社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

方策勾配型強化学習を用いたEDCAにおける送信遅延時間短縮の検討

新﨑　聖峰†　香田　優介†　山本　高至†　西尾　理志†　守倉　正博†

† 京都大学大学院情報学研究科　〒606-8501 京都市左京区吉田本町
E-mail: †kyamamot@i.kyoto-u.ac.jp

あらまし　無線LAN (Local Area Network) における QoS (Quality of Service) 制御を行う方式として，IEEE 802.11e では EDCA（Enhanced Distributed Channel Access）方式が規定されている．本稿では，それぞれの AP（Access Point）が一定の高優先度パケットを送信する状況を考える．このとき，AP 全てが高優先度パケットの送信を完了する時間の最小化を目的とする．そのために EDCA 方式において，高優先度パケットが他のアクセスカテゴリ（AC：Access Category）にも分類可能であるとする．その上で，AP 全てが高優先度パケットの送信を完了する時間を最小化するために最適な分類則を獲得する．ここで，それぞれの AP における高優先度パケットの送信を完了する時間を最小化しても，全ての AP における高優先度パケットの送信を完了する時間を最小化できるとは限らない．そのため，AP 全てが高優先度パケットの送信を完了する時間を最小化するために，全ての AP が協力することを考える．そこで中央制御局が，それぞれの AP に到着したパケットを AC に分類する．このとき，他の AP の情報も考慮する必要があるため，最適な分類則の獲得は困難である．そのため本稿では，最適な分類則を経験的に獲得するために，強化学習を用いることを提案する．更に，強化学習のうち方策勾配法を用いることで学習を進められることを示す．その上で，学習を進めるための方策を設計する．シミュレーション評価では，提案方式が従来の EDCA 方式と比較して送信遅延時間が小さいことを示す．

キーワード　IEEE 802.11e，EDCA，強化学習，方策勾配法

Policy Gradient Reinforcement Learning for Reducing Transmission Delay in EDCA

Masao SHINZAKI†, Yusuke KODA†, Koji YAMAMOTO†,

Takayuki NISHIO†, and Masahiro MORIKURA†

† Graduate School of Informatics, Kyoto University　Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †kyamamot@i.kyoto-u.ac.jp

Abstract　This paper proposes a packet mapping algorithm among Access Categories (ACs) in Enhanced Distributed Channel Access (EDCA) scheme based on policy gradient Reinforcement Learning (RL). In EDCA scheme based on an autonomous distributed control, high priority packets obtain more transmission opportunity than low priority packets. The arrival rate of high priority packets can be higher than that of low priority packets. In such a situation, mapping high priority packets to the AC defined in EDCA scheme is not necessarily the best mapping algorithm to minimize the transmission delay of high priority packets. Therefore, we assume that EDCA scheme can map high priority packets to any AC. Although each AP sends high priority packets early, however, APs can not always send high priority packets early. This paper proposes policy gradient RL to empirically obtain optimal mapping algorithm. By using the mapping algorithm based on RL, simulation results reveal that the transmission delay can be reduced. The average transmission delay of the proposed mapping algorithm is 13.8% smaller than that of the conventional mapping algorithm. Moreover, the average transmission delay of the proposed mapping algorithm is 5.2% smaller than that of the heuristic mapping algorithm.

Key words　IEEE 802.11e, EDCA, reinforcement learning, policy gradient

1. まえがき

無線 LAN（Local Area Network）における QoS（Quality of Service）の向上を目的として，IEEE 802.11e [1] が標準化されている．IEEE 802.11e では優先制御方式として，EDCA（Enhanced Distributed Channel Access）方式が規定されている．EDCA 方式では，到着したパケットを優先度に応じて決められたアクセスカテゴリ（AC：Access Category）に分類し，高優先度パケットが分類される AC ほど送信機会が多く与えられる．

送信局に到着したパケットのうち特定の優先度のパケットの割合が，他の優先度のパケットの割合と比較して大きい状況がある．このような状況では，EDCA 方式に従って AC に分類すると，送信局に到着した割合が大きい優先度のパケットが分類される AC のキューに収容されているパケット数が大きくなる．ここで，送信局に到着した割合が大きい優先度のパケットを他の AC に分類すると，パケットが到着してからパケットの送信が完了するまでの時間を短縮できる [2]．

