタルカーリングを題材として強化学習法の 1 手法である一般化反復方策（GPI）を次の 2 つの観点により検討する．まず，ランダムショットを行うプレイヤから開始した強化学習により導かれたプレイヤの強さを対戦実験に基づき調査し，方策改善の度に性能が向上していくことを確認する．次に，このような強化学習法により，方策改善の度に方策関数が初歩的な行動知識を獲得していく様子を明らかにする．本研究は著者らの研究報告 [11] の内容を発展させたものである．

2. 基礎知識

本章では，まず，2.1 節でデジタルカーリングの概要を述べ，次に，2.2 節で強化学習法の基礎的事項のうち，本研究と関連の深いものを述べる．

2.1 デジタルカーリング

デジタルカーリングは 2 人のプレイヤが交互にストーンをショットしてプレイするゲームである．以下のように進行するエンドを 8 回繰り返し[*2]，その合計得点で勝敗を決める．

(1) 前エンドで得点したプレイヤが先攻
(2) 先攻と後攻が交互にストーンをショット

[*2] カーリングでは 1 ゲーム 8 または 10 エンドが主流．デジタルカーリング大会では 8 エンドが主流であるため，本研究ではこれにならう．

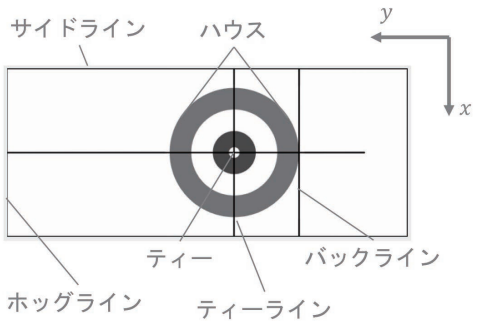


図 1 デジタルカーリングのシート拡大図
Fig. 1 Enlarged view of the digital curling sheet.

(3) 互いに 8 ショットした時点で得点を計算[*3]

デジタルカーリングをプレイするシートの拡大図を**図 1** に示す．ホッグライン，バックライン，サイドラインに囲まれた領域をプレイエリアと呼び，ホッグラインを超えないストーン，サイドラインに触れたストーン，バックラインを超えたストーンはプレイエリア外にあると見なされ除外される．ハウスと重なっているストーンのことを，ハウス内にあるストーンという．シートの座標は，バックラインと平行な x 軸と，サイドラインと平行な y 軸により表現される．

プレイエリアにある各ストーン座標は単精度浮動小数点数の対で表現される．また，ショットの意思決定は初速度の x，y 成分（それぞれ単精度浮動小数点数）および回転方向（左右の 2 値）で表現される．プレイヤは対戦サーバからストーン配置や得点状況などを受信して，意思決定（初速度，回転方向）を対戦サーバに送信する．対戦サーバは受信した初速度に対して乱数を加え，乱数加算後のショットに基づいて物理シミュレーションを行い，ショット後のストーン配置を生成する．なお，本研究においては，乱数加算前の初速度および回転方向の意思決定をプレイヤの行動と呼び，乱数加算後の初速度および回転方向をショットと呼ぶ．

行動の集合 $\mathcal{A}$ は初速度 $\boldsymbol{v} = (v_x, v_y)$ と回転方向 r の対 $a = (\boldsymbol{v}, r)$ すべてからなる集合とする．ただし，本研究では，$\boldsymbol{v}$ の成分 v_x は区間 $[-3.10, 3.10]$ に属し，v_y は区間 $[-33.7, -26.7]$ に属す単精度浮動小数点数とする．この行動集合 $\mathcal{A}$ は，ホッグラインを通過するデジタルカーリングの初速度すべてを含むのに十分な大きさを持つ．また，行動 $a \in \mathcal{A}$ によりホッグラインを超えず除去されるストーンもショット可能である．

2.2 強化学習問題と方策オフ型モンテカルロ法

本節では書籍から [12]，本研究で用いる用語を中心に抜粋して説明する．強化学習問題は，意思決定を行う主体が

[*3] プレイヤの得点は，ハウス内のストーンのうち相手プレイヤの最もティーに近いストーンよりもティーに近いストーンの数．

それ以外すべてから構成される環境と相互作用を行いながら，これが目標を達成するように学習させる問題である．このような相互作用下で主体が経験する状態の遷移規則は有限マルコフ決定過程（有限 MDP）としてモデル化される．有限 MDP は状態と行動の集合 $\mathcal{S}$，$\mathcal{A}$ と 1 ステップダイナミクスから定義される．1 ステップダイナミクスは対 $(s,a) \in \mathcal{S} \times \mathcal{A}$ から次の状態 $s' \in \mathcal{S}$ への遷移確率 $\mathcal{P}^a_{ss'}$ と報酬の期待値 $\mathcal{R}^a_{ss'}$ により主に特徴付けられる．

意思決定を行う主体の目的はエピソード $(s_0 a_0 r_1 \cdots a_{T-1} r_T s_T)$ の各時間ステップ $t \in \{0,\ldots,T-1\}$ の収益 $R_t = r_{t+1} + r_{t+2} + \cdots + r_T$ を最大化することである．確率的方策 $\pi(s,a)$ の値は状態が $s \in \mathcal{S}$ のときにこの主体が行動 $a \in \mathcal{A}$ をとる確率である．決定論的方策 $\pi(s)$ の値は状態が $s \in \mathcal{S}$ のときにこの主体が確率 1 でとる行動である．状態・行動対 (s,a) の行動価値 $Q^\pi(s,a)$ は，状態 s で行動 a を行い，以降方策 π に従った場合の期待収益を表す．状態 s の価値 $V^\pi(s)$ は，その状態以降方策 π に従った場合の期待収益を表す．

本研究で用いる強化学習アルゴリズムは関数近似手法を利用した方策オフ型モンテカルロ法（MC 法）である．方策オフ型 MC 法は一般化方策反復（GPI）の考え方に基づき，方策評価と方策改善の 2 つの過程を繰り返して，期待収益を最大化するように最適方策を求める．方策評価とは，方策 π の行動価値や状態価値の推定値を計算することである．方策改善とは，各状態の価値が大きくなるように方策を更新することである．各状態で推定価値が最大となる行動を選択するような決定論的方策をグリーディ方策と呼ぶ．

方策オフ型手法とは，意思決定を行う確率的方策 π'（挙動方策）と，方策評価で評価される決定論的方策 π（推定方策）が分離される手法である．到達可能な対 (s,a) すべてを探査するためには，挙動方策が与える確率は推定方策によって選ばれうる行動すべてにおいて非ゼロであることが必要である．

