地産地消型電力ネットワークの為の Natural Actor-Critic を用いた自動取引エージェントの構築

Design of autonomous trading agent using natural actor-critic method to realize a locally produced and consumed electric energy network

谷口 忠大　高木 圭太　榊原 一紀　西川 郁子
Tadahiro Taniguchi　Keita Takagi　Kazutoshi Sakakibara　Ikuko Nishikawa
立命館大学
Ritsumeikan University

Abstract: In this paper, we describe an adaptive trading agent which can sell and buy electric power effectively in a locally produced and consumed electric energy network, ECONET (Electric Power Cluster Oriented Network). The trading agents manage the amount of electric power generated by solar panels or other renewable energies and stored in a storage battery in a minimal cluster. The agent learns a trading strategy by maximizing future cumulative reward based on reinforcement learning method. Especially, we build autonomous trading agents based on the natural actor-critic method, which is a type of natural policy gradient methods. Several experiments show that the adaptive trading agents can reduce useless energy consumption and a deficiency in most cases.

1　はじめに

産業革命以来，近代社会は化石燃料の依存度を高め続けてきた．しかし，近年，二酸化炭素の排出に伴う地球温暖化や，石油などのエネルギー資源の枯渇などが徐々に現実的な問題として認識されるようになり，環境への負荷が小さく，原理的に枯渇することのない再生可能エネルギーへの期待が高まっている．これに伴い，各国では多様な新エネルギーの導入に向け様々な政策が実行され始めている [1]．日本でもエネルギー基本計画において「１つのエネルギー源に過度に依存することなく，供給途絶リスクの小さいエネルギーを中心に，エネルギー源の多様化を図る.」「循環型社会の形成に資するための施策を推進する.」としているが [2]，このような喫緊の課題に対し，知能情報処理技術が如何に貢献していけるかは重要な問題である．本研究では重要なエネルギー供給システムの一つである電力システムを扱う．特に，松本らにより提案された ECO ネットと呼ばれる分散型の電力ネットワーク構想における電力融通方式を問題とする [3]．近年，米国から情報通信とエネルギー技術の融合を図る，スマートグリッドについての議論が盛んである [4]．しかし，情報通信を既存の配電網に加えるだけでは，再生可能エネルギー普及を取り巻く問題の本質的解決にはならず，直流配電を含めた分散型の地域電力ネットワークのアーキテクチャの導入と，情報化・知能化を併せて構築する，自律分散型直流スマートグリッドとでも呼ぶべきものの普及が重要となる．

本論文の目的は ECO ネット上での電力取引モデルを構築したうえで，取引戦略を最適化するように学習を進めるエージェントを電力融通装置としての電力ルータ上に設計する手法を構築する事である．これにより再生可能エネルギーを自律分散的に授受するシステムについて検討する．学習エージェントについては強化学習を適用可能なように取引過程の定式化を行い，具体的な学習手法としては方策勾配法の一種である Natural Actor-Critic[5] を採用し，その有効性を検証することを目的とする．

2　研究背景

2.1　再生可能エネルギーに基づく分散型電力ネットワーク

化石燃料は地球上に偏在し，その輸送や発電に大規模なシステムを必要とすることから，経営には膨大な資本と設備・管理技術を必要とする．それ故に発電事業者は独占・寡占化され経済学的にも健全な競争市場を形成することが困難であり，独占経済におけるミクロ経済学的な非効率性等が問題視されてきた．日本でも’90 年代には電力市場の自由化が課題として取り組まれてきた経緯がある [6, 7]．また，配電においても海岸や山奥などの人里離れた発電所で大規模に発電し各住戸．事業所に送電されるという集中管理型システムが形成されてきており，この長距離送電の為に発生する電力ロスも大きい．エネルギー・環境問題は経済のみの問題ではなく化石資源が持つ偏在性は社会・政治システム的にも大きな問題を有していることが指摘されている [8].

これに対し，太陽光，風力等に代表される再生可能エネルギーは地球上に遍く存在する事が特徴である．また，発電とその管理には大規模なインフラを整備する必要が無いため，インフラ産業に多く見られる程の極端な規模の経済は働きにくく，資源国から脈々と続く化石資源の連鎖 [8] からも独立して存立し得る点が特徴的である．その為に，集中管理型の電力ネットワークではなく，自律分散的に多数存在する発電拠点をつなぎ合わせる分散型電力ネットワークの構想が再生可能エネルギー利用には適合していると考えられている [3].

マイクログリッド，ECO ネットといった分散型電力ネットワークは従来の大規模発電・長距離送電方式と比較して，電力の送電ロスの回避や発電による熱エネルギーの有効活用，環境負荷が小さい，また災害による大規模停電を起こりにくくするリスクの分散化ができるなどのメリットがあるが，現在既に敷かれている集中管理型の系統電力網との共存を如何に果たすかが問題で

あり研究が進められている．

2.2 分散型電力ネットワーク

再生可能エネルーギーによる発電は分散的に行うことが可能であり，太陽光発電においても各世帯レベルでの導入が図られている．しかし，一方で太陽光・風力に代表される再生可能エネルギーによる発電は非定常的であり，各世帯レベルの電力消費パターンに合わせて平準化することは困難である．これに対し，主に二つのアプローチがなされて来た．一つは従前の配電網を前提とした電力会社による電力供給と買取りである．日本では RPS 法，ドイツ等では Feed-in Tariff といった法整備がなされ再生可能エネルギーの普及に一役を買っている．しかし，これには交流の系統電力を不安定化させる逆潮流の問題が存在し，再生可能エネルギーが既存の系統電力に対し無視しうる割合である時期は問題ないが，一定の位置を占めようとした場合には問題となる．もう一つは，地域の電力網を一つの単位として管理しようとするマイクログリッドの考え方である．八戸や愛地球博で実証実験が行われてきた．しかし，交流の配電を基礎とし，系統電力との整合性をとろうとするマイクログリッドでは，システムの全体設計が複雑であり，システム設計後の改変が容易ではない．

これに対し，松本らは ECO ネットと呼ばれる太陽光発電と直流主体の自律分散型の電力ネットワークを提案している [3]．ECO ネットでは地域のネットワークを系統電力網からは基本的に独立させ，直流でつなぐ[1]ため，交流特有の同時同量の原則は存在せず適応的なシステム改変が可能であり，システム設計後の新たな発電・消費拠点の追加が可能である．太陽光発電，蓄電池，電力ルータを持てば，電力ネットワークに接続できるという意味で，インターネット類似のシステムとして提案されている．電力会社による買取りのアプローチが電力系統をバッファとして捉えるのに対して，マイクログリッドや ECO ネットでは地域ネットワークに蓄電池を持つ事でバッファとする．しかし，各世帯が単一の蓄電池を設置し全てを賄うのはコスト的に不可能である．よって，発電・消費拠点間で各発電性能や消費パターンの時間的特性に合わせて適切に融通しあう必要がある．

多くのマイクログリッドの実証実験では融通自体を集中管理的に行ってきたが，各世帯・事業所が蓄電池や発電装置を投資により設置しネットワークに接続することで運用することを前提とする ECO ネットでは集中管理の視点では無く各参加者の投資回収可能性の視点から考えなければならない．トップダウンな集中管理により余剰電力を無償で融通されてしまうような電力システムでは，投資回収を困難にし，結果的に再生可能エネルギー機器に対する投資意欲を削ぎ，再生可能エネルギーの普及を妨げてしまう可能性がある．そこで，本研究では ECO ネット上で各発電・需要者間の電力のやり取りを自由化し相互に売買する事の出来る電力融通制度を考える．

[1]広義には DC マイクログリッドという概念に含まれる．

2.3 電力融通の自動化

しかしながら，電力売買を各世帯単位で行う事は新たな問題を生む．各世帯において誰が売買を行うかという問題である．各世帯で誰か一名を電力売買のデイトレードの為に割くという発想は非現実的である．そこで，本研究では電力融通を担う電力ルータを知能化し，自動的に電力売買を行わせることを考える．株取引やオプション取引の領域で自動的に売買を行うシステムトレードの方法は研究されており，その様な研究は人工市場と呼ばれているが，電力取引への適用事例は少ない [9]．さらに，この電力ルータ上のエージェントに経験に基づく学習を行わせる事で，よりその状況に合わせた電力取引戦略を獲得する事を目指す．ECO ネットにおいて各世帯・事業者にとって最適な取引戦略は発電量，消費量，蓄電池の容量，属するネットワークの他の参加者の戦略などの諸条件により明らかに変化する．故に固定的な取引戦略では十分ではなく，電力ルータ上の学習エージェントは接続されたネットワーク，地域に合わせた取引戦略を獲得することが重要となる．

