深層強化学習に基づく トレーディングカードゲーム 環境の構築

創発ソフトウェア研究室

B3　西村 昭賢

発表の流れ

- はじめに
- 要素技術
- 提案手法と実験
- まとめと今後の課題

発表の流れ

- はじめに
- 要素技術
- 提案手法と実験
- まとめと今後の課題

深層強化学習

- 実世界の問題への応用
  ⇒ ロボット制御, 自動運転

- ゲームへの応用
  ⇒ 完全情報ゲーム (囲碁, 将棋) ,
  不完全情報ゲーム (麻雀, ポーカー)

本研究の目的

- 深層強化学習と進化型計算を用いて
  ゲームバランスを最適化する手法を提案

- 独自のトレーディングカードゲーム (TCG) 環境
  における数値実験で提案手法の有効性

発表の流れ

- はじめに
- 要素技術
- 提案手法と実験
- まとめと今後の課題

要素技術

- Deep Q Network (DQN)
  - Q 学習 × 深層学習
  - 代表的な深層強化学習手法
- Genetic Algorithm (GA)
  - 最適化手法の 1 つ
  - 多目的最適化への応用も可能 (NSGA-II)

Mnih, V., Playing Atari with Deep Reinforcement Learning, arXiv e-prints, 2013.

発表の流れ

- はじめに
- 要素技術
- 提案手法と実験
- まとめと今後の課題

提案手法

1. 数値実験用の独自の TCG 環境

2. DQN を用いたあるデッキにおける
   定量的なカードパワー評価手法

3. 調整するカードの枚数を限定した
   GA による TCG 環境のバランス調整手法

提案手法

1. 数値実験用の独自の TCG 環境

2. DQN を用いたあるデッキにおける
   定量的なカードパワー評価手法

3. 調整するカードの枚数を限定した
   GA による TCG 環境のバランス調整手法

独自の TCG 環境


File Edit Assets GameObject Component Window Help
Center Local
Account Layers Layout
Game
Display 1 Full HD (1920x1080) Scale 0.85x
Maximize On Play Mute Audio Stats Gizmos
10
後攻
40
TURN END
10
先攻
40

提案手法

1. 数値実験用の独自の TCG 環境

2. DQN を用いたあるデッキにおける
   定量的なカードパワー評価手法

3. 調整するカードの枚数を限定した
   GA による TCG 環境のバランス調整手法

提案手法 2

手順 1. DQN を用いてデッキにおける
妥当な戦略を持つエージェントを構築

手順 2. エージェント同士で, 先攻後攻それぞれ
カードを 1 種類ずつ除いて対戦し勝率を計算

⇒ 定量的に構築戦略下のカードパワーを評価

提案手法

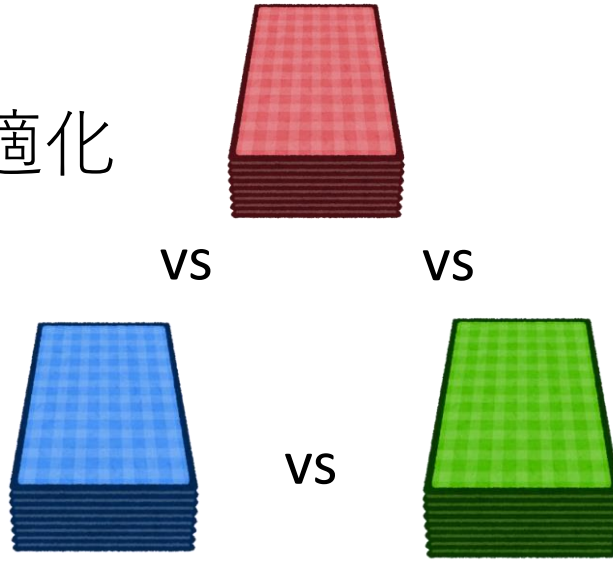
1. 数値実験用の独自の TCG 環境

2. DQN を用いたあるデッキにおける
   定量的なカードパワー評価手法

3. 調整するカードの枚数を限定した
   GA による TCG 環境のバランス調整手法

関連研究

- GA を用いてデッキ間の勝率を最適化



- 単目的 GA で勝率のみ最適化

⇒変更量が多すぎて原型がない

⇒変更量が多いとユーザーへの影響が大きい

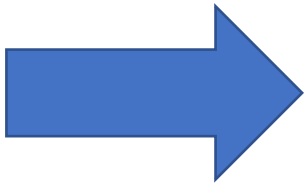
- 多目的 GA によりパラメータの変更量も最適化

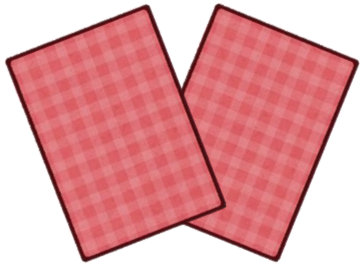
Fernando de Mesentier Silva, Rodrigo Canaan, Scott Lee, Matthew C. Fontaine, Julian Togelius, and Amy K. Hoover. 2019. Evolving the Hearthstone Meta. In 2019 IEEE Conference on Games (CoG). IEEE Press, 1–8. https://doi.org/10.1109/CIG.2019.8847966

