The evaluation of agile demand response : an applied methodology

Citation for published version (APA):
Babar, M., Nguyen, H. P., Cuk, V., Kamphuis, I. G., Bongaerts, M. A., & Hanzelka, Z. (2018). The evaluation of agile demand response : an applied methodology. *IEEE Transactions on Smart Grid*, *9*(6), 6118 - 6127. Article 7926413. https://doi.org/10.1109/TSG.2017.2703643

DOI:
10.1109/TSG.2017.2703643

Document status and date:
Published: 01/11/2018

Document Version:
Accepted manuscript including changes made at the peer-review stage

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2017.2703643, IEEE Transactions on Smart Grid

The Evaluation of Agile Demand Response: An Applied Methodology

M. Babar, P.H. Nguyen, V. Cuk, I.G. Kamphuis, M. Bongaerts, Z. Hanzelka,

Abstract—**This paper formulates an applied methodology for an agile demand response using mathematical micromodels. The optimal strategy chosen by an aggregator is the maximization of social welfare derived from demand flexibility. The notion of complex demand bidding is already given in the literature, however heretofore it is formulated as the relationship of price with both demand elasticity and marginal cost along with temporal and profit constraints. Although the planning of flexible demand is already handled by using advance learning techniques in literature, herein simple Q-learning technique in a decentralized fashion is proposed. Moreover, trade-offs between the proposed complex bidding rules are explored in a day-ahead market context. Due to the given complex bidding rules and principle of learning, the methodology can be easily applied in active distribution network. Several number of houses, equipped with the proposed complex bidding mechanism and decentralized learning capability, has been simulated, thus illustrating the application of methodology formulated herein.**

Index Terms—**Agile methodology, Complex bidding, Demand response, Multi-agent system, Reinforcement learning.**

I. Introduction

A. Overview

In theory, regulator should try to deliver maximum social welfare to a society, i.e. electricity supply demand balance should follow the path of least-cost over long term development. Recently, due to inevitable uncertainty on the growth of the demand and the generation, demand response (DR) has been introduced as one of the potential solutions to the electricity markets and network issues. DR generally refers as shifting of dispatchable loads by the customers in response to market prices or when network reliability is jeopardized [1]. DR performed by the customers at LV network could be the main driver to shift demand peaks which usually occur in the morning and early evening hours of a routine day. Moreover, DR could enhance the economic efficiency of power systems and market price volatility, reduce the carbon footprint and eliminate the need of committed peaking units in the electricity sector.

To further maximize the social welfare to European society by considering DR benefits, Art. 15.8 in EU Directive 2012/27/EU encourages ISOs to accept DR participation alongside supply in the wholesale market [2]. Following the directive, many EU member states have implemented multiple DR programs by adjusting electricity market rules [3].

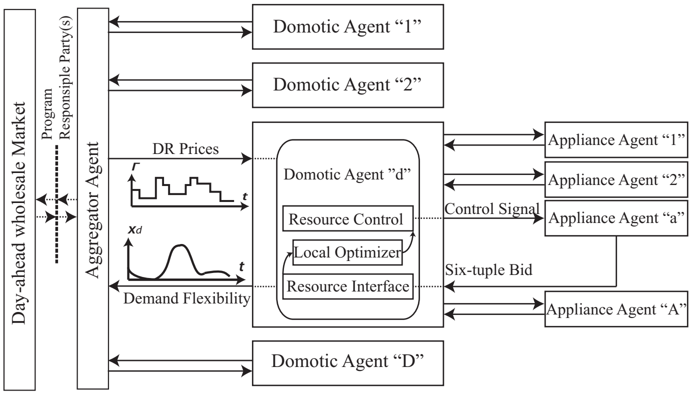

Fig. 1. Block diagram of proposed DR Scheme.

With this notion, explicit DR has been gaining substantial focus for the future utilization of DR benefits. In the literature, the focus on the utilization of DR benefits is limited for a group of stake-holders, for instance; in [4] and [5] DR benefits are determined for retailers and aggregators. On the other hand, [6] and [7] determine DR benefits for distributed system operator (DSO) and transmission system operator (TSO). In general, it is inferred that demand flexibility acts as a social commodity which is a special type of resource attracted by each stakeholder to consume. In this situation, each stakeholder should have fair chance to gain benefit from the demand flexibility. That is why, the concept of agile DR has been developing which does not require significant changes in existing functioning of the stakeholders and would allow fair sharing of DR benefit among all [8]–[10].

Despite the steady stream of market framework for demand dispatch in DR research, very little attention has been paid to developing a mechanism for planning the demand flexibility that also considers fair distribution of DR benefits across the downstream local customers associated with an aggregator [11]. Such a mechanism is important because without it, benefits from demand dispatch may become sub-optimal [10].

Therefore, there is a need for a comprehensive as well as applied methodology that allocates explicit DR benefits across each market players including an aggregator, retailers, DSO, program responsible parties (PRP) and especially the customers. The scope of this paper is the formulation and analysis of agile methodology for explicit DR planning, which ensures the fair distribution of DR benefits to customers and eventually represents their say in the wholesale market.

M. Babar P.H. Nyugen, V. Cuk and I.G. Kamphuis are with Electrical Energy Systems Group of Department of Electrical Engineering at Eindhoven University of Technology, Eindhoven, 5600MB, The Netherlands.
Z. Hanzelka is with the Department of Electrical Engineering in AGH University of Science and TEchnology, Krakow - Poland.
M. Bongaerts is Innovation Manager at Energy Transition Liander, Alliander, Arnhem, The Netherlands.
E-mail: M.Babar@tue.nl

B. Paper Contribution

It has been noted that there are many technical challenges in explicit demand response such as; (i) an adequate DR in peak hours will not only reduce high peak but might also the average day-ahead market price, (ii) overall day-ahead price may also reduce due to efficient use of distributed generation, (iii) unscheduled response (in actual hour of use), thus risking the market players to activate the balancing reserves in that hour of operation. For this reason, without learning demand flexibility for DR planning, the day-ahead price in market may become inappropriate, making the market clearing sub-optimal.

Heretofore these has been no contribution that mathematically formulates the applied methodology for explicit demand response with the concept of agility. In [8], authors presented a concept that fills in the research gap between market-driven bidding model of demand flexibility in multi-agent system and DR planning. For this reason, first time this work formulates an extensive rules for complex bidding in agile DR in combination with the principle of decentralized learning. The mathematical micromodels of multiple type of agents, learning objectives, bidding rules, observations and experiences given in the research would benefit the technical society to explore in the area of agile demand response.

The rest of the article is structured as follows. Section II of the paper provides the analytical description for the application of agile methodology. Section III formulates the optimization problem with an objective to maximize the DR benefit by minimizing the total energy cost of appliance agents associated with domotic agents. Reinforcement technique as an optimization tool is used to learn the DR schedule over day-ahead basis in section IV. Simulation results are given in section V and section VI summarizes the main contribution of the paper in the field of demand response.

II. Agile Methodology

According to EU Directive 2012/27/EU [2] and the proposed recent amendments in the directive [12], the aggregator either independently or mutually with program responsible party(s) can trade demand flexibility to electricity market, as shown in Fig. 1. Although the directive does not comply the aggregator to implant any particular type of explicit demand response for the utilization of demand flexibility at customer level, the system as a whole should ensure maximization of social welfare by considering DR benefits [13]. In general, DR benefit is equal to the difference between the benefit that consumers derive from the reduction of electrical energy and the cost of producing this energy. In specific, when the energy consumption reduces or shifts due to DR during peak hours, electricity supply demand balances at relatively lower cost than without DR.