ECO ネットはこれまでの集中管理型の系統電力網やマイクログリッドと異なり，普及の容易さを確保するために，新規参入者が発電機，蓄電池，電力ルータを持つことで新たなミニマルクラスタを逐次的に自由に追加しうるシステムである点が特徴となっている．そのため，新たな電力装置の設置や電力の売買という意思決定はミニマル・クラスター所有者の所有権の下に行われる必要がある．故に，ミニマル・クラスターの所有者である経済主体の利得最大化を行うのが電力取引を委任されている各電力ルータ上の学習エージェントの目的となるべきである．そこに電力ロスの低減という付加的なペナルティを加える事で，全体としての電力ロス最小化を副次的に目指そうというのが ECO ネット上の自動電力売買システムの目的となる．

本研究では，強化学習に基づき ECO ネット上で電力ルータが適応的に取引を行う全体像をモデル化し，Natural Actor-Critic によって電力ルータの学習則を構築し，その有効性を検証する．

3 地産地消型電力ネットワークと電力市場モデル

3.1 ECO ネットの概要

ECO ネットとは松本らによって考え出された次世代の電力アーキテクチャの構想であり，電力クラスターを指向したネットワーク (Electric Power Cluster Oriented Network) の略である [3]．図 1 に ECO ネットの概念図を示す．ECO ネットには発電・消費する複数のミニマル・クラスターが存在しており，それらが自らの持つ蓄

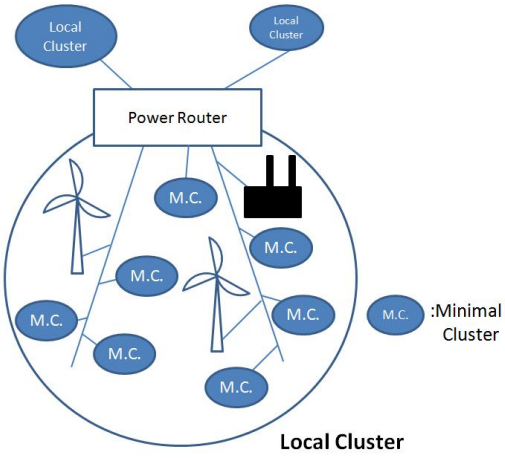

図 1: ECO ネットの概念図

電装置に電力を蓄え，また余剰・不足分については電力ルータを通じて電力を融通しあう．このミニマル・クラスターが地域社会内で各々の家庭や工場に相当する．既存の電力システムでは大規模発電を行う電力系統を頂点とし需要家を底辺に持つ放射状系統が基本であるのに対し，ECO ネットでは分散型を成している．従来の電力ネットワーク構想が，少数の発電事業者と多数の末端消費者という二分法であったのに対して，全ての参加者を基本的に生産消費者 (prosumer)[10] として捉えるのが ECO ネットの特徴の一つである．ここで，ECO ネットに登場する概念の説明を行う [3].

電力クラスター ECO ネット上での電力クラスターとは，巨大な単一システムではなく，複数の電力システム (太陽光発電や，風力発電，燃料電池など) からなる電力のネットワークで構成されたものである．電力クラスターは基本的に電力を自給自足するものとするが，それぞれ発電と需要に過不足が生じても良いとする．また，電力クラスターに存在しうるものは必ずしも発電機能，需要機能を備えていなくても良い．発電だけを行うクラスターや，蓄電のみ，あるいは電力の消費のみを行うクラスターというものも例外的に考えられる．

電力ルータ 電力クラスターをノードとして，電力クラスター間の送電線をリンクとして，電力ネットワークが形成される．そして，電力クラスター間の電力のやり取りを制御する機能が電力クラスターに備わっていなければならない．この機能を持つ装置を電力ルータと呼ぶ．電力ルータの基本機能は電気エネルギーのルーティングであり，どの近隣クラスターを対象とするのか，エネルギーを送り出すのか受け取るのか，電力量の制御などである．本研究では電力売買における資金管理も電力ルータの対象であるとする．

ミニマル・クラスター 電力ネットワークの最小単位を，ミニマル・クラスターと呼ぶ．構成要素は，各種電気機器，給電線，蓄電装置，発電装置，そして電力ルータである．このミニマル・クラスターは一般家庭や，小規模工場，共同住宅などが想定される．従来の電力系統が単一方向の電力エネルギーの流れを想定しているのに対し，ミニマル・クラスター同士の双方向の電気エネルギーのやり取りを基本とする．

ローカル・クラスター ミニマル・クラスターを数個から一万個程度ネットワーク接続した電力クラスターを，ローカル・クラスターと呼ぶ．ローカル・クラスターの構成要素は，中小規模発電装置（火力，水力，風力など）を持ち，そのローカル・クラスター内の電力が不足したミニマル・クラスターを支援し，近隣のローカル・クラスターとの電力の融通を行うために電力ルータを備える．本研究における電力売買では市場として決済情報を集約する役割もローカル・クラスターが担うものとする．

ECO ネットの構想では電力融通の手法については定めていない．本論文では前章に述べたように，各ミニマル・クラスターに適応的な電力取引エージェントを導入し電力売買を通して市場原理に則った電力融通を実現する事を考える．

3.2 電力取引市場の設計

電力市場の自由化は従来の電力系統下でも進められてきた．発電事業者の電力の価格は総括原価主義の規制下でムダなコストまで料金に上乗せされていたのが，電力の自由化が導入されれば競争によって発電コストが下がり，社会的効用の増大とコストを下げた発電者への利潤を増大できること等が期待されてきた [1]．再生可能エネルギーを主とした ECO ネット下では，必ずしも独占企業の存在は前提とされないため，自由な売買市場を需給バランスに則って設計することが基本的には望ましいと考えられる．本節では本論文で採用する電力取引制度及び取引方法について述べる．

3.2.1 取引制度の設計

本研究における ECO ネットでは各ミニマル・クラスターは電力ルータを通じて余剰電力を融通し合う．このとき，各ミニマル・クラスターは，自らの利潤を最大化しようとするという行動基準により取引を行うものとする．北欧４カ国で機能しているノルドプール (NordPool) [1] という電力取引所を参考として取引制度を設計する．この取引制度では発電会社，最終需要家，配電会社，大口のトレーダーやブローカー等の市場参加者が取引市場が開かれる日の１時間ごとの自社が望む需要，供給予定表を前日正午に通知する．取引市場ではこれを集計し，市場全体の需要曲線と供給曲線を描き，両曲線の交点で取引量，均衡価格を決定する．これをシステム価格とい

う．これを参考にしつつ，本研究のモデルでは各ミニマル・クラスターが需要・供給曲線を市場に毎時提出する事により電力取引を行う．

まず，電力市場におけるミニマル・クラスターのエージェント集合を $M = \{A_i : i = 1, \ldots, n\}$ とする．エージェント A_i は自らのミニマル・クラスターの蓄電量や時刻等の諸条件を勘案し，市場に取引条件を出力する．取引条件は売買量 s_i に対する単位電力量あたりの価格 $P_i(s_i)$ という形で市場機能を担うローカル・クラスターの電力ルータに提出される．ここでは購入の場合 s_i を正，販売の場合は負とする．電力取引市場ではすべてのエージェントの取引条件を積み上げる事によってローカル・クラスター内における需要曲線と供給曲線を決定する．その上で，両曲線の交点を計算することで，その時間の市場での電力量の単位当たり価格と各エージェントの総取引量を決定して電力売買の決済を行うものとする．図 4 に入札から決済に至る概略図を示す. 詳細は 4.3 節にて説明する．各市場参加者が注文数量と価格を市場に提出し，そのマッチングをとることにより取引を行う方式を板寄せ方式と呼ぶが，本方式は連続的な価格と注文数量の関係を関数として出力するが広い意味では板寄せ方式の一種と言える．全エージェントが入札を一斉に行った後に市場価格により各エージェントの電力取引量と電力単価が決定される為に，自らが「売り」か「買い」かもこの時点で決定する．これによりエージェントはローカル・クラスター電力市場で売買することが可能となる．具体的な数式については後述する．