提案手法 3

- 調整されるカードを限定し GA の解空間次元を削減
  ⇒調整されるカード枚数を最小限にバランス調整
- 限定は提案手法 2 で得られたカードパワーを参考

カードを限定





GA への
入力変換

{カード 1 パラメータ,
カード 2 パラメータ,
…,
カード n パラメータ}

{カード 1 パラメータ,
カード 2 パラメータ}

問題設定

TCG 環境に新たにデッキを追加し，
デッキ間の勝率を 50 % に近づける

後攻 / 先攻	追加デッキ	アグロ	コントロール
追加デッキ	0.6724	0.8005	0.8859
アグロ	0.3383	0.5255	0.5424
コントロール	0.2771	0.5121	0.5053

追加するデッキ

これらの 15 種類を 2 枚ずつ, 計 30 枚のカードからなる

HP	攻撃力	コスト	特殊効果	HP	攻撃力	コスト	特殊効果
4	4	1	無し	3	3	3	無し
1	3	2	取得	1	1	2	攻撃
2	3	3	召喚	1	1	1	治癒
3	1	3	速攻	2	1	2	速攻
5	4	5	無し	1	1	1	取得
2	2	2	無し	2	3	3	取得
4	3	4	無し	2	2	2	召喚
1	1	5	治癒				

実験内容

調整する種類数を増やしながら GA を適用

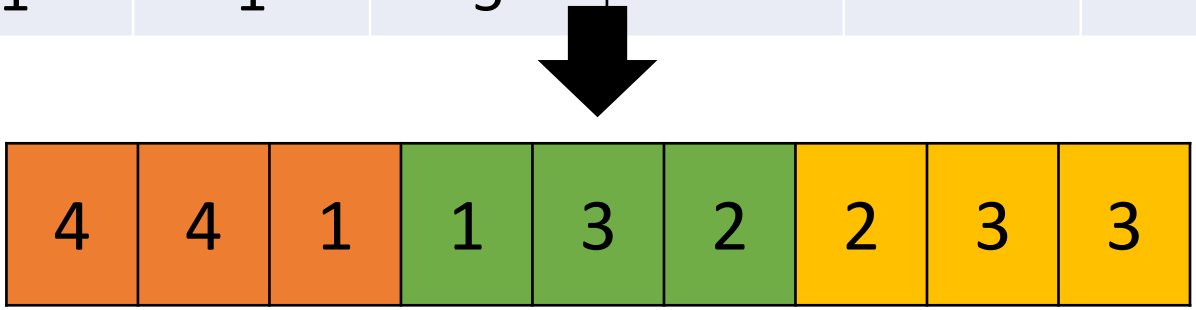
※ GA はデッキ間の勝率を最適化する単目的 GA

比較手法として以下を適用

- デッキ間の勝率を最適化する単目的 GA
- デッキ間の勝率とパラメータの変更量の 2 目的を最適化する多目的 GA

GA の個体

HP	攻撃力	コスト	HP	攻撃力	コスト
4	4	1	3	3	3
1	3	2	1	1	2
2	3	3	1	1	1
3	1	3	2	1	2
5	4	5	1	1	1
2	2	2	2	3	3
4	3	4	2	2	2
1	1	5			



適応度

- 単目的 GA の適応度, 多目的 GA の目的関数として,

  適応度 1 : デッキ間の勝率 f_w

  適応度 2 : パラメータの総変更量 f_p

  適応度 3 : 調整されたカード種類数 f_c

  の 3 つを定義

適応度 1

$$f_w = \exp\left(-\sum_{i=0}^{4}\sqrt{(0.50 - r_i)^2}\right)$$

先攻＼後攻	追加デッキ	アグロ	コントロール
追加デッキ	r_0	r_1	r_2
アグロ	r_3	0.5255	0.5424
コントロール	r_4	0.5121	0.5053