MC 法とは，エピソードを何度も生成して標本収益を平均化することにより方策評価を行う方法である．この方法は，時間ステップ t の価値を $t+1$ から T までの報酬をバックアップして更新するような方法と見なすことができる．MC 法ではエピソードが終わるまで価値の更新を待つ必要があり，T が大きくなりうるような過程では学習効率に期待ができない．

一方で，TD 法では，1 ステップ先の報酬と推定価値のみをバックアップする方法であり，価値の更新は 1 ステップ待つだけでよい．TD(λ) 法は，TD 法から MC 法へと移行するように報酬や推定価値のバックアップを組み合わせる 1 つの手法である．ここで，$\lambda \in [0,1]$ はこれら 2 種のバックアップ法の組合せのバランスをとるパラメタである．TD(0) は TD 法，TD(1) は MC 法のバックアップに相当する．

関数近似手法とは，状態や行動数が大きく，状態・行動対の 1 つに 1 つのエントリが対応するような表形式の価値関数を実装することが困難な場合に有効な手法である．この手法では，価値関数はパラメタ $\boldsymbol{\theta}$ の関数として近似的に表現される．価値を収益の標本平均に近づけることは，関数近似を行った場合には適切な損失関数（収益との平均 2 乗誤差など）を最小化するように $\boldsymbol{\theta}$ を更新していくことでなされる．

3. 目的

本論文では，デジタルカーリングのプレイヤの開発において深層学習と強化学習を組合せる手法の法則性発見の足掛かりとなるような 1 事例研究を行う．本研究では次の 3 つを目標に設定した．

1 つめの目標は，GPI に基づく行動価値の学習を行うことである．デジタルカーリングの行動集合は非常に大きいため，グリーディ方策を計算して方策改善することが困難である．そこで，価値を最大化するグリーディ方策は MC 法により近似的に求める．

2 つめの目標は，初歩的な行動知識獲得の過程を観察することである．おおよそカーリングの予備知識を用いないような行動集合を用いて，この過程を明らかにする．これにより，高度な行動知識獲得を達成して強いプレイヤを開発するときに有効な知見を得たり，現実世界のカーリングとこれを抽象化したデジタルカーリングのゲームとしての戦略性がおおよそ同じであることが確認できたりするようなことが期待される．なお，本研究では，プレイヤの初歩的な行動知識獲得とは，カーリングの書籍 [1], [13] で解説されている様々なドローやテイクアウトなどのショットのうち初歩的なものを行うようになることであると考える．さらに，本研究はデジタルカーリングを題材として行い，カーリング競技の技術的側面は考慮せずにこれの戦術面のみに注目する．3 つめの目標は，グラフィックスプロセッシングユニット（GPU）を搭載した一般的なワークステーション数台で，数カ月程度で遂行できる実験を行うことである．ヒューリスティック木探索や確率的方策の学習は行わず，さらに，1 エンドのみのプレイに対して強化学習法を適用する．

4. 先行研究

本章では，まず，4.1 節においてデジタルカーリングの先行研究を紹介し，本研究との関連性を述べる．次に，4.2 節において人間プレイヤや既存プログラムのプレイ記録を必要としない強化学習の顕著な成功事例と本研究との類似点と相違点を示す．

表 1 先行研究と本研究の比較
Table 1 A comparison between related work and our work.

	ゲームの種類	エピソードの長さ *a	行動集合の大きさ *b	NN 1 個ごとの重みの数	報酬のバックアップ	強さ
Tesauro [8]	二人不確定	10^2	10^1	10^4	TD(λ)	上級者以上
Mnih ら [9]	一人不確定	10^3	10^1	10^6	TD(0)	上級者 *c
Silver ら [10]	二人確定	10^2	10^2	10^7	TD(1)	上級者以上
本研究	二人不確定	10^1	10^{16}	10^5	TD(1)	初級者

*a 2 人ゲームの場合は両方のプレイヤの行動回数の和，Atari2600 では 1 分間の行動数で見積もった．バックギャモンではダブリングキューブを使用しないとして見積もった．

*b デジタルカーリングでは，区間 $[-3.10, 3.10]$ に 10^9 個，区間 $[-33.7, -26.7]$ に 10^7 個程度の単精度浮動小数点数が属する（符号，指数，仮数部それぞれ 1，8，23 ビットとして見積もった）．

*c 49 個中 29 個のゲームで人間と同等以上のスコア．

4.1 デジタルカーリング

ヒューリスティック探索アルゴリズムと，方策や評価関数の機械学習法により強いプレイヤを構成する試みが報告されている．

加藤らは不確定性を考慮してエンドのゲーム木探索を行う Expectimax を適用した [5]．末端節点の評価にはヒューリスティックに設計した評価関数を用いた．また，得点差と残りエンド数から勝率の推定も行った．この手法に基づき開発された「じりつくん」は，2015 年の IEEE の国際会議（Computational Intelligence and Games）で開催された大会で優勝した．加藤らはさらに，「じりつくん」のエンド得点確率を推定し，得点期待値を求めて，事後状態の価値を機械学習する方法も提案した [7]．この確率の推定には，2 層の畳込み層，2 層のプーリング層，2 層の全結合層からなる畳込みニューラルネットワーク（CNN）を用いた．本研究でもこれにならい，CNN を用いてエンドの得点確率を推定することとした．ただし，本研究では，ショットの不確定性も考慮した推定を行うことを目指し，事後状態ではなく，状態・行動対の価値を推定することとした．また，推定の精度を向上させるために，規模のより大きな CNN を用いた．

大渡らは，過去大会上位プログラムのプレイ記録を用いて，重み約 4 万の特徴の線形重み和からなるソフトマックス関数を用いた行動確率の推定を行った [6]．また，連続な状態空間を考慮したモンテカルロ木探索（MCTS）をデジタルカーリングの 1 エンドに適用し，エンドを試行する方策にこの推定された行動確率を用いた．さらに，得点差と残りエンド数などから勝率を推定し，これをエンドの試行結果に用いた．この方法に基づき開発された歩（あゆむ）は，2016 年の国内会議（ゲームプログラミングワークショップ）で開催された大会で 2016 年に優勝した[*4]．

