Negotiated Learning for Smart Grid Agents: Entity Selection based on Dynamic Partially Observable Features

Prashant P. Reddy
Machine Learning Department
Carnegie Mellon University
Pittsburgh, USA
ppr@cs.cmu.edu

Manuela M. Veloso
Computer Science Department
Carnegie Mellon University
Pittsburgh, USA
mmv@cs.cmu.edu

Abstract

An attractive approach to managing electricity demand in the Smart Grid relies on real-time pricing (RTP) tariffs, where customers are incentivized to quickly adapt to changes in the cost of supply. However, choosing amongst competitive RTP tariffs is difficult when tariff prices change rapidly. The problem is further complicated when we assume that the price changes for a tariff are published in real-time only to those customers who are currently subscribed to that tariff, thus making the prices partially observable. We present models and learning algorithms for autonomous agents that can address the tariff selection problem on behalf of customers. We introduce *Negotiated Learning*, a general algorithm that enables a self-interested sequential decision-making agent to periodically select amongst a variable set of *entities* (*e.g.*, tariffs) by negotiating with other agents in the environment to gather information about dynamic partially observable entity *features* (*e.g.*, tariff prices) that affect the entity selection decision. We also contribute a formulation of the tariff selection problem as a *Negotiable Entity Selection Process*, a novel representation. We support our contributions with intuitive justification and simulation experiments based on real data on an open Smart Grid simulation platform.

Introduction

Integration of distributed sustainable energy resources, *e.g.*, wind and solar, into our electricity supply is essential to reducing the environmental impact of our growing energy demand (Gomes 2009). However, such resources often introduce significant volatility to the level of supply and thus make it difficult to match supply with demand. Smart Grid systems must include the ability to actively manage electricity demand so that grid operators are better able to balance supply and demand. An attractive approach to managing demand relies on the use of real-time pricing (RTP) tariffs, where customers are made aware of the dynamic cost of electricity supply through rapidly changing prices and incentivized to adapt their consumption behavior accordingly (Strbac 2008). However, directly conveying prices from wholesale electricity markets to retail customers subjects them to excessive risk (Barbose, Goldman, and Neenan 2005). Therefore, alternate tariff structures that mitigate the price volatility risk, such as fixed-rate, time-of-use (TOU), and critical peak-pricing (CPP) tariffs, have seen wider field deployment to date (Hammerstrom 2008).

Further innovation is needed to enable wider adoption of RTP tariffs (Faruqui and Palmer 2012). The introduction of supplier competition in liberalized retail electricity markets is a key enabler because it encourages novel tariff structures and provides customers with a wider array of tariff choices so that they can select tariffs best suited for their specific consumption behavior and risk appetite (Block, Collins, and Ketter 2010). However, the resulting *tariff selection* problem, *i.e.*, periodically selecting amongst the set of competitive tariffs, is difficult when prices are allowed to change rapidly. The problem is further complicated when we assume that the price changes for a particular tariff are published in real-time only to those customers that are currently subscribed to that tariff, thus making the prices partially observable when selecting amongst the tariffs.

Introducing *autonomous customer agents* that can select tariffs on behalf of a customer and control the customer's demand in response to changes in tariff prices can alleviate the decision-making burden on Smart Grid customers (Reddy and Veloso 2012). We present several models and learning algorithms for tariff selection by such agents. Most significantly, we introduce *Negotiated Learning*, a general algorithm that allows a self-interested sequential decision-making agent to periodically select amongst a variable set of *entities* (*e.g.*, tariffs) by negotiating with other agents in the environment to gather information about dynamic partially-observable entity *features* (*e.g.*, tariff prices) that affect the entity selection decision. This algorithm allows an agent to exploit the multiagent structure of the problem to control the degree of partial observability in the environment.

In following sections, we describe the tariff selection problem in more detail and contribute a formulation of the problem as a *Negotiable Entity Selection Process* (NESP), a novel representation for the type of multiagent partial-observability problem that we address here. We then describe how a Negotiated Learning agent uses *Attractions* and a *Negotiation Model* to determine when to acquire which information from which other agents to help make its entity selection decisions. We include intuitive justification for the algorithms and support our contributions with experimental results from simulations based on real data using Power TAC, a large open Smart Grid simulation platform.

Variable Rate Tariff Selection

We assume liberalized retail electricity markets where *suppliers* compete to acquire portfolios of customers. Each supplier offers one or more tariffs, which are published contracts that customers can accept or reject without modification. A tariff contract includes various terms and fees including one or more rate specifications. A *variable rate* specification says that the dynamic price to be charged for a given metering time period is conveyed to the customer at the start of some *advance notice window* before the metering period starts. Metering periods and advance notice windows vary widely amongst real-world tariffs, from months or days for residential customers to hours or minutes for commercial customers. Without loss of generality, we assume an hourly metering period with no advance notice for our experiments and in the following discussion.

