Vol.2021-DPS-186 No.29
Vol.2021-CSEC-92 No.29
2021/3/15

# 密集無線LAN環境におけるQ学習を用いた送信電力・信号検知閾値制御の検討

武松 未来[1,a)]　坂井 渉太[2,b)]　重野 寛[2,c)]

**概要**：無線 LAN の密集化によるスループット性能の低下を解決するために送信電力・信号検知閾値制御を初めとする周波数資源を有効活用する空間再利用に関する技術の採用が検討されている．既存研究 RTOT アルゴリズムでは，1 つの変数 M により送信電力と信号検知閾値を同時にコントロールすることが可能となった．しかし，シナリオに応じた適切な変数 M を使用する必要があるが，既存研究では変数 M の決定に関しては考慮していない．そこで本稿では，Q 学習により RTOT アルゴリズムの変数 M を決定することを提案する．また，Q 学習についての既存研究で多く使用されている報酬の計算方法では公平性を考慮していないものが多く，スループットは高いが，公平性が低いという問題があったため，本稿では，公平性を向上させる報酬の計算方法を提案する．シミュレーションによって，提案報酬を使用した Q 学習を用いた RTOT アルゴリズムでは，既存手法に比べてスループットと公平性ともに向上することを確認した．

# A Study of Power Control and Dynamic Sensitivity Control using Q-Learning in dense Wireless LAN

## 1. はじめに

近年，無線 LAN は利便性や設置の容易さ，コストの良さから広く導入されるようになったことで無線 LAN の急激な高密度化が進み [1]，1 つのアクセスポイント（AP）と複数のステーション（STA）から構成される Basic Service Set（BSS）間の干渉が増加している．このような環境では，送信フレームの衝突や送信機会の喪失によるスループット性能の低下が発生している [2] ．前述のような密集環境の問題解決に向けて，周波数資源を有効活用する空間再利用（Spatial Reuse）に関する技術の採用が検討された．空間再利用の技術にもいくつかの種類が存在し，信号検知閾値制御（DSC: Dynamic Sensitivity Control），送信電力制御（TPC: Transmit Power Control）などが挙げられる．DSC は CST を調整することで送信を促進する技術で，TPC は送信電力をできる限り抑えることで他の端末への干渉を抑制する技術である．

既存研究の空間再利用技術 RSSI To OBSS Threshold（RTOT）アルゴリズム [3] では，ビーコンの受信信号強度（RSSI）を用いて CST，送信電力を計算することで DSC と TPC を実装している．ここで RSSI 値空間から CST 値空間へ変換するために変数 M を導入しており，この変数 M の値は送信電力・信号検知閾値を同時に制御できるため重要な値である．しかし，変数 M の値の決定方法は述べられていない．また，既存研究では，CST，送信電力のようなスループットや公平性などの全体の性能に影響を及ぼす値は手作業で決定することが多いが，近年，空間再利用に機械学習を導入する手法への関心が高まっている [4] ．

そこで本稿では，RTOT アルゴリズムの変数 M を強化学習の一手法である Q 学習によって決定することを提案する．RTOT アルゴリズムの変数 M を Q 学習により最適化する目的は，従来の空間再利用技術よりも良い性能とすること，シナリオごとに適切な変数 M を決定することである．また，多くの既存研究では報酬において公平性を考慮しておらず，スループットは高いが，公平性が低い問題 [5] があったため，本論文では，公平性を向上させることを目

[1] 慶應義塾大学理工学部情報工学科
Faculty of Science and Technology, Keio University,Yokohama, Kanagawa, 223-8522, Japan
[2] 慶應義塾大学大学院理工学研究科
Graduate School of Science and Technology, Keio University,Yokohama, Kanagawa, 223-8522, Japan
[a)] takematsu@mos.ics.keio.ac.jp
[b)] sakai@mos.ics.keio.ac.jp
[c)] shigeno@mos.ics.keio.ac.jp

的とした報酬の計算方法を提案する．

以下，本稿の構成について述べる．まず第 2 章で関連研究について紹介し，第 3 章では本研究の提案を説明，第 5 章でシミュレーションを用いた評価を行う．そして最後に第 6 章で結論を述べる．