これらの先行研究では膨大な状態・行動空間を洗練された方法で粗視化してヒューリスティック探索を行い，強化学習を利用せずに強いプレイヤを構成した．一方で本研究では，デジタルカーリングの予備知識をおおよそ用いないような行動集合を仮定する強化学習を行った．

4.2 ゲームプレイヤの GPI を行う強化学習

強化学習法は，意思決定を行う主体が 1 人ではなかったり，状態・行動対すべてを網羅的に生成し記録することがおおよそ無理であったりするようなゲームにおいても適用可能な方法である．2 人ゲームの 2 人の行動両方を制御したり，状態・行動価値をニューラルネットワーク（NN）で関数近似したりした場合において，最適価値や最適方策を得る一般的な条件や方法は知られていない．それでもなお，人間と同等かそれ以上の強さを持つプレイヤが強化学習法により生成されたという成功事例が複数のゲームで報告されている（**表 1** 参照）．本研究の行動集合は顕著に大きく強化学習が困難となることが予想される．なお，表にまとめてはいないが，本研究の状態集合もこれらの事例より大きいと考えられる[*5]．また，本研究のエピソードは表の他の事例と比較して短いため，ブートストラップを利用しない TD(1) 法を用いることとした．

バックギャモンでは NN と強化学習を組み合わせた研究の成功事例がいち早く報告されている．Tesauro は，ゲーム状況から直接的に得られる特徴を入力とする NN を用いて，TD(λ) 法により事後状態の価値推定を行った．十万回程度のゲームプレイで訓練された TD-Gammon は，高度にチューニングされた Neurogammon と同等の性能を持っていた．このゲームは二人零和有限不確定完全情報ゲームであることに加えて，偶然手番とプレイヤ手番がおおよそ交互に起きる点においてもデジタルカーリングの性質と類似している．

Mnih らは，深層学習と行動価値の強化学習を組み合わせた Deep Q-Learning（DQN）を提案した．Atari2600 の複数のゲームに対してゲームの画面を CNN に入力し，ゲームから得られる報酬を最大化するように DQN を適用したと

[*4] http://minerva.cs.uec.ac.jp/curling/wiki.cgi

[*5] ただし，Atari2600 の各ゲームの状態数の適切な数え方は不明である．

ころ，29 のゲームで人間の上級プレイヤとおおむね同等かそれ以上のスコアを記録した．さらに，深層学習と MCTS を組合せた方法も Guo らによって提案されている [14]．

Silver らは，囲碁の確率的方策と状態価値両方を推定する CNN を構成した．このネットワークも，MCTS と組合せて学習された．また，MCTS と組合せることなく関数近似された価値関数のみでプレイしても，当時高度にチューニングされた囲碁プログラムよりも強いことが示された [15]．さらに，同様の手法を将棋・チェスに適用し，既存の最上位プログラムより高い性能を得たと報告している [16]．本研究の実験規模は，Silver らの研究と比較すると小規模ではあるが，計算内容を厳選し，ヒューリスティック木探索は行わず，グリーディ方策単体の性能を検証することとした．また，Silver らは確率的方策と価値両方を推定しているが，本研究では同じ種類に分類されるゲームを研究した Tesauro にならい，価値の方のみを推定することとした．

5. 提案手法

本章では，デジタルカーリングにおいて，初歩的な行動知識を獲得するための強化学習法について述べる．5.1 節では CNN を用いてエンドの行動価値を推定する方法を述べる．5.2 節では関数近似手法を利用した方策オフ型 MC 法を適用する方法を述べる．

5.1 CNN を用いた行動価値の推定

本節では，CNN を用いたエンド得点確率の推定方法について説明する．まず，CNN の構成について述べる[*6]．エンド得点確率の推定値はショット番号 $m \in \{0, \ldots, 15\}$，ストーン配置 X，行動の初速度 $\boldsymbol{v} = (v_x, v_y)$，行動の回転方向 $r \in \{\text{右}, \text{左}\}$，エンド得点インデックス $i_R \in \{1, \cdots, N_{\text{out}}\}$ に対して定まるものであると考える．ただし，N_{out} は行動の回転方向ごとの CNN の出力数（奇数），R はエンド得点，i_R は

$$i_R = \begin{cases} 1 & (2R < -N_{\text{out}} + 1) \\ N_{\text{out}} & (2R > N_{\text{out}} - 1) \\ R + (N_{\text{out}} + 1)/2 & (\text{otherwise}) \end{cases}$$

である．CNN は，ショット番号 m 各々に対応する重みベクトル $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{15})$ を持つとして，入力 $(X, \boldsymbol{v})$ と重みベクトル $\boldsymbol{\theta}_m$ から各回転方向 r に対応する N_{out} 個の値 $\boldsymbol{z}_r(X, \boldsymbol{v}, \boldsymbol{\theta}_m)$ を出力するように構成する（**図 2**，図 2 中の畳込み部の詳細に関しては**表 2** を参照）[*7]．そして，期待エンド得点の推定値 $\overline{z}_r(X, \boldsymbol{v}, \boldsymbol{\theta}_m)$ は

[*6] CNN に関する機械学習の基礎知識は，たとえば，書籍 [17] を参照．

[*7] 出力はベクトル値 $\boldsymbol{z}_r = (z_{r1}, \ldots, z_{rN_{\text{out}}})$ とする．

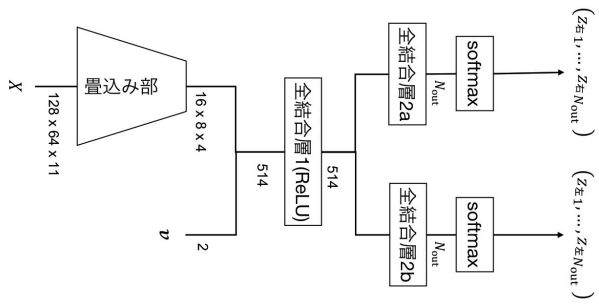


図 2 CNN 概形
Fig. 2 An outline of CNN.

表 2 畳込み部詳細．ストライドは畳込み 1，プーリング 2．
Table 2 Details of convolution part. Strides are 1 (convolution) and 2 (pooling).