A customer must always be subscribed to one tariff in order to maintain electricity supply. The customer is allowed to choose an alternate tariff at any time, effective starting at the next metering period, for a fixed *switching cost*. The prices conveyed through a variable rate specification are a key component of the uncertainty in evaluating which tariff is best suited for a particular customer. Since prices evolve over time, the customer benefits from reevaluating the tariffs continuously and thus tariff selection is better described as a decision process rather than a singular event.

We define the resulting *tariff selection decision process* (TSDP) over the discrete time sequence, $\mathcal{T} = 1..T$. Given a set of tariffs, $\mathcal{X}$, the policy, π, of the decision process for a customer, D, is a map of tariff *subscriptions* over time:

$$\pi_D : \mathcal{T} \rightarrow \mathcal{X}$$

We assume that the TSDP is given a set of *demand forecasts*, $\mathcal{Y}$, which represent the possible consumption patterns over a tariff evaluation horizon, H. Thus, each forecast $\hat{y} \in Y$ is a map $\mathcal{T}_t^{t+H} \rightarrow \mathbb{R}^+$.

At time t, tariff, $x \in \mathcal{X}$, specifies a price $p_x(t)$. Then, let $p_D^\pi(t)$ be the price specified by the tariff $x_D^\pi(t)$, *i.e.*, the tariff to which the customer is subscribed at time t. The goal of the agent is to minimize the lifetime cost of electricity over the sequence of observed demand levels, $y(t)$, given the demand forecasts $\mathcal{Y}$:

$$\min_\pi \sum_{t=1..T} p_D^\pi(t) y(t)$$

This simple definition of the problem is similar to the nonstochastic or adversarial multi-armed bandit problem, where a gambler must choose one of several slot machines–*bandits*–to play at each timeslot under no statistical distribution assumptions for the rewards from each bandit (Auer et al. 1995). For this problem, the Exp3 family of algorithms provides strategies for balancing exploration and exploitation using exponential-weighting to achieve optimal performance bounds. However, as we show in experimental results, our Negotiated Learning algorithm produces significantly better results than Exp3/Exp3.P/Exp3.S when applied to the tariff selection problem because our approach exploits the specific multiagent structure of the problem.

Negotiable Partial Observability

Fundamentally, the uncertainty in the attractiveness of tariff choices is due to three reasons:

1. **Price Imputation Uncertainty**: When prices in variable rate tariffs are published only to customers that subscribe to the tariff, it is possible that for some tariffs the only historical price information available to the customer agent is some initial or reference price. Then, the agent must apply an *imputation model* to estimate any missing prices.
2. **Price Prediction Uncertainty**: Even if perfect information about past prices is available, the agent must still apply a *prediction model* to estimate how the prices will evolve in the future, over some *tariff evaluation horizon*.
3. **Demand Prediction Uncertainty**: Forecasts of customer demand typically increase in uncertainty as the time span of the forecast increases. Moreover, if the demand for a certain period is very low, switching to a better tariff is not as compelling during that period.

Since tariff selection is a forward-looking optimization, only the uncertainty in predicted prices and demand forecasts affect the decision. We assume here that the demand prediction uncertainty is difficult to mitigate as it stems from factors that the agent cannot observe or control. However, price predictions are often highly dependent on price histories, which raises the question of whether the agent can improve its price predictions, and therefore its tariff selection decisions, by mitigating the price imputation uncertainty.

We observe that since tariffs are published contracts, the prices for a particular variable rate specification are the same for all potential customers. So, even though the prices for tariffs that the customer is not subscribed to are hidden from the customer agent, the agent has the ability to potentially acquire current *price samples* or entire *price histories* from other customers who are subscribed to those tariffs. Thus, it is possible for the population of customers to *cooperatively* pool their information and decrease the amount of hidden information for each of them. However, we assume a more realistic model where each customer is *self-interested* and *semi-cooperative*; i.e., each customer needs to be incentivized to share their information. Incentives can take many forms such as in-kind exchange of information, credits for future use, or cash payments. If our decision-making agent wants to acquire information from another customer, it must *negotiate* with that customer for that information. We can intuitively expect, as we also demonstrate in our experiments, that *learning from this negotiated information* can significantly reduce the price imputation uncertainty.

We can draw a parallel between this insight and *oracles* in POMDPs (Armstrong-Crews and Veloso 2007). The customer agent can view the population of other customers as a *multiagent oracle*, albeit an incomplete one since some information is hidden from all customers. We refer to this semi-cooperative multiagent structure as *negotiable partial-observability*. In following sections, we will enrich the formal definition of the tariff selection problem to explicitly represent this structure and also describe in detail how our Negotiated Learning algorithm addresses the price imputation, price prediction and demand prediction uncertainties.

Negotiable Entity Selection Process

Let D be a sequential decision-making agent that chooses one *entity* from a variable set of entity choices, $\mathcal{X}(t)$, at each time step t. The optimal choice at each t depends on the values of dynamic partially observable *features*, $\mathcal{F}$, that characterize each entity. The environment includes a set of neighboring agents, $\mathcal{I}$, a set of agent classes, $\mathcal{K}$, and an *agent classification model*, $\mathbf{K} = \mathcal{I} \rightarrow \mathcal{K}$.

