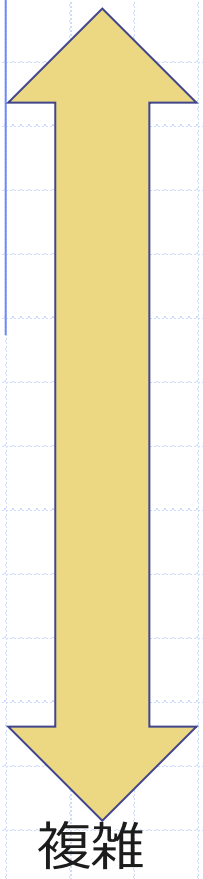
HOKKAIDO UNIVERSITY

講義「人工知能」
第9回 強化学習3
自己学習で強くなる TD Gammon

北海道大学大学院情報科学研究院
情報理工学部門 複合情報工学分野
調和系工学研究室 准教授 山下倫央
http://harmo-lab.jp
tomohisa@ist.hokudai.ac.jp
2024年5月7日(火)

ゲーム理論の位置付け



- 戦略形ゲーム
  - 1回or繰り返し：囚人のジレンマ
- 展開形ゲーム
  - 多段階：ゲーム木の探索(ミニマックス法、αβ法)
- 展開形ゲーム(バックギャモン)
  - 多段階＋確率事象：TD(λ)
- TVゲーム(Atari2600)
  - Deep Q-Network

ゲームの分類

- 非協力ゲーム…拘束力のある合意なし
  - 戦略形ゲーム…一度きりの行動決定を同時におこなう
    - ゼロ和ゲーム…全員の利得の総和がゼロ　e.x.) ジャンケン
    - 非ゼロ和ゲーム…全員の利得の総和が非ゼロ
  - 展開形ゲーム…行動決定を時間の経過とともにおこなう

    e.x.) 将棋、チェス
  - 繰り返しゲーム…同じゲームを繰り返しおこなう
    - 無限繰り返しゲーム…戦略形ゲームを無限回繰り返す
    - 有限繰り返しゲーム…戦略形ゲームを有限回繰り返す
- 協力ゲーム…拘束力のある合意あり

  e.x.) 費用分配、遺産分配、投票力分析

展開形ゲームの木表現

- ゲーム木の始点：根（root）
  - 現在の局面が根に対応
- ゲーム木の途中途中の分岐点：ノード（node）
- 子ノードがないノード：葉（leaf）
- あるノード以下の分岐：枝（branch）

囚人のジレンマの利得行列(双行列表現)

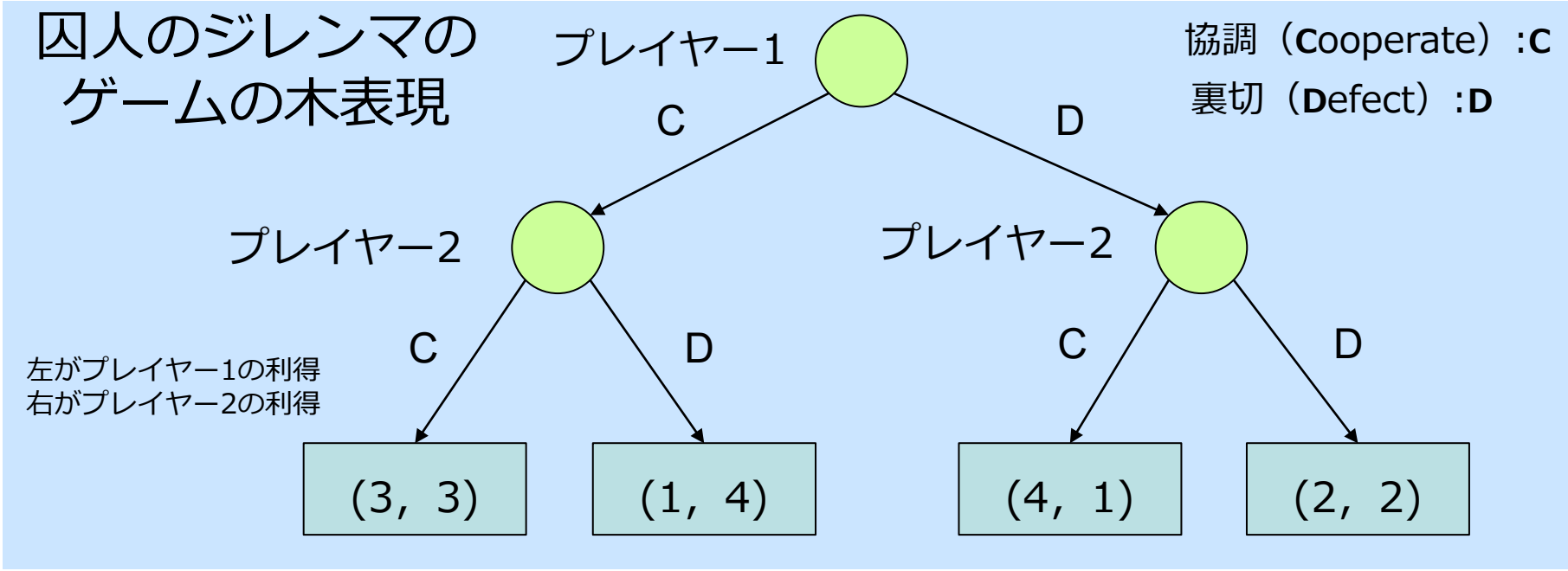
		プレイヤー2 C	プレイヤー2 D
プレイヤー1	協調（Cooperate）:C	3,3	1,4
プレイヤー1	裏切（Defect）:D	4,1	2,2