層（フィルタサイズ）	出力サイズ	ReLU
入力層	$128 \times 64 \times 11$	
畳込み（5×5）	$128 \times 64 \times 32$	✓
畳込み（3×3）	$128 \times 64 \times 32$	
最大プーリング（2×2）	$64 \times 32 \times 32$	
畳込み（3×3）	$64 \times 32 \times 64$	✓
畳込み（3×3）	$64 \times 32 \times 64$	
最大プーリング（2×2）	$32 \times 16 \times 64$	
畳込み（3×3）	$32 \times 16 \times 96$	✓
畳込み（3×3）	$32 \times 16 \times 96$	
最大プーリング（2×2）	$16 \times 8 \times 96$	
畳込み（3×3）	$16 \times 8 \times 128$	✓
畳込み（3×3）	$16 \times 8 \times 128$	
畳込み（1×1）	$16 \times 8 \times 4$	

表 3 ストーン配置の表現法
Table 3 The representation of stone placement.

特徴	チャネル数
すべてのストーン	1
各プレイヤのストーン	2
最もティーに近いストーン	1
各プレイヤの最もティーに近いストーン	2
ハウス内のティーからの距離（不変）	1
プレイエリアの端からの距離（不変）	4

$$\overline{\boldsymbol{z}}_r(X, \boldsymbol{v}, \boldsymbol{\theta}_m) = \sum_{R=-(N_{\text{out}}-1)/2}^{(N_{\text{out}}-1)/2} R\, z_{r\, i_R}(X, \boldsymbol{v}, \boldsymbol{\theta}_m)$$

のように計算する．

次に，ストーン配置 X の表現方法を述べる．配置 X は，11 のチャネルからなる画像により表現される．各チャネルは，区間 $[0, 1]$ の輝度を持つ 128×64 の画素からなる．この画素は格子状に区切られたプレイエリアの各マスに対応し，輝度はこのマスにおけるある特徴の値を表す（**表 3**）．ストーンに関する特徴の値は，**図 3** に示すようにその画素とストーンが重なる面積の割合とする．距離に関する特徴の値は，その画素の中心までのシートの xy 座標平面上の

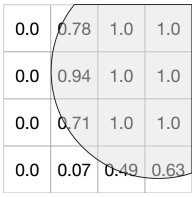


図 3 あるストーン配置の特徴抜粋．格子は画素，灰色の円はストーン，各画素の数値はストーンと重なった面積を表す．各画素の面積は 1 とする

Fig. 3 An extraction of features of stone placements. Grids represent pixels, the gray circle represents a stone, each pixel shows a value that represents the overlap area with the stone. The area of each pixel is one.

距離を d として，$\exp(-d)$ とする．また，行動の初速度 $\boldsymbol{v}$ の各成分も，これがとりうる値が区間 $[-1,1]$ になるように線形変換する．縦横比を 2 にしたのは，デジタルカーリングのプレイエリアがおおよそそうであるからである．

さらに，計算資源の利用方法について述べる．CPU はグリーディ方策を MC 法により近似的に計算するときに用いる．このとき，畳込み部すべてと全結合層 1 の 512 個の入力に関する部分は，計算結果を保存し再利用する．これにより，同じストーン配置 X に対してとられる様々な初速度 $\boldsymbol{v}$ の順伝播計算が容易になる．GPU は，誤差逆伝播法をこの CNN で行うときに用いた．

さらに，重みベクトルの調整法について述べる．この重み調整法にはミニバッチを用いる慣性項付きの確率的勾配降下法（SGD）と Adam [18] を用いる．慣性項の係数は 0.9，Adam のパラメタは文献 [18] で推奨されるものと同じとする．損失関数は CNN が推定する得点確率分布 $\boldsymbol{z}_r$ と実際の得点 R の交差エントロピー誤差 $E(\boldsymbol{z}_r, R) = -\ln z_{r\,i_R}$ を用いて，

$$L(m, X, \boldsymbol{v}, r, R, \boldsymbol{\theta}) = E(\boldsymbol{z}_r(X, \boldsymbol{v}, \boldsymbol{\theta}_m), R)$$

とする．ミニバッチは同じ回転方向 r を持つ 32 組から構成し，重みベクトルは Glorot らの手法にならって初期化する [19]．重みベクトル $\boldsymbol{\theta}$ の調整と出力値 $\boldsymbol{z}_r(X, \boldsymbol{v}, \boldsymbol{\theta}_m)$ の計算は Caffe 1.0 を用いて行う [20]．

5.2 関数近似手法を利用した方策オフ型 MC 法の適用

本研究で用いた強化学習法の大枠を Algorithm 1 に示す．8 行目において，R はエンド先攻から見た得点であり，符号 $\pm$ は s がエンド先攻の手番ならば $+$，後攻の手番ならば $-$ の意である．本研究ではデジタルカーリングの 1 つのエンドを有限 MDP の 1 つのエピソードと見なして扱う．状態の集合 $\mathcal{S}$ は，ショット番号 m とストーン配置の対 $s = (m, X)$ により表される非終端状態すべてと，終端状態すべてからなる集合とする．なお，本論文では以降，非終端状態 s の対 (s, a) を組 (m, X, a) と書いたり，組 $(m, X, \boldsymbol{v}, r)$ と書いたりする．デジタルカーリングでは対

Algorithm 1

1: 学習データメモリ $\mathcal{B} \leftarrow$ 空集合
2: $\pi \leftarrow$ ある決定論的推定方策
3: **loop**
4: $\quad \pi' \leftarrow$ ある挙動方策
5: $\quad \pi'$ でエンド $(s_0 a_0 \ldots s_{15} a_{15} R)$ をプレイ
6: $\quad \tau \leftarrow a_\tau \neq \pi(s_\tau)$ を満たす最も大きい τ，なければ 0
7: $\quad$ **for each** 時間 τ から N_{T} に出現した (s, a) **do**
8: $\quad\quad \mathcal{B}$ に $(s, a, \pm R)$ を追加
9: $\quad$ **if** $|\mathcal{B}| \geq N_{\mathrm{th}}$ **then**
10: $\quad\quad \boldsymbol{\theta} \leftarrow$ ある重み
11: $\quad\quad L_{\mathcal{B}} \leftarrow \sum_{(s,a,R)\in\mathcal{B}} L(s, a, R, \boldsymbol{\theta})$
12: $\quad\quad L_{\mathcal{B}}$ をある程度小さくするように $\boldsymbol{\theta}$ を調整
13: $\quad\quad \mathcal{B} \leftarrow$ 空集合
14: $\quad\quad \pi \leftarrow$ MC 法で計算される $\arg\max_a Q(\cdot, a, \boldsymbol{\theta})$

(s, a) から定常的な確率 $P^a_{ss'}$ に従い次の時間ステップの状態 s' が生成されると仮定する．非終端状態 s はショット番号 m をあらわに含むため，1 つのエンドのプレイで 2 回以上同じ状態が現れることはない．