Let $\mathbf{S}$ be D's *state model* which includes an occluded view of the dynamic features of each entity. Given a state $s \in \mathbf{S}$, let $\varphi(s, \mathcal{F})$ be the set of *state transforms* that can be reached by obtaining more information about any of the occluded features in $\mathcal{F}$; *i.e.*, each state in $\varphi(s, \mathcal{F})$ mitigates the uncertainty in one or more occluded features in s.

We can then define a *negotiation* as a pair (s', k), $s' \in \varphi(s, \mathcal{F})$, and $k \in \mathcal{K}$, which describes the action of gathering information from any agent in k about the features that transform s to s'. Thus, $\mathcal{A}_1(t) = \varphi(s(t)) \times \mathcal{K}$ defines the set of *negotiation actions* available at t.

We then define a *negotiation model*, $\mathbf{N}$, which maps each possible negotiation to a triple of *negotiation parameters*, (c, t, p), $c \in \mathbb{R}$, $t \in \mathbb{I}^+$, and $p \in \mathbb{R}[0, 1]$:

$$\varphi(\mathbf{S}, \mathcal{F}) \times \mathcal{K} \rightarrow \{(c, t, p)_i\}_{i=1}^{m}$$

where c is the *cost of information*, t is *interval to information*, and p is the *probability of information*, with the intuitive understanding that if information is made available quickly (low t) with high reliability (high p), then the cost, c, of the negotiation is likely to be higher.

Note that defining negotiations on agent classes instead of individual agents offers better scalability of the model for large $|\mathcal{I}|$. As illustrated in Figure 1, the negotiation model can also be viewed as a bipartite graph from the state transforms, $\varphi(\mathbf{S}, \mathcal{F})$, to the agent classes, $\mathcal{K}$, with each of the m edges of the graph carrying a (c, t, p) instance.

Let $\mathcal{A}_2(t)$ be the size $|\mathcal{X}(t)|$ set of *entity selection actions* available at t and $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$. The agent's policy, π, chooses one action from $\mathcal{A}_2(t)$ along with any combination from $\mathcal{A}_1(s(t))$ at each t. Finally, let $\mathbf{T}$ be a transition model and $\mathbf{R}$ a reward model as they are usually defined in Markov decision processes. A *negotiable entity selection process*, Z, for agent D is then defined as:

$$Z_D = \langle \mathbf{K}, \mathcal{X}, \varphi(\mathbf{S}, \mathcal{F}), \mathbf{N}, \mathcal{A}, \mathbf{T}, \mathbf{R} \rangle$$

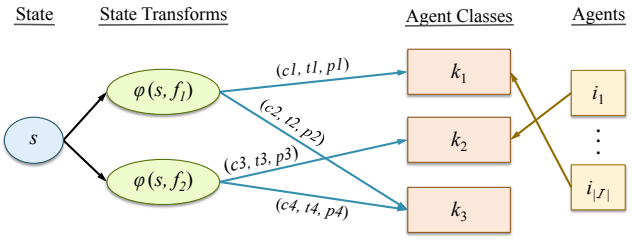


Figure 1: An example *negotiation model* with $|\mathcal{F}|$=2 and $|\mathcal{K}|$=3 as a bigraph from *state transforms* to *agent classes* with each *negotiation* edge carrying a triple of *parameters*.

We can then formulate the variable rate tariff selection problem as an NESP by translating 'entities' as 'tariffs' and casting the dynamic prices as entity features. Concretely, we define two features: (i) *price sample*, and (ii) *price history*. A request for a price sample, $p_x(t)$, on a tariff, x, can be fulfilled by a neighbor only if that neighbor is currently subscribed to x, whereas a request for price history yields all the price samples known to the neighbor for x.

We also formulate a simple set of agent classes, $\mathcal{K} = \{$*Desirable, Undesirable*$\}$, which represent the relative bias for obtaining information from agents in those classes based on a weighted combination of (c, t, p) values.

Negotiated Learning

We can now describe the Negotiated Learning algorithm. While the algorithm is generally applicable for any problem that can be defined as a Negotiable Entity Selection Process, we will continue to use the terminology of tariffs/prices instead of entities/features for clarity.

The algorithm forms a three-layered learning process:

1. *Learning from negotiated information*: Price samples and histories obtained through negotiation are used in the agent's price imputation models. The resulting imputed price series are combined with price prediction models and demand prediction models to compute *Attractions* for each tariff, which determine the tariff chosen at time t. An Attraction is defined by the triple (μ, β^+, β^-), which can be interpreted as the mean, upper and lower confidence bounds on some measure of attractiveness.
2. *Learning the negotiation model*: If $\mathbf{N}$ is hidden, the agent's history of negotiations is used to estimate the negotiation parameters, (c, t, p), for the edges in the bipartite graph representing the negotiation model.
3. *Learning the agent classification model*: If $\mathbf{K}$ is hidden, the neighboring agents in $\mathcal{I}$ are dynamically mapped into the $\mathcal{K}$ classes based on past negotiations.