（協調＝黙秘、裏切＝自白）

左がプレイヤー1の利得
右がプレイヤー2の利得

展開形ゲームの木表現

- ゲーム木の始点：根（root）
  - 現在の局面が根に対応
- ゲーム木の途中途中の分岐点：ノード（node）
- 子ノードがないノード：葉（leaf）
- あるノード以下の分岐：枝（branch）



バックギャモン

バックギャモン (Backgammon) は基本的に二人で遊ぶボードゲームの一種で、盤上に配置された双方15個の駒をどちらが先に全てゴールさせることができるかを競う。

バックギャモンは世界最古のボードゲームとされるテーブルズ（英語版）の一種である。日本には奈良時代（飛鳥時代との説もある）に伝来し、平安時代より雙六・盤双六の名で流行したが、その後賭博の一種として朝廷に禁止され、一度廃れている。

サイコロを使用するため、運が結果に対する決定因子の一つであるものの、長期的には戦略がより重要な役割を果たす。プレイヤーはサイコロを振るたびに膨大な選択肢の中から、相手の次の可能性のある手を予測しながら手を選択し、自分の駒を移動させる。20世紀初頭のニューヨークを起源とする現代のバックギャモンは、ゲーム中に賭け金をレイズする（上げる）ことができる。

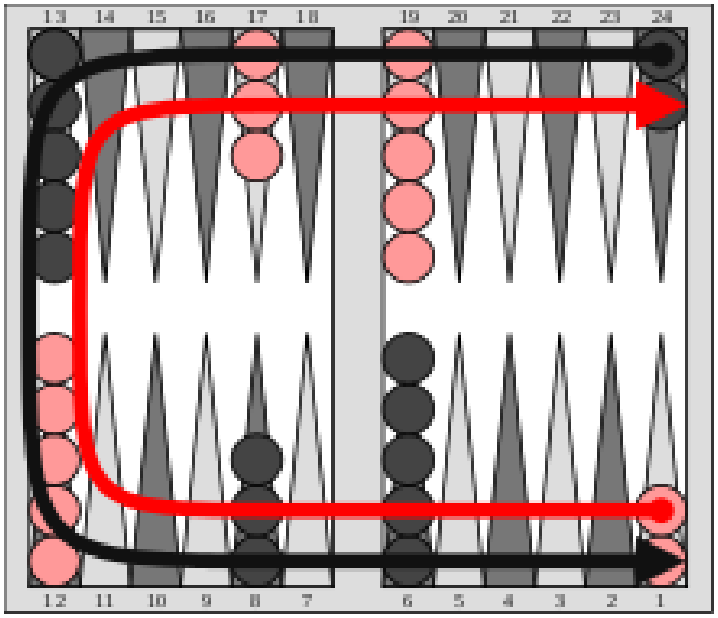
チェスと同様に、バックギャモンは計算機科学者の興味の対象として研究がなされてきた。この研究によって、バックギャモンソフトは人間の世界チャンピオンを破る程に発展している。

バックギャモンのボードと駒

Wikipediaバックギャモン
https://ja.wikipedia.org/wiki/%E3%83%90%E3%83%83%E3%82%AF%E3%82%AE%E3%83%A3%E3%83%A2%E3%83%B3

バックギャモンのルール1

- コマの配置
  - 赤と黒の丸いコマを受け持つ二人が対戦
  - コマをおくマスは図の水色と黄色の部分
  - お互い15個ずつコマを並べた状態から開始
- コマの移動
  - 2個のサイコロの目に応じて交互にコマを進めて，全部のコマをアガリまで先に動かした方が勝ち
    - 赤は反時計回り（24→1の方向）
    - 黒は時計回り（1→24の方向）



- その他のルール
  - ヒット
    - 相手の駒が一枚あるところに自分の駒を進めると、相手の駒はふりだしにもどる
    - ヒットされた駒はバー（中央のしきり）の上に置く
  - ブロック
    - 移動先のポイントまたは再配置しようとしたポイントに敵の駒が2つ以上存在する場合、そこには移動できない
- TDギャモン (TD-Gammon) https://www.ai-gakkai.or.jp/whatsai/AItopics4.html

バックギャモンのルール2

- 基本的なゲームポイント
  - ポイントとは、勝ち点のことである。バックギャモンのゲームのポイントはその勝ち方によって3通りに分かれる。

  1. 相手の駒がゴールし始めている状態で勝利した場合、勝者は1ポイントを獲得する。これをシングルという。
  2. 相手の駒が1つもゴールしていない状態で勝利した場合、勝者は2ポイントを獲得する。これをギャモンという。
  3. 上記の場合でさらに相手の駒がバーもしくは勝者側のインナーに残っていた場合、勝者は3ポイントを獲得する。これをバックギャモンと呼ぶ。