本稿では，それぞれの AP（Access Point）が一定の高優先度パケットを送信する状況を考え，AP 全てが高優先度パケットの送信を完了する時間の最小化を目的とする．AP 全てが高優先度パケットの送信を完了する時間を最小化するために，高優先度パケットを他の AC に分類可能であるとする．文献 [3] で提案された分類則では，キューに収容されているパケット数と，到着したパケットの優先度により分類する AC を決定する．しかしながら，1 つの AP の送信頻度を大きくしても，送信パケットの衝突が起こる可能性が高くなり，他の AP の送信成功確率が小さくなる．そのため，それぞれの AP が自身の高優先度パケットの送信を完了する時間を最小化しても，全ての AP が高優先度パケットの送信を完了する時間を最小化できるとは限らない．AP 同士が協力するために，中央制御局がそれぞれの AP に到着したパケットを AC に分類する．

高優先度パケットが到着した AP の情報だけでなく，他の AP の情報も考慮するため，最適な分類則の獲得は困難である．そのため，最適な分類則を経験的に獲得するために，強化学習を用いることを提案する．更に，強化学習のうち方策勾配法を用いることで学習を進められることを示し，その後学習を進めるための方策を設計する．シミュレーション評価により，提案方式が従来の EDCA 方式と比較して，全ての AP が高優先度パケットの送信を完了する時間が小さいことを示す．更に，学習した分類則についての考察を行う．

本稿の構成は以下の通りである．2. ではシステムモデルについて述べる．3. では，システムモデルを強化学習問題として定式化する．4. では，提案方式のシミュレーション評価を行う．5. で本稿の総括をする．

2. システムモデル

図 1 にシステムモデルを示す．2 組の AP-STA（Station）間の下り回線において，2 つの AP は，EDCA 方式を用いてパケットを送信する．ここで，中央制御局は，AP 1，AP 2 それぞれに到着した VO パケットの分類制御を行う．また，AP 1，AP 2 はそれぞれ，STA 1，STA 2 にパケットを送信する．ただし VO パケットは，従来の EDCA 方式において AC_VO（Voice）に分類されるパケットとし，VO キューは VO パケットが分類される AC のキューとする．同様に VI パケットは，従来の EDCA 方式において AC_VI（Video）に分類されるパケットとし，VI キューは VI パケットが分類される AC のキューとする．

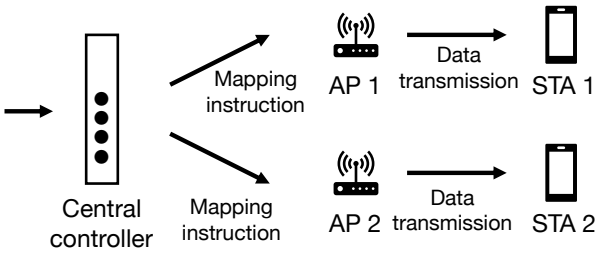

図 1　システムモデル

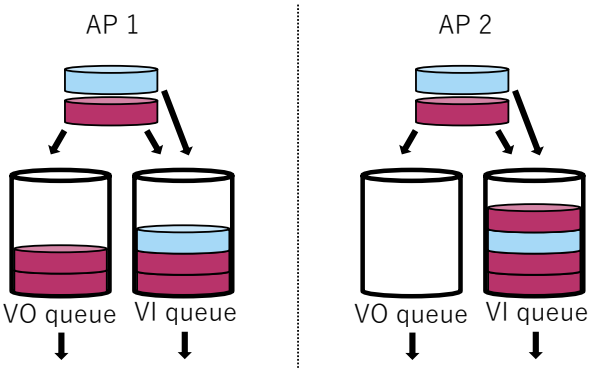

図 2　優先度に応じたパケットの分類制御

図 2 に本システムで提案する分類制御を示す．簡単のため，AC が AC_VO と AC_VI の 2 種類のみを考える．また，VO パケット，VI パケットの到着は，到着率が一定で，ポアソン分布に従うものとする．ここで，到着した VO パケットは AC_VO，AC_VI どちらにも分類可能とする．一方，到着した VI パケットは必ず AC_VI に分類する．

本稿の目的は，システム全体の高優先度パケットの送信遅延時間の期待値を最小化する分類制御の獲得とする．ただし，送信遅延時間は，2 つの AP どちらかに最初の VO パケットが到着してから，全ての AP が N 個の VO パケットの送信を完了するまでの時間とする．

3. 強化学習による分類制御

本章では，強化学習 [4] を用いて，送信遅延時間の期待値を最小化する方法を述べる．そこで，目的関数を送信遅延時間の期待値の符号反転として，目的関数の最大化を考える．まず，状態・行動・報酬を設計し，報酬が過去の状態と行動の系列に依存することを示す．ここで，マルコフ決定過程（MDP：Markov decision processes）を前提とする強化学習は即時報酬が現在の状態，現在の行動，次の状態のみによって決定される．そのた