The first layer is summarized in Algorithms 1-3. The second and third layers are included in Algorithm 3. Algorithm 1 is activated at each t with the negotiable entity selection process Z, the currently selected tariff x_t^π, a deferred tariff selection x_τ^π, the current *Attractions* for each tariff $\mathcal{V}_t$, the map of *ongoing negotiations* $\mathcal{N}$ keyed by tariff, learning-rate parameters $\Omega = \{\omega_e, \omega_b, \omega_\alpha\}$ all in $\mathbb{R}[0, 1]$, a confidence bounds growth parameter $\lambda \in \mathbb{R}^+$, and a negotiation budget parameter $\gamma \in \mathbb{R}[0, 1]$.

We first initialize the tariff selection decision x_{t+1}^π to the current tariff, and then obtain a set of current demand forecasts from the environment. We then update the Attraction for the current tariff according to Algorithm 2, which we will describe shortly. Then for all other tariffs x in the current tariff choices $\mathcal{X}$, we first check to see if x is included in the ongoing negotiations $\mathcal{N}$. If it is and the corresponding negotiation completed successfully, then the price samples or histories obtained through negotiation are incorporated into the state model $\mathbf{S}$. That negotiated information is also used to recompute the Attraction of x using the *experience update* weight, ω_e, instead of the *belief update* weight, ω_b.

Algorithm 1 NLActivate$(t, Z, x_t^\pi, x_\tau^\pi, \mathcal{V}_t, \mathcal{N}, \Omega, \gamma, \lambda)$

$x_{t+1}^\pi \leftarrow x_t^\pi$
$\eta \leftarrow 0$
$\mathcal{Y} \leftarrow$ **Env.GetDemandForecasts**(t, Z)
$v^\pi \leftarrow$ **NLUpdateAttraction**$(x^\pi, \mathcal{V}[x_t^\pi], \mathcal{Y}, \Omega.\omega_e, \lambda)$
$\mathcal{V}_{t+1} \leftarrow \emptyset \cup v$
for x in $Z.\mathcal{X} \setminus x_t^\pi$ **do**
  $\omega \leftarrow \Omega.\omega_b$
  if x in *keys*$(\mathcal{N})$ **then**
    $n \leftarrow \mathcal{N}[x]$
    if n.*success* = TRUE **then**
      NLUpdateTariffPrices$(Z.\mathbf{S}, x, n)$
      $\omega \leftarrow \Omega.\omega_e$
  $v \leftarrow$ **NLUpdateAttraction**$(x, \mathcal{V}[x], \mathcal{Y}, \omega, \lambda)$
  $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_{t+1} \cup v$
  if $v.\beta^+ > v^\pi.\beta^+ \parallel v.\beta^- > v^\pi.\beta^-$ **then**
    $\mathcal{U} \leftarrow \mathcal{U} \cup x$
    if $(v.\beta^+ - v^\pi.\beta^+) > (\eta/\gamma)$ **then**
      $\eta \leftarrow \gamma * (v.\beta^+ - v^\pi.\beta^+)$
  if $v.\mu > v^\pi.\mu$ **then**
    if $x = x_\tau^\pi$ **then**
      $x_{t+1}^\pi \leftarrow x$
    else
      $x_\tau^\pi \leftarrow x$
$\mathcal{N} \leftarrow \mathcal{N} \cup$ **NLInvokeNegotiations**$(\mathcal{N}, \mathcal{U}, \eta, \Omega.\omega_\alpha)$

If the upper or lower confidence bound, β^+ and β^-, for x's Attraction is higher than that of the current tariff, it is added to the set of *uncertain tariffs*, $\mathcal{U}$, to be considered for negotiation. If the *mean* of x's Attraction is greater than that of the current tariff, then it is saved as the deferred tariff x_τ^π, unless it is already the deferred tariff in which case x is chosen to replace the current tariff for time $t+1$. Finally, the uncertain tariffs are evaluated for possible negotiation by invoking Algorithm 3. The negotiations are constrained by a budget, η, that is a γ-fraction of the best bounded benefit over all $x \in \mathcal{X} \setminus x_t^\pi$ according to the computed Attraction values.

Attractions are a key component of our approach because they effectively capture the uncertainties in price imputation, price prediction, and demand prediction. Algorithm 2 describes their computation. We assume that we have a library of domain-dependent *price imputation models*, Γ_i, that fill in missing historical prices. We also assume a similar library of *price prediction models*, Γ_p. We generate one imputation per imputation model, and then generate a set of price predictions for each imputation using each price prediction model. We recognize that demand forecasts have higher uncertainty farther into the future, so we give more importance to forecast values for the near future. We do this by choosing a set of *lookahead thresholds*, $\mathcal{L}$, all less than the tariff evaluation horizon. For each lookahead threshold we compute the average charge over the set of demand forecasts for each price prediction. We thus collect $|\Gamma_i| \times |\Gamma_p| \times |\mathcal{Y}| \times |\mathcal{L}|$ real-valued charges, the mean of which is used to update the Attraction's mean. The standard deviation is used for the upper and lower confidence bounds along with λ, a parameter which allows the bounds to diverge over time to trigger exploration.