Two facts are important from the point of evaluating social welfare in traditional demand response. Firstly traditional DR permits a larger demand to be marketed at the lower price during off-peak hours. Consequently, consumers would unquestionably gain from DR. Secondly the net social welfare is indeterminate due to the absence of real-time magnitude of the shifts in demand for different prices and the elasticity of demand. Therefore, in [8], authors proposed agile demand response as an alternative to traditional demand response schemes, as it endeavors to increase social welfare by branching the aggregation of DR benefits into two independent hierarchical nodes, namely domotic and aggregator. As shown in Fig. 1, the former node is capable to respond to unpredictable environment through incremental learning as well as submits complex bids, thus systematically reflects resident DR benefits. Each domotic agent is assumed to be an economic agent that acts in a greedy manner. This means domotic agents only attempt to maximize their own benefits by minimizing the total cost of energy consumed under the economic constraints. On the other hand, appliance agent is a representation of smart appliance and based on the assumption in [14], there is a defined pattern for what appliance agent is willing to bid during different time of day as well as for different use of electricity. Accordingly, the price elasticity of demand can easily be estimated over time for appliances at domotic agent [15].

The later node acts as an aggregated information agent, who generates DR price to lead the downstream domotic agents. DR price is a relative price to an actual sum of network tariff and day-ahead hourly pricing incentive that may reflect hourly spot-price (plus taxes, which is not included here for simplicity). Accordingly, the aggregator with relative elastic demand and different price offers to the market, results in demand dispatch that increases net social welfare. Authors in [16] assessed the agile demand response along with a traditional demand response (i.e. price based demand response), and found that for the majority of the domotic agents, agile DR generates the most social welfare. Moreover, it is also found that DR prices can reflect the actual situation in the electricity market, thus demand flexibility by the consumers via the aggregator would be included in the wholesale market to improve market responsiveness. Therefore, later in this section, we provide micromodels for multiple types of agent and rules for complex bidding. Based on this multi-agent system, later this paper formulates design optimization problem in section III. Since this paper mainly focuses on the evaluation of agile demand response by the aggregator, finding of pareto optimal market clearing mechanism is out of the scope of this research.

A. Aggregator Agent

Here, an aggregator, as shown in Fig. 1, is a collective representation of all downstream agents to the wholesale market. The aggregator sends DR prices to domotic agents and aggregates the received demand flexibility from each domotic agent, and then eventually the given demand flexibility is traded in the market.

B. Domotic Agent

Domotic Agent represents a smart customer who is participating in demand response program.

Let $\mathcal{D}$ denote the set of Domotic Agents (DAs), where the number of DAs is $D \triangleq |\mathcal{D}|$. For each agent-$d \in \mathcal{D}$, let l_d^k denote the total demand at interval $k \in \mathcal{K} \triangleq$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2017.2703643, IEEE Transactions on Smart Grid

$\{1, 2, \ldots, k, \ldots, K\}$, where $K = 24$. Without the lose of generality, we assume that an interval is one hour. The daily total demand of a customer represented by agent-d is denoted by $\mathbf{l}_d \triangleq [l_d^1, \ldots, l_d^K]$.

Remember, total demand curve $\mathbf{l}_d$ is a combination of total flexible demand $\mathbf{x}_d$ as well as non-flexible demand $\mathbf{b}_d$ (referred as base-load). Although agent-d does not consider base-load $\mathbf{b}_d$ for scheduling as it is non-flexible load, it is considered to build the final aggregated demand curve which is traded in the market, as shown in Fig. 1.

C. Appliance Agent

For a residential customer, there are various types of appliances. Four different types of appliances such as buffer, time shifter, uncontrollable and unconstraint have been defined in [17] as an energy flexible platform and interface (EF-Pi). EF-Pi proposes control spaces that contains general control model for each type of appliance. Herein, control space of every appliance is represented by an appliance agent. Hence for each agent-$d \in \mathcal{D}$, let $\mathcal{A}_d$ denote the set of appliance agents as every customer has different mix of appliance agents. Let $\mathcal{A}_d^b$ denote the set of buffer appliances like plug-in vehicles and storage. Similarly, $\mathcal{A}_d^s$ denotes the set of time-shifter appliances like dish-washer or washer. However, $\mathcal{A}_d^u$ denotes the set of uncontrollable appliances like photvoltaics or windmills and $\mathcal{A}_d^t$ denotes unconstrained appliances like freezers. In particular, $\mathcal{A}_d$ is a family of sets $\{\{\mathcal{A}_d^b\}, \{\mathcal{A}_d^s\}, \{\mathcal{A}_d^u\}, \{\mathcal{A}_d^t\}\}$ which are mutually disjoint. Hence, for every $a \in \mathcal{A}_d$, the vector of demand scheduling $\mathbf{x}_{d,a}$ is define as

$$\mathbf{x}_{d,a} \triangleq [x_{d,a}^1, \ldots, x_{d,a}^k, \ldots, x_{d,a}^K] \tag{1}$$

D. Bid Modeling

Authors developed a prioritization model referred as six-tuple bid model $\eta_{d,a}$ for a dispatchable load. Mathematically it is written as:

$$\eta_{d,a} = \langle \alpha_{d,a}, \beta_{d,a}, \omega_{d,a}, \theta_{d,a}, \boldsymbol{\gamma}_{d,a}, \Lambda_{d,a} \rangle$$

Wherein it is given that appliance agent-a of the dispatchable load needs to select the beginning $\alpha_{d,a} \in \mathcal{K}$ and the end $\beta_{d,a} \in \mathcal{K}$ of a time interval such that the dispatchable load can be scheduled. Moreover, the agent selects the cycle duration $\omega_{d,a} \in \mathbb{N}$ for which agent-a should remain "ON" and $\boldsymbol{\gamma}_{d,a}$ denotes the demand profile of the dispatchable load in "kW".

Rule 1. *Certainly, $\beta_{d,a} > \alpha_{d,a}$. Furthermore, the set of time intervals $\mathcal{K}_{d,a} \triangleq \{\alpha_{d,a}, \ldots, \beta_{d,a}\}$ provided by an agent to commit dispatch should be greater or equal to cycle duration $\omega_{d,a}$, required for process. Thus, $\omega_{d,a} \leq (\beta_{d,a} - \alpha_{d,a} + 1)$.*

The model also segregates agents in to two groups by using a boolean variable $\theta_{d,a}$; first group of agents is referred as atomic and second group of agents is referred as in-atomic. The only difference between two groups is that an atomic agent completes its cycle duration $\omega_{d,a}$ once it starts working. However, an in-atomic agent can distribute its usage through out the active duration $(\alpha_{d,a}, \beta_{d,a})$ for the given cycle duration $\omega_{d,a}$. Thus, $\theta_{d,a} = 1$ if an agent is atomic and $\theta_{d,a} = 0$ if an agent is in-atomic.