適応度 2, 3

- パラメータの総変更量 p

$$f_p = \exp\left(-\frac{p}{200}\right)$$

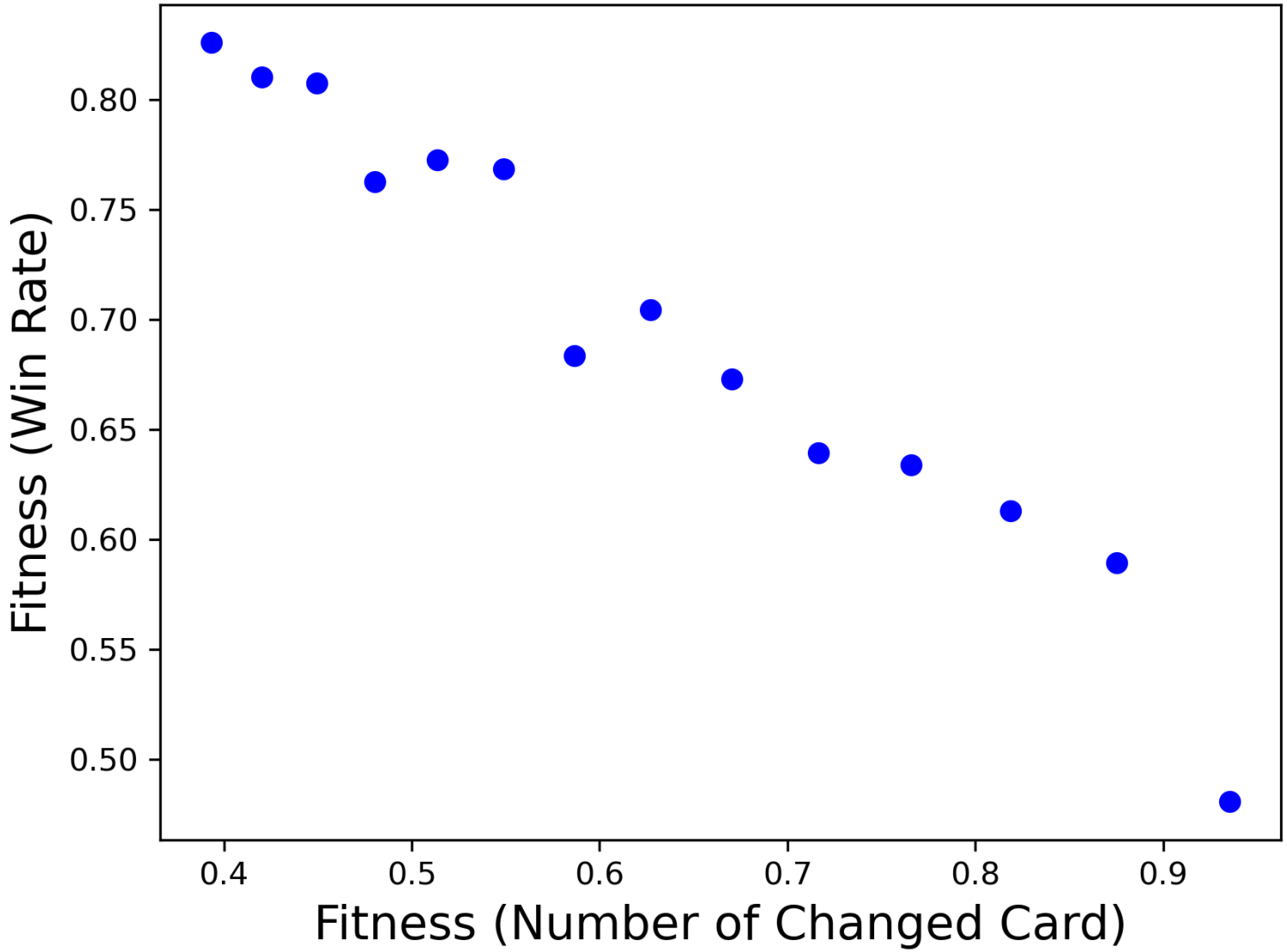
- 調整されたカード種類数 c

$$f_c = \exp\left(-\frac{c}{15}\right)$$