Algorithm 2 NLUpdateAttraction$(x, v, \mathcal{Y}, \omega, \lambda)$

Charges $\leftarrow \emptyset$
for i in Γ_i **do**
  $\overrightarrow{h} \leftarrow$ **GeneratePriceImputation**(x, i)
  for j in Γ_p **do**
    $\overrightarrow{p} \leftarrow$ **GeneratePricePrediction**$(j, \overrightarrow{h})$
    for $\overrightarrow{y}$ in $\mathcal{Y}$ **do**
      for l in $\mathcal{L}$ **do**
        Charges $\leftarrow$ *Charges* $\cup$ $(\overrightarrow{p}[1..l] \cdot \overrightarrow{y}[1..l])/l$
$v.\mu \leftarrow (1-\omega) * v.\mu + \omega *$ **Mean**$($*Charges*$)$
$\delta \leftarrow (1+\lambda) * 2 *$ **StdDev**$($*Charges*$)$
$v.\beta^+ \leftarrow (1-\omega) * v.\beta^+ + \omega * (v.\mu + \delta)$
$v.\beta^- \leftarrow (1-\omega) * v.\beta^- + \omega * (v.\mu - \delta)$

In addition to the negotiation model, $\mathbf{N}$, which includes *negotiation parameters* for each possible negotiation, the agent maintains a *neighbor model*, $\mathbf{B}$, which includes its beliefs about which tariffs its neighbors are subscribed to and about the agent classification model, $\mathbf{K}$; *i.e.*, (i) $\mathbf{B}.\mathcal{X} = \mathcal{I} \rightarrow \mathcal{X}$, and (ii) $\mathbf{B}.\mathcal{K} = \mathcal{I} \rightarrow \mathcal{K}$. Algorithm 3 first uses information from each completed negotiation in $\mathcal{N}$ to update the tariff mapping for the neighbor involved in the negotiation. It then applies a weighted update to the negotiation parameters for the neighbor's agent class. The cost c of the negotiation is determined by the neighbor and p is 0 or 1 to indicate negotiation failure or success. If the agent classification model is unknown, then we reclassify the neighbor based on the negotiation's (c, t, p) and domain-specific heuristics. Then, let $\mathcal{O}$ be the set of *negotiation options* derived as the cross product of the possible state transforms on the uncertain tariffs, $\varphi(\mathcal{U}, \mathcal{F})$, with the agent classes, $\mathcal{K}$. We then obtain a set of desired negotiations by solving a zero-one program over $\mathcal{O}$ with the goal of maximizing the expected information and the constraints that (i) the total cost of the negotiations is under the budget η, and (ii) no more than one option is chosen for a particular state transform. For each desired negotiation, with probability $1-\epsilon$ we find the neighboring agent that was most recently mapped to that tariff in $\mathbf{B}.\mathcal{X}$, and randomly otherwise, and initiate negotiation with that neighbor.

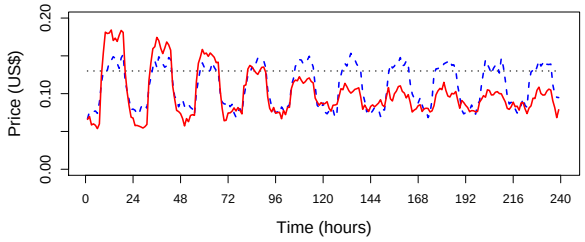
Algorithm 3 NLInvokeNegotiations$(\mathcal{N}, \mathcal{U}, \eta, \alpha)$

for n in $\mathcal{N}$ **do**
  if n.*status* = Completed **then**
    $\mathbf{B}.\mathcal{X}[n.i] \leftarrow n.x$
    old.(c,t,p) $\leftarrow (1-\alpha) * \mathbf{N}[\mathbf{K}(n.i)].(c,t,p)$
    $\mathbf{N}[\mathbf{K}(n.i)].(c,t,p) \leftarrow$ *old.*$(c,t,p) + \alpha * n.(c,t,p)$
  $\mathbf{B}.\mathcal{K}[n.i] \leftarrow$ **NLReclassifyNeighbor**$(n.i)$
  $\mathcal{N} \leftarrow \mathcal{N} \setminus n$
$\mathcal{O} \leftarrow \varphi(\mathcal{U}, \mathcal{F}) \times \mathcal{K}$
$\mathcal{N}^* \leftarrow$ **ZeroOneProgram**$(\mathcal{O}, \eta)$
$\mathcal{N} \leftarrow \mathcal{N} \cup N^*$
for n in $\mathcal{N}^*$ **do**
  $i \leftarrow$ **NLSelectNeighbor**$(n.k)$
  Env.InitiateNegotiation(i, n)