12 行目において，推定方策 π の評価精度を高めるために，$\boldsymbol{\theta}$ を調整して，損失関数 $L_{\mathcal{B}}$ の値を小さくする．方策が改善される様子の分析が容易になることを期待して，CNN の重みの調整は毎回独立に行う．すなわち，重みベクトル $\boldsymbol{\theta}$ を調整するときには毎回 $\boldsymbol{\theta}$ を初期化し（10 行目），メモリ $\mathcal{B}$ を空にする（13 行目）．

14 行目において，方策 π が改善される．行動価値 $Q(s, a, \boldsymbol{\theta})$ は期待エンド得点 $\overline{\boldsymbol{z}}_r(t, X, \boldsymbol{v})$ により推定される．最大値を与える行動 $a \in \mathcal{A}$ を求めるときには，これを正確に求めるのは困難なため，N_{gd} 点の初速度を一様ランダムに生成して，そのなかから行動価値の推定値が最も大きい行動を求める．このように近似計算されたグリーディ方策を推定方策と見なす．

6. 実験設定

本章では実験設定について述べる．6.1 節では，強化学習実験の設定を述べる．6.2 節で，性能評価に関する対戦実験の設定を述べる．

6.1 強化学習

Algorithm 1 の 14 行目の方策改善を 2 回行う．1 回目で用いる推定方策 π と挙動方策 π' は，それぞれ π_0，π'_0 と書く．これにより得られたグリーディ方策は π_1 と書く．同様に，2 回目は，それぞれ π_1，π'_{1h} であり，これにより得られた方策は π_2 である．

方策改善 1 回目の推定方策 π_0 は一様ランダムに値が初期化された決定論的方策とする．以降，これをランダム方策と呼ぶ．挙動方策 $\pi'_0(m, X, a)$ は，$m = 0$ ならば一様ランダムな行動 $a \in \mathcal{A}$ を生成し，$m > 0$ ならば推定方策と同じ行動 a に確率 1 を与える．N_{th} は 2.6×10^8，N_{T} は 15 とする．また，簡単のために行動 $\pi_0(m = 0, X)$ が挙動方策の行動と

同じになることは実験をとおしてないと仮定して，τ はつねに 0 としてプログラムを実装する．重み $\boldsymbol{\theta}$ の調整は，学習率 10^{-3} の慣性項付き SGD で行う．ミニバッチの処理回数は，各ショット番号 m に対してそれぞれ 50 万回とする．

方策改善 2 回目では，挙動方策 $\pi'_{1h}(m, X, a)$ は，$m < h$ ならば一様ランダムに初速度を 16 点生成した中から行動価値の推定値が最も高い行動 a，$m = h$ ならば一様ランダムな行動 $a \in \mathcal{A}$，$m > h$ ならば行動 $a = \pi_1(m, X)$ を生成する．h には 0 から 15 までの整数をランダムに与える．N_{th} は 3.2×10^8（各 h あたり約 2,000 万），N_{T} は h とする．重みの調整は，学習率 10^{-5} の Adam を用いて，各ショット番号 m に対しそれぞれミニバッチを 1,000 万回処理し，さらに，学習率 10^{-6} にして，それぞれさらにミニバッチを 1,000 万回程度処理して行う．また，CNN が扱う行動集合 $\mathcal{A}$ の大きさをおおよそ半分にするため，センタラインでストーン配置 X を反転させるテクニックを用いて，0 未満の初速度 v_x の値を CNN に入力しない．なお，実験をとおして，$N_{\mathrm{out}} = 15$，$N_{\mathrm{gd}} = 4{,}096$ とする．

実験は Xeon X5690 × 2 相当のワークステーションを 3〜9 台，Geforce GTX 1080 相当 1〜4 枚を 3 カ月程度占有して行った．台数や枚数が一定でないのはその時々で空いているものを用いたためである．

本実験方法・設定は，最も効率が良い方法とは考えにくいが，これは著者らが予備実験を行った範囲においては良好な効率を示したものである．11 チャネルを入力する CNN は，各プレイヤのストーン 2 チャネルのみを入力するものよりも価値推定精度が高かった．また，CNN の推定精度は，同程度の重み数とバッチ処理回数からなる多層全結合 NN の精度よりも良かった．さらに，2 回目の方策改善では Adam の学習効率は，SGD の効率よりも良かった．行動の初速度の各成分を区間 $[-1, 1]$ に収まるように線形変換するのも，そうしない場合と比較して，推定精度が良くなるためである．方策改善 2 回目においても N_{T} を 15 としなかったのは，メモリ $\mathcal{B}$ にグリーディ方策の行動ばかりが追加されて，これに CNN が適合して他の行動に対する適合が悪くなる現象を回避するためである．

6.2 対戦実験による性能評価

グリーディ方策の性能を対戦実験により評価する．対戦相手にはデジタルカーリングのサンプルプレイヤとして配布されている CurlingAI を用いる．このプレイヤは見本として実装されていて高度にチューニングされているとはいいにくいが，本研究で生成するグリーディ方策の性能を測るには十分な性能を持つ．

さらに，様々な性能を持つ対戦相手を用意するため，確率 ϵ で一様ランダムに $\mathcal{A}$ の行動をとり，確率 $1 - \epsilon$ で CurlingAI と同様に行動する CurlingAI(ϵ) も用意する．対戦は 8 エンドからなるゲームで行い，第 1 エンドの先攻と後攻は順番に入れ替える．

性能評価の指標には Elo レーティングを用いる．これは，プレイヤ i がプレイヤ j に勝利する確率を

$$P_{ij} = \frac{1}{1 + 10^{(R_j - R_i)/400}}$$

と推定するような R_i をプレイヤ i の強さ（レート）と考える方法である．R_i の値は最尤推定により求める．

7. 実験結果

本章では，方策改善と方策評価の実験結果と（7.1 節），得られたグリーディ方策が行う行動の分析結果（7.2 節）を示す．なお，本章の表や期待エンド得点の誤差はすべて標準誤差を用いた 95%信頼区間で見積もった．

7.1 方策改善と方策評価

1 つのエンドのみからなるゲームの対戦実験の結果を述べる．対戦相手を π_0 とし，先攻・後攻交互に合計 2,000 エンドプレイして，方策 π_1 の性能を評価した．π_1 が得たエンド得点 R の標本平均は 2.92 ± 0.08 であった．π_0 の R の期待値は 0（理論値）であるから，π_0 から π_1 の方策の更新が，期待収益を大きくするという意味で改善になっていたと考えられる．同様に，π_2 が π_1 に対して得た R の標本平均は 2.94 ± 0.12 であったことから，π_1 から π_2 の方策更新も改善になっていたと考えられる．