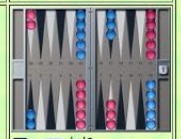
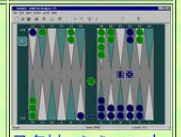
- ダブルがなされている場合には、ダブリングキューブが表示する倍率をこれに乗じたものとなる。
- 競技会ルールでは、5以上の奇数ポイントを統一して設定し、そのポイントを先取した者の勝利としてゲームを行うことが普通である。

バックギャモン

- https://ja.wikipedia.org/wiki/%E3%83%90%E3%83%83%E3%82%AF%E3%82%AE%E3%83%A3%E3%83%A2%E3%83%B3

バックギャモンのソフト

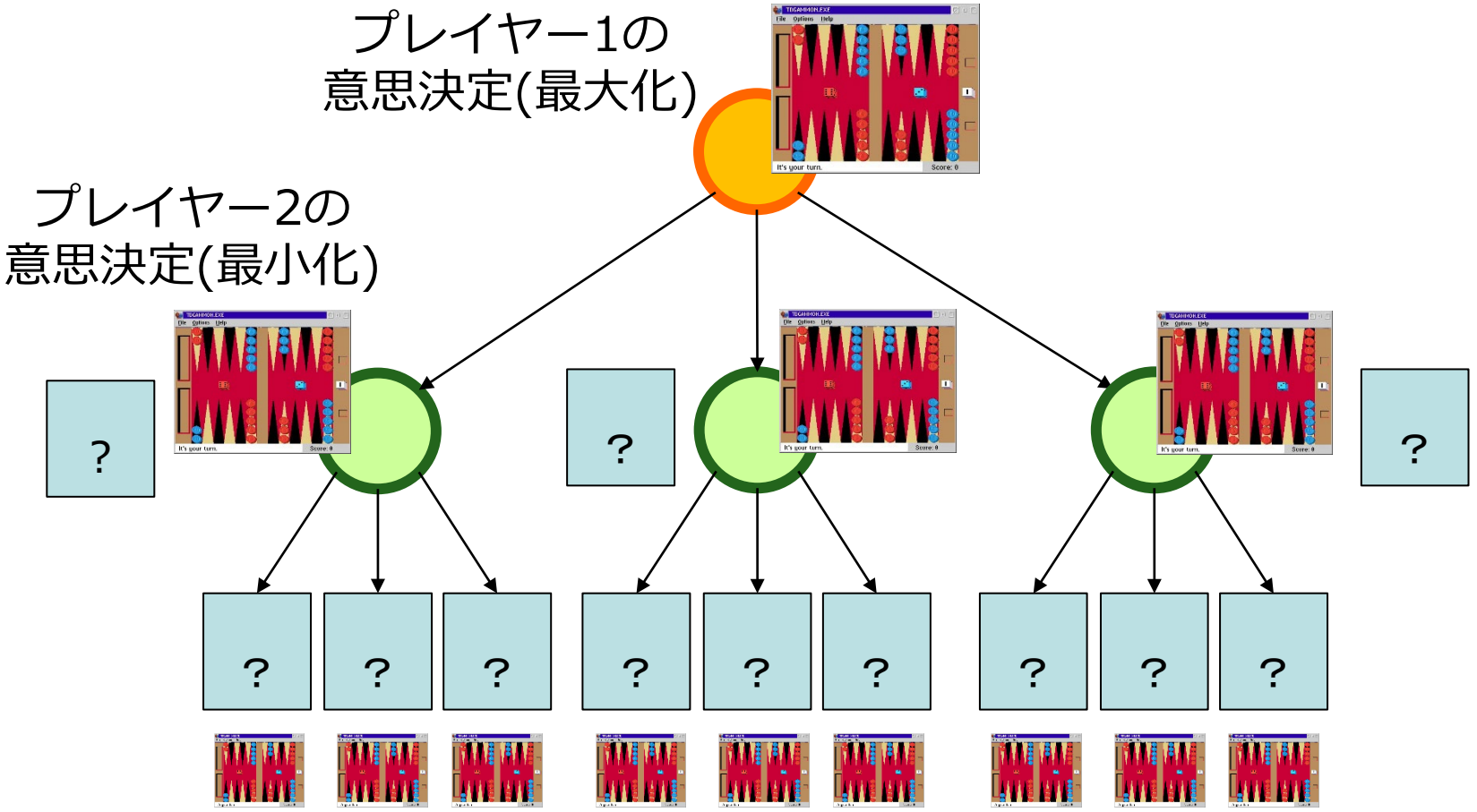
- バックギャモンソフト
  - http://seki.webmasters.gr.jp/gammon/software.html