## 2. 関連研究

### 2.1 RTOT アルゴリズム

RTOT アルゴリズム [3] とは，802.11ax で提案されている Overlapping Basic Service Set Physical Detection-based SR（OBSS_PD-base SR）技術を実装するためのアルゴリズムである．また，OBSS_PD-based SR とは主に，2 つの信号検知閾値（CST）を受信したフレームが同 BSS 内の端末からの送信なのか，他 BSS からなのかで使い分ける信号検知閾値制御（DSC）と送信電力を制御する送信電力制御（TPC）の 2 つの制御からなり，RTOT アルゴリズムでは OBSS_PD-based SR を実装することによりチャネルの使用効率をあげている．

実際の RTOT アルゴリズムは以下の通りに進行する．まず, 検知した RSSI と定義された変数 M から $OBSS\_PD_{Thr}$ を計算する．

$$OBSS\_PD_{Thr} = Beacon_{RSSI} - M \qquad (1)$$

ここでの $OBSS\_PD_{Thr}$ とは受信したフレームが他 BSS の端末からの場合に使用する CST である．次に $TXPWR$ の計算と $OBSS\_PD_{Thr}$ の調整を行う．計算した $OBSS\_PD_{Thr}$ がその最大値より大きい場合は $OBSS\_PD_{Thr}$ を $OBSS\_PD_{ThrMax}$ に，TXPWR（送信電力）は $TXPWR_{Min}$（送信電力の最小値）に設定する．$OBSS\_PD_{Thr}$ がその最小値より小さい場合は，$OBSS\_PD_{ThrMin}$ に，TXPWR は $TXPWR_{Max}$（送信電力の最大値）に設定する．また，$OBSS\_PD_{Thr}$ が最大値と最小値の間である場合，$OBSS\_PD_{Thr}$ は計算で算出した値を使用し，$TXPWR$ は式で計算する．

$$TXPWR = OBSS\_PD_{ThrMin} + TXPWR_{ref} - OBSS\_PD_{Thr} \qquad (2)$$

RTOT アルゴリズムの目的は，STA が送信先の AP に近い場合は小さい $OBSS\_PD_{Thr}$ と大きい TXPWR にし，逆に STA が送信先の AP に遠い場合は大きい $OBSS\_PD_{Thr}$ と小さい TXPWR を採用することでチャネルの使用効率をあげることである．また，ビーコン RSSI 値は距離を表現することができるが，$[OBSS\_PD_{ThrMin}, OBSS\_PD_{ThrMax}]$ の範囲である可能性は限りなく低い．そのため，変数 M を導入することでビーコン RSSI 値空間から $OBSS\_PD_{Thr}$ 値空間への変換を可能としている．ここで導入されている変数 M は性能を決定する重要な値であるが，目的やシナリオによって適切な値が異なる．そのため，シナリオに応じて異なる変数 M を使用する必要があるが，既存研究では具体的な変数 M の決定方法について述べられておらず，適切な変数 M を決めることは難しい．

### 2.2 強化学習を用いた空間再利用技術

本節では，強化学習によって送信電力・信号検知閾値制御を行っている研究をいくつか紹介する．

F.Wilhelmi ら [5] は分散型 Q 学習による空間再利用を提案している．この研究では，分散型シナリオに焦点を当てているため，状態を考慮しない Q 学習により送信端末は送信電力とチャネルを最適化している．Q 学習により各端末がスループットを向上させる最良のアクションを探索することで高いスループットとなる端末を生み出すことができた．しかし，端末同士でスループットの値に差が発生している．これは個々の端末が周りの状況を考慮にいれず，自身の性能をあげることだけを考えて学習をしているためだと考えられる．そのため周囲の状況を考慮した学習方法を考える必要がある．

F.Wilhelmi ら [6] は Multi-Armed Bandits（MAB）をベースとした強化学習により使用チャネル，送信電力，信号検知閾値を最適化することを提案している．このとき，利己的な学習と環境を考慮した学習の 2 つの戦略を使用，評価している．1 つ目の利己的な学習では，各端末自身の獲得したスループットのみから報酬を計算する．2 つ目の環境を考慮した学習では，ネットワーク全体の性能を考慮するために，報酬はネットワーク全体が獲得した max-min スループットより報酬を計算する．本研究の利己的な学習では，公平性は低下してしまうがシステム全体のスループットを上げることに成功している．また，環境を考慮した学習では，利己的な学習での課題である公平性の向上に成功しているが，高いスループットが得られそうな端末も一番スループットが低い端末に性能を合わせてしまっているため，スループット性能を制限している．