Unlike any prioritization modeling in [18], [19], the model uses bidding strategy which means agent a can propose its flexible demand $x_{d,a}^k$ as a function of an opportunity cost (referred as bid-price) $\lambda_{d,a}^k \in \mathbb{N}$ for a given interval $k \in \mathcal{K}$. So, as per empirical economics, current bid $\lambda_{d,a}^{\bar{k}}$ reflects the marginal cost of invoking a DR services and can be calculated as:

$$\lambda_{d,a}^k = \Gamma^k \times (1 + \frac{1}{\varepsilon_{d,a}}) \tag{2}$$

Thus, $\Lambda_{d,a} \triangleq [\lambda_{d,a}^1, \ldots, \lambda_{d,a}^K]$ denotes the vector of bids which are proposed by every agent a for any given day as a function of demand schedule. Where $\varepsilon_{d,a}$ is price elasticity of demand. The existing model of price elasticity of demand for the evaluation of demand response, as discussed in [20], is deterministic and usually studied for estimating aggregated demand response over long term. Therefore, in case of deterministic model, current bid $\lambda_{d,a}^k$ would only reflect average DR effect due to appliance agent but not its marginal cost. However, price elasticity of demand based on the assessment in [15] depends on the DR Prices Γ^k invoked by the aggregator agent and the current state of the appliance agent , and new bid is then derived by comparing expected payment when scheduling a demand with current bid. It implies that the evaluation of DR by using existing demand elasticity model falls behind the approach in [15], and thus the incorporation of the approach in appliance agent would be more realistic estimation of $\lambda_{d,a}^k$. Although in detail explanation of the phenomenon is out of the scope of this work, authors argued it in [21].

Rule 2. *For all agent-$a \in \{\{\mathcal{A}_d^s\}, \{\mathcal{A}_d^u\}, \{\mathcal{A}_d^t\}\}$, $\theta_{d,a} = 1$. However, For all agent-$a \in \{\mathcal{A}_d^b\}$, $\theta_{d,a} = 0$. Given $\lambda_{d,a}^k \leq \Gamma^k$, agent-$a$ would schedule during interval $\mathcal{K}_{d,a}$ as far as other constraints in $\eta_{d,a}$ are satisfied. However, given $\lambda_{d,a}^k > \Gamma^k$, agent-$a$ would commit dispatch without any delay from $k = \alpha_{d,a}$ to $k = \alpha_{d,a} + \omega_{d,a} - 1$.*

It should not be overlooked that if $\omega_{d,a} \leq (\beta_{d,a} - \alpha_{d,a} + 1)$, then agent-$a$ has some intervals to shift itself through out $\mathcal{K}_{d,a}$. Let $\tau_{d,a} \subseteq \mathcal{K}_{d,a} : |\tau_{d,a}| = \omega_{d,a}$ denote a set of variables which presents intervals to outset dispatchable load. Last but not least, $\mathcal{U}_{d,a} \triangleq \{u_{d,a}^1, \ldots, u_{d,a}^K\}$ is a set which keeps status of agent-a in three states i.e. $\{available \rightarrow 0, fuctioning \rightarrow 1, unavailable \rightarrow \infty\}$.

All these factors impose certain constraints on the vector of demand scheduling $\mathbf{x}_{d,a}$. In fact, the cycle duration $\omega_{d,a}$ for which agent-a can be scheduled equals to its predetermined daily consumption $E_{d,a}$ in "kWh", that is

$$\sum_{k=\alpha_{d,a}}^{\beta_{d,a}} x_{d,a}^k = E_{d,a} \tag{3}$$

Rule 3. *For all agent-$a \in \{\{\mathcal{A}_d^s\}, \{\mathcal{A}_d^b\}\}$; $x_{d,a}^k = 0, u_{d,a}^k \in \{0, \infty\} \; \forall \; k \in \mathcal{K} \setminus \tau_{d,a}$. Moreover, for all agent-$a \in \{\mathcal{A}_d^u\}$; $x_{d,a}^k = 0, u_{d,a}^k \in \{0, \infty\} \; \forall \; k \in \mathcal{K} \setminus \mathcal{K}_{d,a}$, which means* *uncontrolled loads always function between $\alpha_{d,a}$ and $\beta_{d,a}$. On the other hand, for all agent-$a \in \{\mathcal{A}^t_d\}$, $\alpha_{d,a} = 1$ and $\beta_{d,a} = K$, which means it has strict energy consumption pattern.*

The energy consumption of agent-a may have strict demand pattern like washing machine or it may have strict scheduling constraints like refrigerator or PV generation. Therefore, herein $\gamma^{min}_{d,a}$ defines the minimum power level and $\gamma^{max}_{d,a}$ defines the maximum power level for each appliance. However, $\gamma^{st}_{d,a}$ refers to standby power which is the power consumed while agent-a is in standby mode. Hence, it can be inferred that $\boldsymbol{\gamma}_{d,a}$ is a vector that either contains strictly/expected demand pattern i.e. $\boldsymbol{\gamma}_{d,a} \triangleq \{\gamma^1_{d,a} \cdots \gamma^{\omega_{d,a}}_{d,a}\}$ or has possible power levels i.e. $\boldsymbol{\gamma}_{d,a} \triangleq \{\gamma^{min}_{d,a}, \gamma^{st}_{d,a}, \gamma^{max}_{d,a}\}$.

Lastly, let $\mathbf{x}_d$ refer as demand flexibility which is formed by summing up the vector of demand scheduling $\mathbf{x}_{d,a}$ for all appliance agents corresponding to agent-$d \in \mathcal{D}$. Hence, a set of demand flexibility for agent-d is defined as

$$\mathcal{X}_d = \left\{ \mathbf{x}_d \mid \sum_{k=\alpha_{d,a}}^{\beta_{d,a}} x^k_{d,a} = E_{d,a}, \begin{array}{c} x^k_{d,a} = 0, \forall\, k \in \mathcal{K} \setminus \mathcal{K}_{d,a} \\ \gamma^{min}_{d,a} \leq x^k_{d,a} \leq \gamma^{max}_{d,a}, \forall\, k \in \mathcal{K}_{d,a} \end{array} \right\} \tag{4}$$

Rule 4. *The demand flexibility $\boldsymbol{x}_d$ is valid only if $\boldsymbol{x}_d \in \mathcal{X}_d$. Moreover, demand flexibility $\boldsymbol{x}_d$ is always less than total demand $\boldsymbol{l}_d$, such that the base load $\boldsymbol{b}_d = \boldsymbol{l}_d - \boldsymbol{x}_d$.*

III. Optimization Problem

It has been discussed that there are different types of appliance agents, corresponding to each appliance agent, a domotic agent should decide the intervals during which a particular appliance agent has to be committed. Hence, the optimization problem is formulated as for energy cost minimization by the agent-d and appliance agents. Thus, the problem could be formally defined as follows:

A. Players

An aggregator is a software agent presenting energy service provider, agent-$d \in \mathcal{D}$ represents smart system at home. Appliance agents given by the set $\mathcal{A}_d$ represents smart appliance committed for demand response.

B. Bidding Rules

In the giving settings, as discussed in sec. II, an increasing number of appliance agents run in a multi-agent based system (MAS). Any large or medium aggregator might run a hundreds of domotic agents that are composed of few distributed appliance agents. This hierarchical strucuture results in distributed combinatorial configuration [11], thus a single aggregator can entail a bottleneck in demand planning. Although many researchers solved the problem at aggregated level (simple centralized approach) by using complex programming techniques like batch reinforcement learning [22], mix-integer programming [23] and deep learning [24], this paper purposes an integration of intelligence at an intermediate agent-d (simple decentralized approach). Although advanced learning techniques can also be implemented at agent-d to increase sensitivity and accuracy, the main purpose of this study is to mathematically formulate the comprehensive problem that can be solved by using different learning techniques easily.

Furthermore, mostly simple bidding rules are applied for demand dispatch to ensure that the critical loads are served first with an objective to minimize delay-related costs. That is why, simple bidding rules can only be optimal for one particular load (i.e. for any particular type of appliance agent). Therefore, the case of different types of appliance agents can lead to very inefficient solutions as low-priority appliance agents (like time-shifters) are often delayed significantly and avoidable starvation of demand instances can be observed. Herein a rules for complex bidding are introduced that automatically proposes bid to changes in agent environment as well as changes in demand response. This strategy is considered market-driven as it aims at minimizing the market impact in terms of costs resulting from peak demand.