め，報酬が過去の状態と行動の系列に依存する場合には，MDP を前提とする強化学習は学習を進められるとは限らない．しかし，強化学習の手法のうち方策勾配法を用いることで学習を進めることができる [5]．そのため，本稿では，報酬を最大化するために，方策勾配法を用いる．方策勾配法を用いる中で方策をパラメータ $\boldsymbol{\theta}$ の関数で表すとき，目的関数が $\boldsymbol{\theta}$ の関数となることを示す．目的関数が $\boldsymbol{\theta}$ の関数であることから，パラメータ $\boldsymbol{\theta}$ の更新式を示す．更に，パラメータを $\boldsymbol{\theta}$ とする方策の具体的な設計を行う．

3.1 状態・行動・報酬

状態集合 $\mathcal{S}$ は，次式のように定義する．

$$\mathcal{S} \coloneqq \mathcal{S}_x \times \mathcal{S}_{\mathrm{AP1}} \times \mathcal{S}_{\mathrm{AP2}} \tag{1}$$

ただし，VO パケットが到着した AP を表す集合を $\mathcal{S}_x$ とし，$k = 1, 2$ に対して AP k の状態を表す集合を $\mathcal{S}_{\mathrm{AP}k}$ とする．ここで，集合 $\mathcal{S}_x, \mathcal{S}_{\mathrm{AP}k}$ を次式のように表す．

$$\begin{aligned}
\mathcal{S}_x &= \{1, 2\} \\
\mathcal{S}_{\mathrm{AP}k} &= \mathcal{S}_A \times \mathcal{S}_{Q_{\mathrm{VO}}} \times \mathcal{S}_{Q_{\mathrm{VI}}} \times \mathcal{S}_{C_{\mathrm{VO}}} \times \mathcal{S}_{C_{\mathrm{VI}}}
\end{aligned}$$

ただし，過去に到着した VO パケットの数を表す集合 $\mathcal{S}_A$，VO キューに収容されているパケット数を表す集合 $\mathcal{S}_{Q_{\mathrm{VO}}}$，VI キューに収容されているパケット数を表す集合 $\mathcal{S}_{Q_{\mathrm{VI}}}$，AC_VO のバックオフカウントを表す集合 $\mathcal{S}_{C_{\mathrm{VO}}}$，AC_VI のバックオフカウントを表す集合 $\mathcal{S}_{C_{\mathrm{VI}}}$ は次式のように表す．

$$\begin{aligned}
\mathcal{S}_A &= \{0, 1, \ldots, N-1\} \\
\mathcal{S}_{Q_{\mathrm{VO}}} &= \{0, 1, \ldots, \mathrm{Q_{VOmax}}\} \\
\mathcal{S}_{Q_{\mathrm{VI}}} &= \{0, 1, \ldots, \mathrm{Q_{VImax}}\} \\
\mathcal{S}_{C_{\mathrm{VO}}} &= \{0, 1, \ldots, \mathrm{CW_{max}}[3]\} \\
\mathcal{S}_{C_{\mathrm{VI}}} &= \{0, 1, \ldots, \mathrm{CW_{max}}[2]\}
\end{aligned}$$

ここで，$\mathrm{Q_{VOmax}}, \mathrm{Q_{VImax}}$ はそれぞれ VO キュー，VI キューそれぞれに収容できる最大のパケット数を表す．$\mathrm{CW_{max}}[3], \mathrm{CW_{max}}[2]$ は AC_VO，AC_VI それぞれのコンテンション・ウィンドウ（CW：Contention Window）の最大値を表す．状態は，VO パケットが到着したときに観測し，$n = 1, 2, \ldots, 2N$ に対して，n 個目の VO パケットが到着したときの状態を $s_n \in \mathcal{S}$ と表すこととする．

行動は，AC として AC_VO と AC_VI のみを考えているため，到着した VO パケットをその AP の AC_VO に分類するか，AC_VI に分類するかの 2 通りである．そのため，行動集合 $\mathcal{A} = \mathcal{A}_{\mathrm{AP1}} \cup \mathcal{A}_{\mathrm{AP2}}$ とする．ただし，$k = 1, 2$ に対して，$\mathcal{A}_{\mathrm{AP}k}$ は次式のように定義する．ここで，行動 $\mathrm{VO}_{\mathrm{AP}k}$，行動 $\mathrm{VI}_{\mathrm{AP}k}$ はそれぞれ，到着した VO パケットを AP k の AC_VO，AC_VI のへの分類制御を表す．

$$\mathcal{A}_{\mathrm{AP}k} \coloneqq \{\mathrm{VO}_{\mathrm{AP}k}, \mathrm{VI}_{\mathrm{AP}k}\} \tag{2}$$