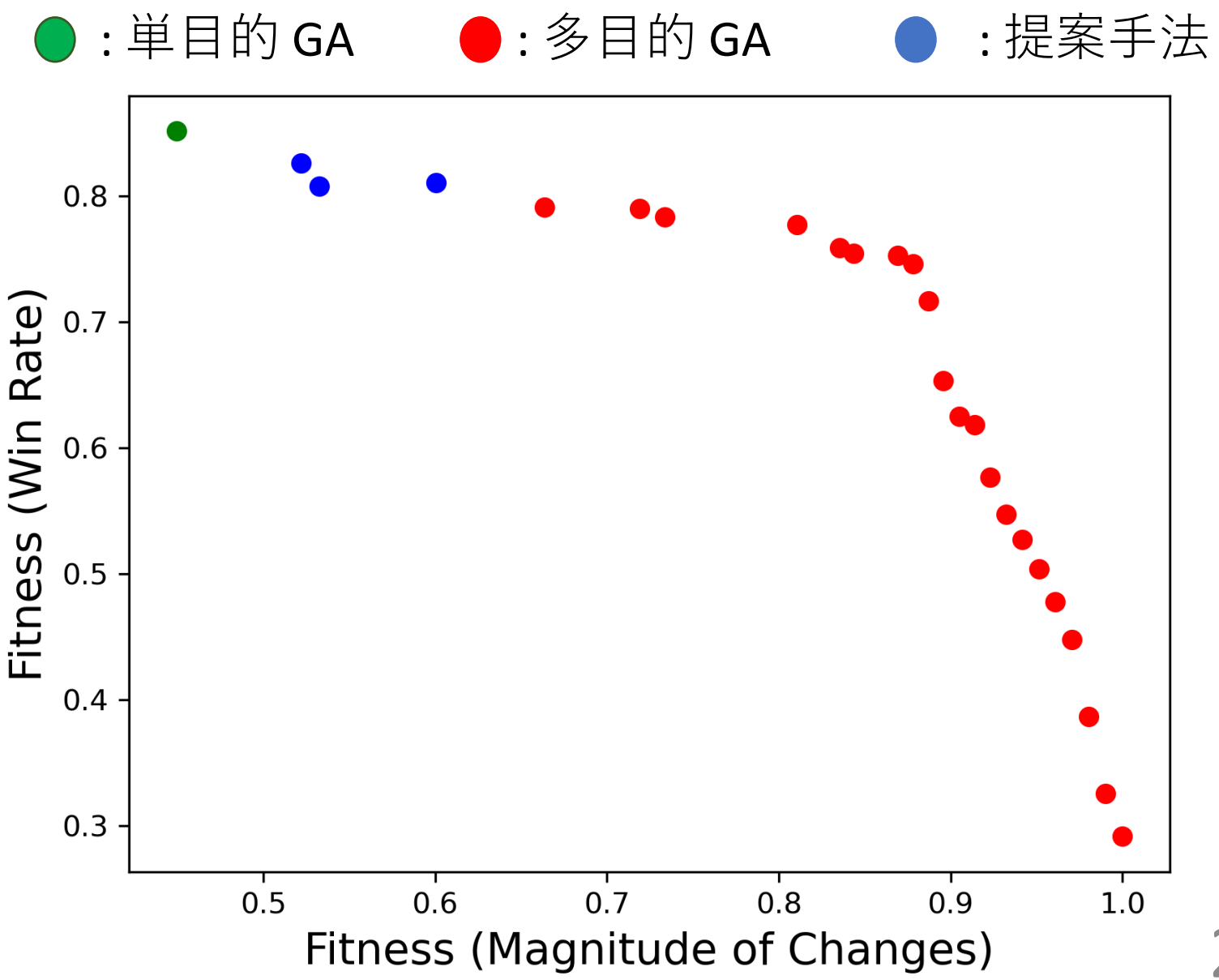
GA のパラメータ

パラメータ	値
世代数	50
個体数	50
遺伝子長	調整するカード種類数 × 3
交叉率	0.4
交叉の種類	2 点交叉
個体ごとの突然変異率	0.2
選択	1 個体だけエリート保存 その他は トーナメント方式
トーナメントサイズ	3
多目的 GA のアルゴリズム	NSGA - II