C. Energy Cost

The energy cost of an appliance agent depends on its bid-model $\eta_{d,a}$, and its daily schedule $\mathbf{x}_d \in \mathcal{X}_d$, thus energy cost function $f(\mathbf{x}_d)$ of agent-d is the collection of energy cost functions of appliance agents, as follows:

$$f(\mathbf{x}_d) = \sum_{a=1}^{A_d} f(\mathbf{x}_{d,a}) = \sum_{a=1}^{A_d} \sum_{k=\alpha_{d,a}}^{\beta_{d,a}} \Gamma^k x^k_{d,a} \tag{5}$$

D. Objective function

The objective of agent-d is to find optimum schedule $\overset{*}{\mathbf{x}}_d \triangleq \{\overset{*}{\mathbf{x}}_{d,1}, \ldots, \overset{*}{\mathbf{x}}_{d,A_d}\}$. Moreover, it is recalled that $\mathcal{U}_{d,a}$ stores the status of agent-a, thus for optimum schedule agent-d has to find the sequence of decisions $\mathcal{U}_{d,a}\ \forall\ a \in \mathcal{A}_d$ such that the total energy cost of all appliance agents is minimized. Mathematically, the objective problem can be stated as

$$\underset{u^k_{d,a} \in \mathcal{U}_{d,a}\ \forall\ a \in \mathcal{A}_d}{\text{minimum}} \left\{ \sum_{a=1}^{A_d} \sum_{k=\alpha_{d,a}}^{\beta_{d,a}} \Gamma^k x^k_{d,a} u^k_{d,a} \right\} \tag{6}$$

Recall from section I that $\Lambda_{d,a}$ represents the vector of bids for agent-a per interval. Even though the optimization problem in this case is similar to the problem described in eq. (6), the objective function will have an additional term dependent on bids $\Lambda_{d,a}$. Let $q(\lambda^k_{d,a}, u^k_{d,a})$ denote a qualitative function that captures the bid $\lambda^k_{d,a}$ with respect to the availability $u^k_{d,a}$ of agent-a during the given interval k. Moreover, function $q(\lambda^k_{d,a}, u^k_{d,a})$ have an indirect relation with a boolean variable $\theta_{d,a}$. Hence, the optimization problem for the proposed setting can be stated as:

$$\underset{u^k_{d,a} \in \mathcal{U}_{d,a}\ \forall\ a \in \mathcal{A}_d}{\text{minimum}} \left\{ \sum_{a=1}^{A_d} \sum_{k=\alpha_{d,a}}^{\beta_{d,a}} \Gamma^k x^k_{d,a} u^k_{d,a} + q(\lambda^k_{d,a}, u^k_{d,a}) \right\} \tag{7}$$

Observation 1. *It is noticed that the energy cost function $f(\boldsymbol{x}_d)$ is monotonically increasing and strictly convex $[0, \infty)$. Hence, in this case, the existence of best possible schedule $\overset{*}{\boldsymbol{x}}_d \triangleq \{\overset{*}{\boldsymbol{x}}_{d,1}, \dots, \overset{*}{\boldsymbol{x}}_{d,A_d}\}$ by each appliance agent denotes the global optimal solution of the problem* (6).

It can be noted that the optimization problem in (7) focuses in particular on the minimization of total energy cost of all appliances' agents. However, other demand response benefits like reducing risk of system outages (reliability benefits), congestion and over-voltage network issues, avoided (or deferred) T&D infrastructure upgrades, reduced greenhouse gas emission etc. can be included into (7) . These other benefits will also increase the "social welfare". Moreover, it could be noted that objective function is formulated in terms of energy cost, so all other DR benefits should be formulated in terms of energy cost before adding therein.

IV. Learning Mechanism

Recalled from section I, agent-d should decide the schedule of the corresponding appliance agents. For this reason, agent-d has a learning capability to make a sequence of decisions. This section provides an algorithm to be implemented in each domotic agent to reach optimum solution and achieve optimal system performance. Later in the section, proof for convergence and optimality of the proposed algorithm are shown.

A. Principle of the Learning

It can be noticed that problem (6) and (7) have the same objective functions. However, the problem (7) has some local constraints for each appliance agent corresponding to agent-d. Hence, problem in Eq. (7) can be viewed as multi-stage decision making problem. As, the next state of the system depends on the current state and the current decision, it can be solved by reinforcement learning technique. Moreover, if each appliance agent is scheduled for best response in an asynchronous fashion, then there will be no coupling constraints and problem (7) could be reordered as:

$$\sum_{a=1}^{A_d} \underset{u^k_{d,a} \in \mathcal{U}_{d,a}}{\text{minimum}} \left\{ \sum_{k=\alpha_{d,a}}^{\beta_{d,a}} \Gamma^k x^k_{d,a} u^k_{d,a} + q(\lambda^k_{d,a}, u^k_{d,a}) \right\} \tag{8}$$

Agent-d can solve problem (8) as long as it knows price Γ^k for all $k \in \mathcal{K}$ as well as $\hat{\eta_{d,}}$, the vector containing the bid-model for all corresponding appliance agents. In order to solve problem (8) by using reinforcement learning (RL) technique, agent-d should define its state space, action space, state transition function and the reward function. Following are the definitions of parameters for the proposed RL based control algorithm.

B. State Space and Action Space

The state of the agent captures the information that depends on the energy cost as per time of use and the cycle time $\omega_{d,a}$ of the appliance agent. Hence, the state is represented by two dimensional vector $\hat{s}_a = (s_1, s_2)$; where s_1 is current interval and s_2 is the number of intervals in which agent-a is functioning, as stated in rule 3. Mathematically it is written as:

$$\mathcal{S}_a = \left\{ \hat{s}^1_a, \hat{s}^2_a, \dots \mid \hat{s}^i_a = (s_1, s_2) \right\} \begin{array}{l} \forall\, s_1 \in \{\alpha_{d,a}, \dots, \beta_{d,a}\} \\ \forall\, s_2 \in \{0, 1, \dots, \omega_{d,a}\} \end{array} \tag{9}$$

It can be implied the number of pairs of state space I for each appliance agent consists of $(\beta_{d,a} - \alpha_{d,a}) \times \omega_{d,a}$ elements. Moreover, an action δ_a to be taken by agent-a at each state is either to be ON (i.e. committed) or OFF, simply $\delta_a \in \{0, 1\}$

Observation 2. *If rules 1 and 2 hold, then the process of learning by the domotic agent for each appliance agent always starts at $\hat{s}^1_a = (\alpha_{d,a}, 0)$ and terminates at $\hat{s}^i_a = (\beta_{d,a}, \omega_{d,a})$.*

C. State Transition Function

Since the state of an agent in is the representation of time, thus the agent obtains next state from the current state as per following state transition function:

$$(s^{k+1}_1, s^{k+1}_2) = (s^k_1 + 1, s^k_2 + \delta^k_a) \;\forall\; \delta^k_a \in \{0, 1\} \tag{10}$$