ソフト名 (ホームページにリンク)	ライセンス	言語	プラットフォーム	補足説明	スクリーンショット
GNU Backgammon (gnubg)	フリーソフト (GPL)	日本語対応、多言語	Windows, Linux, Mac OS X 等。Mac OS X へのインストールは、Darwinportsを使うと良い。Leopard用のパッケージもある。	**最もおすすめ。** たれなり流「牛さんの飼い方」に、GNU Backgammon のインストールの手順、使い方が解説されている。 機能、操作性ともに申し分ないが、まだアルファ版であり、ソフトの安定性は完璧ではない。 もしも、動作が不安定に感じたら、他のソフトを試してみるといい。 ボードについては、2D, 3Dのボードを色々変えることができる。	 スクリーンショットより
BGBlitz	フリー版と有料版 ($23,$46)がある	日本語対応、多言語	Windows, Mac OS 9, Mac OS X, Linux 版がある。	**特にマックユーザーにおすすめ。** BGBlitzの使い方を書きました。 Mac OS X へのインストールは、GNU Backgammon よりも簡単なので、GNU Backgammon のインストールがうまく出来ない人はどうぞ。	 テーマより
JellyFish (ジェリフィッシュ; くらげ)	フリー版と有料版 ($35,$100,$220)がある	英語	Windows	スノーウィーが有名になる前は、「ギャモンソフトといえばジェリフィッシュ」という時代が長く続いた。	 スクリーンショットより
Snowie (スノーウィー)	有料($100,$380)	英語	Windows	**出費を厭わない人には、最もおすすめ。** プロレベルのプレーヤーから高く評価されている。 日本バックギャモン協会のShopから購入可能。	