表 4 に，8 エンドからなるゲームの対戦の勝敗をもとに推定されたレートを示す．この結果から，GPI に基づく 2 回の方策改善が有効なものであり，π_0 から π_1，π_1 から π_2 と方策改善を行うたびに性能が向上していったことが分かる．グリーディ方策 π_2 の性能は CurlingAI より劣っているが，2 回の方策改善でレートが約 2,000 も上昇したことから，さらなる方策改善によって得られるグリーディ方策のレートは CurlingAI を上回ることが期待される．ランダム方策 π_0 を改善した条件とグリーディ方策 π_1 を改善した条件を比較すると，後者の方がより規模の大きい計算機実験を要するものである．このような観測から，グリーディ

表 4 Elo レーティング
Table 4 Elo ratings.

プレイヤ	レート
π_0	0
CurlingAI($\epsilon = \frac{15}{16}$)	287
CurlingAI($\epsilon = \frac{14}{16}$)	450
CurlingAI($\epsilon = \frac{12}{16}$)	787
π_1	1170
CurlingAI($\epsilon = \frac{8}{16}$)	1402
CurlingAI($\epsilon = \frac{4}{16}$)	1652
CurlingAI($\epsilon = \frac{1}{16}$)	2029
π_2	2053
CurlingAI	2169

表 5 学習データメモリ $\mathcal{B}$ に追加した組 $(m, X, \boldsymbol{v}, r, R)$ が，$m = 15$ かつ $2R \notin [-n+1), n-1]$ であった割合（%）．括弧内の値は誤差を表す

Table 5 Probabilities that a tuple $(m, X, \boldsymbol{v}, r, R)$ added to training data memory $\mathcal{B}$ satisfied $m = 15$ and $2R \notin [-n+1, n-1]$ in percentage. Values in parentheses represent errors.

n	方策改善 1 回目	方策改善 2 回目
17	0	0
15	$5(7) \times 10^{-6}$	$2.2(7) \times 10^{-4}$
13	$4(2) \times 10^{-5}$	$3.81(9) \times 10^{-2}$
11	$1.9(2) \times 10^{-3}$	$7.78(4) \times 10^{-1}$
9	$3.48(6) \times 10^{-2}$	4.956(10)
7	$4.33(2) \times 10^{-1}$	14.79(2)

方策 π_2 を改善するためには，さらに大規模な実験をセットアップしたり，より効率の良い強化学習法を適用したりすることが求められるのではないかと考えられる．

表 5 に，学習データのエンド得点が区間 $[(-N_{\text{out}}+1)/2, (N_{\text{out}}-1)/2]$ から外れた確率を示す．表には $m = 15$ の組の得点分布のみを示したが，他の組もおおむね同様であった．学習データの得点分布は方策改善 1 回目よりも 2 回目の方が広いことが分かった．本実験で生成した各 h につき 2,000 万程度の学習データではデータ点が少なく，8 点と -8 点のショットの学習は困難である．N_{out} の値は，大きいほど得点期待値を高精度に推定可能になるが，学習に必要な十分なデータ数が得られないため 15 で十分であることが分かった．

表 6 に方策改善 2 回目の期待エンド得点の推定精度を示す．学習データと同様に，テストデータの組 $(m, X, \boldsymbol{v}, r, R)$ も先攻後攻両方が π'_{1h} でエンドをプレイして生成した．データサイズは各 h ごとに 10 万とした．期待エンド得点 $\boldsymbol{z}_r(X, \boldsymbol{v}, \boldsymbol{\theta}_m)$ の重み $\boldsymbol{\theta}_m$ は，方策改善 2 回目が終わったときのものを用いた．推定精度は，決定係数と，相関係数を用いて比較した[*8]．各 m ごとに，テストデータの組 $(X, \boldsymbol{v}, r)$ すべてにわたる得点 R の標本平均を推定値とした場合，決定係数は 0 となる．また，推定精度が理想的な場合（推定値すべてが実測値と一致），決定係数と相関係数は 1 となる．相関係数は，予測を一様ランダムに行う場合 0 となる．さらに，ショット前のストーン配置で計算した得点をエンド得点として予測する手法の性能とも比較した．表より，m が大きくなるにつれて，おおむね推定精度も向上していく傾向がみられた．また，いずれの m についても CNN による推定精度が比較手法より高いことが分かった．なお，$m = 15$ において比較手法の推定精度がかなり高かった．ショット番号 m の行動が一様ランダムに生成されるために，ハウス内のストーンの配置に影響を与えないショットが多いことがその一因だと考えられる．

[*8] 決定係数の定義には，文献 [21] の式 (1) を用いた．標本 n 点から計算した相関係数 x の誤差は，分散を $(1-x^2)/\sqrt{n}$ により近似的に計算して見積もった [22].

表 6 CNN と比較手法が方策 π_1 を評価する精度．精度は推定期待得点 $\boldsymbol{z}_r$ と実測得点 R の決定係数と相関係数（大きいほど高精度）で測った．括弧内の値は誤差を表す．m はショット番号を表す

Table 6 Accuracies of policy π_1 evaluations by means of CNN and the comparative method. Accuracies were measured by the coefficient of determination and correlation coefficient (greater value means higher accuracy) between an inferred expected score $\boldsymbol{z}_r$ and actual score R. Values in parentheses represent errors. m represents a shot number.

m	決定係数		相関係数	
	CNN	比較手法	CNN	比較手法
0	0.00	−0.02	0.048(7)	-
1	0.05	0.00	0.224(6)	0.061(7)
2	0.07	−0.16	0.270(6)	0.163(7)
3	0.18	0.04	0.428(6)	0.280(6)
4	0.14	−0.13	0.371(6)	0.266(6)
5	0.19	−0.04	0.436(6)	0.334(6)
6	0.24	−0.05	0.492(5)	0.389(6)
7	0.31	0.12	0.562(5)	0.466(5)
8	0.32	0.04	0.570(5)	0.476(5)
9	0.38	0.18	0.617(4)	0.539(5)
10	0.43	0.21	0.657(4)	0.585(5)
11	0.55	0.39	0.740(3)	0.685(4)
12	0.58	0.40	0.760(3)	0.705(4)
13	0.69	0.59	0.831(2)	0.793(3)
14	0.81	0.72	0.898(2)	0.866(2)
15	0.94	0.91	0.9717(4)	0.9574(6)