行動は VO パケットが到着したときに行い，$n = 1, 2, \ldots, 2N$ に対して，n 個目の VO パケットが到着したときの行動を $a_n \in \mathcal{A}$ と表すこととする．

報酬は，最大化を考えるため，送信遅延時間の符号反転とする．ここで，送信遅延時間は，N 個の VO パケットそれぞれが到着したときの状態と行動全てに依存する．つまり，状態と行動の系列を $\tau \coloneqq (s_1, a_1, s_2, a_2, \ldots, s_{2N}, a_{2N})$ としたとき，送信遅延時間は $t_N(\tau)$ と表される．したがって，報酬を $r_N(\tau)$ とすると，$r_N(\tau) = -t_N(\tau)$ となる．さらに目的関数は，報酬の期待値であるため，$\mathbb{E}_\tau[r(\tau)]$ と表される．

3.2 方策勾配法における更新式

τ はパラメータを $\boldsymbol{\theta}$ とする方策 $\pi_{\boldsymbol{\theta}}(a \,|\, s)$ に依存するため，目的関数 $\mathbb{E}_\tau[r(\tau)]$ は $\boldsymbol{\theta}$ の関数として，次式のように表される．

$$J(\boldsymbol{\theta}) = \mathbb{E}_\tau[r(\tau)] \tag{3}$$

また，パラメータ $\boldsymbol{\theta}$ の更新式は，勾配法を用いて次式のように表される．ただし，η は学習率である．

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \eta \nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\theta}^{(k)}\right) \tag{4}$$

ここで，$J(\boldsymbol{\theta})$ の勾配は，ベースライン $b \in \mathbb{R}$ を導入して次式のように表される [6]．

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_\tau\left[(-t(\tau) - b)\left(\sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_n \,|\, s_n)\right)\right] \tag{5}$$

更に，$(-t(\tau) - b)\left(\sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_n \,|\, s_n)\right)$ の分散が 0 となるようなベースライン b は，モンテカルロ近似すると次式となる [7]．

$$b = \frac{-\sum_{m=1}^{M} t(\tau^m)\left(\sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_n^m \,|\, s_n^m)\right)^2}{\sum_{m=1}^{M}\left(\sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_n^m \,|\, s_n^m)\right)^2} \tag{6}$$

ただし，m 回目のエピソードにおいて，n 個目のパケットが到着したときの状態と行動をそれぞれ $s_n^m \in \mathcal{S}$，$a_n^m \in \mathcal{A}$ とし，m 回目のエピソードにおける状態と行動の系列を $\tau^m = \{s_1^m, a_1^m, s_2^m, a_2^m, \ldots, s_{2N}^m, a_{2N}^m\}$ とする．式 (5) における期待値は，解析的に求めるのが困難である．そのため，モンテカルロ近似により求めたうえで，パラメータ $\boldsymbol{\theta}$ の更新式を導く．式 (5) は再度モンテカルロ近似することで，式 (7) となる．

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx -\frac{1}{M}\sum_{m=1}^{M} t(\tau^m) \sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_n^m \,|\, s_n^m) \tag{7}$$

式 (7) を式 (4) に代入することにより，次式となる．

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \frac{\eta}{M}\sum_{m=1}^{M} t(\tau^m) \sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}^{(k)}}(a_n^m \,|\, s_n^m) \tag{8}$$

3.3 方策の設計

方策 $\pi_{\boldsymbol{\theta}}(a \,|\, s)$ は，ソフトマックス関数を用いて次式のように

表す [8]．ここで，$\boldsymbol{\theta}$ と $\boldsymbol{\phi}(s,a)$ の次元数は等しい．また，$\boldsymbol{\phi}(s,a)$ は特徴ベクトルであり，任意に設計できる．

$$\pi_{\boldsymbol{\theta}}(a\,|\,s) \coloneqq \frac{\exp\Bigl(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(s,a)\Bigr)}{\displaystyle\sum_{b\in\mathcal{A}}\exp\Bigl(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(s,b)\Bigr)} \tag{9}$$

このとき，$\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a\,|\,s)$ は次式のように表される [8]．

$$\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a\,|\,s) = \boldsymbol{\phi}(s,a) - \sum_{b\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(b\,|\,s)\boldsymbol{\phi}(s,b) \tag{10}$$