  - App Store に多数あり
    - iPhone、iPod touch、iPad向け

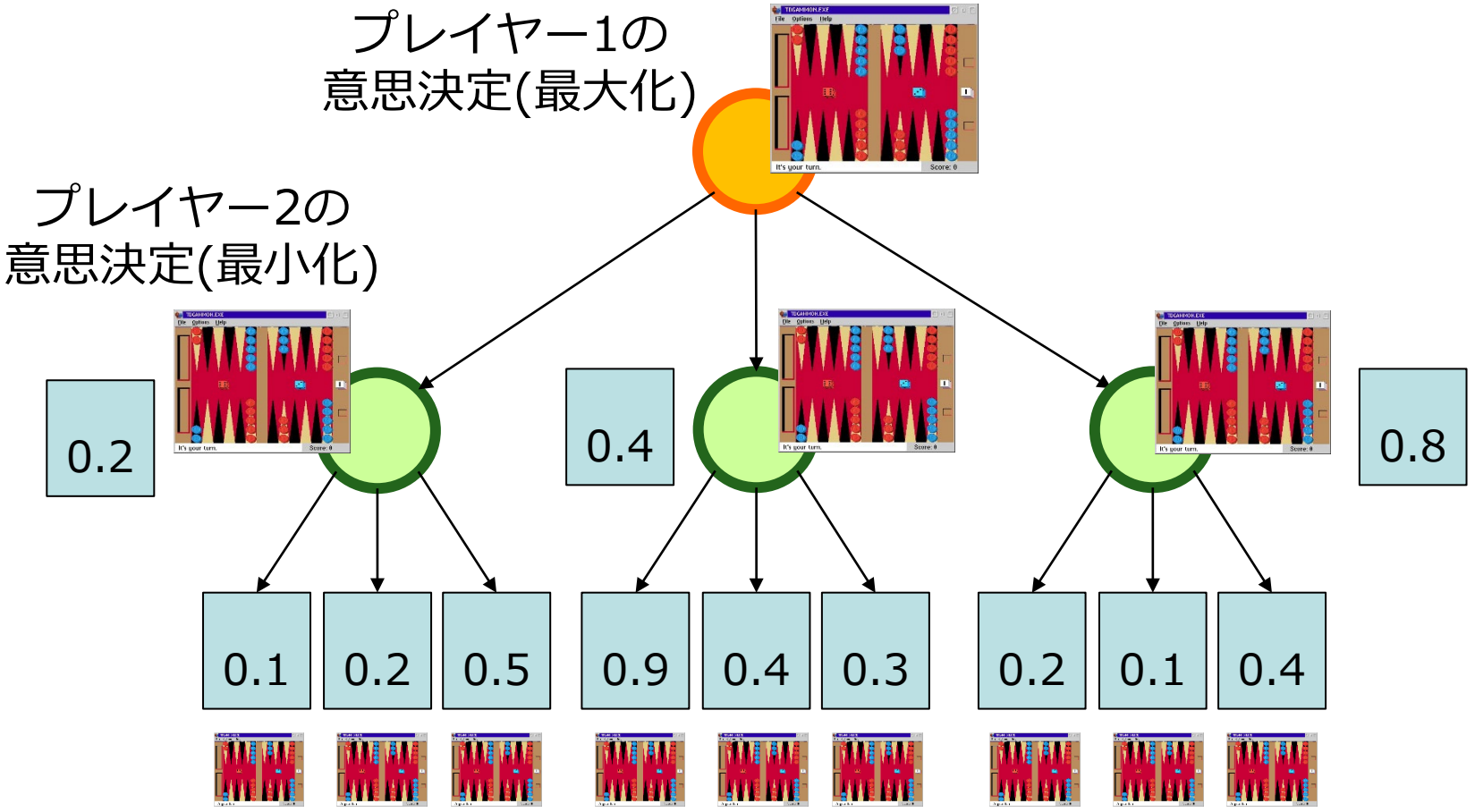
従来手法：評価関数の適用

- ゲームの局面の状態を静的に評価し数値に変換
  - 今後このゲームに勝利する確率などを評価値とする



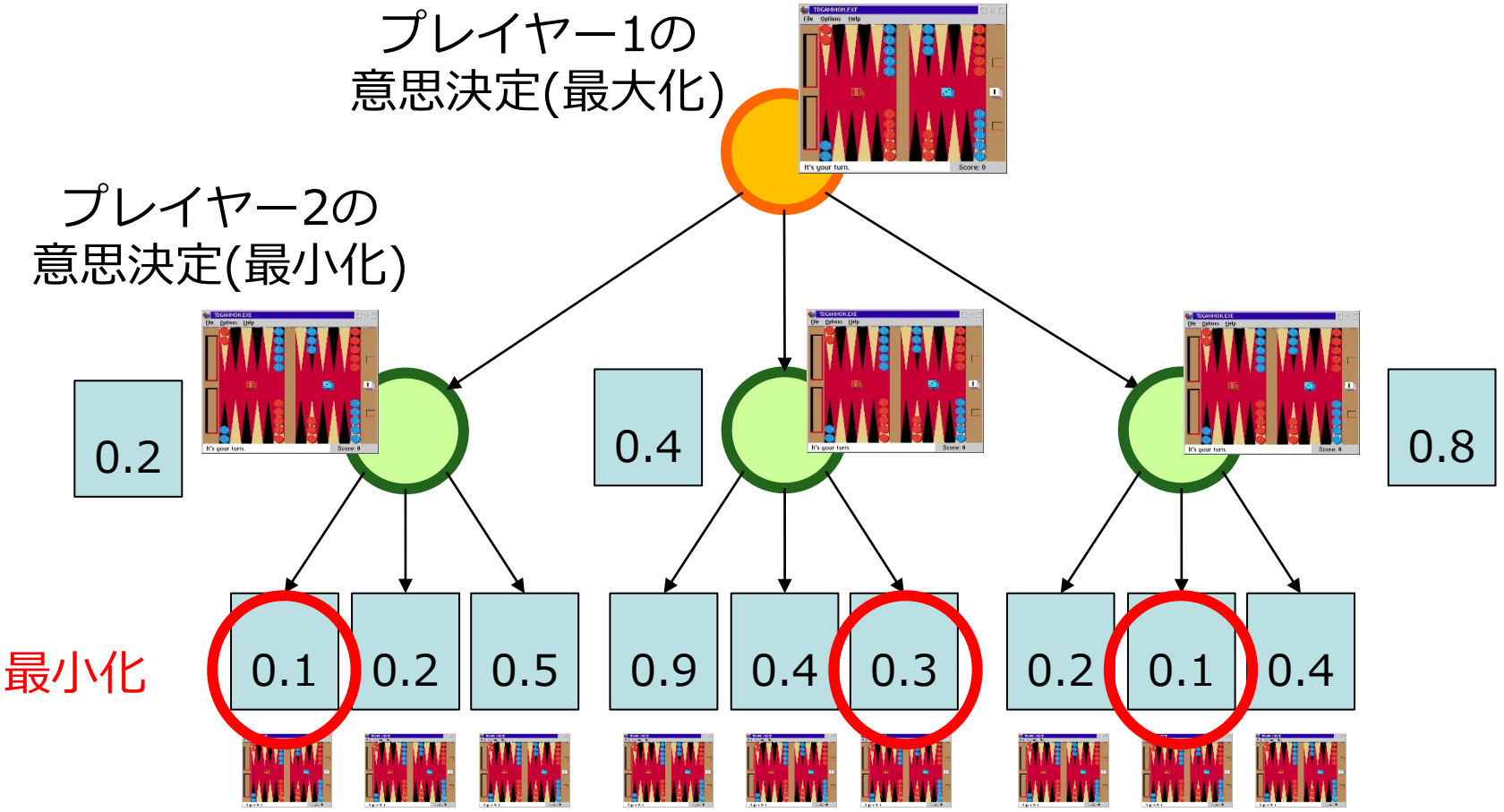
従来手法：評価関数の適用

- ゲームの局面の状態を静的に評価し数値に変換
  - 今後このゲームに勝利する確率などを評価値とする



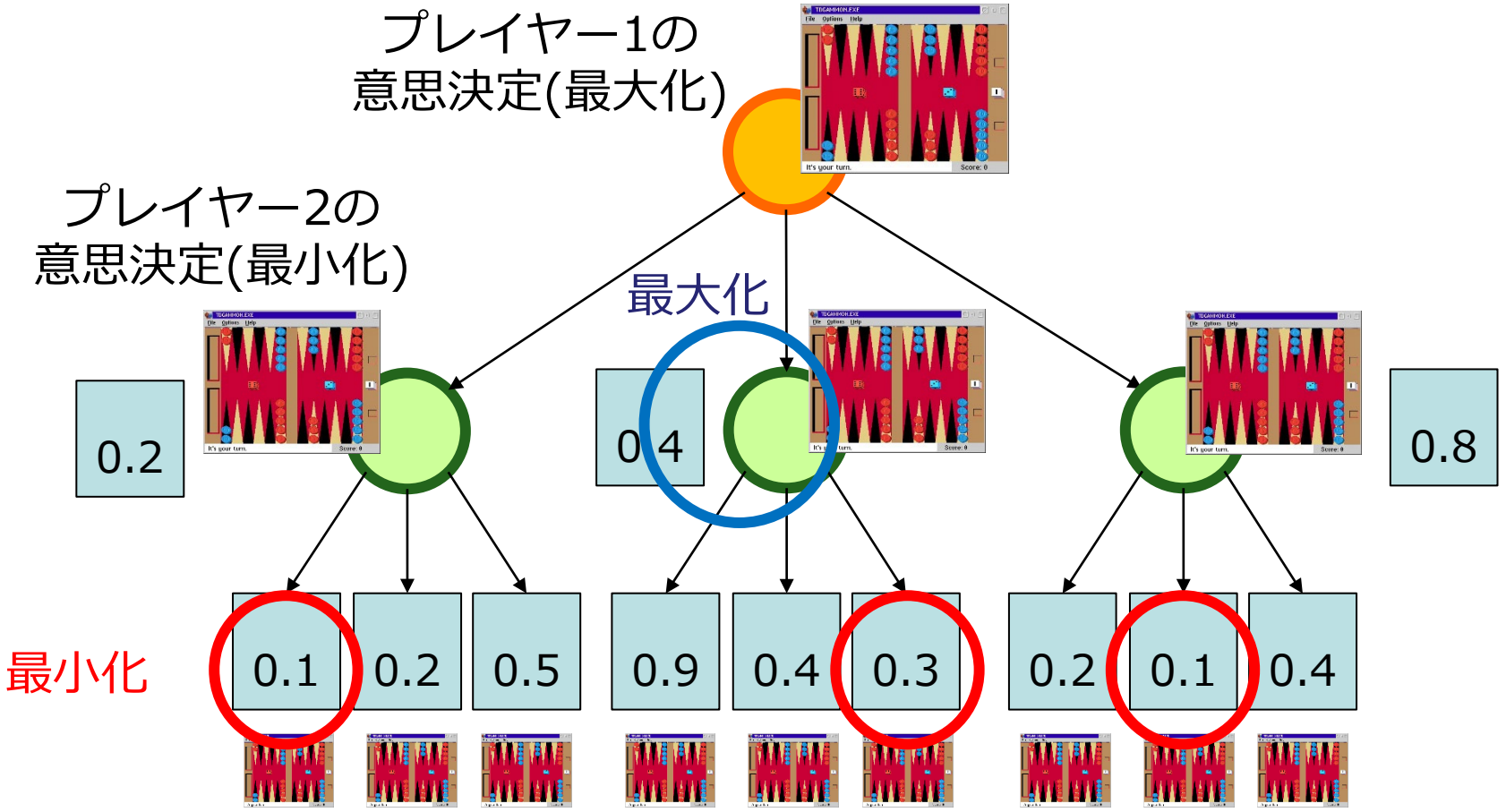
従来手法：評価関数の適用

- ゲームの局面の状態を静的に評価し数値に変換