S.Maghsudi ら [7] はインフラレスネットワークにおける送信電力，チャネル選択のための MAB アプローチを提案している．チャネル配分問題を解決するために，本研究では失敗する度合いを表す後悔を最小化するための 2 つのアプローチ方法を提案した．1 つ目は指数関数ベースの加重平均戦略による学習で，失敗するアクションほど選択しないようになる方法である．2 つ目は過去に最も失敗しなかったアクションを選択する方法である．どちらのアプローチ方法でも後悔が最小となるように学習できているが，評価で使用した端末数とアクション数が少ない．既存の学習ベース空間再利用の手法では性能を向上させることができることを確認できるが，密集環境で高い公平性を獲得することが課題である．特に，利己的な報酬による学習では，高いスループットを獲得できる反面，周囲の環境を考慮していないため公平性が低下している．

Vol.2021-DPS-186 No.29
Vol.2021-CSEC-92 No.29
2021/3/15

**Algorithm 1** 提案手法

**Input:** set of possible action(=M) in {1,...,K}, $Beacon_{RSSI}$
  **Initialize** : $t = 0$,$\hat{Q}(a_k) = 0$,$r_{a_k} = 0$
  **while** true **do**
    **Select action**
    **if** prob $= 1 - \varepsilon$ **then**
      $a_k = \arg\max_{1,\ldots,K} \hat{Q}(a_k)$
    **else**
      $a_k = i\,\mathcal{U}(1, K)$
    **end if**
    **Calculate RTOT value**
    $OBSS\_PD_{Thr} \leftarrow Beacon_{RSSI} - a_k$
    **if** $OBSS\_PD_{Thr}$>$OBSS\_PD_{ThrMax}$ **then**
      $OBSS\_PD_{Thr} \leftarrow OBSS\_PD_{ThrMax}$
      $TXPWR \leftarrow TXPWR_{Min}$
    **else if** $OBSS\_PD_{Thr}$<$OBSS\_PD_{ThrMin}$ **then**
      $OBSS\_PD_{Thr} \leftarrow OBSS\_PD_{ThrMin}$
      $TXPWR \leftarrow TXPWR_{Max}$
    **else**
      $TXPWR \leftarrow OBSS\_PD_{ThrMin} + TXPWR_{ref} - OBSS\_PD_{Thr}$
    **end if**
    **Observe reward**
    **if その時点までのスループット総和と**
    **現在のスループットが上位 N 番目 then**
      $r_{a_k} = \frac{\Gamma_{i,t}}{\Gamma_i^*}(-1)$
    **else**
      $r_{a_k} = \frac{\Gamma_{i,t}}{\Gamma_i^*}$
    **end if**
    $\hat{Q}(a_t) \leftarrow (1 - \alpha_t)\hat{Q}(a_t) + \alpha_t(a_t + \gamma(\max_{a'} \hat{Q}(a')))$
    $\varepsilon_t \leftarrow \frac{\varepsilon_0}{\sqrt{t}}$
    $t \leftarrow t + 1$
  **end while**

## 3. 公平性向上を目的とした報酬による Q 学習を導入した RTOT アルゴリズム

本章では，既存研究 RTOT アルゴリズムの変数 M を決定するために Q 学習を導入し，さらに公平性改善のための新しい報酬の計算方法を提案する．

### 3.1 概要

本制御では RTOT アルゴリズム内の変数 M を Q 学習によって最適化する．本稿の提案手法のフローチャートを図 1 に示す．

学習者である STA は，各端末のスループットをチャネルを観察することによって推定する．ここでチャネルを観察することで各端末のスループットを完全に推定できるものとする．さらに推定した各端末のスループットと自身のスループットにより報酬を計算し，学習を進めていく．アクション選択戦略として $\varepsilon - greedy$ 法を用いることで，強化学習の課題の 1 つである，探索と利用のジレンマを解決する．また，本稿の提案手法のアルゴリズムをアルゴリズムアルゴリズム 1 に示す．