次に，特徴ベクトル $\boldsymbol{\phi}(s,a)$ の設計を行う．$S_1,\ldots,S_{10}\in\mathbb{N}\cup\{0\}, S_0\in\mathcal{S}_x$ を用いて，状態 $s=(S_0,S_1,\ldots,S_{10})\in\mathcal{S}$，行動 $a\in\mathcal{A}_{\mathrm{AP}S_0}$ とする．ここで，a,S_0 と $S_1,\ldots,S_{10}$ は性質が異なる．a,S_0 はそれぞれ，到着した VO パケットを分類した AC と，VO パケットが到着した AP の種別を表す．一方 $S_1,\ldots,S_{10}$ は，パケット数やバックオフカウントの大きさを表す．

$\boldsymbol{\phi}(s,a)$ は，s,a の関数であるため，a,S_0 が変化すると $\boldsymbol{\phi}(s,a)$ の値が変化することを考える．a,S_0 はそれぞれ，到着した VO パケットを分類した AC と，VO パケットが到着した AP の種別を表すため，a,S_0 がそれぞれ $\alpha\in\mathcal{A}_{\mathrm{AP}S_0}, \chi\in\mathcal{S}_x$ に一致するときにだけ $\mathbf{0}$ にならない関数 $\boldsymbol{\phi}_{\alpha,\chi}(s,a)\in\mathbb{R}^d$ を $\boldsymbol{\varphi}(S_1,S_2,\ldots,S_{10})\in\mathbb{R}^d$ を用いて次式のように表す．ただし，$d\in\mathbb{R}$ である．

$$\boldsymbol{\phi}_{\alpha,\chi}(s,a) = \boldsymbol{\varphi}(S_1,S_2,\ldots,S_{10})\mathbb{1}(S_0=\chi\wedge a=\alpha) \tag{11}$$

$\boldsymbol{\theta}\in\mathbb{R}^{4d}, \boldsymbol{\phi}(s,a)\in\mathbb{R}^{4d}$ は，$\boldsymbol{\theta}_{\alpha,\chi}\in\mathbb{R}^d$，$\boldsymbol{\phi}_{\alpha,\chi}(s,a)\in\mathbb{R}^d$ を用いて次式のように表す．ただし，$\mathbb{1}(\cdot)$ は指示関数である．

$$\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_{\mathrm{VO}_{\mathrm{AP}S_0},\mathrm{AP1}}\\ \boldsymbol{\theta}_{\mathrm{VO}_{\mathrm{AP}S_0},\mathrm{AP2}}\\ \boldsymbol{\theta}_{\mathrm{VI}_{\mathrm{AP}S_0},\mathrm{AP1}}\\ \boldsymbol{\theta}_{\mathrm{VI}_{\mathrm{AP}S_0},\mathrm{AP2}}\end{bmatrix}, \boldsymbol{\phi}(s,a) = \begin{bmatrix}\boldsymbol{\phi}_{\mathrm{VO}_{\mathrm{AP}S_0},\mathrm{AP1}}(s,a)\\ \boldsymbol{\phi}_{\mathrm{VO}_{\mathrm{AP}S_0},\mathrm{AP2}}(s,a)\\ \boldsymbol{\phi}_{\mathrm{VI}_{\mathrm{AP}S_0},\mathrm{AP1}}(s,a)\\ \boldsymbol{\phi}_{\mathrm{VI}_{\mathrm{AP}S_0},\mathrm{AP2}}(s,a)\end{bmatrix} \tag{12}$$

ここで，$\boldsymbol{\varphi}(S_1,S_2,\ldots,S_{10})$ は，$S_1,S_2,\ldots,S_{10}$ の関数であるため，$S_1,S_2,\ldots,S_{10}$ が変化すると $\boldsymbol{\varphi}(S_1,S_2,\ldots,S_{10})$ の値が変化することを考える．$S_1,S_2,\ldots,S_{10}$ は 3 種類以上の値をとるため，$S_0=\chi, a=\alpha$ のとき，$\boldsymbol{\theta}_{\alpha,\chi}^{\mathrm{T}}\boldsymbol{\phi}_{\alpha,\chi}(s,a)$ を D 次多項式で表す．そのため，$\boldsymbol{\varphi}(S_1,S_2,\ldots,S_{10})$ は式 (13) のように表し，$\varphi_{\sigma_1,\ldots,\sigma_{10}}(S_1,S_2,\ldots,S_{10})\in\mathbb{R}$ は式 (14) のように表す．

$$\boldsymbol{\varphi}(S_1,S_2,\ldots,S_{10}) = \begin{bmatrix}\varphi_{0,0,\ldots,0}(S_1,S_2,\ldots,S_{10})\\ \varphi_{1,0,\ldots,0}(S_1,S_2,\ldots,S_{10})\\ \vdots\\ \varphi_{0,0,\ldots,D}(S_1,S_2,\ldots,S_{10})\end{bmatrix} \tag{13}$$

$$\varphi_{\sigma_1,\ldots,\sigma_{10}}(S_1,S_2,\ldots,S_{10}) = \prod_{k=1}^{10}\left(S_k'\right)^{\sigma_k} \tag{14}$$