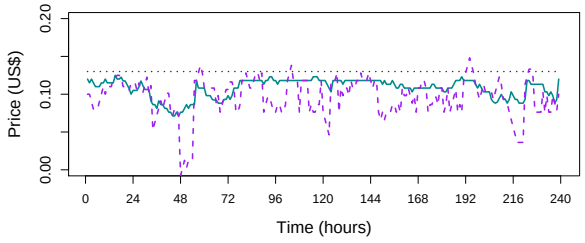
Experimental Results

We present results from simulation experiments to demonstrate how our variable rate tariff selection approach exploits favorable tariff price movements to generate cost savings for customers. We simulate 60 agents of various capabilities in 10-day episodes, using the Power TAC simulation environment (Ketter, Collins, and Reddy 2012). We generate demand forecasts using noise-added subsets of real hourly consumption data for homes in California (San Diego Gas & Electric 2012).

We use a combination of heuristic and reference simulation data to generate tariff prices. Figure 2a shows hourly prices over 10 days for 3 tariffs: (i) the dotted black line represents a fixed default utility tariff, (ii) the dashed blue line represents a stable *dynamic TOU* tariff where each cycle of the pattern represents a day and prices are generally higher 8am-8pm, and (iii) the solid red line represents another similarly structured but non-stationary tariff whose prices are more volatile and less attractive initially but then stabilize over time into a more attractive option. Figure 2b is an illustration of two additional tariffs that are drawn for each simulation episode from a reference set of simulated tariff prices offered by competitive electricity suppliers who employ various pricing strategies that include variable rates that are indexed to a wholesale electricity market (IESO of Ontario 2011), and related adaptive and learning-based pricing strategies that optimize for the supplier's profit maximizing goals (Reddy and Veloso 2011).



(a) Generated prices for fixed and *dynamic TOU* tariffs.



(b) Sample variable rate prices from reference data set.

Figure 2: We use heuristically generated prices along with reference prices offered by simulated competitive suppliers.

We use four price imputation models to estimate hidden values in the price history of each tariff using known prices, which may have been observed directly as a subscriber to the tariff or obtained through negotiation:

1. **GlobalMean**: Set equal to the mean of the known values.
2. **CarryForward**: Set equal to the prior known value.
3. **BackPropagation**: Set equal to the next known value, or to the prior known value if all later prices are hidden.
4. **Interpolation**: Each contiguous sequence of hidden values is assigned using interpolation from the prior known value to the next known value.

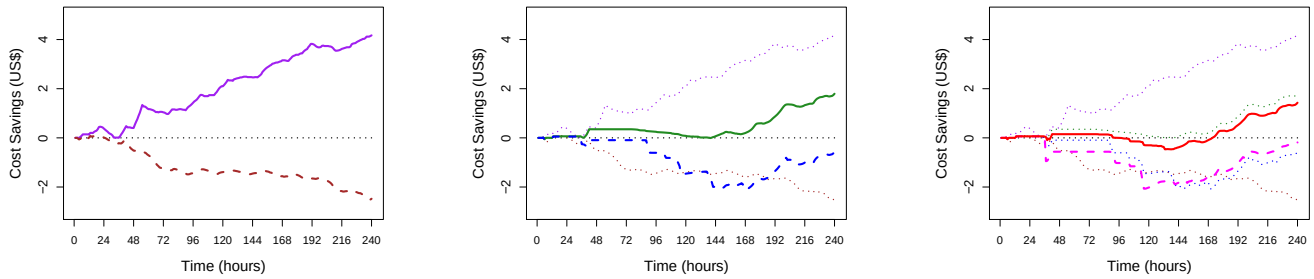
We also use four price prediction models, based on domain knowledge that wholesale market prices for a given hour are well correlated with prices at the previous hour and at the same hour the previous day:

1. **Lag24**: $p_t = p_{t-24} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$
2. **AR(1)**: $p_t = \mu + \phi\, p_{t-1} + \varepsilon$
3. **ARMA(1,1)**: $p_t = \mu + \phi\, p_{t-1} + \varepsilon + \theta\, \varepsilon_{t-1}$
4. **Seasonal ARMA(1,1)$\times$(1,1)$_{24}$**: $p_t = \mu + \phi\, p_{t-1} + \Phi\, p_{t-24} + \varepsilon + \theta\, \varepsilon_{t-1} + \Theta\, \varepsilon_{t-24} + \Theta\, \theta\, \varepsilon_{t-25}$

Figure 3a demonstrates the value of exploiting the multiagent structure of the problem. The y-axis shows the cost savings for a given algorithm relative to a *baseline* algorithm (dotted black line) that explores the available tariffs using the imputation models and prediction models but does not negotiate for information, thus missing opportunities to select better tariffs. The dashed brown line below the baseline, which holds with insignificant variations for Exp3, Exp3.P, and Exp3.S, shows negative savings, *i.e.*, higher cost than the baseline. The solid purple line demonstrates the significant opportunity for cost savings if the algorithm had full information about the dynamic prices for all tariffs. Each line represents averaged results for 10 agents of that type.