7.2 獲得した行動知識

π_1 どうし（**表 7** 参照），π_2 どうし（**表 8** 参照）それぞれ 1 万エンド分の対戦記録から，ショットを分類および集計して傾向を調査した．各列が表す各項目とショットの分類法は以下のとおりである．なお，本論文で用いたショット分類の方法は，主にショットしたストーンの座標に関する条件からなる，簡易なものである．

ドロー ドロー（プレイエリアにストーンを止めるショット）の割合．左から，ハウスへのドロー[*9]，ガード[*10]，カムアラウンド[*11]，フリーズ[*12]を表す．

TO1 ショットしたストーンがプレイエリア内に残り，かつストーン 1 つを弾き出すテイクアウト[*13]の割合．

[*9] ショットしたストーンがハウス内にあり，かつ，他のストーンがプレイエリア外へ移動しない．

[*10] ショットしたストーンがフリーガードゾーン（ティーラインに達しないハウス外のプレイエリア）にとどまった．

[*11] ショットしたストーンの手前に x 座標の差がストーン半径 (r_{stone}) 未満のストーンがある．

[*12] ショットしたストーンの奥に x 座標の差が r_{stone} 未満かつ距離が $3r_{\text{stone}}$ 以内のストーンがある．

[*13] ショットしたストーンではないストーンがプレイエリア外へ移動する．

表 7 π_1 のショット傾向．誤差は 1.3 未満である．
Table 7 Shot tendencies of π_1. Errors are less than 1.3.

m	ドロー	TO1	TO2	その他
0	100/0/-/-	-/-	-/-/-	0/100
1	98/0/0/0	-/2	-/-/-	0/84
2	99/0/0/4	0/0	-/0/-	0/49
3	76/0/2/2	8/3	-/0/0	10/16
4	72/1/1/3	3/3	0/1/0	16/16
5	26/3/1/1	5/3	0/0/0	50/3
6	13/1/2/1	9/13	1/3/1	47/18
7	40/5/2/2	9/13	0/1/1	25/14
8	34/15/2/2	7/6	0/1/0	30/15
9	41/9/2/2	6/10	0/0/0	30/13
10	20/10/2/1	9/10	1/2/1	38/14
11	9/6/3/1	8/12	1/3/2	50/12
12	14/15/2/1	8/9	1/2/1	45/12
13	26/13/3/1	7/14	0/1/1	32/15
14	18/16/2/1	5/7	0/1/0	49/14
15	29/12/3/2	6/7	1/2/1	39/18

表 8 π_2 のショット傾向．誤差は 1.7 未満である．
Table 8 Shot tendencies of π_2. Errors are less than 1.7.

m	ドロー	TO1	TO2	その他
0	100/0/-/-	-/-	-/-/-	0/100
1	56/0/0/9	-/44	-/-/-	0/77
2	74/0/0/8	6/18	-/0/-	0/60
3	26/0/0/2	15/52	-/2/1	1/72
4	78/0/1/4	4/14	0/1/0	1/55
5	14/1/2/2	14/55	0/3/2	5/61
6	86/1/1/4	1/8	0/0/0	2/51
7	18/1/3/3	15/50	1/5/3	2/56
8	84/1/2/5	2/9	0/0/0	2/51
9	11/1/4/2	17/42	2/9/4	5/53
10	79/1/2/4	3/10	0/1/1	3/60
11	18/1/5/3	16/46	1/5/3	4/55
12	70/4/4/4	5/14	0/1/1	3/55
13	31/2/6/3	13/35	1/6/3	3/59
14	73/3/3/4	4/14	0/1/1	3/69
15	42/1/5/4	12/33	1/5/3	1/74

ヒットアンドステイとヒットアンドロールはこの条件を満たす．左から，手番プレイヤのストーン 1 つ，相手のストーン 1 つのテイクアウトを表す．

TO2 ショットしたストーンがプレイエリア内に残り，かつストーン 2 つを弾き出すテイクアウトの割合．ダブルテイクアウトはこの条件を満たす．左から，手番プレイヤのストーン 2 つ，手番プレイヤと相手のストーン 1 つずつ，相手のストーン 2 つのテイクアウトを表す．

その他 左から，スルーショット[*14]の割合，手番プレイヤがナンバーワンストーン（ティーに最も近いハウス内のストーン）を持たないという条件下において，ショットによってナンバーワンストーンを得た割合を表す．

まず，カーリングにおいて最も初歩的な行動の 1 つであるハウスへのドローを行う割合を考察する．ここで，表には示さないが，π_0 のそれは 10%以下である．π_1 の $m = 0$ のショットはすべてハウスへのドローであった．また，すべてのショット番号でこの割合が π_0 より高かった．さらに，π_0 のナンバーワンストーンを得た割合は 10%以下であることに対し，π_1 のそれはおおむね 12%以上であった．したがって，1 回目の方策改善によってグリーディ方策が初歩的なドローを行うようになったことが分かった．

次に，ストーン 1 つのテイクアウト（TO1）を行う割合を考察する．ここでは，手番プレイヤのストーンのテイクアウトと相手ストーンのテイクアウトの 2 種の割合を比較する[*15]．π_0 ではどちらも 5%以下である．π_1 では π_0 と比較してこれらの割合が増加したものの，2 種の割合に大きな違いはみられなかった．また，π_2 でも π_1 と比較してこれらの割合が増加していた．さらに，π_2 ではこれら 2 種の割合に有意な差がみられ，相手ストーンを選択的にテイクアウトしていたことが分かった．さらに，π_2 では先攻（m が偶数）はハウスへのドローを好み，後攻はテイクアウトを好むことが分かった．このことは，不利な先攻が攻撃的なドローを行い，有利な後攻が守備的なテイクアウトを行うようになったとして理解することができる[*16] [13]．さらに，π_1 のナンバーワンストーンを得た割合よりも，π_2 のそれはおおむね高かった．これらのことから，2 回目の方策改善によってグリーディ方策が初歩的なテイクアウトを行うようになったことが分かった．