ここで，$k=1,2,\ldots,10$ に対して S_k' は，学習を進めるための定数 γ_k,δ_k を用いて次式のように S_k を変数変換したものである．

$$S_k' = \gamma_k(S_k+\delta_k) \tag{15}$$

表 1　シミュレーション諸元

送信遅延時間を求める際の試行回数	1000
パラメータ $\boldsymbol{\theta}$ の更新回数 K	100
各 AP が送信する VO パケットの数 N	10
エピソード数 M	1,000
学習率 η	10^{-4}
VO パケットの到着率 λ_{VO}	$5\times10^5\,\mathrm{s}^{-1}$
VI パケットの到着率 λ_{VI}	$2.5\times10^5\,\mathrm{s}^{-1}$
スロットタイム σ_s	$9\,\mu$s
データの送信時間 t_{DATA}	$248\,\mu$s
ACK 信号の送信時間 t_{ACK}	$24\,\mu$s
ある AC における AIFS の時間 AIFS[AC]	$16\,\mu$s
式 (13) における次元 D	2
式 (14) における定数 γ	1/5
式 (14) における定数 δ	1

表 2　EDCA 方式におけるパラメータ

	$\mathrm{CW}_{\min}$	$\mathrm{CW}_{\max}$	AIFSN
AC_VO	3	7	2
AC_VI	7	15	2

4. シミュレーション評価

4.1 シミュレーション諸元

表 1 にシミュレーション諸元を示す．ここで，表 1 のうち，データフレームの送信時間 t_{DATA}，ACK（Acknowledgement）信号のフレームの送信時間 t_{ACK}，AIFS（Arbitration Inter Frame Space）時間を次のように求める．PHY（Physical）レートを 54 Mbit/s，MSDU（MAC Service Data Unit）を 1500 B とすると，$t_{\mathrm{DATA}}=248\,\mu\mathrm{s}, t_{\mathrm{ACK}}=24\,\mu\mathrm{s}$ となる [9]．また，EDCA 方式において，各 AC における CW の最小値と最大値 $\mathrm{CW}_{\min},\mathrm{CW}_{\max}$ と AIFSN（AIFS Number）は表 2 の通りである．更に，SIFS（Short Inter Frame Space）は，$\mathrm{SIFS}=16\,\mu$s とすると，AIFS は次式となる [10]．

$$\mathrm{AIFS[AC]} = \mathrm{SIFS} + \sigma_s\times\mathrm{AIFSN[AC]} = 34\,\mu\mathrm{s} \tag{16}$$

4.2 比較方式の方策

比較方式として，2 つの方式を示す．従来の EDCA 方式は全て AC_VO に分類するため，従来の EDCA 方式で用いる方策 $\pi^{(1)}(a\,|\,s)$ は次式の通りとする．

$$\pi^{(1)}(a\,|\,s) = \mathbb{1}(a=\mathrm{VO}_{\mathrm{AP}S_0}) \tag{17}$$

VO パケットが到着した AP のキューに収容されているパケット数を用いたヒューリスティックな方式を考える．ヒューリスティックな方式は，次式で示す方策 $\pi^{(2)}(a\,|\,s)$ に従う方式とする．方策 $\pi^{(2)}(a\,|\,s)$ は，パケット数が少ない方の AC に分類す