3.2.2 マルチエージェント系における適応的電力取引

ミニマル・クラスターは発電装置を備えた一般の家庭を想定し，秒から時間オーダーで変動する非定常な発電を行う再生可能エネルギーを電力源として想定する．また，電気二重層コンデンサなどの蓄電装置を持つ事を想定するが，蓄電池はまだまだ大きく，かつ高価であるため，一家庭が必要十分な蓄電を行う事は出来ないと考えられる．そのため適切なタイミングで頻繁に効率的な取引を行い，ローカル・クラスター内で平準化する事が期待される．頻繁な人手による取引を行うことは現実的ではないため，前述したように自動取引エージェントの導入を検討する．

多様な金融市場については市場を計算機上でモデル化し，結果を生み出す要因やプロセスを把握しようとする研究が盛んに行われるようになってきている．このような市場モデルを構築する手法は人工市場と呼ばれ株式や先物市場など多くの市場が研究の対象とされている [11, 12, 9, 13]．また，実用的な自動取引についてもシステムトレードについての一連の研究開発がなされてきている [14]．電力市場についての研究では特に既存の電力系統における電気事業者の価格決定を議論する場合が多く [15, 16, 17]，ECO ネットのように生産消費者の集合として電力市場を捉えた場合の研究や，時々刻々と発電・消費量が変化する中での時間的な取引戦略の獲得についての研究は未だ少ない．

定常的な発電を行う原子力発電など大規模な事業者による発電が主と捉えられたシステムでは時間単位での状態変化はクリティカルな問題では無い．しかし，再生可能エネルギーの非定常性の特質および，蓄電装置の容量制約などを考慮すると，状態変化を有する系としてモデル化する必要がある．本研究ではマルコフ決定過程 (MDP: Markov Decision Process) として各エージェントにとってのダイナミクスを捉える事で強化学習に基づく取引戦略の獲得を行う．

市場のような各々が利益を最大化しようと強化学習を行う集団はマルチエージェント強化学習の系として捉えられる．しかし，この場合，複数のエージェントが同時進行的に学習することにより各エージェントにとっての環境の MDP 性が崩れ，学習が困難となる同時学習問題の存在が知られている．例えば，マルチエージェント強化学習では，他のエージェントの行動目標や行動・状態などを観測できないことが多く，部分観測マルコフ決定過程 (Partially Observable MDP : POMDP) となっている [18]．これらの問題に対応するために，profit-sharing を行う方法やタスク構造に合わせた報酬分割を行う手法などが提案されている [19, 20].

これに対し，通常の MDP に対する強化学習手法であっても，方策を直接探索する方策勾配法 [21, 22] と呼ばれる手法は Q-learning や SARSA などの価値関数から間接的に方策を決定する手法に比べ POMDP 下でも頑強であると言われている．方策勾配法の中でも近年では方策勾配の方向にパラメータ空間の情報幾何的な特性を考慮した自然方策勾配法が注目されている [23, 24]. 本研究では，特に，自然方策勾配法の一種として Peters らが提案した Natural Actor-Critic[5] に基づいて取引戦略の適応的改善を実装する．Natural Actor-Critic は旧来の方策勾配法よりも学習が効率的であることが知られている．

4 Natural Actor-Critic による適応的取引エージェントの構築

本章では ECO ネットにおける適応的取引エージェントの構築を行う．まず，用いる強化学習手法である Natural Actor-Critic の導入を行い，次に ECO ネットでの取引過程を強化学習で扱いうる形に定式化する．最後に Natural Actor-Critic を取引エージェントに適用する．

4.1 方策勾配法

MDP に基づく強化学習では，エージェントが時間 t において環境の状態 $x_t \in \boldsymbol{X}$ を観測した後に方策 $\pi(x, u)$ に従い行動意思決定を行い行動 $u_t \in \boldsymbol{U}$ を環境に出力した後に遷移後の環境の状態 x_{t+1} を観測し，同時に報酬 r_t がエージェントに与えられる [25] という一連の流れを前提にしてエージェントが学習を行う．エージェン

トは将来に亘る獲得報酬の期待値を最大化するように学習を行う．ここで，MDP とは次状態の生起確率と報酬の期待値が現在の状態と行動のみに依存する事を指す．本節では本論文でエージェントが行動戦略を獲得するための学習方法である Natural Actor-Critic の基礎となる方策勾配法の枠組みについて述べる．強化学習には主に Q-learning や SARSA 等といった行動価値関数を推定することで最適方策を求める価値に基づく方法と，Actor-Critic 法のように最適方策を直接推定する方策に基づく方法がある．後者の方策勾配法では確率的な方策を陽に表現し最適方策を直接近似するため，Q-learning などの価値に基づく方法のように行動価値関数の変化に対して急激に方策が変化する様なことが無い．そのため評価関数を徐々に増加させるように，逐次的に方策自体を直接的に改善していけるため，ある程度の POMDP 下でも学習を進める事が出来ると言われている．また，方策勾配法では直接方策を近似するため，行動価値に基づく方法に比べ方策の出力値を連続値として設計する事が容易である．本論文が問題とする系においては行動出力が電力取引の価格指定に相当し，この値は連続的に変化する．これも本論文が方策勾配法を採用する理由の一つである．

方策勾配法では次のように収益の期待値を $J(\boldsymbol{\theta})$ で表す．

$$J(\boldsymbol{\theta}) = E\{(1-\gamma)\sum_{t=0}^{\infty}\gamma^t r_t \mid \boldsymbol{\theta}\} \quad (1)$$

$$= \int_{\boldsymbol{X}} d^{\pi_\theta}(x)\int_{\boldsymbol{U}}\pi_{\boldsymbol{\theta}}(x,u)r(x,u)dxdu \quad (2)$$

$$d^{\pi_\theta} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(x_t = x) \quad (3)$$

$J(\boldsymbol{\theta})$ は将来に亘って得られる累積報酬の割引和になっており，この評価値を最大化するように方策を最適化する事が強化学習の課題となる．ここで，方策 $\pi_{\boldsymbol{\theta}}(x,u) = p(u \mid x;\boldsymbol{\theta})$ はパラメータベクトル $\boldsymbol{\theta}$ によって表される方策であり，d^{π} は割り引かれた状態分布である．次に J をパラメータベクトル $\boldsymbol{\theta}$ に関して偏微分し $J(\boldsymbol{\theta})$ の勾配を求める．

方策勾配法では，この勾配ベクトル $\nabla_{\boldsymbol{\theta}}J$ を用いて方策のモデルパラメータ $\boldsymbol{\theta}$ を次のように更新する事で局所最適な方策を探索する．

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}) \quad (4)$$

ここで $\alpha(>0)$ は学習率，勾配 $\nabla_{\boldsymbol{\theta}}J$ は収益の期待値におけるパラメータ $\boldsymbol{\theta}$ の最急上昇方向を表すので方策勾配法は少なくとも局所最適解を得ることができる事が知られている．

4.2 Natural Actor-Critic

収益の期待値をパラメータ θ に関して偏微分した $\nabla_{\boldsymbol{\theta}}J$ は収益の期待値における θ の最急上昇方向を表すとされるが，パラメータ空間が歪んでいるときにはプラトー現象などにより学習性能が悪化する事が知られている．このプラトー問題を回避する有効な手段として提案された手法が自然勾配法である．自然勾配法では通常の勾配にフィッシャー情報行列 F の逆行列を乗ずる事で自然勾配を求めるが，強化学習においてはこの自然勾配が比較的容易に計算しうることが示されている [24]．Natural Actor-Critic とは自然勾配法を方策勾配法に適用した自然方策勾配法の一つである．自然勾配はアドバンテージ関数を $\nabla_{\boldsymbol{\theta}}\log\pi$ を基底関数として推定したときの係数ベクトルに一致することが知られている．アドバンテージ関数とは行動価値と状態価値の差分によって表される関数であり，状態価値の寄与を除いた行動の価値を表現する関数である [26]．具体的な実装としては価値関数の推定方法として LSTD-Q(λ) を用いた無限区間タスクでの Peters らによって提案された Natural Actor-Critic[5] を用いる．LSTD-Q(λ) は Bradtke らによって提案された価値関数の推定手法であり，旧来の TD 学習が TD 誤差に基づいて逐次的に価値関数を更新するのに対し，価値関数が基底関数の線形和で表されることを利用し最小自乗法に基づき解く手法である [27, 28]．その解は TD 学習により得られる最適解と共通であることが知られている．Natural Actor Critic では行動価値関数の一部であるアドバンテージ関数を正確に求める必要があるため，統計量から最適な価値関数を直接求める事の出来る LSTD-Q(λ) を用いている[2]．アルゴリズムは以下である．

