Reinforcement Learning for Optimal Energy Management of a Solar Microgrid

R Leo
SSN College of Engineering
leor@ssn.edu.in

R S Milton
SSN College of Engineering
miltonrs@ssn.edu.in

S Sibi
Student, SSN College of Engineering
sibi.vasank@gmail.com

***Abstract*—In an optimization based control approach for solar microgrid energy management, consumer as an agent continuously interacts with the environment and learns to take optimal actions autonomously to reduce the power consumption from grid. Learning is built in directly into the consumer's behaviour so that he can decide and act in his own interest for optimal scheduling. The consumer evolves by interacting with the influencing variables of the environment. We consider a grid-connected solar microgrid system which contains a local consumer, a renewable generator (solar photovoltaic system) and a storage facility (battery). A model-free Reinforcement Learning algorithm, namely three-step-ahead Q-learning, is used to optimize the battery scheduling in dynamic environment of load and available solar power. Solar power and the load feed the reinforcement learning algorithm. By increasing the utility of battery and the solar power generator, an optimal performance of solar microgrid is achieved. Simulation results using real numerical data are presented for a reliability test of the system. The uncertainties in the solar power and the load are taken into account in the proposed control framework.**

***Index Terms*—Solar microgrid; Reinforcement learning; Q-learning; Battery scheduling; Optimization.**

I. Introduction

Renewable energy plays a significant role in building a green and sustainable environment. The world economy is largely dependent on quality power. Solar and wind are the only solution to the growing energy crisis in the world. We require a "smartgrid" approach in order to deal with distributed production, intermittent variations of renewable energy and optimal scheduling of the demand [1]. Microgrid is the building block of smart grid. Integrating renewable energy into microgrid is the way forward for economic and environmental optimization, generating clean and green energy and, thereby, providing solution to the global warming [2]. The importance of having more reliable, efficient, smart systems is getting more public attention. In coming years, consumer wants smart machines and expect machines to think and operate autonomously and optimally. Energy management of microgrids using fuzzy logic is discussed in [2]. Smart energy management of microgrids using genetic algorithm is discussed in [3] and [4]. Energy management of hybrid renewable energy generation using constrained optimization was proposed in [5]. Expert system and other classical and heuristic algorithms for energy management of microgrids are discussed in [6],[7] and [9]. An agent-based modelling approach is used to model microgrids and, by simulation, the interaction between individual intelligent decision-makers are analysed in [8]. In all these methods, the concept of learning is not emphasized and interactive learning is not feasible. Reinforcement Learning (RL) has been shown to achieve excellent performance in dynamic power management of wind micrgrid systems in [10]. Provably convergent algorithms are available using interactive learning to solve single agent learning tasks, whether the environment state space is small [11], large or continuous [15]. Strategic bidding using reinforcement learning in microgrid is discussed in [14]. In the current work, we propose a three-step-ahead Q-learning method for optimal energy management of solar microgrid where intelligence or learning ability is embedded in the agents so the agents learn as humans learn by trial and error. After enough interactive learning the agent learns to optimally act looking at long term objectives. The agent not only replaces the human intervention but can also take action to impact its long term objectives. This approach introduces the concept of autonomous optimization.

The rest of the paper is organized as follows. Section II presents the modelling framework of the solar microgrid and the details of solar photovoltaic system. Section III provides a comprehensive framework of consumer energy management with reinforcement learning. In section IV, case studies and simulation results are analysed. The performance of the solar microgrid system in the long run is given in section V. Conclusion and possible improvements are given in the last section.

II. Model of the Solar Microgrid

A microgrid is a localized grouping of electricity sources and loads that normally operates connected to and synchronous with the traditional centralized grid (macrogrid) but can disconnect and function autonomously as physical and/or economic conditions dictate [2]. The urban solar microgrid involves a consumer with a dynamically varying load D_t, a transformer providing electricity power from the external grid, a solar generator (solar photovoltaic system) with available power output P_{sp} and a storage facility with a level of battery charge R_t. The architecture of the considered microgrid is shown in Fig. 1. The consumer can cover his demand partly by using the electricity produced by the local renewable (solar) generator, store electricity in the battery when the solar source is available and can discharge the storage when needed. The consumer has the possibility to control the storage and the

978-1-4799-4097-4/14/$31.00 ©2014 IEEE

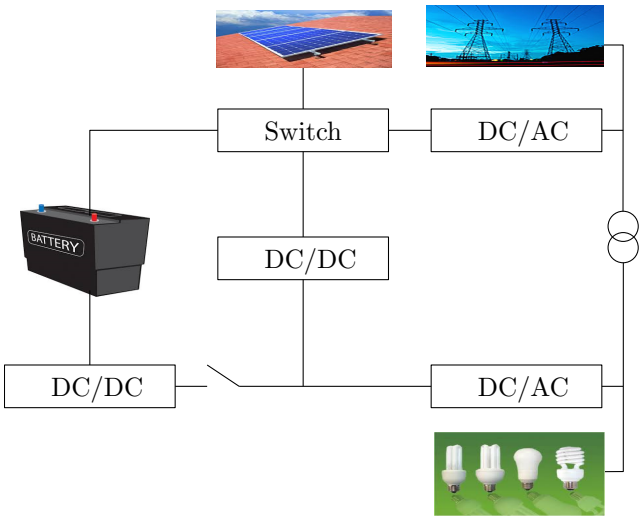

Fig. 1. Solar Microgrid

solar power generator. The main aim of solar microgrid is to satisfy the requirements of the load qualitatively while maximizing the utilization of the solar power and optimizing the operation of the battery and solar photovoltaic (PV) system.