D. Reward Function

As per the problem (8), the objective is to find a sequence of actions $\hat{\delta_a} = \{\delta^1_a, \delta^2_a, \dots, \delta^k_a, \dots, \delta^K_a\}$ such that the total energy cost is minimized while meeting the constraints and reducing the delay in dispatch depending on bids $\Lambda_{d,a}$. Thus, the reward function should be designed in such a way that the agent-d learns best sequence of actions for each appliance agent without violating constraints defined in bid-model $\eta_{d,a}$. Suppose all the constraints are satisfied, then reward function is equal to the objective function in (8) for a particular sequence of actions $\hat{\delta_a}$. Mathematically it is written as;

$$\sum_{k=\alpha_{d,a}}^{\beta_{d,a}} g(\hat{s}^k_a, \delta^k_a, \hat{s}^{k+1}_a) = \sum_{k=\alpha_{d,a}}^{\beta_{d,a}} \Gamma^k x^k_{d,a} u^k_{d,a} + q(\lambda^k_{d,a}, u^k_{d,a}) \tag{11}$$

Let for all $k \in \mathcal{K}$,

$$g(\hat{s}^k_a, \delta^k_a, \hat{s}^{k+1}_a) = \Gamma^k x^k_{d,a} u^k_{d,a} + q(\lambda^k_{d,a}, u^k_{d,a}),\; \forall\, a \in \mathcal{A}_d$$

Given observation 3 and 4 hold, then the formal definition of g function to capture all the objectives and constraints can be stated as:

$$g(\hat{s}^k_a, \delta^k_a, \hat{s}^{k+1}_a) = \begin{cases} \Gamma^k x^k_{d,a} u^k_{d,a}, & \text{if } \lambda^k_{d,a} = 0 \\ \infty, & \text{if } (\hat{s}^{k+1}_1 = \beta_{d,a}) \wedge (\hat{s}^{k+1}_2 = \omega_{d,a}) \\ \infty, & \text{if } (1 \leq \hat{s}^k_2 \leq \omega_{d,a}) \wedge (\delta^k_a = 1) \\ \lambda^k_{d,a} x^k_{d,a} u^k_{d,a}, & \text{if } \delta^k_a = 0 \end{cases} \tag{12}$$

Observation 3. *Let the unique DR price Γ^k and $\lambda^k_{d,a}$ as per observation 2 is the optimal solution of energy cost*

```
ALGO 1: Executed by each agent-$d \in \mathcal{D}$ forall interval $k \in \mathcal{K}$
Receive bid $\eta_{d,a}$ from all $a \in \mathcal{A}_d$
Repeat (for each appliance agent $a \in \mathcal{A}_d$)
  Initialize $\mathcal{Q}(\hat{s}, \delta)$ arbitrarily
  Initialize $\epsilon$
  Repeat (for each iteration)
    Initialize $i = 1$, such that $\hat{s}_a^i \leftarrow \hat{s}_a^1$
    Repeat (for each pair of state space)
    Choose $\delta_a^k$ using $\epsilon$-greedy selection action.
    Obtain $\Gamma^k$ and $\hat{s}_a^{i+1}$ by using (10)
    Take action $\delta_a^i$, observe $g(\hat{s}_a^i, \delta_a^i, \hat{s}_a^{i+1})$ by using (12)
    Update $\mathcal{Q}(\hat{s}_a^i, \delta_a^i)$ by using (14)
    $\hat{s}_a^i \leftarrow \hat{s}_a^{i+1}$
    Until $\hat{s}_a^i \leftarrow \hat{s}_a^I$ is terminated
  Update $\epsilon$
  Until iteration is terminated
  Observe $\overset{*}{\mathbf{x}}_{d,a}$
Until $a$ is terminated
```

minimization problem (8)*. Then, it is concluded from* (12) *that the domotic agent will learn to schedule the appliance agent without any delay provided its bid $\lambda_{d,a}^k > \Gamma^k$, otherwise the appliance agent can contribute more flexibility in scheduling i.e. $\boldsymbol{x}_{d,a}$.*

Agent-d can simply achieve the objective problem (8) through replacing the given reward function by the formulated reward function (12). One of the most important breakthrough in the solution of multi-stage problem is advancement in Reinforcement Learning technique, which is a model free approach for optimization. Reinforcement learning that involves learning of state-action function i.e. $\mathcal{Q}(\hat{s}, \delta)$, referred as $\mathcal{Q}$-value. In general, state-action function $\mathcal{Q}(\hat{s}, \delta)$ corresponds to an optimal solution is the one that holds an action δ^* with minimal $\mathcal{Q}$-value for all $\hat{s} \in \mathcal{S}$, such as

$$\delta^* = \arg\min_{\delta} \mathcal{Q}(\hat{s}, \delta) \tag{13}$$

Hence, the overall idea is to learn the schedule $\mathcal{Q}(\hat{s}^k, \delta^k)$ repeatedly over time by taking the old schedule and then makes an update based on new information. Mathematically it is written as;

$$\begin{aligned} \mathcal{Q}(\hat{s}^k, \delta^k) &\leftarrow \mathcal{Q}(\hat{s}^k, \delta^k) + \sigma \, [\; g(\hat{s}^k, \delta^k, \hat{s}^{k+1}) \\ &+ \min_{\delta'} \mathcal{Q}(\hat{s}^{k+1}, \delta') - \mathcal{Q}(\hat{s}^k, \delta^k) \;], \quad \forall \; a \in \mathcal{A}_d \end{aligned} \tag{14}$$

For an explanation of (14), let $\mathcal{Q}^i(\hat{s}^k, \delta^k)$ be the initial guess, such that during a time interval k, the process is in state $\hat{s}^k$ and agent-d takes an action δ^k for agent-a based on the current estimate $\mathcal{Q}(\hat{s}^k, \delta^k)$. Remember, agent-$d$ randomly explores an action δ' with ϵ-greedy action selection, $\epsilon = 0.1$. Moreover, any ϵ-greedy action selection with respect to state-action function is an improvement over any further selection as it determines which state-action pair $(\hat{s}^k, \delta^k)$ are visited and updated. Moreover, σ is a learning co-efficient that modifies $\mathcal{Q}$-values in each iteration. Under the assumptions of usual stochastic approximation conditions throughout the learning process, state-action function will converge to optimal state-action function $\mathcal{Q}^*$.

In model-based value iteration, and more generally in mix-integer programming [23] and DP [25], an algorithm is said to be accurate, consistent and convergence is guaranteed for offline. However, for online, the approximate value function converges to the optimal one as the accuracy of approximation technique increases, correspondingly increases cost of computation aggressively to the number of appliances and iterations [26]. In model-free value iteration, and more generally in Reinforcement Learning, accuracy is sometimes understood as the convergence to a well-defined solution, means stronger convergence results in accurate, consistent and optimal solution. In the area of online approximate value iteration, the approximation technique increases accuracy and convergence [27], correspondingly increases cost of computation as well as care must be taken when selecting approximation parameters to prevent possible expansion and divergence. So, if the learning agent is independent and distributed, it is difficult to adjust the approximation parameters of each agent-a, which closely depends on its local decision problem. Therefore, besides the advanced techniques in [22] and [24], an uncomplicated $\mathcal{Q}$-learning technique that typically rely on non-expansive approximation, as shown in procedural form in ALGO 1, is considered for the illustration of the methodology. Hence, the given algorithm for each domotic agent-d is found suitable in the light of literature, as it takes relatively less computation time as well as handles problem of decision-making for demand dispatch in distributed fashion.