まず，パラメータ表現された方策 $\pi(x,u) = p(u \mid x,\boldsymbol{\theta})$(初期パラメータ $\boldsymbol{\theta} = \boldsymbol{\theta}_0$)，その導関数 $\nabla_{\boldsymbol{\theta}}\log\pi(x,u)$，状態価値関数 $V^{\pi}(x)$ のパラメータ化のための基底関数 $\boldsymbol{\phi}(x)$ とおく．

1. 初期状態 $x_0 \sim p(x_0)$, パラメータを選択 $\mathbf{A}_{t+1} = \mathbf{0}, \mathbf{b}_{t+1} = \mathbf{z}_{t+1} = \mathbf{0}$.
2. t=0,1,2,... で以下 (3.~8.) 繰り返し
3. 行動 $u_t \sim \pi(x_t,u_t)$ を決定，次の状態 $x_{t+1} \sim p(x_{t+1} \mid x_t,u_t)$ とそれにともなう報酬 $r_t = r(x_t,u_t)$ を観測.
4. LSTD-Q による状態価値評価 (Critic 器):
   価値関数を

$$V^{\pi}(x_t) = \boldsymbol{\phi}(x_t)^{\top}\boldsymbol{\nu}_t \quad (5)$$

   として，Actor に適合するアドバンテージ関数の近似式を

$$f^{\pi}_{\boldsymbol{\omega}}(x_t,u_t) = \nabla_{\boldsymbol{\theta}}\log\pi(u_t \mid x_t)^{\top}\boldsymbol{\omega}_t \quad (6)$$

   とする．

[2] 自然方策勾配法の定式化において価値関数を逐次的に求める形のものも試されているがそのような実装は学習が不安定化する事が多い [29].

5. 基底関数の更新:

$$\tilde{\phi}_t = [\phi(x_{t+1})^\top, \mathbf{0}^\top]^\top \quad (7)$$

$$\hat{\phi}_t = [\phi(x_t)^\top, \nabla_\theta \log \pi(x_t, u_t)^\top]^\top \quad (8)$$

6. 十分統計量の更新:

$$\mathbf{z}_{t+1} = \lambda \mathbf{z}_t + \hat{\phi}_t \quad (9)$$

$$\mathbf{A}_{t+1} = \mathbf{A}_t + \mathbf{z}_{t+1}(\hat{\phi}_t - \gamma \tilde{\phi}_t)^\top \quad (10)$$

$$\mathbf{b}_{t+1} = \mathbf{b}_t + \mathbf{z}_{t+1} r_t \quad (11)$$

7. Critic のパラメータ更新

$$[\boldsymbol{\nu}_{t+1}^\top, \boldsymbol{\omega}_{t+1}^\top]^\top = \mathbf{A}_{t+1}^{-1} \mathbf{b}_{t+1} \quad (12)$$

8. Actor のパラメータ更新

アドバンテージ関数のパラメータベクトルで表される自然勾配 ω_t が時間枠 W_h で収束したとき (例えば $\forall \tau \in [0, ..., W_h]$ に対し, $\cos(\angle(\boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}_{t-\tau})) \geq \epsilon$) 方策パラメータを更新した上で，十分統計量を減衰させる．

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \boldsymbol{\omega}_{t+1} \quad (13)$$

$$\mathbf{z}_{t+1} \leftarrow \beta \mathbf{z}_{t+1} \quad (14)$$

$$\mathbf{A}_{t+1} \leftarrow \beta \mathbf{A}_{t+1} \quad (15)$$

$$\mathbf{b}_{t+1} \leftarrow \beta \mathbf{b}_{t+1} \quad (16)$$

9. 終了

本研究ではこの Natural Actor-Critic に基づく方策勾配法を用いて自動取引エージェントを構築する．

また，マルチエージェント系に於いては方策更新を非同期にする事で同時学習問題を緩和できるという研究もなされているが，Peter らの手法を用いると，方策更新は価値関数の収束に伴い非同期的に行われる為，このような非同期的な方策更新を人為的に作り込むことなく自動的に生み出す事が出来る．この点からも Natural Actor-Critic は本論文が問題にするようなマルチエージェント強化学習課題に有効であると考えられる．

4.3 ECO ネットへの適用

本節では実験のための ECO ネットにおける取引制度と各ミニマル・クラスターのモデル化，及びエージェントの学習器のモデル化について述べる．

4.3.1 モデルの全体像

図 2 に ECO ネットのミニマル・クラスターの発電・消費からエージェントの取引条件の意思決定，市場取引，そしてその結果からの収益と電力量の流入，に至る情報の流れを簡単に示す．まず i 番目のエージェント A_i の時刻 t での発電量，消費量を $E_i(t), C_i(t)$ とする．ここでその差分 $D_i(t) = E_i(t) - C_i(t)$ がミニマル・クラスター

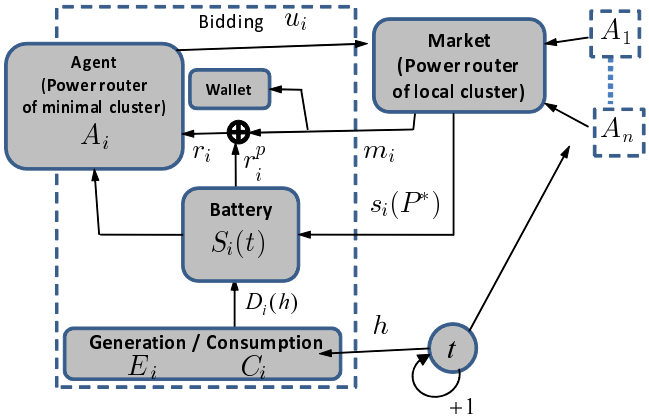

図 2: 本稿のモデルにおける ECO ネットの電力取引全体プロセス

内部での蓄電池への流入量として重要になる．また，時刻 t における売買による流入は A_i の購買量 $s_i(t)$ と等しくなる[3]．各ミニマル・クラスターは蓄電池を有しており時刻 t において蓄電残量を $S_i(t)$ とする．また，その最大，最小容量を S_i^{max}, S_i^{min} とする．これらの流入により蓄電残量 $S_i(t)$ は次式に基づいて毎時刻変化する．

$$\bar{S}_i(t) = S_i(t-1) + s_i(t-1) + D_i(t) \quad (17)$$

$$S_i(t+1) = \max((\min(\bar{S}_i(t), S_i^{max}), S_i^{min}) \quad (18)$$

取引は 2 時間に一度行われるものとし 2 時間毎に t は 1 加算される．h は t の周期成分を取り出した 24 時間周期の値であり，具体的には h は t を 12 で割った余りである．本研究では太陽光発電の下で一定の生活パターンで居住者が生活する系を想定し，系はノイズ程度の変化を除き 24 時間周期の周期性を持つものとする．これらに基づき，エージェントは毎時刻，状態変数 $x_i = (S_i, h)$ を観測出来るものとする．この状態変数に基づきエージェントは取引条件の意思決定を行い，行動出力として $u_i = (P_i^{buy}, P_i^{sell})$ を出力する．

P_i^{buy}, P_i^{sell} は市場に提出する需要・供給曲線の内それぞれ最大限買う，最大限売る場合の価格であり，このときの価格 p に対する購入量 s を決定する入札曲線は図 3 の様になる．購入量 s の負の部分は売却量を表す．

$$s_i(p) = \begin{cases} S_i^{min} - S_i(t) & (p > P_i^{sell}) \\ \frac{S_i^{min} - S_i^{max}}{P_i^{sell} - P_i^{buy}}(p - P_i^{sell}) + (S_i^{min} - S_i(t)) & \\ \qquad (P_i^{sell} \geq p \geq P_i^{buy}) & \\ S_i^{max} - S_i(t) & (p < P_i^{buy}) \end{cases} \quad (19)$$

エージェントは毎時刻 P_i^{buy}, P_i^{sell} を市場に出力する事で，図 3 に示すような入札曲線を生成する．入札曲線は価格から数量への関数であるが，関数を上下限を持つ一次関