A fully-informed algorithm sets the upper bound on cost savings, but it is unrealistic in our setting. Figure 3b shows approximate bounds for algorithms that acquire information through negotiation. The flat line represents the same baseline as before. The solid green line establishes an upper bound on cost savings for a negotiating agent, given accurate negotiation and agent classification models. Conversely, the dashed blue line shows the negative savings if a negotiation model is not used, *i.e.*, the agent chooses a neighbor to negotiate with randomly. The gap between the two new lines illustrates the value of the negotiation model.

We set up the experiments such that agents in the different classes, $\{$*Desirable, Undesirable*$\}$, exhibit different (c, t, p) attributes, *i.e.*, charge different prices, require varying amounts of time to respond, and vary in reliability, depending on the agent class they belong to. The dashed magenta line in Figure 3c illustrates performance slightly better than random negotiation when the agent classification model and the negotiation model are initially unknown to the agent. The agent then learns the models over multiple episodes, with each episode using a possibly different set of tariff prices from the reference set and different real data subsets for the demand forecasts. As learning progresses, the agent's performance approaches the benchmark performance obtained with known negotiation and agent classification models, as illustrated by the solid red line.



(a) *Fully informed agents outperform agents based on exploration-exploitation.*

(b) *Agents with a negotiation model outperform agents who negotiate randomly.*

(c) *Agents can approach known-model performance through model learning.*

Figure 3: *Our experiments demonstrate the efficacy of Negotiated Learning in (a) extracting the value of negotiated information, (b) extracting the value of an informed negotiation model, and (c) incorporating the learning of the negotiation model.*

Related Work

Research on Smart Grid agents explores the reliability of grid infrastructure, *e.g.*, (Gellings, Samotyj, and Howe 2004), integration of distributed generation sources, *e.g.*, (Kok 2010), electric vehicles as micro storage (Vytelingum et al. 2010), and adaptive policies for distributed control of demand based on dynamic prices (Ramchurn et al. 2011). Our current research builds upon our previous work (Reddy and Veloso 2012), which introduces autonomous decision-making agents for Smart Grid customers in liberalized markets; we complement the focus on algorithms for demand management in that work with our current focus on the customer's tariff selection decision process.

Exploiting structure in multiagent problems is studied extensively in machine learning, planning, and game theory. (Busoniu, Babuska, and De Schutter 2008) review several multiagent reinforcement learning algorithms. Other related examples of exploiting structure in MDPs/POMDPs include soft-state aggregation (Singh, Jaakkola, and Jordan 1994), semi-Markovian options (Sutton, Precup, and Singh 1999), layered Q-learning (Melo and Veloso 2009), and oracular POMPDs (Armstrong-Crews and Veloso 2007). Partially-observable stochastic games (POSGs) offer the most general representation but corresponding algorithms generally do not scale well (Emery-Montemerlo et al. 2004). A key aspect of our Negotiable Entity Selection Process representation is that it trades off some generality to expose elements of multiagent structure that are lost in other representations.

No-regret online learning (Foster and Vohra 1999) is closely related to the tariff selection problem and our approach is related to fictitious play (Fudenberg and Levine 1999) (Hart and Mas-Colell 2000). The Exp3 algorithms that we compare with in the multiarmed adversarial bandit setting are detailed in (Auer et al. 1995). Note that while we pursue the acquisition of partially hidden information, our setting does not match that of full-information no-regret learning (Littlestone and Warmuth 1994) where the payoffs of each *expert* are revealed after the current time step.

Extensive work also exists in dynamic coalition formation (Sandholm and Lesser 1995) but our Negotiated Learning approach varies in that each negotiation is a point-in-time transaction and there are no joint payoffs. (Crawford and Veloso 2008) demonstrate the elicitation of hidden attributes about neighbors through semi-cooperative negotiations, but they do not consider cost/payments or quality of information. Our use of *Attractions* is based largely on (Camerer and Ho 1999) who propose a behavioral framework that combines reinforcement learning and no-regret learning to learn from own experiences as well as beliefs about other agents' experiences. Cognitive hierarchies as a partition of agent populations with different levels of reasoning capability, which rationalizes our use of agent classes, is also due to (Camerer 2008). Our use of upper and lower confidence bounds in the Attractions draws upon model-based interval estimation (Strehl and Littman 2008) and the principle of communicating only when acquired information may change the agent's policy action (Roth 2003). However, none of these works combine negotiating for paid information and simultaneously learning a negotiation model to address partial observability.

Conclusion

We have contributed general models and algorithms for entity selection based on dynamic partially observable features, and have applied them towards variable rate tariff selection by Smart Grid customer agents. Our Negotiated Entity Selection Process is a novel representation, which captures the multiagent structure that enables the development of our Negotiated Learning algorithm. We have demonstrated through experimental results (i) the value of negotiated information, (ii) the importance of a well-informed negotiation model, and (iii) learnability of negotiation models. Future work could explore other negotiation models (*e.g.*, bipartite multigraphs) and negotiation selection approaches that constrain the budget for specific negotiations.