E. Convergence and Optimality

In this section,the convergence and optimality properties of the proposed algorithm is proved by policy improvement theorem that states "*for any non-optimal state-action function $\mathcal{Q}(\hat{s}, \delta)$, the state-action function $\mathcal{Q}(\hat{s}, \delta')$ should be a strict improvement such that the action δ' optimizes the given state-action function*" [28]. Suppose, $(\hat{s}, \delta)$ and $(\hat{s}, \delta')$ be any pair of deterministic state-action functions, then for all $\hat{s} \in \mathcal{S}$.

$$\mathcal{Q}(\hat{s}, \delta') \leq \mathcal{Q}(\hat{s}, \delta) \tag{15}$$

Together, from observations 1 and 3, the best response from each appliance agent would be equivalent to solving optimization problem 7. Hence, if agent-d finds the best state-action function $\mathcal{Q}(\hat{s}, \delta^*)$ for agent-a subsequently through running algorithm in a asynchronous fashion, the total energy cost either decrease or remained unchanged every time agent-a updates its bid-model $\eta_{d,a}$.

Observation 4. *If the updates of the bid-model $\eta_{d,a}$ of an appliance agent are asynchronous among the appliance agents, then the algorithm takes an action starting from any randomly selected initial condition that looks best in the short term according to its current state-action function. In this way, the algorithm always converges to its best response due to the optimal convergence policy* (15).

From observations 2 and 4, if all the bid-model $\eta_{d,a}$ of appliance agents remains unchanged within the given constraints, then algorithm schedules appliance agents over a day-ahead basis. Otherwise, if agents change their needs frequently, the algorithm schedules agents in a more real-time fashion [29] by predicting expected future prices. In this paper, the focus will remain on DR scheduling over day-ahead basis, although by coupling the concept given in [29] would transform it into real-time scheduling.

V. Illustrative Study

Artificial problem sets consist of $\{5, 10, 25, 50, 100, 250, 500, 1000\}$ houses are constructed based on specifications (in section II) for the scheduling problem in (8). Each house is represented by a domotic agent, which has maximum upto ten appliance agents. In each experimentation, the scheduling problem is solved for 100 days in an object oriented way using MATLAB. Herein, each agent-a is implemented by using proposed complex bid model in sec II. Furthermore, the agent-a offers flexibility within 24 hours sliding window depending over its defined bid-model $\eta_{d,a}$, with a granularity of an hour. On the other hand, other parameters are generated randomly within boundaries of operating specifications in [30] throughout the experimentation. Moreover, for aggregator and its agent-d, it is assumed that DR Prices strictly follows the local retail day-ahead market price.

A. Experimentation For Q-iteration

In this experiment, an approximate policy iteration (presented in (14)) is applied at every agent-d. As per (2) and discussed in [29], two scenarios are simulated. First self-elastic case which can be referred as open-loop case of bid strategy because herein appliance agent only uses self-elasticity of demand for bid generation. Second, cross-elastic case which can be referred as close-loop case of bid strategy because herein appliance agent also uses cross-elasticity of demand for bid generation.

Both cases perform policy improvement and explore the best policy, thus the solutions by both the cases are of a similar quality. Specifically, it was observed in Fig. 2, 3 and 4 that exploration in self-elastic case is more than cross-elastic case which is obvious because cross-elastic case drives new bid by comparing expected payment when scheduling a demand with current bid. Moreover, it was also observed that the convergence to an optimal schedule in self-elastic case is guaranteed by the theory of improvement. On the other hand, for cross-elastic case the optimal schedule might not be similar as self-elastic case, however the optimal schedule will be the best response for the cluster under given constraints, thus providing numerical proof of observation 1 and observation 2.

B. Convergence Simulation

This section illustrates the convergence of algorithm. As, the convergence is given by (12) which depends on bid $\lambda^k_{d,a}$, so five different scenarios are simulated such that elasticities of demand are increased during each respective scenarion. For an explanation purposes, an artificial set of five houses (agents-d) in selected, such that $d = 1 \rightarrow \{|\mathcal{A}^b_1|, |\mathcal{A}^s_1|, |\mathcal{A}^u_1|, |\mathcal{A}^t_1|\} = \{1, 2, 4, 0\}$. Similarly, $2 \rightarrow \{0, 2, 2, 0\}$, $3 \rightarrow \{1, 3, 2, 0\}$, $4 \rightarrow \{0, 3, 1, 0\}$, $5 \rightarrow \{1, 3, 2, 0\}$. Thus, in total 27 appliance agents are under consideration. For sake of simplicity, consider demand elastic factor $Fr = \frac{\lambda^k_{d,a}}{\Gamma^k}$; where $Fr = 0$ means appliance agents are unit elastic, $Fr < 1$ means agents-a are elastic, $Fr = 1$ means agents-a are relatively elastic, however $Fr > 1$ means agents-a are relatively inelastic.

1) Offline Learning: Fig. 2 represents the impact of offline learning for 100 days on the convergence of reward function, as shown in (12), under different Fr. Impact is illustrated by considering the concept of resource dilation factor (RDF), given in [31]. Herein RDF is a ratio of exploration over exploitation, that it simply an aggregated ratio between the final decisions over first estimations during a learning process for every day by all agents. Moreover, if RDF is found equal to unity, there is no or little exploration that implies agents have deterministic behavior and thus algorithm has learned it absolutely.

Firstly it can be observed that in general the learning behavior of agents under all Fr is exponential. Secondly for $Fr > 1$, RDF moves faster toward unity which means every agent has too defined scheduling routine and economic constraints, consequently algorithm performs exploitation more than exploration around the state space which satisfies the fact that agents are tightly constrained. Moreover, even in this case RDF never found to be unity because of ϵ-greedy action selection that always maintain some form of exploration. Thirdly when $Fr = 1$ agents are loosely constrained, so agents converged faster towards better initial estimations but did not moves towards unity. It is also because of ϵ-greedy action selection maintaining balance between the exploration and the exploitation of actions across the state-space through out the learning process. Lastly, for Fr within the flexible range each agent continuously finds the balance between the exploration and exploitation of actions. Although it mostly appears higher than $Fr = 0$, it satisfies the fact that the bid constraint makes reward function more stochastic in nature, thus requires more exploration.

Hence, the discussion provides the numerical proof of observation 3 and onservation 4, such that for all $\lambda^k_{d,a} \leq \bar{\Gamma}^k$ DR appears to be the best response.

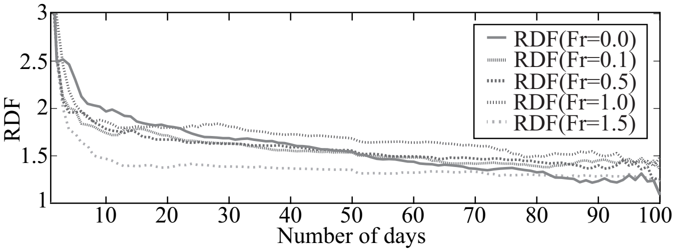

Fig. 2. Resource dilation factor calculated over 100 days of simulation.

2) Online Learning: Fig. 3 shows the convergence of reward function for a day during online excution of algorithm after offline learning for 100days. As shown in Fig. 3(b) , when Fr is either less than or equal to 1, average reward is almost equal to average reward obtained when $Fr = 0$. On the other hand, Fr greater than 1 causes the reward to diverge from its optimal point which satisfies the fact that agents are least flexible.