[3]簡単の為，今回は送電ロスや蓄電ロスは考慮しない．流入に係数などをかけることにより拡張は可能である．

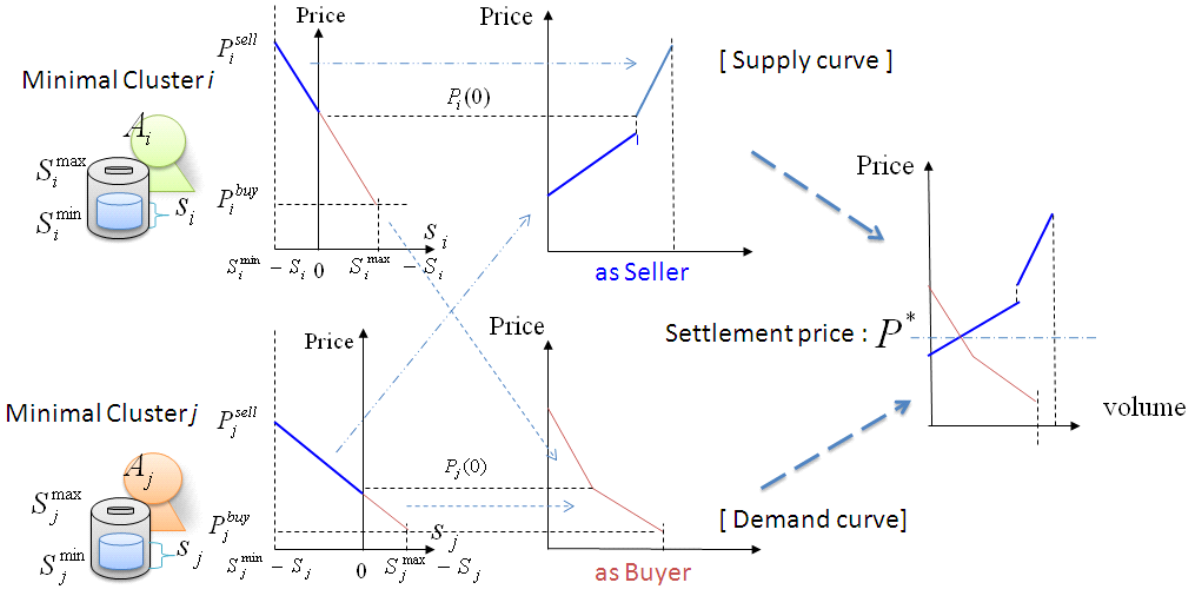

図 4: 市場決済の概念図

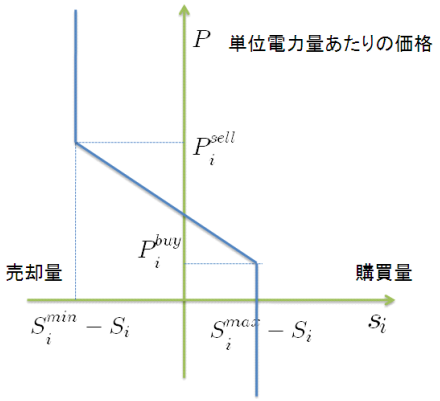

図 3: 入札曲線の例

数に制約することで，端点二点の出力で入札曲線の出力に替える事ができる．板寄せ方式の市場では取引数量と価格を各取引主体が出力し，これを合わせる事で全体での需要・供給曲線が形成されるが，当市場でも同様の事を行う．図 4 に示すように，各エージェントは P_i^{buy},P_i^{sell} を出力する事で取引条件を市場に出力するが，取り引き数量 s_i について正の部分が買い手として，負の部分が売り手としての取引条件となる，これらを数量の方向に足し合わせる事で需要曲線・供給曲線が生成される．これらの曲線の交点を求める事で決済される際の価格 P^* が求まる．具体的には数式 (20) を満たす均衡価格 P^* を決済価格とする．

$$\sum_{s.t.A_i \in M} s_i(P^*) = 0 \tag{20}$$

時刻 t での市場での決済価格を $P^*(t)$ とすると，エージェント A_i の売買収支は $m_i(t) = -s_i(t) * P^*(t)$ となる．決済後，各エージェントに購買量 $s_i(t)$ と売買収支 $m_i(t)$ が流入し，一サイクルが終了する．

報酬 r_t は売買収支 m_i と電力超過・不足ペナルティ r_i^p の和とし，r_i^p は

$$r_i^p(t) = \begin{cases} -(m^+ + \zeta^+ |\bar{S}_i(t) - S_i^{max}|) & (\bar{S}_i(t) > S^{max}) \\ -(m^- + \zeta^- |S_i^{min} - \bar{S}_i(t)|) & (\bar{S}_i(t) < S_i^{min}) \end{cases}$$

ここでそれぞれの定数は超過・不足自体に係るペナルティ $m^+, m^- > 0$，単位電力あたりのペナルティ $\zeta^+ > 0, \zeta^- > 0$ である．2 章で記述したように，あくまで電力ルータは個々のミニマル・クラスターの所有の下に有るため，個々のミニマル・クラスターの利得最大化に基づいてエージェントの学習が進行する事を前提としている．

このように系をモデル化する事で，他のエージェントの方策が一定であるとしたとき，この問題は他のエージェントの蓄電残量を隠れ状態とし，周期的に変化する時間と自らの蓄電残量を状態とした強化学習課題として定式化される．これにより様々な強化学習手法を ECO ネット上の自動取引課題に適用する事が可能となる．

4.3.2 学習器の設定

入札曲線の学習について述べる．1 エージェントについて述べるので，簡単の為，エージェントの添え字 i は省略する．方策関数 $\pi(x, u; \boldsymbol{\theta})$ を方策パラメータ $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2\}$ とガウス関数を用いて下の様にモデル化した．

$$\pi(x, u; \boldsymbol{\theta}) = \prod_{k=1,2} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(\frac{-1}{2} \frac{(u_k - f_k(x, \boldsymbol{\theta}_k))^2}{\sigma_k^2}\right) \tag{21}$$

f_1, f_2 は行動出力分布の中心を表わし，それに標準偏差 σ_1, σ_2 のガウスノイズが足される事で，探索を行いながら価格決定がなされる．ここで $u_1 = P^{buy}$，$u_2 = P^{sell}$ とする．ここで，価値関数 $V^{\pi}(x)$ および f_k を共通の基底関数 ϕ を用いて下記のように表わす．

$$V^{\pi}(x) = \phi(x)^{\top}\nu \tag{22}$$

$$f_k(x;\theta_k) = \phi(x)^{\top}\theta_k \tag{23}$$

また，アドバンテージ関数の係数ベクトル ω については，上記方策関数の定義に従い，式 (6) に基づき定義する．ここで基底関数 $\phi(x)$ の第 j 成分 $\phi_k(x)$ はクロネッカーのデルタ δ_{ij} を用いて，

$$\phi_k(x) = \begin{cases} \delta_{kh} & (0 \le k < 12) \\ \rho^{k-12} & (12 \le k \le 15) \end{cases} \tag{24}$$

$$\rho = \frac{S_{max} - S}{S_{max} - S_{min}} \tag{25}$$

と表わされるとする．上段のクロネッカーのデルタで表される部分は周期的に変化する時間に依存する項，下段の項は電池残量への依存性を三次の多項式で表現している．この前提の下で自然勾配計算の為に $\nabla_{\theta}\log\pi(x,u)$ を計算する．θ_k の第 j 成分を θ_{kj} として，

$$\frac{\partial}{\partial\theta_{kj}}\log\pi = \frac{\phi_j}{\sigma_k^2}\varepsilon_k \tag{26}$$

$$\frac{\partial}{\partial\sigma_k}\log\pi = \frac{(\varepsilon_k^2 - \sigma_k^2)}{\sigma_k^3} \tag{27}$$

となる．ただし $\varepsilon_1 = P^{buy} - f_1, \varepsilon_2 = P^{sell} - f_2$ である．

よって，Natural Actor-Critic における各の基底関数の更新については

$$\tilde{\phi}_t = [\phi(x_{t+1})^{\top}, \mathbf{0}^{\top}]^{\top} \tag{28}$$

$$\hat{\phi}_t = [\phi(x_t)^{\top}, \nabla_{\theta}\log\pi(x_t,u_t)^{\top}]^{\top} \tag{29}$$

より，$\phi' = \phi(x_{t+1}), \phi = \phi(x_t)$ として，

$$\tilde{\phi}_t = [\phi'^{\top}, \mathbf{0}^{\top}]^{\top} \tag{30}$$

$$\hat{\phi}_t = [\phi^{\top}, \frac{\varepsilon_1}{\sigma_1^2}\phi^{\top}, \frac{\varepsilon_2}{\sigma_2^2}\phi^{\top}, \frac{\varepsilon_1^2 - \sigma_1^2}{\sigma_1^3}, \frac{\varepsilon_2^2 - \sigma_2^2}{\sigma_2^3}]^{\top} \tag{31}$$