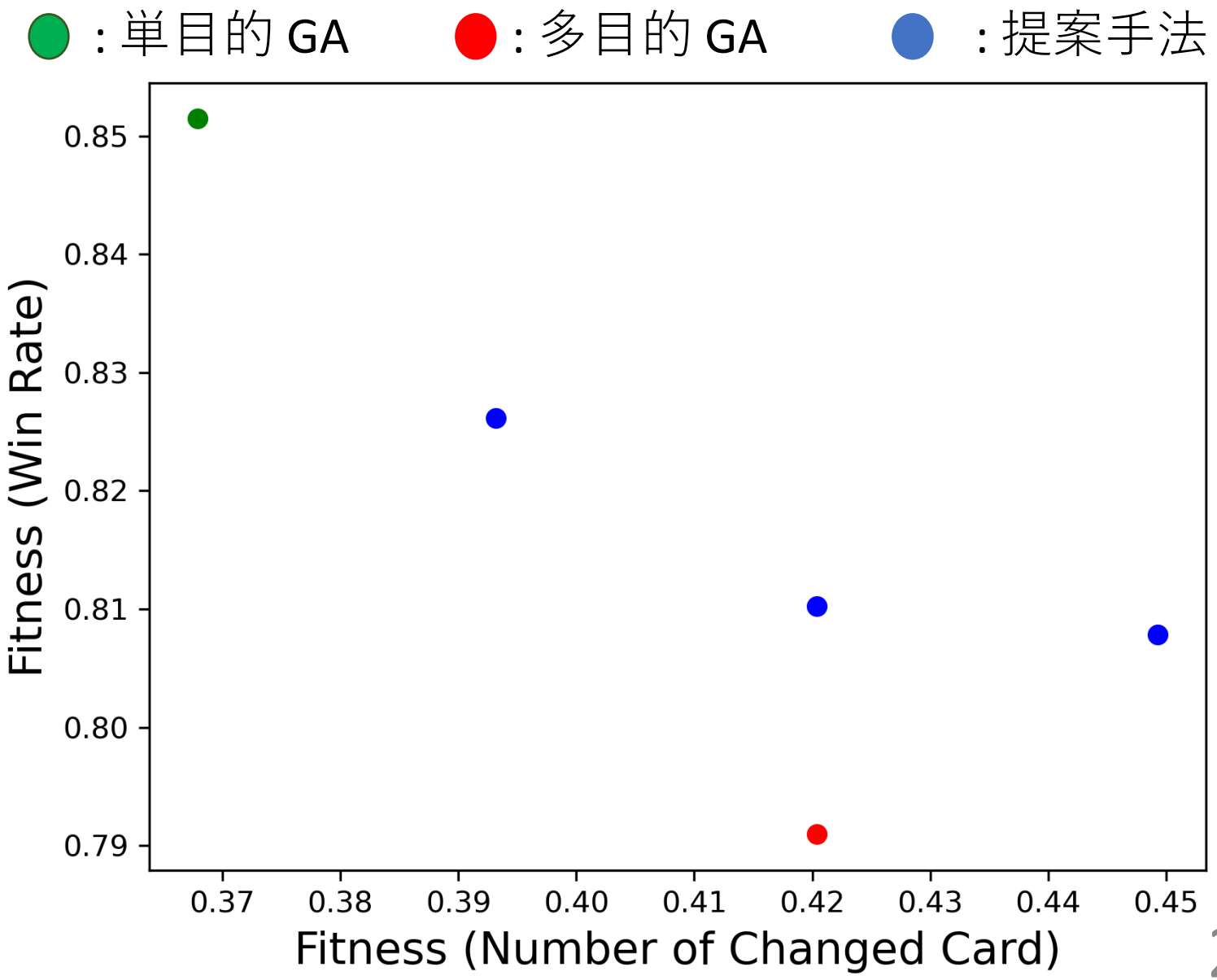
調整カード種類数を限定したGA



既存手法との比較


: 単目的 GA
: 多目的 GA
: 提案手法
Fitness (Win Rate)
0.8
0.7
0.6
0.5
0.4
0.3
0.5
0.6
0.7
0.8
0.9
1.0
Fitness (Magnitude of Changes)

既存手法との比較


: 単目的 GA
: 多目的 GA
: 提案手法
Fitness (Win Rate)
0.85
0.84
0.83
0.82
0.81
0.80
0.79
0.37
0.38
0.39
0.40
0.41
0.42
0.43
0.44
0.45
Fitness (Number of Changed Card)

既存手法との比較

適応度 / 手法	f_p	f_w	f_c
単目的 GA	0.44933	**0.85146**	0.36788
多目的 GA	**0.66365**	0.79097	0.42035
提案手法	0.53259	0.80783	**0.44933**

f_p：パラメータの変更量に関する適応度
f_w：デッキ間の勝率に関する適応度
f_c：調整されたカード枚数に関する適応度

発表の流れ

- はじめに
- 要素技術
- 提案手法と実験
- まとめと今後の課題

まとめ

- 数値実験用の独自の TCG 環境を構築した

- 深層強化学習を基にした提案手法を独自の TCG 環境において数値実験し, 有効性を確かめた

今後の課題

- DQN 以外の深層強化学習手法の適用
- GA 以外の最適化アルゴリズムの適用
- 最適なハイパーパラメータの発見
- デッキ間の相性まで考慮したバランス調整の検討

ご清聴ありがとうございました.

状態空間

パラメータの説明	次元	値域 (離散値)
各プレイヤーの HP	2	0 ～ 20
各プレイヤーの残りマナ	2	0 ～ 5
自手札 1 ～ 9 の HP , 攻撃力, コスト, 特殊効果	36	0 ～ 5
自盤面 1 ～ 5 の HP , 攻撃力	10	0 ～ 5
敵盤面 1 ～ 5 の HP , 攻撃力	10	0 ～ 5
自盤面 1 ～ 5 が 行動可能かどうか	5	0 ～ 1
両デッキ残り枚数	2	0 ～ 30
計	**67**	

行動空間

パラメータの説明	次元
手札 1 ～ 9 を盤面に出す	9
盤面 1 で敵盤面 1 ～ 5 を攻撃 or 敵プレイヤーを攻撃	6
盤面 2 で敵盤面 1 ～ 5 を攻撃 or 敵プレイヤーを攻撃	6
盤面 3 で敵盤面 1 ～ 5 を攻撃 or 敵プレイヤーを攻撃	6
盤面 4 で敵盤面 1 ～ 5 を攻撃 or敵プレイヤーを攻撃	6
盤面 5 で敵盤面 1 ～ 5を攻撃 or敵プレイヤーを攻撃	6
ターンエンド	1
計	**40**

報酬の定義

- 1 ステップ終了時

$$r = 0.0$$

- 1 エピソード終了時

$$r = \begin{cases} 1.0 & \text{(学習側勝利)} \\ -1.0 & \text{(対戦相手勝利)} \end{cases}$$

トレーディングカードゲーム (TCG)

マジック：ザ・ギャザリング.新たな旗のもとで.
2017. https://mtg-jp.com/reading/publicity/0019775/

- 2 人のプレイヤーからなる
- 先攻と後攻に分かれ,
  ターン制で進行する
- 各プレイヤーは異なる複数のカードからなる
  デッキを持つ
- 相手プレイヤーの手札など一部の情報は観測できない
  (不完全情報ゲーム)