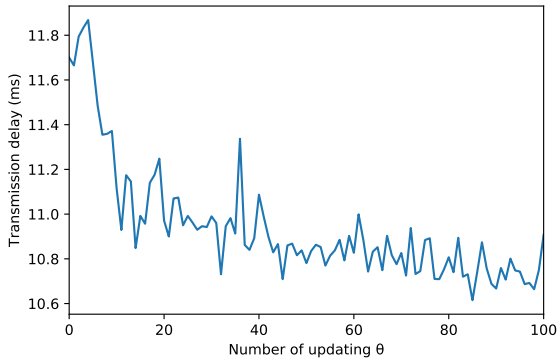

図 3　提案方式における学習曲線．パラメータ $\boldsymbol{\theta}$ を更新するにつれて送信遅延時間を短縮した．

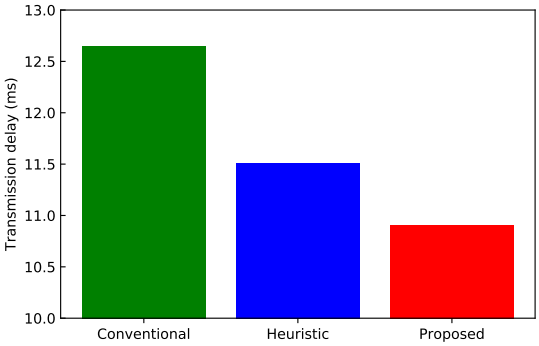

図 4　各方式における送信遅延時間

る方策である．ただし，AC_VO と AC_VI のパケット数が同じとき，送信される機会が多い AC_VO に分類する方策である．

$$\pi^{(2)}(a\,|\,s) = \begin{cases} \mathbb{1}(Q_{\mathrm{VO}} \leq Q_{\mathrm{VI}}), & a = \mathrm{VO}_{\mathrm{AP}S_0}; \\ \mathbb{1}(Q_{\mathrm{VO}} > Q_{\mathrm{VI}}), & a = \mathrm{VI}_{\mathrm{AP}S_0}. \end{cases} \tag{18}$$

4.3　計 算 手 順

方策勾配法の計算手順は次の通りである．

（1）　$\boldsymbol{\theta}^{(0)} = \mathbf{0}$ と初期化する．

（2）　式 (9) に従って，方策 $\pi_{\boldsymbol{\theta}}(a\,|\,s)$ を計算する．

（3）　方策 $\pi_{\boldsymbol{\theta}}(a\,|\,s)$ を用いてシミュレーションを行い，$t(\tau^m)$，$\varpi_{\boldsymbol{\theta}}(\tau)$ を求める．

（4）　式 (6) に従って，b を計算する．

（5）　式 (7) に従って，$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ を求める．

（6）　式 (4) に従って，$\boldsymbol{\theta}$ を更新する．

（7）　手順 (2) から手順 (6) を K 回繰り返す．

ただし，$\varpi_{\boldsymbol{\theta}}(\tau) = \sum_{n=1}^{2N} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_n\,|\,s_n)$ とする．

4.4　シミュレーション結果

図 3 に学習曲線を示す．方策勾配法を用いて学習を進めるにつれて，送信遅延時間を短縮した．

図 4 に提案方式と比較方式のシミュレーションの結果を示す．提案方式は，従来の EDCA 方式と比較して 13.8% 送信遅延時間を短縮した．更に，提案方式は，ヒューリスティックな方式と比較して 5.2% 送信遅延時間を短縮した．

図 5 は，AP 1 に到着した VO パケットを AC_VO に分類する確率 $\pi_{\boldsymbol{\theta}}(\mathrm{VO}_{\mathrm{AP1}}\,|\,s)$ と VI キューに収容されているパケット数 S_3 の関係を示している．図 5 から，AC_VO，AC_VI それぞれにおいて，AC のキューに収容されているパケット数が大きいほど，その AC に分類される確率は小さくなることがわかる．分類した VO パケットを送信するまでの時間は，他のパケットを送信する数が少なくなることで，小さくなる．そのため，送信遅延時間を短縮できたと考えられる．

図 6 は，AP 1 に到着した VO パケットを AC_VO に分類する確率 $\pi_{\boldsymbol{\theta}}(\mathrm{VO}_{\mathrm{AP1}}\,|\,s)$ と AP 2 に到着したパケット数 S_6 の関係を示している．図 6 から，AP 1，AP 2 それぞれにおいて，

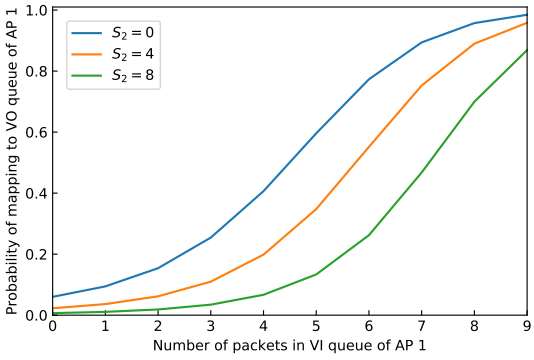

図 5　AP 1 に到着した VO パケットを AC_VO に分類する確率 $\pi_{\boldsymbol{\theta}}(\mathrm{VO}_{\mathrm{AP1}}\,|\,s)$ と AP 1 の VI キューに収容されているパケット数 S_3 の関係．AP 1 の VO キューに収容されているパケット数 S_2 が $S_2 = 0, 4, 8$ の 3 通りの場合を示した．ただし，$k = 1, 4, 5, \ldots, 10$ に対して $S_k = 3$ とする．キューに収容されているパケット数に応じて方策が決定する．