It can also be inferred from Fig. 3(c) that for all $Fr \leq 1$, i.e. referred as flexible range, convergence happens relatively close to each other. Moreover, $Fr = 0$ is counted as most flexible scenario, because therein agent-d has maximum freedom to schedule with least economic constraints. Average reward found to be the most optimal solution, however it takes most time to converge that is evident due to the fact that therein agent-d has relatively broader state space to explore. On the other hand, $Fr > 1$ is counted as limiting case for demand flexibility because therein agent-a somehow limits the scheduling towards its starting time. Thus, convergence happens fast but way far from an optimal point. Furthermore, for all Fr between 0.1 and 1 convergence happens relatively faster than $Fr = 0$ and slower than $Fr > 1$, it is also due to the fact that by increase in Fr each agent-a limits its scheduling window along with economic constraints. Optimum Fr in flexible range is between 0.5 and 1, because it converges faster than rest as well as closest to optimal solution.

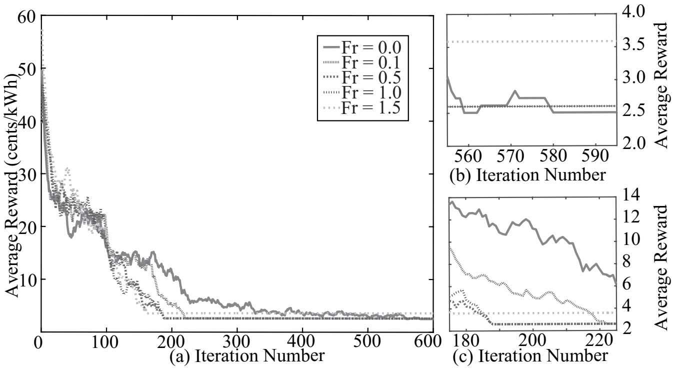

Fig. 3. Convergence of reward function for an initial day while learning process

C. Performance Results

This section presents results for self-elastic and cross-elastic scenarios with similar initial conditions. The performance measure is average cost per allocation decision summed over all appliance agents associated with particular domotic agent. For a variety of policy learning, we compare the policy performance of two scenarios with increasing number of domotic agents per aggregator.

Our general observations regarding the experiments in Fig. 4 are as follows. First, the RL approach to policy improvement clearly works quite well for scalability. Second, the results are generally in accordance with prior studies of policy iteration, where one typically finds large improvement starting from weak initial policies, and progressively smaller improvement starting from stronger initial policies. Third, the average revenue remains slightly higher in case of cross-elasticity than self-elasticity as explained earlier. Lastly, it has been inferred that the increase in cost for the cross-elastic scenario in given model due to sub-optimal allocations, however the RL trained for self-elastic sets are able to do better due to more flexibility as per theorem 4.

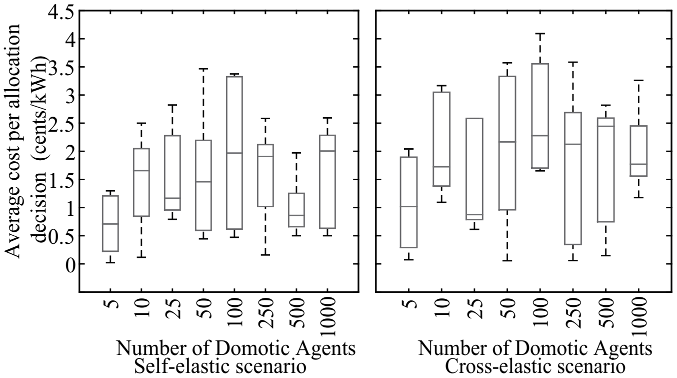

Fig. 4. Performance of various number of domotic agents in self-elastic as well as cross-elastic scenarios.

VI. Conclusion

This paper contributes to the technical literature as follows:

- A comprehensive applied methodology for agile demand response. The methodology taps the potential demand flexibility from the participating customers and maximizes social welfare of the society in an unbundled electricity market.
- The DR services by the agents are explicitly modeled in terms of complex bid model $\eta_{d,a}$. The proposed complex bidding rules integrates appliance characteristics and preferences into the decisions for DR planning. Moreover, the deployment of the proposed micromodels of multiple agents would enable the system to monitor, learn and schedule the demand flexibility with more precision, as well as optimize and customize it for DR prices.
- The proposed micromodels outperforms the existing price elasticity based models, thus the proposed methodology is of high practical relevance in modern Multi-agent system.
- The principle of distributed learning, instead of centralized, introduces a procedure to automatically align the bid to state changes and adjust maximum DR services to mitigate model inefficiencies during state transition.
- Due to distributed learning, an agent does not require deep and complex learning techniques for DR planning of even heterogeneous appliances. Thus, it saves computation time, agent processing requirements, results in simple theoretical approach as well as require reasonable amount of data for learning.

Lastly, simulation results evaluate in detail the methodology and conclude the fact that an aggregators applying this methodology for DR planning would have more stability in trading of demand flexibility in the day-ahead market. Moreover, the DR scheduling is verified by measuring the demand dispatch of the number of agents for couple of days by assessing actual load shift and curtailment quantities at individual appliance agent.

References

[1] U. D. of Energy, "Benefits of Demand Response in Electricity Markets and Recommendations for Achieving Them - A Report to the United States Congress Pursant to Section 1252 of the Energy Policy Act of 2005," no. February, p. 122, 2006.

[2] E. E. Directive, “Directive 2012/27/eu of the european parliament and of the council of 25 october 2012 on energy efficiency, amending directives 2009/125/ec and 2010/30/eu and repealing directives 2004/8/ec and 2006/32,” *Official Journal, L*, vol. 315, pp. 1–56, 2012.

[3] C. F. Covrig, M. Ardelean, J. Vasiljevska, A. Mengolini, G. Fulli, E. Amoiralis, M. Jiménez, and C. Filiou, “Smart grid projects outlook 2014,” *Joint Research Centre of the European Commission: Petten, The Netherlands*, 2014.

[4] W. Wei, F. Liu, and S. Mei, “Energy Pricing and Dispatch for Smart Grid Retailers Under Demand Response and Market Price Uncertainty,” *IEEE Trans. Smart Grid*, vol. 6, no. 3, p. 1, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6994285

[5] M. G. Vaya and G. Andersson, “Optimal bidding strategy of a plug-in electric vehicle aggregator in day-ahead electricity markets,” *International Conference on the European Energy Market, EEM*, 2013.

[6] A. a. S. Algarni and K. Bhattacharya, “A generic operations framework for discos in retail electricity markets,” *IEEE Transactions on Power Systems*, vol. 24, no. 1, pp. 356–367, 2009.

[7] F. Rahimi and A. Ipakchi, “Demand response as a market resource under the smart grid paradigm,” *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 82–88, 2010.

[8] M. Babar, P. H. Nyugen, V. Cuk, I. G. R. Kamphuis, M. Bongaerts, and Z. Hanzelka, “The rise of AGILE demand response: Enabler and foundation for change,” *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 686–693, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.rser.2015.11.084

[9] M. Negnevitsky, T. D. Nguyen, and M. De Groot, “Novel business models for demand response exchange,” *IEEE PES General Meeting, PES 2010*, pp. 1–7, 2010.

[10] Smart Energy Collective, “An introduction to the Universal Smart Energy Framework,” 2013.

[11] K. Kok, “The PowerMAtcher: Smart Coordination for the Smart Electricity Grid,” *PhD Thesis*, p. 293, 2013. [Online]. Available: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf

[12] “Commission proposes new rules for consumer centred clean energy transition,” nov 2016. [Online]. Available: http://ec.europa.eu/energy/en/news/commission-proposes-new-rules-consumer-centred-clean-energy-transition

[13] H. Wu, M. Shahidehpour, A. Alabdulwahab, and A. Abusorrah, “Demand response exchange in the stochastic day-ahead scheduling with variable renewable generation,” *IEEE Transactions on Sustainable Energy*, vol. 6, no. 2, pp. 516–525, 2015.