カードゲーム型対戦環境

- 2 人のプレイヤーからなる
- プレイヤーは複数のカードからなるデッキを持つ
- プレイヤーは HP, マナの 2 つのパラメータを持つ

用語説明

- 手札, 盤面

各プレイヤーがカードを保有する領域

- ドロー

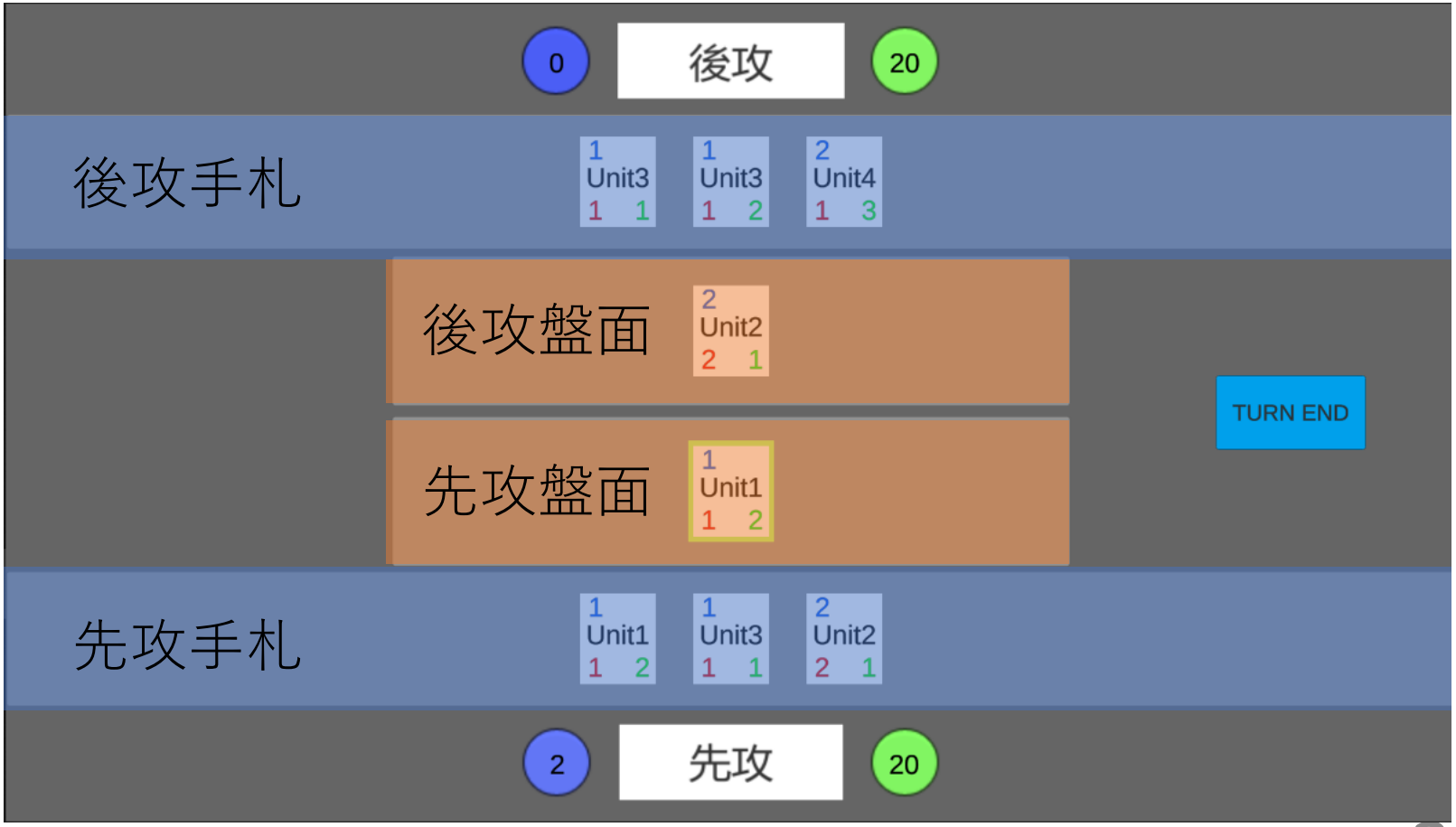
デッキからカードを取り出し, 手札に加える操作

- プレイ

手札から盤面にカードを出す操作

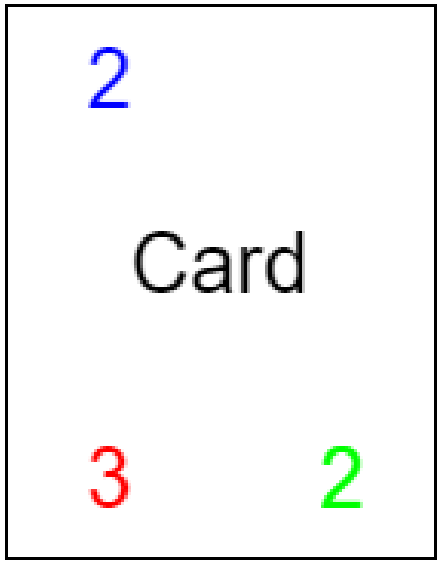
用語説明

- 各プレイヤーは手札, 盤面を持つ



用語説明

- カードはそれぞれ
  攻撃力, HP, コストを持つ
- カードの中には以下のような
  特殊効果を持つものがある



特殊効果	説明
召喚	盤面に出したら (攻撃力 , HP) = (1 , 1)のユニット追加で出す
治癒	盤面に出したら自プレイヤーの HP を 2 回復する
攻撃	盤面に出したら敵プレイヤーの HP を 2 減らす
取得	盤面に出したら 自プレイヤーは 1 枚カードをドロー
速攻	盤面に出たターンに攻撃できる