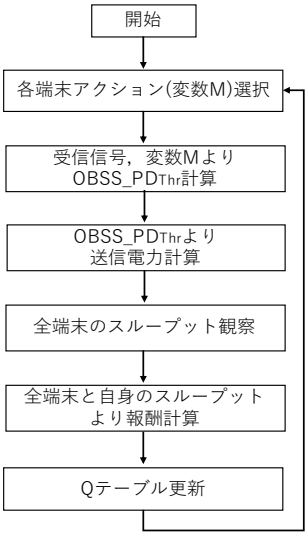


**図 1** 提案手法のフローチャート
**Fig. 1** Flowchart of the proposed scheme

### 3.2 アクションの選択

各端末は選択肢の中からアクションである変数 M を選択する．アクションの選択戦略としては，$\varepsilon - greedy$ 法を用いる．また，一回学習が終わるごとに $\varepsilon_t$ を $\varepsilon_t = \frac{\varepsilon_0}{\sqrt{t}}$ によって更新する．$\varepsilon_t$ は時刻 t における $\varepsilon$ のこと，$\varepsilon_0$ は $\varepsilon$ の初期値，$t$ は時刻のことを表す．更新によって，序盤はランダムにアクションを選択することが多くなり，終盤に進むにつれて経験したアクションの中で報酬が一番高くなるアクションを選択する確率が増加する．このように探索メインと利用メインの変化によって，強化学習の課題の 1 つである，探索と利用のジレンマ，具体的には，学習をする際に新しく高い報酬となるアクションを発見するためには探索することが必要だが，システム性能を向上させるためには確実に報酬が高くなると分かっているアクションを利用した方がよいという 2 つの考え方によるジレンマを緩和することができる．

### 3.3 報酬の計算

既存研究で使用している報酬はスループット性能は向上させることができるものの，公平性が低下してしまうものが多い．よって本稿の報酬では，密集環境における端末のスループット性能の公平性を考慮する．その時点までで獲得したスループットの総和が上位 N 番目，かつ現在獲得したスループットも全端末中で上位 N 番目の場合，報酬をマイナスにする方法である．まず，スループットのみから報酬を計算する．次に，各端末でその時点までで獲得したスループットの総和を求める．求めた総和と現在獲得したスループットが両方とも全端末のうち，上位 N 番目の場合は報酬をマイナスにする．

その時点までで高いスループットを獲得してきた端末が低いスループットを獲得するアクションを選択すれば，自

端末が占有していた送信時間を開放し，周囲の端末に分け与えるため，報酬が良くなかった他の端末の性能が上がり，公平性が向上する．また，本稿で使用するアクション選択戦略 $\varepsilon - greedy$ 法では，アクションはランダムか期待累積報酬が最大となるアクションを選択するため，報酬がマイナスになったアクションは選択されにくくなる．よって，その時点までで高いスループットを獲得してきた端末が，高いスループットを獲得したアクションを選択しにくくするために，高いスループットを獲得した際（全端末の中で上位 N 番目以内のスループットを獲得した場合）は報酬をマイナスにする．そして，スループットの総和が高い端末が多くのスループットを獲得しないアクションを選択しやすくなることによって，その端末が送信時間を占有することがなくなり，結果として公平性が上がる．報酬は式 3 を用いて計算する．

$$r_{i,t} = \begin{cases} \frac{\Gamma_{i,t}}{\Gamma_i^*}(-1) & (\text{スループットが上位 n 番目}) \\ \frac{\Gamma_{i,t}}{\Gamma_i^*} & (\text{上記以外}) \end{cases} \quad (3)$$

### 3.4 モデルの更新

Q 学習では，Q テーブルと呼ばれる表を使用してモデルを更新する．Q テーブルとは，各端末の各状態とアクションにおける期待累積報酬を格納しているテーブルのことを指し，このテーブル内の期待累積報酬を用いて次のアクションを選択する．今回，分散型のシナリオに焦点を当てているため，各端末の状態は考慮しないため，Q テーブルには各端末の各アクションにおける期待累積報酬が格納されている．また，Q テーブルは式 4 によって更新される．

$$\hat{Q}(a_t) \leftarrow (1-\alpha_t)\hat{Q}(a_t) + \alpha_t(a_t + \gamma(\max_{a'}\hat{Q}(a'))) \quad (4)$$

ここで，$\alpha_t$ は時刻 t における学習率，$\max_{a'}\hat{Q}(s_{t+1}, a')$ は次の状態 $s_{t+1}$ における最大値である．