最後に，上述したショットよりも高度であると考えられるショットを π_2 が行う割合を考察する．ガード，カムアラウンド，フリーズを行う割合は小さいため，これらの行動の分析は困難であった．ダブルテイクアウトを行う割合もまた小さく，相手のストーン 2 つを選択的にテイクアウトする様子も確認できなかった．スルーを行う割合も小さかった．表には示されていないが，ピールのように，ショットしたストーンがプレイエリア外に移動するテイクアウトの割合はすべての m で 4%未満であり小さかった．これらのショットを行う割合は小さく，これらの分析は困難であり，高度なショットの学習は確認できなかった．π_0，π_1 がナンバーワンストーンを得る確率は $m = 3$ 以降は 20%以下と小さく，これらの方策の改善は，ナンバーワンストーンを容易に獲得するハウスへのドローや相手のストーン 1 つのテイクアウトを行うことで十分達成可能であったこと

[*14] ショットしたストーンがプレイエリアになく，かつ，他のストーンすべての座標に変化がない．

[*15] 一般に，手番プレイヤのストーンのみをテイクアウトすることが有効な状況は稀である．

[*16] 1 エンドのみのゲームプレイは，複数エンドからなるゲームの同点で迎えた最終エンドのプレイに近いと考えられる．このようなエンドでは基本的に，不利な先攻は攻撃的な試合運びを狙い，有利な後攻は守備的な試合運びを狙う．

が伺える．

8. まとめ

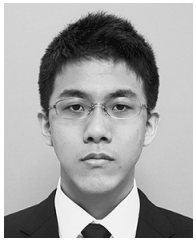
デジタルカーリングを題材に，ランダム方策から開始するGPIを行う強化学習法を検討した．行動集合$\mathcal{A}$にはおおよそカーリングの予備知識を用いないものを仮定して，この巨大な行動集合のグリーディ方策をMC法により近似的に計算した．

本研究の実験により，CNNの約70万×16個の重みの値を約2,000万×16本のエピソードを用いて調整して，サンプルプログラムとして公開されているCuringAIよりもやや弱いグリーディ方策が得られることを明らかにした．1回目の方策改善ではグリーディ方策は主に，初歩的なドローであるハウスへのドローの知識を獲得した．そして，2回目の方策改善では初歩的なテイクアウトであるストーン1つのテイクアウトの知識を獲得し，先攻ならばより多くのハウスへのドローを，後攻ならばテイクアウトを行うようになった．2回の方策改善で強さが順調に向上したことから，より大規模な実験をセットアップしたり，より効率の良い強化学習法を適用したりすることにより，ガードやダブルテイクアウトなどのより高度なショット知識を獲得して，グリーディ方策はさらに強くなるのではないかとの感触を得た．

謝辞 本研究はJSPS科研費JP16K00503，JP18H03347の助成を受けたものです．

参考文献

[1] Coleman, G.: *Introduction to Curling Strategy Black & White Edition* (2014).
[2] 北清勇磨，伊藤毅志：デジタルカーリングシステムの提案と構築，第9回E&Cシンポジウム，pp.13–16 (2015).
[3] 岡田 彰：ゲーム理論 新版，有斐閣 (2011).
[4] Stuart, R. and Peter, N.: *Artificial Intelligence: A Modern Approach 3rd Edition*, Pearson (2010).
[5] 加藤 修，飯塚博幸，山本雅人：不確定性を含むデジタルカーリングにおけるゲーム木探索，情報処理学会論文誌，Vol.57, No.11, pp.2354–2364 (2016).
[6] 大渡勝己，田中哲朗：カーリングAIに対するモンテカルロ木探索の適用，ゲームプログラミングワークショップ2016論文集，Vol.2016, pp.180–187 (2016).
[7] 加藤 修，加藤博幸，山本雅人：デジタルカーリングにおける局面評価関数の学習，第21回知能メカトロニクスワークショップ講演論文集，pp.215–217 (2016).
[8] Tesauro, G.: TD-Gammon, a Self-teaching Backgammon Program, Achieves Master-level Play, *Neural Comput.*, Vol.6, No.2, pp.215–219 (1994).
[9] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D.: Human-level control through deep reinforcement learning, *Nature*, Vol.518, pp.529–533 (2015).
[10] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D.: Masterinng the game of Go without human knowledge, *Nature*, Vol.550, pp.354–359 (2017).
[11] 松井亮平，保木邦仁：畳込みニューラルネットワークを用いたデジタルカーリングAI作成の試み，技術報告12 (2017).
[12] Sutton, R.S. and Barto, A.G.: 強化学習, 森北出版 (2000).
[13] 小川豊和：公益社団法人 日本カーリング協会 オフィシャルブック 新みんなのカーリング，学研教育出版 (2014).
[14] Guo, X., Singh, S., Lee, H., Lewis, R.L. and Wang, X.: Deep Learning for Real-Time Atari Game Play Using Offline Monte-Carlo Tree Search Planning, *Advances in Neural Information Processing Systems 27*, Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D. and Weinberger, K.Q. (Eds.), pp.3338–3346, Curran Associates, Inc. (2014).
[15] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol.529, pp.484–489 (2016).
[16] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K. and Hassabis, D.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, ArXiv e-prints (2017).
[17] 岡谷貴之：深層学習，講談社 (2015).
[18] Kingma, D.P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol.abs/1412.6980 (2014).
[19] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proc. 13th International Conference on Artificial Intelligence and Statistics*, Teh, Y.W. and Titterington, M. (Eds.), *Proc. Machine Learning Research*, Vol.9, pp.249–256, PMLR (2010).
[20] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding, arXiv preprint arXiv:1408.5093 (2014).
[21] Kvalseth, T.O.: Cautionary Note about R^2, *The American Statistician*, Vol.39, No.4, pp.279–285 (1985).
[22] Bowley, A.L.: The Standard Deviation of the Correlation Coefficient, *Journal of the American Statistical Association*, Vol.23, No.161, pp.31–34 (1928).



松井 亮平 （学生会員）

1994年生．2017年電気通信大学情報理工学部情報・通信工学科卒業．同年同大学大学院情報理工学研究科情報・ネットワーク工学専攻博士前期課程入学．

保木 邦仁 （正会員）

1998年東北大学理学部卒業以降，化学領域での研究活動に従事．2000年東北大学大学院理学研究科博士前期課程修了．2003年同研究科博士後期課程修了．2003～2006年トロント大学博士研究員．2006年東北大学大学院理学研究科研究支援者．2007～2009年同研究科助手．2009年東北大学高等教育開発推進センターへ移動．2010年電気通信大学先端領域教育研究センター特任助教，2015年より電気通信大学大学院情報理工学研究科准教授，現在に至る．