用語説明

- ドロー

デッキからカードを取り出し, 手札に加える操作

- プレイ

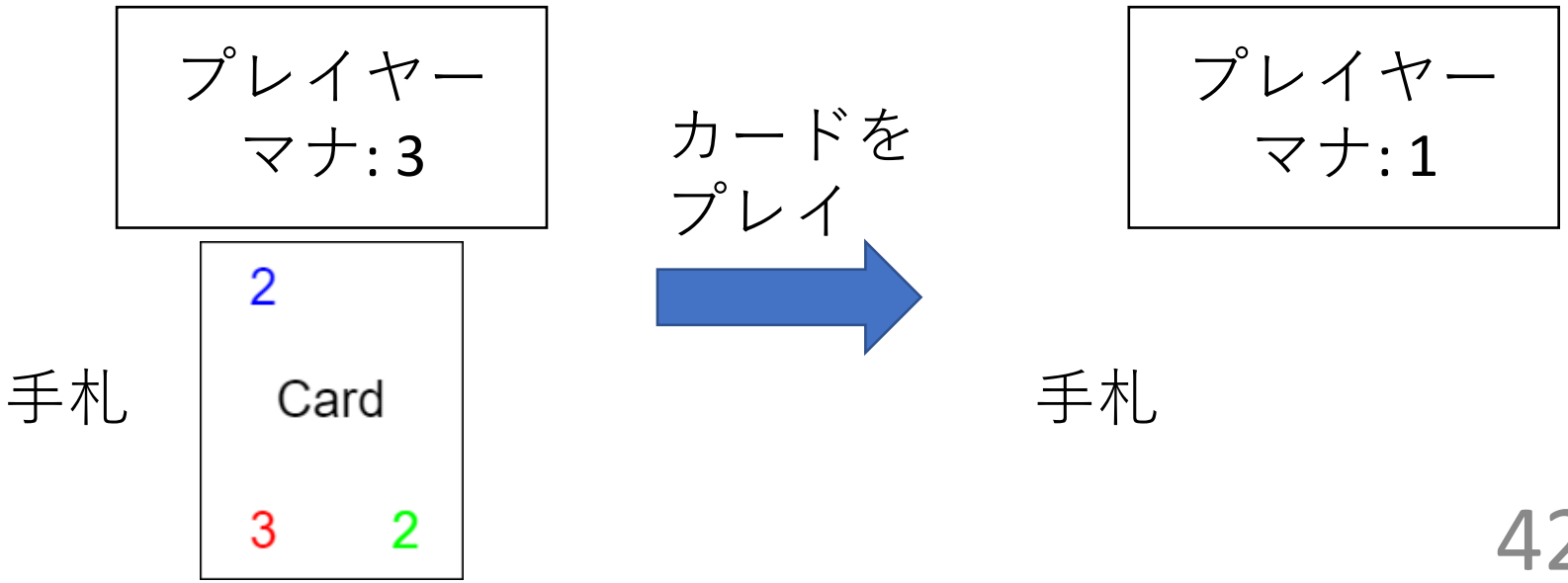
手札から盤面にカードを出す操作

- デッキ切れ

ゲーム中にデッキのカードが無くなる状態

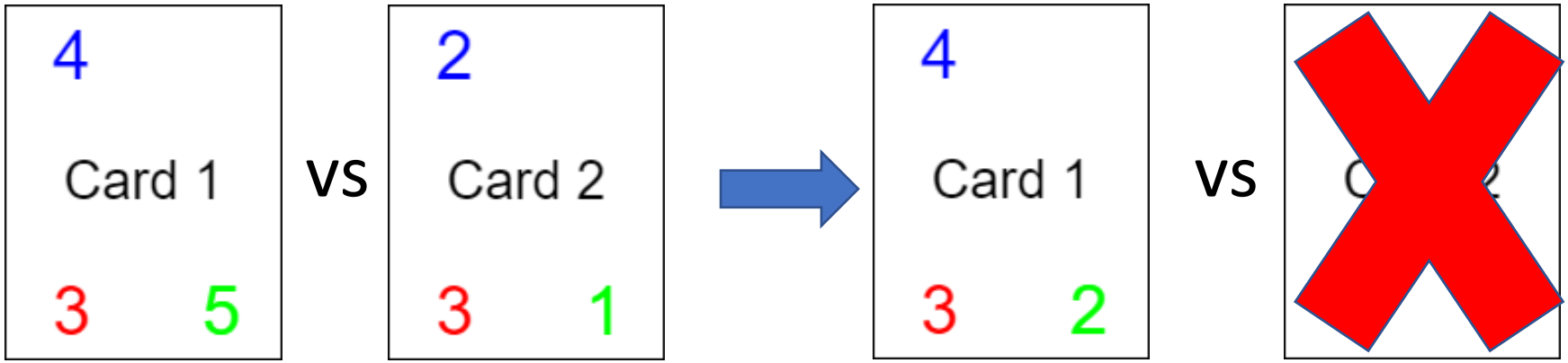
マナコスト処理

- プレイヤーの保有するマナ ≥ カードのコスト
  ⇒カードを盤面にプレイ可能
- ターンが回ってくるたびに回復
- 最大値 5 となるまでマナの上限も更新



カードの攻撃処理

- 盤面にあるカードは相手の盤面のカード,
  相手プレイヤーに攻撃することができる
- プレイされた次のターンから攻撃できる
- 攻撃したカードは攻撃対象から反撃を受ける



ゲームフロー

1. 各プレイヤーはデッキをシャッフル
2. 各プレイヤーは初期手札として 5 枚ドロー
3. 先攻プレイヤーはドローステップをスキップして行動
4. 後攻プレイヤーはカードを 1 枚ドローして行動
5. 先攻プレイヤーはカードを 1 枚ドローして行動
6. 4 , 5 の繰り返し ターンプレイヤーは行動前にマナ回復
7. デッキ切れの状態でドローしようとした, あるいは HP が 0 となったプレイヤーが敗北, その時点でゲーム終了