## 4. シミュレーション評価

### 4.1 シミュレーション環境

提案手法を ns-3 [8] を用いて，アパートメントシナリオ [9] により評価した．今回使用するアパートメントシナリオは，図 2 のような，1 部屋 10m×10m×3m の立方体で，各フロア $10 \times 2 = 20$ 部屋 1 階建ての建物である．図 2 のように，1 部屋に 1 台の AP と 1 台の STA から構成される BSS が配置される．AP の位置と STA の位置は，すべての部屋において $xy$ 平面上でランダムである．また，床からの高さは AP，STA ともに $z = 1.5$m に固定されている．シミュレーションは AP，STA の位置をランダムには変えながら 5 回行った．また，表 1 にアパートメントシナリオにおいて使用したパラメータを示す．表 1 にあるように，伝搬損失については residential path model [9] を使用した．

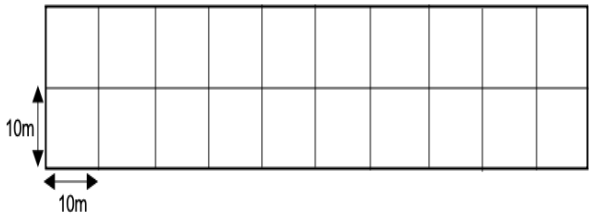


**図 2** アパートメントシナリオの概略図
**Fig. 2** Scenario layout

**表 1** シミュレーションで使用した主なパラメータ
**Table 1** Simulation parameters

| | |
|---|---|
| シミュレータ | ns-3 |
| シミュレーション時間 | 1000 秒 |
| タイムステップ | 0.05 秒 |
| モビリティ | なし |
| 無線通信規格 | IEEE802.11ac |
| 使用周波数帯 | 5 GHz |
| 周波数帯域幅 | 20MHz |
| MCS | 7 |
| トラフィックモデル | CBR |
| トラフィックロード | UL |
| アンテナゲイン | AP: +0 dBi, STA: 0 dBi |
| ノイズ指数 | 7 dBm |
| 伝搬損失モデル | Residential path loss model [9] |
| Fading/Shadowing | 無効 |
| 最大アグリゲーション | 64 |
| 最大再送回数 | 7 回 |
| RTS/CTS | 無効 |
| STA $TXPWR$ | 最大値:15dBm，最小値:3dBm |
| $TXPWR_{AP}$ | 20dBm |
| $OBSS\_PD_{Thr}$ | 最大値:-62dBm，最小値:-82dBm |
| 変数 M | 25〜45 |

### 4.2 評価概要

比較手法は Spacial Reuse 技術を使用しておらず送信電力 23dBm，信号検知閾値-82dBm と固定されている legcy，既存報酬 1（スループットのみより計算） [5] による Q 学習を用いた RTOT アルゴリズム [3] ，既存報酬 2（公平性考慮）[6] による Q 学習を用いた RTOT アルゴリズム，提案報酬による Q 学習を用いた RTOT アルゴリズムの 4 つである．また今回，提案報酬内の $N$ は $N = 5$ に設定した．また，評価項目はシステム全体の総スループット，Jain's Fairness Index [7] を使用した．総スループットは式 5 によって計算される．

$$\Gamma_{total} = \sum_{i}^{n} \Gamma_i \quad (5)$$

ここでの $\Gamma_i$ は $STA_i$ の総スループットであり，$n$ は STA の数である．また Jain's Fairness Index [7] とは，各 STA の総スループットは公平かどうかを示す指標であり，式 6 より計算される．