となる．その後，Natural Actor-Critic アルゴリズムにそって状態価値関数，アドバンテージ関数の近似式のパラメータベクトルを更新し，ある方向に収束したと判定されれば方策を更新する．収束判定は時間 $t+1$ と時間 $t-\tau$ の間のそれぞれのパラメータベクトルのなす角の cos 値が十分 1.0 に近い閾値 ϵ に対し $\cos\theta = \frac{\langle\omega_{t+1}, \omega_{t-\tau}\rangle}{\|\omega_{t+1}\|\|\omega_{t-\tau}\|} > \epsilon$ を満たせば収束したとみなし，方策パラメータを更新する．

5 シミュレーション実験

実験では，仮想的なローカル・クラスターを計算機環境に構築し，本手法の有効性を検証する．

5.1 実験条件

本モデルに於ける学習エージェントにとって発電量 E_i と消費量 C_i は直接的には意味が無く，蓄電池への流入量 D_i のみが問題となる．本モデルでは一日毎の発電・消費の周期性を仮定しているので，ミニマル・クラスター i の時間 t における流入量 $D_i(t)$ を以下の様に sin 波形を用いて近似する．

$$D_i(t) = -g_i * cos(2\pi\frac{t + a_i}{12}) + b_i + n_i(t) \tag{32}$$

ここで g_i，a_i，b_i，$n_i(t)$ はそれぞれゲイン，パターンの時間方向のずれ，発電が消費に勝る度合い，ノイズ項となっている．本実験では g_i に対して 5% のノイズ項を一様乱数で生成し付加する．

また，学習しないエージェントの方策，及び学習エージェントの初期方策は残存蓄電量を参照しながら，十分に残量がある際には安く，少なくなった際には高く値付けするよう下式に設定した．

$$P^{sell} = 40 - 20\rho \tag{33}$$

$$P^{buy} = 20 - 10\rho \tag{34}$$

つまり $\theta_{1,12} = 20, \theta_{1,13} = -10, \theta_{2,12} = 40, \theta_{2,13} = -20$ とした．初期方策では時間に対する依存性は自明でないので設計しなかった．さらに価格の出力の変動幅（ガウス分布に従う）は $\sigma_1 = \sigma_2 = 0.5$ としている．また出力する取引条件について $P_i^{sell} > P_i^{buy} > 0$ を制約条件として与える．なお，実験結果は状態遷移，行動等にノイズを含んでいるので各条件に対して 5 回実験を行い，その平均を示している．

実験条件として 1 日の取引回数は 12 回，学習率 $\alpha = 1 \times 10^{-4}$，適格度トレースの割引率 $\lambda = 0.98$，割引率 $\gamma = 0.98$，十分統計量の保持率 $\beta = 0.8$，方策勾配の収束判定のウィンドウ幅 $W_h = 12 \times 7 = 1$[week] 及び収束判定の閾値 $\varepsilon = 0.99$ とした．また，報酬関数の係数 $m^+ = m^- = 1.0 \times 10^2$，$\zeta^+ = 1.5 \times 10^4, \zeta^- = 1.0 \times 10^4$ とした．勾配については更新ベクトルのノルム最大値を 1.0 として，急激な変更がなされないよう制約を与えた．他のパラメータは $g_i = 20, b_i = 0, S_i^{max} = 100, S_i^{min} = 25$ とした．また，初期蓄電量は $S_i(0) = 50$ とする．

各ミニマル・クラスターの電力流入の特性については，操作可能なパラメータが多い．本実験では焦点を絞るために式 (32) の a_i を変化させることで各ミニマル・クラスター毎の違いを設計した．ローカル・クラスターの条件設定としては g_i や b_i に多様性を持たせる事も検討の上では重要である．しかし，一方で a_i が同一で，各ミニマル・クラスターにおいて発電消費パターンが同期してしまった場合，あるミニマルクラスタで余剰が発生するときには他のミニマルクラスタでも余剰が発生し，あるミニマルクラスタで不足が発生した場合には他のミニマルクラスタでも不足が発生すると言うことが起きる．この場合，電力融通において適切な需給関係が構築でき

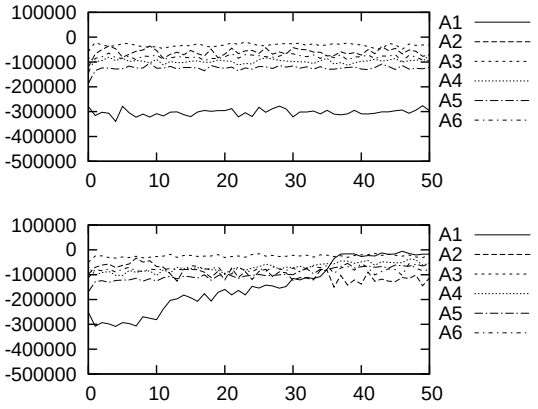

図 5: （上）学習者の存在しない系での各エージェントの獲得報酬遷移,（下）A_1 のみが学習する系での各エージェントの獲得報酬遷移

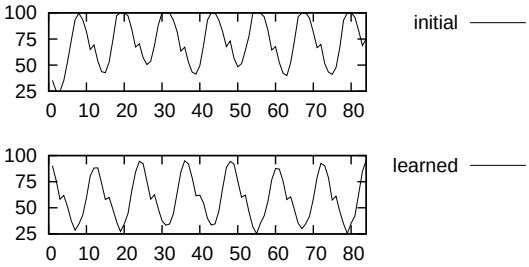

図 6: 上：学習初期のバッテリー残量変化，下：学習終期でのバッテリー残量変化 (ともに横軸は時間ステップ)

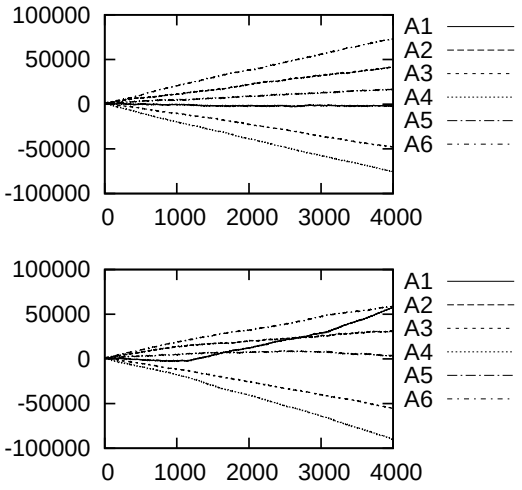

図 7: 各ミニマル・クラスターの所持金の変化,（上）学習エージェントが存在しない場合,（下）エージェント A_1 のみが学習した場合.

ず，如何に学習エージェントが適切な電力売買を行っても電力ロスや停電を防ぐことは出来なくなる．a_i にずれが存在し，発電消費パターンが非同期であることはネットワーク化により平準化が行える前提条件となる．

本実験では 6 つのミニマル・クラスターにより構成されるローカル・クラスターを考える．各ミニマル・クラスターの発電・消費の特性は $(a_0, a_1, a_2, a_3, a_4, a_5) = (0, 2, 4, 6, 8, 10)$ とし，発電消費パターンに十分な非同期生を持つように設定した．この条件下ではローカル・クラスター全体としては全時刻で電力の流入と流出はバランスするため，適切に電力融通が行われれば電力ロスと停電を無くす事が出来る．

5.2 単一エージェント学習環境での実験

まず，提案手法がエージェント 1 体のみが学習する環境下で正常に作動する事を確認するためにエージェント A_1 のみを学習させた．全てのエージェントが学習しない場合の各エージェントの獲得報酬の遷移と，エージェント A_1 のみが学習した場合の各エージェントの獲得報酬の遷移を図 5 に示す．図には 1 日の獲得報酬の各週毎の平均を示している．学習を通して A_1 が停電，電力ロスによる無駄を避けつつ有利な取引を行う事で，より多くの報酬を得ている事が分かる．次に，学習初期と学習終期でのエージェント A_1 の蓄電量の時間変化を図 6 に比較する．初期には $S_1^{max} = 100$ にしばしば達する事で電力ロスを生じていたのが，適切な取引戦略を得ることにより，過不足無く蓄電量をコントロール出来るようになっている事がわかる．これより，本論文で示した定式化により強化学習を通じて取引戦略を学習可能なエージェントが構築できている事がわかる．

また，学習エージェントの存在しない場合と，エージェント A_1 のみが学習している場合での各ミニマル・クラスターの所持金の変化を図 7 に示す．エージェント A_1 が有利な取引を行う事で，電力の無駄をなくすのみならず，自らの所持金をも増大させている事がわかる．