  - 今後このゲームに勝利する確率などを評価値とする



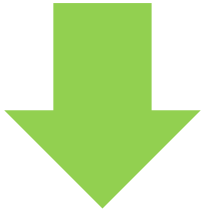
従来手法：評価関数の適用

- ゲームの局面の状態を静的に評価し数値に変換
  - 今後このゲームに勝利する確率などを評価値とする



評価関数の導入

- ゲームをプレイする人工知能を作成する際、評価関数が必要
- 以降互いに最善のプレーをすると仮定した際の勝率が理想＝状態価値
  - 理論的には機械学習が可能



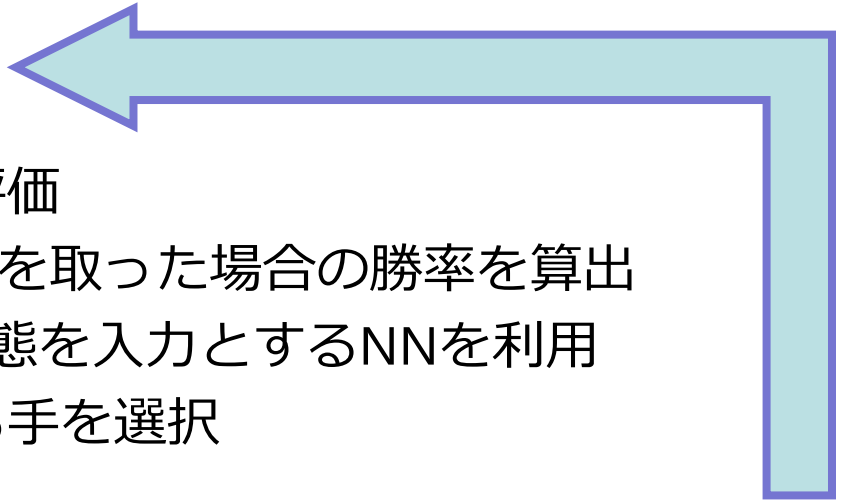
- バックギャモンの局面数は10^144
  - すべての状態を列挙して評価値を割り当てるのはメモリ、計算時間の観点からほぼ不可能