Q学習

- 代表的な価値ベースの強化学習手法の 1 つ

- Q 値を以下の式に従って 1 ステップごとに更新していく

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\{r_{t+1} + \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\}$$

TD誤差

α : 学習率 (Q 値の更新の度合い)

γ : 割引率 (将来の価値の割引度合い)

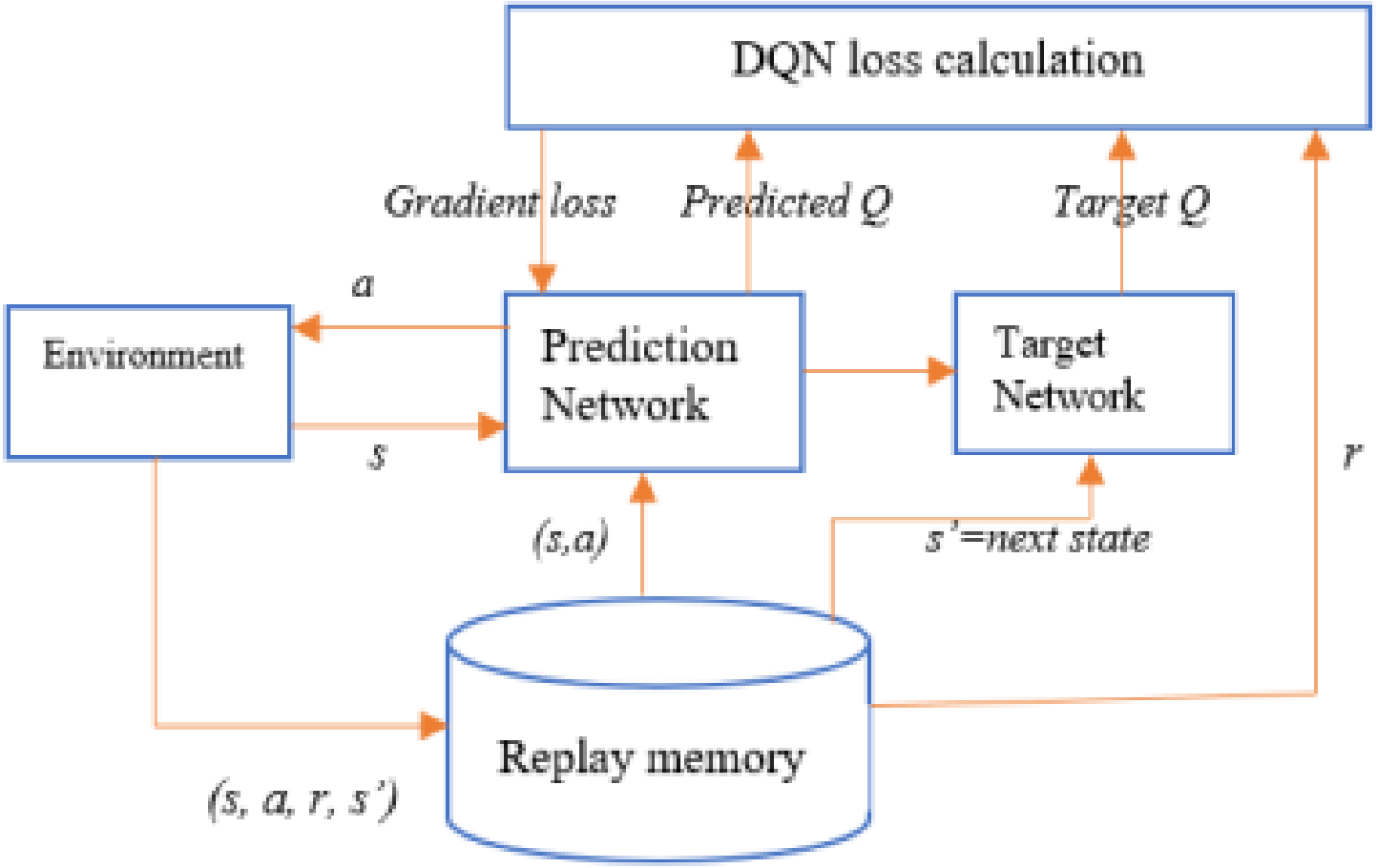
Deep Q Network (DQN)

- Q 学習では状態や行動の次元数が増えると
  現実的に計算ができなくなる
  ⇒ 深層学習を用いることで学習可能に

- Experience Replay や Fixed Target Network により
  安定した学習が可能になる

Mnih, V., Playing Atari with Deep Reinforcement Learning, arXiv e-prints, 2013.

DQN



Arwa, Erick & Folly, Komla. (2020). Reinforcement Learning Techniques for Optimal Power Control in Grid-Connected Microgrids: A Comprehensive Review. IEEE Access. 8. 1-16. 10.1109/ACCESS.2020.3038735.

DQN

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

Mnih, V., Playing Atari with Deep Reinforcement Learning, arXiv e-prints, 2013.