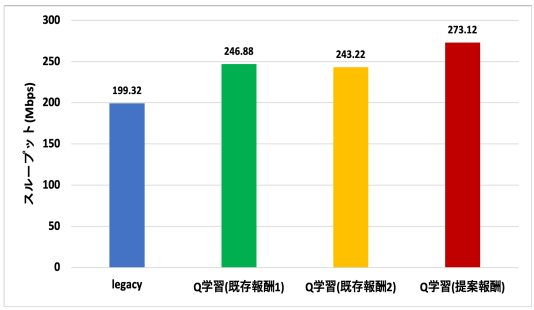


**図 3** 各手法の総スループット
**Fig. 3** Aggregate throughput of each method

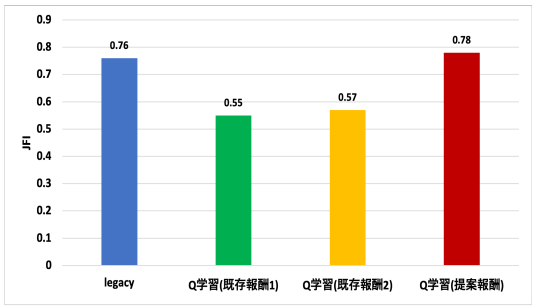


**図 4** 各手法の公平性 (JFI)
**Fig. 4** JFI of each method

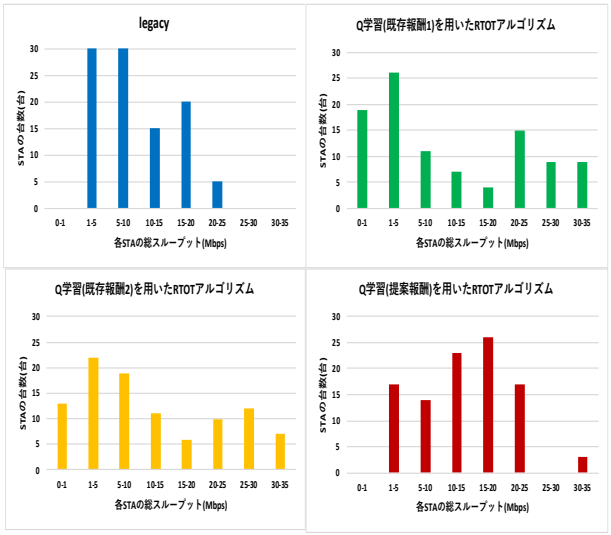


**図 5** 各手法の各 STA の総スループットにおけるヒストグラム
**Fig. 5** Histogram of aggregate throughput for each method

$$JFI = \frac{(\sum_{j=1}^{n} x_i)^2}{n \sum_{j=1}^{n} x_i^2} \tag{6}$$

ここでの $x_i$ は $STA_i$ の獲得したスループット，$n$ は STA の数を表す．値は $\left[\frac{1}{n}, 1\right]$ の範囲内で，値が大きいほど STA のスループット性能に差がなく公平性が高い．

### 4.3 総スループット

図 3 にシミュレーションでの各手法の総スループットを示す．Q 学習を導入した RTOT アルゴリズムのすべての手法は legacy の総スループット 199.3Mbps と比較して高いスループットを実現していることがわかる．既存報酬 1 による手法では，総スループットが 246.88Mbps と legacy よりも 23.8%増加していることがわかる．また，既存報酬 2 による手法では，総スループットが 243.22Mbps と legacy よりも 22.0%増加していることがわかる．このことから，Q 学習によって RTOT アルゴリズムにおける適切な変数 M を決定することができ，適切に送信電力・信号検知閾値制御されていると考えられる．次に，提案報酬による Q 学習を導入した RTOT アルゴリズムについて見ていく．提案報酬による手法の総スループットは 273.12Mbps となっており，legacy よりも 37.0%増加していた．また，提案報酬による手法の総スループットは既存報酬 1 による手法と比較して 10.6%増加していた．よって，提案した報酬により既存報酬よりもスループット性能を向上させることができたことがわかる．

### 4.4 公平性

図 4 にシミュレーションでの各手法のスループット性能の公平性を示す．また，図 5 に各手法の STA ごとの獲得した総スループットのヒストグラムを示す．legacy の JFI は 0.76 であり，提案報酬による Q 学習を導入した RTOT アルゴリズムのみ 2.5%増加していることが分かる．legacy の場合，全ての端末のスループット性能が低いため公平性が高いと考えられる．