[14] E. Klaassen, M. Reulink, A. Haytema, J. Frunt, and J. Slootweg, “Integration of in-home electricity storage systems in a multi-agent active distribution network,” in *PES General Meeting— Conference & Exposition, 2014 IEEE*. IEEE, 2014, pp. 1–5.

[15] R. Fonteijn, M. Babar, and I. Kamphuis, “An assessment of the influence of demand response on demand elasticity in electricity retail market,” in *Power Engineering Conference (UPEC), 2015 50th International Universities*. IEEE, 2015, pp. 1–6.

[16] M. Nijhuis, M. Babar, M. Gibescu, and J. Cobben, “Demand response: Social welfare maximisation in an unbundled energy market-case study for the low-voltage networks of a distribution network operator in the netherlands,” *IEEE transactions on industrial electronics*, vol. 53, no. 1, pp. 32–38, 2016.

[17] G. Coene, M. J. Konsman, and J. Adriaanse, “FlexiblePower Application Infrastructure - Detailed Functional Design,” 2013.

[18] S. Vandael, B. Claessens, M. Hommelberg, T. Holvoet, and G. Deconinck, “A scalable three-step approach for demand side management of plug-in hybrid vehicles,” *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 720–728, 2013.

[19] F. Ruelens, B. J. Claessens, S. Vandael, S. Iacovella, P. Vingerhoets, and R. Belmans, “Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning,” in *Power Systems Computation Conference (PSCC), 2014*. IEEE, 2014, pp. 1–7.

[20] S. D. Maqbool, M. Babar, and E. A. Al-Ammar, “Effects of demand elasticity and price variation on load profile,” in *Innovative Smart Grid Technologies-Middle East (ISGT Middle East), 2011 IEEE PES Conference on*. IEEE, 2011, pp. 1–5.

[21] M. Babar, P. Nguyen, V. Cuk, and I. Kamphuis, “The development of demand elasticity model for demand response in the retail market environment,” in *PowerTech, 2015 IEEE Eindhoven*. IEEE, 2015, pp. 1–6.

[22] F. Ruelens, B. Claessens, S. Vandael, B. De Schutter, R. Babuska, and R. Belmans, “Residential demand response applications using batch reinforcement learning,” *arXiv preprint arXiv:1504.02125*, 2015.

[23] N. Blaauwbroek, P. H. Nguyen, M. J. Konsman, H. Shi, R. I. Kamphuis, and W. L. Kling, “Decentralized resource allocation and load scheduling for multicommodity smart energy systems,” *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1506–1514, 2015.

[24] B. J. Claessens, P. Vrancx, and F. Ruelens, “Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control,” *arXiv preprint arXiv:1604.08382*, 2016.

[25] M. Babar and E. A. Alammar, “The consumer rationality assumption in incentive based demand response program via reduction bidding,” *Journal of Electrical Engineering & Technology*, vol. 10, no. 1, pp. 64–74, 2015.

[26] M. S. Santos and J. Vigo-Aguiar, “Analysis of a numerical dynamic programming algorithm applied to economic models,” *Econometrica*, pp. 409–426, 1998.

[27] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton, “Toward off-policy learning control with function approximation,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 719–726.

[28] V. Functions and T. D. Learning, *Reinforcement Learning*, 2014.

[29] M. Babar, P. H. Nguyen, V. Cuk, and I. G. Kamphuis, “The development of demand elasticity model for demand response in the retail market environment,” *2015 IEEE Eindhoven PowerTech, PowerTech 2015*, 2015.

[30] P. Samadi, S. Member, H. Mohsenian-rad, V. W. S. Wong, S. Member, and R. Schober, “Real-Time Pricing for Demand Response Based on,” vol. 5, no. 2, pp. 789–798, 2014.

[31] W. Zhang and T. G. Dietterich, “A Reinforcement Learning Approach to Job-shop Scheduling,” *Ijcai*, pp. 1114–1120, 1995.

Muhammad Babar received his Bachelor in Electronic Engineering from NED University of Engineering and Technology, Pakistan in 2010, and Master in Manufacturing System Design and Management Engineering from National University of Science and Technology, Pakistan in 2013. He started his research career as Research Engineer at Wavetec Pvt. Ltd., Pakistan and then worked as a Researcher in Saudi Aramco Chair, Saudi-Arabia and High Voltage Laboratory of King Saud University, Saudi-Arabia for around two and a half year. From September 2013, he is working as a PhD student in Electrical Power Systems (EES) Group, Technology University of Eindhoven, Netherlands. He is also visiting researcher at AGH University of Science and Technology, Krakow - Poland. He is also Research Internee at Alliander, the second largest distribution company in the Netherlands. His interests are focused on but not limited to Agile Demand Response, Energy Management, Big Data and Analytics, Smart Grid and Machine Learning.

Phuong H. Nguyen (M‘06) received the Ph.D. degree from the Eindhoven University of Technology, the Netherlands, in 2010. He is an Assistant Professor in the Electrical Energy Systems (EES) group at the Eindhoven University of Technology, The Netherlands. He was a visiting researcher with the Real-Time Power and Intelligent Systems (RTPIS) Laboratory, Clemson University, USA, in 2012 and 2013. At EES, he is leading a research line of using computational and distributed intelligence to enable active and intelligent distribution grids. His research of interests also includes data analytics with deep learning, real-time system awareness using (IoT) data integrity, as well as predictive and corrective grid control functions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2017.2703643, IEEE Transactions on Smart Grid

Vladimir Ćuk received his Dipl. Ing. degree in Electrical Engineering from the School of Electrical Engineering, University of Belgrade in 2005. From 2006 until 2009 he was working at the Electrical Engineering Institute "Nikola Tesla" in Belgrade, at the Electrical Measurements Department. In 2013 he received a PhD diploma at the Electrical Energy Systems group of the Eindhoven University of Technology, the Netherlands where he is now working as an Assistant Professor. His research is focused on modeling and analysis of power quality phenomena.

I.G. (René) Kamphuis is a senior technologist at TNO, the Dutch organization for technological research, and part-time Professor at Eindhoven Technical University, the Netherlands. He did his Ph. D. in chemical physics at Groningen State university in 1983. He has been involved during the last 15 years in developing smart grids technology and ICT-architectures for setting up smart grid field tests. His interests are in agent-based software engineering and distributed systems computation.

Martijn Bongaerts From 1998 onwards he was involved in electric grid innovation in several roles at Alliander, the second largest distribution company in the Netherlands. Ir. Bongaerts is a part of the innovative projects and successfully has set-up a large number of innovative projects in the Netherlands. He is involved in many Dutch (and also some international) expert- and working groups about Smart Grids and the influence of the Energy Transition on the energy system focusing on electricity grids. He was chair of the project group Smart Grids from the joint Dutch grid-companies. He was one of the people to take the initiative of developing the Dutch innovation program for Smart Grid field trials. At present, he is a part of a EU-expert group on Flexible Generation of ETIP-SNET (European Technology and Innovation Platform Smart Netowkrs for Energy Transition). He is also involved in other fora and working groups for developments of smart grid technology in the Netherlands especially for small distributed electricity systems.

Zbigniew Hanzelka Professor at University of Science and Technoly, Krakow (Poland) , head of the laboratory for power quality and Center for Photovoltaic Research. Author and co-author of many scientific papers, raports and books. Editor of magazines: Electrical Power Quality and Utilization. Area of research interests include quality of electricity supply, including the co-operation of renewable energy power system and issues related to technology platform of smart grids. Contractor of many research projects and work for the industry.