過去に到着した VO パケットの数が大きいほど，到着した VO パケットが AC_VO に分類される確率は小さくなることがわかる．到着したパケット数が小さい AP は，到着したパケット数が大きい AP と比べて，10 個の VO パケットの送信を完了するまでの時間が大きいと考えられる．そのため，到着したパケット数が小さい AP が多くの送信機会を得られるようにするために，過去に到着した VO パケットの数が大きいほど，到着した VO パケットが AC_VO に分類される確率は小さくなったと考えられる．

5.　む　す　び

中央制御局により AP に到着した VO パケットの分類制御を行い，分類則の獲得方法として強化学習を提案した．強化学習における状態・行動・報酬を設計し，本システムにおける方策勾配法の更新式を示した．更に，方策勾配法において，学習を進めるうえで適切な特徴ベクトルを設計した．シミュレーション評価により，提案方式が比較方式と比べて送信遅延時間が小さくなることを示した．

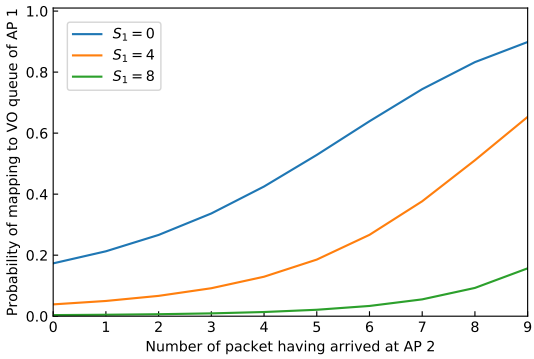

図 6　AP 1 に到着した VO パケットを AC_VO に分類する確率 $\pi_{\boldsymbol{\theta}}(\mathrm{VO}_{\mathrm{AP1}} \mid s)$ と AP 2 の過去に到着したパケット数 S_6 の関係．AP 1 の過去に到着したパケット数 S_1 が $S_1 = 0, 4, 8$ の 3 通りの場合を示した．ただし，$k = 2, \ldots, 5, 7, \ldots, 10$ に対して $S_k = 3$ とする．過去に到着したパケット数に応じて方策が決定する．

謝　　辞

本研究の一部は JSPS 科研費 18H01442 によるものである．

文　　献

[1] “Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Medium access control (MAC) enhancements for quality of service (QoS),” IEEE Std. 802.11e-2007.

[2] R. Sadek, A. Youssif, and A. Elaraby, “MPEG-4 video transmission over IEEE 802.11e wireless mesh networks using dynamic-cross-layer approach,” Natl. Acad. Sci. Lett., vol.38, no.2, pp.113–119, Feb. 2015.

[3] C.H. Lin, C.K. Shieh, C.H. Ke, N.K. Chilamkurti, and S. Zeadally, “An adaptive cross-layer mapping algorithm for MPEG-4 video transmission over IEEE 802.11e WLAN,” TeleCommun. Syst., vol.42, no.3–4, pp.223–234, July 2009.

[4] R.S. Sutton and A.G.Barto, Reinforcement Learning: An Introduction, MIT, 1998.

[5] J. Peters and S. Schaal, “Policy gradient methods for robotics,” Proc. IROS 2006, pp.2219–2225, Beijing, China, Oct. 2006.

[6] J. Peters and S. Schaal, “Reinforcement learning of motor skills with policy gradients,” Neural netw., vol.21, no.4, pp.682–697, Feb. 2008.

[7] M.P. Deisenroth, G. Neumann, J. Peters, et al., “A survey on policy search for robotics,” Found. Trends Robot, vol.2, no.1–2, pp.1–142, Aug. 2013.

[8] R.S. Sutton, D.A. McAllester, S.P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” Proc. NIPS 2000, pp.1057–1063, Denver, CO, USA, Nov. 2000.

[9] 守倉正博，久保田周治，改訂版 802.11 高速無線 LAN 教科書，インプレス R&D，東京，2010.

[10] S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz, and L. Stibor, “IEEE 802.11 e wireless LAN for quality of service,” Proc. European Wireless 2002, pp.32–39, Florence, Italy, Feb. 2002.