TDギャモンの概要

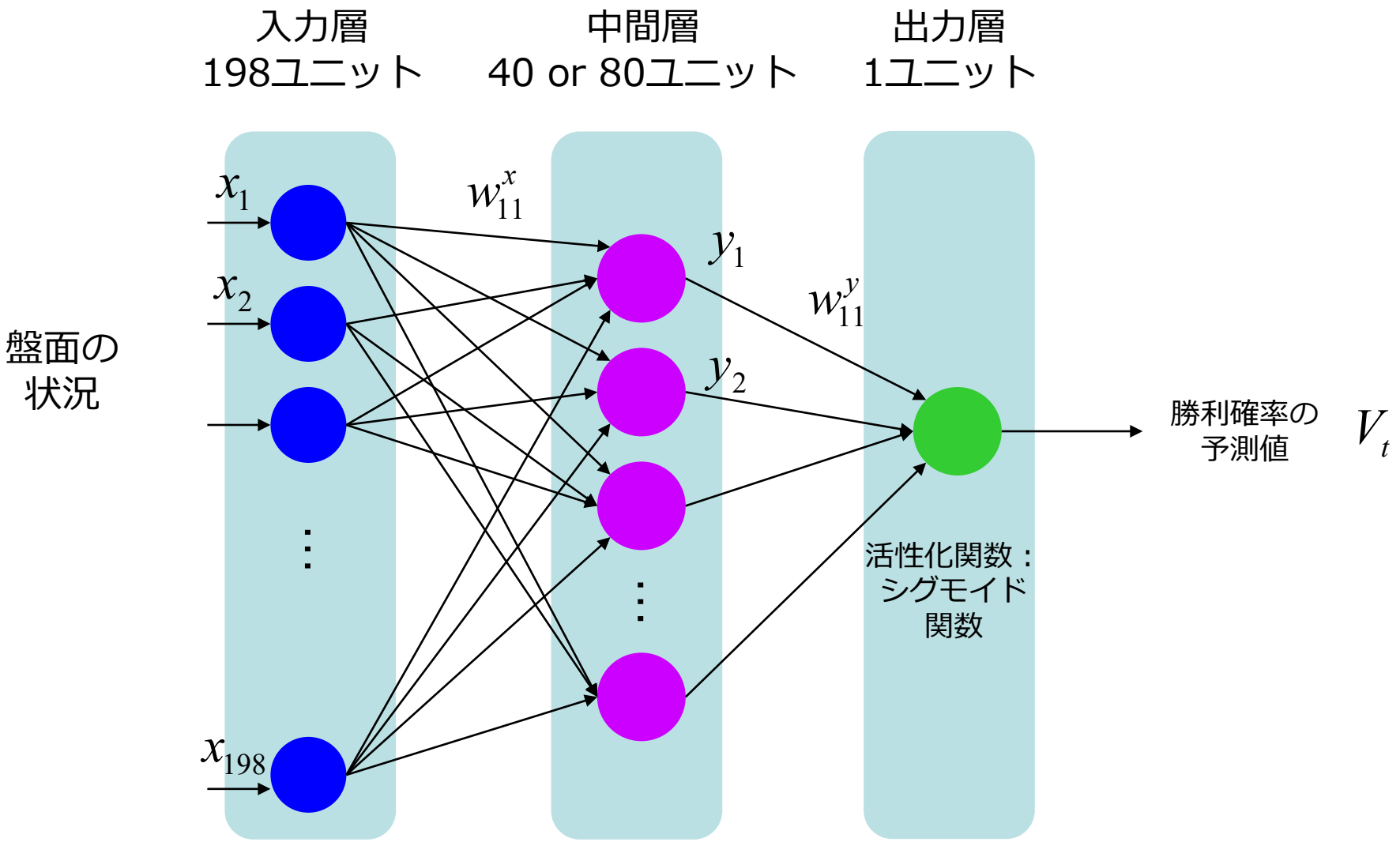
Tesauro, Gerald (March 1995). "Temporal Difference Learning and TD-Gammon". Communications of the ACM 38 (3)