5.3 多数エージェントの同時学習環境での実験

次に，より多くのエージェントが同時に学習する条件下で実験を行った．1 体以上のエージェントが同時に学習する場合は，しばしば部分観測や同時学習の問題が発生する為，学習が正常に進行しない可能性がある．しかし，前述の理由により方策勾配法，特に Natural Actor-Critic はこれらの問題に対して頑健性があると考えられ，多数のエージェントが同時に学習した場合でも強化学習が適切に進行する事が期待される．実験を行った結果，多くの場合で問題なく，学習を行ったエージェントは獲得報酬を増大させる事が出来た．図 8 に 6 体のエージェントが学習を行った場合の各エージェントの獲得報酬の変化を示す．ここでは 6 体ものエージェントが同時学習を行っているにも関わらず，学習はほぼ安定的に進行している．これより Natural Actor-Critic はこのようなローカルな市場取引というマルチエージェント強

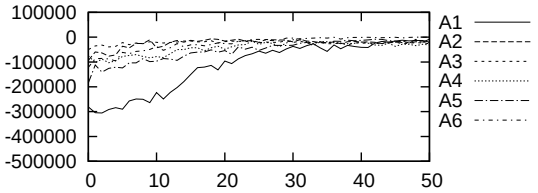

図 8: 全エージェントが同時に学習した場合の各エージェントの獲得報酬変化

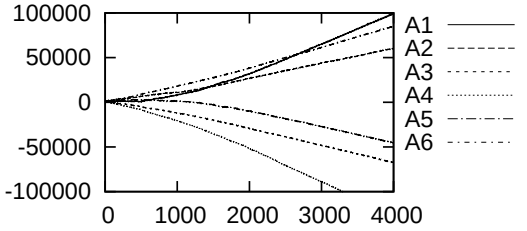

図 9: 全エージェントが学習した場合の各ミニマル・クラスターの所持金の変化

化学習系において有効な手法であることが示されたと言える．

本タスクは相互的な取引環境で競合的に学習を行って居るために，一見全員の報酬が同時に向上する事のないゼロサム・ゲームとなるように感じられるが，実際には全エージェントが獲得報酬を増大させている．これは各エージェントが得る報酬の内，ゼロサムであるのは金銭取引に関わる部分のみであり，電力ロスと停電に関わるペナルティについてはゼロサムでは無いためである．６体のエージェントが学習した場合の各ミニマル・クラスターの所持金の変化を図 9 に示す．学習エージェントが存在しない場合と比べ金額の配分は変化しているが，金銭的報酬についてはゼロサム・ゲームで有る故に全エージェントを通じて増加しているということは無いことが分かる．しかしながら，系の内１体づつエージェントを学習ありに切り替えて行った時には金銭的収益が増加する事が観察された．これは，実際に適応的取引エージェントを組み込んだ電力ルータを市場投入する際に，ミニマル・クラスター所有者が電力ルータの導入を検討するインセンティブになると考えられ重要な点となる．

次に発電した電力のロスに着目して比較を行った．エージェント $A_1 \sim A_k$(k は横軸の値) が学習した際のローカル・クラスター全体での１日の電力ロス和の最終 100 日間平均値を条件毎に図 10 に示す．横軸には何体のエージェントが学習しているかを示しており，縦軸は１日当たりの平均電力ロスを標準偏差と共に示している．これから，各エージェントが学習を行う事でミニマル・クラスターの電力ロスは低下し，その結果，ローカル・クラスター全体としても電力ロスが低減されていっている事がわかる．5 体が学習した場合と 6 体が学習した場合で 6 体の時の方がロスが増大しているように見えるが有意差はなかった．

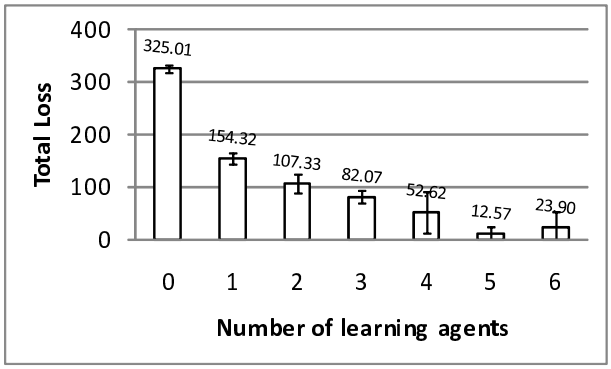

図 10: 学習エージェント数毎のローカル・クラスター全体における学習後の電力ロス（横軸は 6 エージェント中で学習したエージェントの数，縦軸が 1 日間の電力ロス）

これより，各ミニマル・クラスターが電力ロスを低減しようとしながら，自らの利得が最大化されるように取引を行うことで，系全体としても電力ロスを低減することが出来た事が分かる．

この系のようなマルチエージェント課題ではシングルエージェント用に開発された強化学習手法は必ずしもよい学習結果を得る事は出来ないと考えられている．これに対し，Natural-Actor Critic が頑健であった理由としては，方策を直接探索している点，及び，LSTD により勾配方向を決定する為，バッチ的な更新にならざるを得ず，学習過程で各エージェントが断続的な定常状態を繰り替えす為に非同期な更新が暗黙的に実現されているために同時学習問題を回避できているなどの点が考えられる．しかし，解析的な視点から保証が得られている訳ではないので，実際の系に適用する場合にはその動作保証についての検討が必要である．また，本論文では割愛するが，方策勾配法は局所探索手法であるため，方策の初期値によって学習が進行しない場合が存在した．

6 結論と展望

本論文では，再生可能エネルギーの利用を目的とした自律分散型の次世代電力ネットワーク構想 ECO ネットを対象とし，各ミニマル・クラスター間で電力を自動売買するための枠組みについて議論した．その為に，ノルドプールに倣った電力売買を決済する市場を設計し，ローカル・クラスター内に於いて，各ミニマル・クラスターを代表するエージェントが自らの取引条件を提示することにより，自動的に売買を行う事の出来るプロセ

スをモデル化した．また，その取引条件を方策勾配法により学習する枠組みについて提案し，具体的には自然方策勾配法の一種である Natural Actor-Critic を用いて電力自動取引エージェントを構築した．また，仮想的な周期的発電・消費パターンの下でシミュレーション実験を行い，マルチエージェント環境下であっても学習が進行し，学習エージェントを備えたミニマル・クラスターにおける電力ロスの削減と，収益が実現される事が示された．さらにそれを通してローカル・クラスター全体としても電力ロスの低減を図れることが示された．

しかしながら，本論文では扱いきれなかった論点も多くある．本研究の設定では系全体としては発電・消費を外生変数として捉え，余剰・不足に対する変動可能な発電[4]や外部からの電力購買[5]を考えなかった．この点について盛り込んで行くことも今後の課題であると言える．また，各ミニマル・クラスターの住人も，これまでの一方的な消費者の立場から生産消費者へと立場を変え，より明示的に電力利用に基づく収支を認識することで，生活行動パターンを変化させる事も考えられる．これは，数理モデルの中では金銭収支 m_i が，消費パターン C_i に人間の学習を通して変化を与える事として捉えられる．つまり，本論文で扱った学習より一段上位レベルの学習が，当該システムを人間を含んだ系 (Human in the Loop System) として捉えた際には重要となる．実際，環境問題・エネルギー問題を直視した際には，利用の効率化以上に，消費の低減は本質的な解決となり，その際には人間の活動自体を系に含んだ議論はシステム研究の上で不可避となる．

また，本論文の目的は強化学習の定式化が可能なようにローカル・クラスターにおける電力売買をモデル化し，有効な取引戦略の学習手法を提案する事であったため，実際の電力システムの動特性は殆ど考慮せず簡略化して取り扱った．しかしながら，蓄電池における蓄電ロスや時間特性，また取引時の送電ロスなどを考慮することは現実の系への適用時には本質的であり無視する事は出来ない．さらに，今回の系は発電消費の結果としての流入量 D_i を周期性の正弦波で近似できるものとして捉えた，しかしながら，現実の消費・発電パターンはそのように整った形をしていない．これらの実観測データを用い検討を行う事が今後の課題である．また，実際に系が強化学習の対象となる MDP として捉えられる為には，エージェントの状態変数として天候や住人の行動計画など，どこまでを変数として含めるか，また含めるように設計するかが実用上は重要となっている．さらに，電力融通の問題を考えた際には，実際の送電にかかる時間や，その方式についての議論も重要な要素となってくる．本論で検討したのは，既存の電力ネットワークではなく，今後，普及が検討される自律分散型の地産地消型電力ネットワークであり，実際の系と今回のモデル化の差異を明確に示す事は困難である．ゆえに，その将来的な仕様の決定にも参画しつつ，具体的な制約を今後のモデル化に継続的にフィードバックする事が重要だと考えている．