次に，提案報酬による Q 学習を導入した RTOT アルゴリズムについて見ていく．今回，既存報酬による手法のスループットは高くなるが公平性は低くなるという問題を解決するために，提案報酬を提案しているため，提案報酬を既存報酬 1，既存報酬 2 と比較して評価する．提案報酬による手法の JFI は 0.78 と全ての手法の中でも最も高く，既存報酬 1 による手法と比較して 41.8%，既存報酬 2 による手法と比較して 36.8%増加していることがわかる．また，各 STA の総スループットは 15Mbps-20Mbps の階級を中心に山なりとなっており，さらに 0Mbps-1Mbps の階級の STA は存在しない結果となった．これは，今までのステップで大きいスループットを獲得してきた STA が大きいスループットを獲得するアクションを選択しにくくなるため，極端に大きいスループットを獲得する STA が減り，複数の STA がチャネルを独占しなくなったことで，チャネルの使用機会が他の手法よりも平等に与えられているためであると考えられる．そして平等に送信機会が与えられることにより，どの STA も報酬を獲得できるアクションを見つけることができ，公平性も高くなっている．また，スループットが 0Mbps-1Mbps の STA が存在しないことから，提案報酬では性能が低い STA の性能を上げることができたことが分かる．

## 5. おわりに

近年，密集環境の問題解決に向けて，送信電力・信号検知閾値制御による周波数資源を有効活用する空間再利用に関する技術の採用が検討されている．既存研究 RTOT アルゴリズムでは，変数 M により送信電力・信号検知閾値制御を実装している．しかしシナリオに応じた変数 M を決定する必要があるが，既存研究では変数 M の決定方法は記載されていない．

そこで本稿では，公平性向上を目的とした報酬による Q 学習による RTOT アルゴリズム中の変数 M の決定を行うことを提案した．提案手法による目的は，シナリオごとに適切な変数 M を決定すること，従来の空間再利用技術よりもスループット性能，公平性ともに向上させることである．アパートメントシナリオにおいて評価を行い，スループット性能，公平性について比較検討した．提案報酬による手法の総スループットは 273.12Mbps であり従来手法より 37.0%増加した．このことから，Q 学習によって RTOT アルゴリズムの中の変数 M を適切な値に設定することが可能であることがわかった．また，提案報酬による手法の公平性は既存報酬 1 による手法と比較して 2.6%増加したことから，スループット性能を損なうことなく公平性向上を達成した．

今後の課題として，全端末のスループットを報酬の計算に使用するため，自端末が獲得したスループットを他の端末に知らせる制御を考える必要がある．また，アパートメントシナリオ以外のシナリオでもシミュレーションすることを予定している．

**謝辞** 本稿の一部は，東北大学電気通信研究所における共同プロジェクト研究の支援によって行われた．

### 参考文献

[1] Afaqui, M Shahwaiz and Garcia-Villegas, Eduard and Lopez-Aguilera, Elena and Smith, Graham and Camps, Daniel,Evaluation of Dynamic Sensitivity Control Algorithm for IEEE 802.11 ax, 2015 IEEE Wireless Communications and Networking Conference (WCNC), pp.1060-1065(2015)

[2] K. Nishide and H. Kubo and R. Shinkuma and T. Takahashi, Detecting Hidden and Exposed Terminal Problems in Densely Deployed Wireless Networks,IEEE Transactions on Wireless Communications, Vol.11,pp.3841-3849(2012).

[3] Ropitault, Tanguy,Evaluation of RTOT algorithm: a first implementation of OBSS_PD-based SR method for IEEE 802.11 ax,2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp.1-7(2018).

[4] Sutton, Richard S and Barto, Andrew G, Reinforcement learning: An introduction,MIT press(2018).

[5] Wilhelmi, Francesc and Bellalta, Boris and Cano, Cristina and Jonsson, Anders, Implications of Decentralized Q-learning Resource Allocation in Wireless Networks, 2017 ieee 28th annual international symposium on personal, indoor, and mobile radio communications (pimrc),pp.1-5(2017).

[6] Wilhelmi, Francesc and Barrachina-Muñoz, Sergio and Bellalta, Boris and Cano, Cristina and Jonsson, Anders and Neu, Gergely, Potential and Pitfalls of Multi-Armed Bandits for Decentralized Spatial Reuse in Wlans,Journal of Network and Computer Applications,Vol.127,pp.26-42(2019).

[7] S.Maghsudi and S.Stańczak, Joint Channel Selection and Power Control in Infrastructureless Wireless Networks: A Multiplayer Multiarmed Bandit Framework, IEEE Transactions on Vehicular Technology, Vol64,Number10,pp.4565-4578(2015).

[8] ns-3(online), 入手先 ⟨https://www.nsnam.org/⟩, （参照 2021-01-20）.

[9] Merlin, Simone and Barriac, G and Sampath, H and others, TGax simulation scenarios, IEEE802,pp.11-14(2015).