- 自己学習の導入
  - NNプレイヤー同士での対戦
    - 両プレイヤーの行動選択をニューラルネットワークで実装

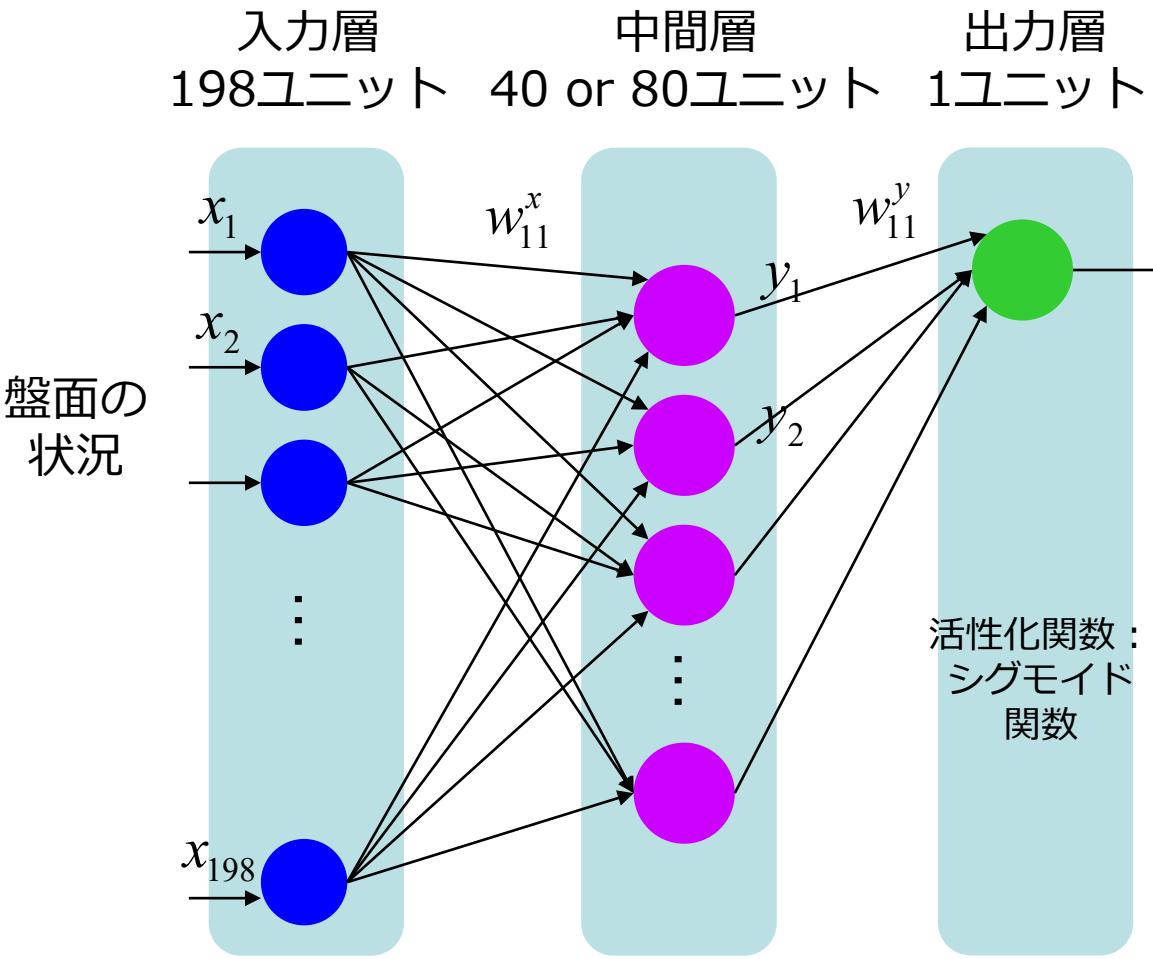
- 各エピソードにおける対戦
  - 自分の番の行動
    - (サイコロを振る)
    - 選択可能な手の評価
      - 選択可能な手を取った場合の勝率を算出
        - » 盤面の状態を入力とするNNを利用
    - 勝率が最大となる手を選択
    - NNの重みを更新
  - 相手も同様の行動を行い、再び自分の番が来る



勝利確率推定するニューラルネットワーク



勝利確率推定するニューラルネットワーク



勝利確率の予測値 V_t

誤差関数の設定

$$E = \frac{1}{2}(z - V_t)^2$$

z :ゲームの結果

- 1: 勝ち
- 0: 負け

誤差逆伝播法の適用
・重みの更新式

$$\Delta w_t = \alpha(V_{t+1} - V_t)\sum_{k=1}^{t}\lambda^{t-k}\frac{\partial V_k}{\partial w}$$

λ ：今回の報酬が次のステップで
どの程度影響を与えているかを示
すパラメータ $(0 \leq \lambda \leq 1)$

入力情報の構成

- 198入力の内訳

ユニット数	詳細

192	各ポイントの白黒それぞれに4ユニットを利用
	• 4（ユニット）× 2（白黒）× 24（ポイント数）
	• 例）ある1ポイントに対する白の数について
	  • 0個 : 0000
	  • 1個 : 1000 (最初の1ユニットが1)
	  • 2個 : 1100 (最初から2ユニットが1)
	  • 3個 : 1110 (最初から3ユニットが1)
	  • 4個以上 : 111(n-3)/2 （n: 駒の数)

2	バー上にある白と黒の駒数をコード化 • n/2 （n: バー上の駒の数）
2	盤面から除かれた白と黒の駒数 • n/15 （n: 取り除かれた駒の数）
2	白黒いずれの番

TDギャモンの使い方

- サイコロを振る
- 出た目に対して、ルールで許されている可能なコマの動かし方全てに対して勝率を試算
  - 勝利確率推定するニューラルネットワークの利用
- 勝算が最も大きいコマの動かし方を選択

  → その動かし方を自分の手とする

自己学習の導入結果

- ランダムな重みから開始
  - 自分対自分を何万回も繰り返す中でNNの重みを更新

- 学習の結果

プログラム	中間層数	訓練ゲーム数	対戦相手	結果
TD-1.0	80	300,000	Robertie,Davis, Magriel	-13点/5ゲーム
TD-2.0	40	800,000	Goulding,Woolsery, Snellings, Russell, Sylvester	-7点/38ゲーム
TD-2.1	80	1,500,000	Robertie	-1点/40ゲーム

  - 人間のトッププレイヤーに拮抗する実力を獲得