エネルギー問題や環境問題は現代社会が抱える喫緊の課題である．それを解決する未来のシステムを支える技術の一つに強化学習のような知能情報処理技術が組み込まれることは意義深いことであり，知能情報システムにおける新たなマーケットを形成する潜在性を有している．米国で進むスマートグリッドの考え方には，このような知能情報技術との融合は現在のところ含まれておらず，自律分散型の直流地域配電網である ECO ネットとスマートグリッドの考え方を融合させ，さらに知能情報技術を適用する事で構成される，自律分散型直流スマートグリッドとでも呼ぶべき電力ネットワークの構築は大きな可能性を秘めていると考えられる．また，そのシステムの作動は計算機実験を通して構成的に検証可能であり，市場の仕組みや補助金，規制などを含めた制度設計の視点からも重要性を持つ．本研究におけるモデル化を基礎としながら，より発展させることが重要であると考える．

謝辞

本研究は 科学研究費補助金 基盤研究 (B)「不便の効用を活用したシステム論の展開」21360191 の 一部支援を受けた．また，当研究のきっかけを戴いた，ＮＰＯ法人ＫＧＣの柴田有三氏，前京都市副市長の山崎一樹氏，及び多くの助言を下さった松本吉彦氏，また，自律分散型直流スマートグリッド i-Rene プロジェクトの関係諸氏に感謝の意を表す．

参考文献

[1] 八田達夫. 電力競争市場の基本構造. RIETI：独立行政法人　経済産業研究所,2004 年 03 月,No.04-J-029.

[2] 経済産業省 資源エネルギー庁エネルギー戦略推進室. エネルギー基本計画, 2007.

[3] 松本吉彦, 柳父悟. 新世代に向けた電力システム構造のビジョン. 電気学会論文誌Ｂ（電力・エネルギー部門誌）, Vol. 123, No. 12, pp. 1436–1442, 2003.

[4] 山家公雄. オバマのグリーン・ニューディール. 日本経済新聞出版社, 4 2009.

[5] J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots*, 2003.

[6] 電力政策研究会. 図説電力の小売自由化. 電力新報社, 2000.

[4]電力不足に対して各世帯でエンジンを回して発電する事や，エコキュートなど.

[5]主には大規模な電力会社からの系統電力網を通した買電.

[7] 桑原鉄也. 電力ビジネスの新潮流. エネルギーフォーラム, 2008.

[8] ヘルマンシェーア. ソーラー地球経済. 岩波書店, 2001.

[9] 和泉潔. 人工市場 (相互作用科学シリーズ). 森北出版, 2003.

[10] ドン・タプスコット/アンソニー・D・ウィリアムズ. ウィキノミクス マスコラボレーションによる開発・生産の世紀へ. 日経 BP 社, 2007.

[11] 喜多一, 出口弘, 寺野隆雄. U-MART: 経済学と工学をエージェントが結ぶ. 第 10 回マルチ・エージェントと協調計算ワークショップ (MACC2001)(2001.11. 16-17), 2001.

[12] 佐藤浩, 久保正男, 福本力也, 廣岡康雄, 生天目章. 人工市場のシステム構造. 人工知能学会誌, Vol. 15, No. 6, pp. 974–981, 2000.

[13] 吉本昌弘, 藤森成一, 佐々木将士. AHP を導入した Profit Sharing エージェントによる株式売買に関する研究 (セッション 2). 情報処理学会研究報告. MPS, 数理モデル化と問題解決研究報告, Vol. 2006, No. 56, pp. 37–40, 2006.

[14] 松井藤五郎. カブロボへの招待: 人工知能を用いた株式取引. 人工知能学会誌, Vol. 22, No. 4, pp. 540–547, 2007.

[15] G.R. Gajjar, S.A. Khaparde, P. Nagaraju, and S.A. Soman. Application of actor-critic learning algorithm for optimal bidding problem of a Genco. *Power Systems, IEEE Transactions on*, Vol. 18, No. 1, pp. 11–18, 2003.

[16] 下村貴裕, 最所祐一, 藤井康正, 山地憲治. マルチエージェント モデル を 用 いた 電力 市場 における 価格 形成過程 の 分析. 電気 学会 論文 誌 B (電力・エネルギー 部門 誌), Vol. 124, No. 2, pp. 281–290, 2004.

[17] V. Nanduri and T.K. Das. A reinforcement learning model to assess market power under auction-based energy pricing. *IEEE Transactions on Power Systems*, Vol. 22, No. 1, pp. 85–95, 2007.

[18] 荒井幸代. マルチエージェント強化学習: 実用化に向けての課題・理論・諸技術との融合 (¡ 特集¿「マルチエージェント技術における新しい可能性」). 人工知能学会誌, Vol. 16, No. 4, pp. 476–481, 2001.

[19] 宮崎和光, 荒井幸代, 小林重信. Profit Sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察. 人工知能学会誌, Vol. 14, No. 6, pp. 1156–1164, 1999.

[20] 田渕一真, 谷口忠大, 堀口由貴男, 中西弘明, 椹木哲夫. マルチエージェント強化学習における報酬分割による分業形成機構. 第 35 回知能システムシンポジウム, pp. 37–42, 2008.

[21] 木村元, 小林重信. Actor に適正度の履歴を用いた Actor-Critic アルゴリズム: 不完全な Value-Function のもとでの強化学習. 人工知能学会誌, Vol. 15, No. 2, pp. 267–275, 2000.

[22] R.S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. Vol. 12, pp. 1057–1063, 2000.

[23] S. Kakade. A natural policy gradient. *Advances in neural information processing systems*, Vol. 2, pp. 1531–1538, 2002.

[24] J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, Vol. 71, No. 7-9, pp. 1180–1190, 2008.

[25] R.S. Sutton, A.G. Barto, 三上貞芳, 皆川雅章共. 強化学習. 森北出版, 東京, 2000.

[26] L.C. Baird. Advantage updating. 1993.

[27] S.J. Bradtke and A.G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, Vol. 22, No. 1, pp. 33–57, 1996.

[28] J.A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, Vol. 49, No. 2, pp. 233–246, 2002.

[29] 木村元. 適性度の履歴を用いた自然勾配 actor-critic 法. 第 19 回自律分散システムシンポジウム, pp. 67–72, 2007.

谷口忠大（正会員）

1978 年 6 月 24 日生．2006 年京都大学工学研究科博士課程修了．2005 年より日本学術振興会特別研究員(DC2)，2006 年より同(PD)．2007 年より京都大学情報学研究科にて(PD)再任．2008 年より現在，個体と組織に於ける記号過程の計算論的な理解や共生社会に向けた知能情報学技術の応用研究にについての研究に従事．京都大学博士（工学)．計測自動制御学会学術奨励賞，システム制御情報学会学会賞奨励賞，論文賞，砂原賞など受賞．計測自動制御学会，日本人工知能学会，システム制御情報学会，日本神経回路学会などの会員．

高木圭太（非会員）

1984 年 8 月 29 日生．2009 年立命館大情報理工学部知能情報学科卒業．マルチエージェント系における学習についての研究に従事．2009 年より TIS 株式会社にてシステムコンサルタントとして勤務．現在に至る．SCM のシステムコンサルティング事業に関わる．

榊原一紀（非会員）

1999 年神戸大学工学部電気電子工学科卒業, 2004 年神戸大学院自然科学研究科博士後期課程修了，同年立命館大学理工学部助手，2005 年同大学情報理工学部知能情報学科助手を経て，2008 年同講師，現在に至る．博士(工学)．スケジューリング問題のモデル化と解法，進化・学習アルゴリズムの理論と応用に関する研究などに従事．電気学会，システム制御情報学会，スケジューリング学会などの会員．

西川郁子（非会員）

1984 年京都大学理学部物理学科卒業, 1990 年京都大学大学院理学研究科物理学第一専攻博士後期課程単位取得退学．同年神戸大学大学院自然科学研究科助手，1993 年立命館大学理工学部情報工学科助手，1995 年同情報学科助教授，2003 年同教授を経て，2004 年同情報理工学部知能情報学科教授，現在に至る．博士（理学)．知能システムの研究に従事．システム制御情報学会，計測自動制御学会，日本神経回路学会，日本物理学会，電子情報通信学会などの会員．