Acknowledgements

This research was partially sponsored by the Portuguese Science and Technology Foundation. The views/conclusions contained in this document are those of the authors only.

References

Armstrong-Crews, N., and Veloso, M. 2007. Oracular Partially Observable Markov Decision Processes: A Very Special Case. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science* 68(68):322–331.

Barbose, G.; Goldman, C.; and Neenan, B. 2005. Electricity in real time–a survey of utility experience with real time pricing. *Energy* 30.

Block, C.; Collins, J.; and Ketter, W. 2010. A Multi-Agent Energy Trading Competition. Technical report, Erasmus University Rotterdam.

Busoniu, L.; Babuska, R.; and De Schutter, B. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems Man and Cybernetics* 38(2):156–172.

Camerer, C. F., and Ho, T. H. 1999. Experience-Weighted Attraction Learning in Normal Form Games. *Econometrica* 67(4):827–874.

Camerer, C. F. 2008. Behavioral Game Theory and the Neural Basis of Strategic Choice. In *Neuroeconomics: Formal Models Of Decision-Making And Cognitive Neuroscience*. Academic Press. 193–206.

Crawford, E., and Veloso, M. 2008. Negotiation in Semi-Cooperative Agreement Problems. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

Emery-Montemerlo, R.; Gordon, G.; Schneider, J.; and Thrun, S. 2004. Approximate Solutions for Partially Observable Stochastic Games with Common Payoffs. In *International Conference on Autonomous Agents*, 136–143.

Faruqui, A., and Palmer, J. 2012. Dynamic Pricing and Its Discontents. Technical report, The Brattle Group.

Foster, D. P., and Vohra, R. 1999. Regret in the On-Line Decision Problem. *Games and Economic Behavior* 29(1-2):7–35.

Fudenberg, D., and Levine, D. K. 1999. Conditional Universal Consistency. *Games and Economic Behavior* 29(1-2):104–130.

Gellings, C.; Samotyj, M.; and Howe, B. 2004. The future power delivery system. *IEEE Power & Energy* 2(5):40–48.

Gomes, C. 2009. Computational Sustainability: Computational Methods for a Sustainable Environment. *The Bridge, National Academy of Engineering* 39.

Hammerstrom, D. 2008. Pacific Northwest GridWise Testbed Demonstration Projects; Part I. Olympic Peninsula Project. Technical report, PNNL-17167, Pacific Northwest National Laboratory.

Hart, S., and Mas-Colell, A. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68:1127–1150.

IESO of Ontario. 2011. http://www.ieso.ca.

Ketter, W.; Collins, J.; and Reddy, P. 2012. Power TAC: A Competitive Simulation of the Smart Grid for Autonomous Trading Agents. In submission.

Kok, K. 2010. Multi-Agent Coordination in the Electricity Grid, from Concept towards Market Introduction. (Aamas):1681–1688.

Littlestone, N., and Warmuth, M. K. 1994. The weighted majority algorithm. *Information and Computation* 108(2):212–261.

Melo, F. S., and Veloso, M. 2009. Learning of Coordination: Exploiting Sparse Interactions in Multiagent Systems. *Autonomous Agents and MultiAgent Systems* 773–780.

Ramchurn, S. D.; Vytelingum, P.; Rogers, A.; and Jennings, N. 2011. Agent-Based Control for Decentralised Demand Side Management in the Smart Grid. In *Autonomous Agents and Multiagent Systems AAMAS 11*.

Reddy, P., and Veloso, M. 2011. Learned Behaviors of Multiple Autonomous Agents in Smart Grid Markets. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI-11)*.

Reddy, P., and Veloso, M. 2012. Factored Models for Multiscale Decision-Making in Smart Grid Customers. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*.

Roth, M. 2003. *Execution-time communication decisions for coordination of multi-agent teams*. Ph.D. Dissertation.

San Diego Gas & Electric. 2012. http://www.sdge.com.

Sandholm, T., and Lesser, V. R. 1995. Coalition Formation among Bounded Rational Agents. *International Joint Conference on Artificial Intelligence* 14(1):662–669.

Singh, S.; Jaakkola, T.; and Jordan, M. I. 1994. Reinforcement Learning with Soft State Aggregation. In *Advances in Neural Information Processing Systems (NIPS)*.

Strbac, G. 2008. Demand side management: benefits and challenges. *Energy Policy 36 (12), 44194426*.

Strehl, A. L., and Littman, M. L. 2008. An Analysis of Model-Based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences* 74(8):1309–1331.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112(1-2):181–211.

Vytelingum, P.; Voice, T. D.; Ramchurn, S. D.; Rogers, A.; and Jennings, N. R. 2010. Agent-based Micro-Storage Management for the Smart Grid. (Aamas):10–14.