A. Solar photovoltaic (PV) module

A photovoltaic (PV) module is the basic element of each photovoltaic system. A photovoltaic system is an arrangement of components designed to supply usable electric power for a variety of purposes, using the sun as the power source. A PV module is a clean energy source used in power systems which absorb solar irradiation and converts it into electric energy [12]. PV generation fits the load demand very well since solar irradiation is higher in daytime. Due to the low voltage of an individual solar cell, several cells are wired to form modules. A solar PV module consists of a number of solar cells connected in series or parallel based on the requirement. Modules may then be strung together into a photovoltaic array. There are various factors which affect the solar power like solar irradiance, temperature, partial shading, cloud, arrangement of the cells and the angle of tilt of the panel. We use mathematical modelling to get electrical parameters from solar irradiance (G) and temperature (T) of the cell. Then we use a simple feed forward neural network and train it with G and T to get equivalent circuit parameters. Once the neural network is trained with sufficient number of examples then we can determine the current and the voltage of the solar module for untrained values of G and T by generalisation. The maximum power P_{sp} is found by Maximum Power Point Tracking (MPPT) algorithm. The MPPT refers to the point with a maximum power output in the curve under specific external temperature and solar irradiation [13].

B. Model for the battery storage

A simplified model of battery storage dynamics is adopted by implementing a discrete time system for the power flow dynamics over the time step interval D_t

$$R_t = R_{t-1} + R_t^{\text{store.charge}} + R_t^{\text{store.discharge}}$$

where R_t and R_{t-1} are the levels of energy stored in the battery at time t and $t-1$ respectively, and $R_t^{\text{store.charge}}$, $R_t^{\text{store.discharge}}$ are the power flows over the time step interval δt from solar generator to battery, and from battery to consumer load respectively [10].

III. Modelling of the Consumer Agent

The dynamic variations of load, solar power and the battery are considered to constitute the external environment. The consumer is modelled as an individual agent who makes use of reinforcement learning for its decision-making, action-taking and moving towards its goal. Reinforcement learning deals with learning in sequential decision making in the problems with limited feedback. MDP has become standard formalism for learning in sequential decision making. In an MDP the environment modelled as set of states, actions can be performed to control the system state. The effect of an action taken in a state is dependent only on that state and not on the prior history. The goal is to control the system in such a way that some performance criterion is maximized. This section presents the reinforcement learning algorithm used by the consumer agent to interact, adapt, and take decisions towards its goal defined in the form of reward functions in the MDP environment, characterized by the available solar power output P_{sp}, the load D_t and the level of battery charge R_t.

Reinforcement learning algorithm is used to model the consumer's adaptation to a dynamically changing environment by performing actions of battery scheduling in an MDP environment [10]. The agents observe the environment and take an action. It gets a reward or punishment from the environment. The agent takes the next action to optimize the reward in the long run. After a number of interactions, the agent finds the optimal policy to achieve long term objective. The goal of an

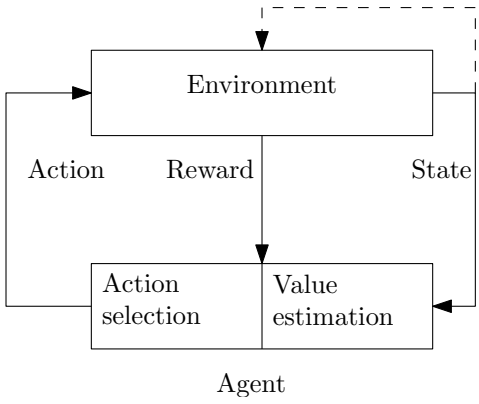

Fig. 2. Reinforcement Learning

agent is to find the optimal policy based on interactive learning with the environment. Fig. 2 shows a simple reinforcement learning scheme. The environment can be characterised by the values of a certain number of its features, collectively

called its state, denoted by $S(t)$ at time t. Each state has an intrinsic value, dependent upon a certain immediate reward or cost, denoted by $R(t)$ at time t, which is generated when the state is entered. At each discrete moment in time the agent may take one of a number of possible actions, $A(t)$, which affects the next state of the system, $S(t+1)$, and therefore the next reward/cost experienced, according to certain transition probabilities. The agent's choice of action, given the current state of the system, is modified by its experience, i.e., it uses its past experience of action taken in a certain state and the concomitant reward/cost experienced, to update its decision making process for future actions [11].

A. Markov Decision Process

Markov Decision Process (MDP) is a way to model a sequential decision making under uncertainty. We formalize an MDP, considering discrete states and actions. The initial state is s_0 and each state will have a reward r associated with it. The transition function $T(s'|a, s)$ indicates the probability of transitioning from state s to s' when action a is taken. A discount factor γ in the range $0 \dots 1$ is applied to future rewards. This represents the notion that a current reward is more valuable than one in the future. If it is near zero, future rewards are almost ignored; a near one places great value on future reward. The reward from a policy is the sum of the discounted expected utility of each state visited by that policy. The optimal policy is the policy that maximizes the total expected discounted reward.

B. Q learning

Q learning is a model-free reinforcement learning where the agent explores the environment and finds the next reward plus the best the agent can do from the next state. In Q learning, the agent does not need to have any model of the environment. It only needs to know what states exist and what actions are possible in each state. We assign each state an estimated value, called a Q value [11]. When we visit a state and take an action we receive a reward. We use this reward to update our estimate of the value of that action in the long run. We visit the states infinitely often and the action values (Q values) are continuously updated till it becomes convergent. The Q learning algorithm is outlined in Algorithm 1 [11]. In the algorithm, γ is the discount factor and α, learning rate.

Algorithm 1: Q-learning

```
Set γ and rewards matrix R.
Initialize Q(s, a) arbitrarily.
foreach episode do
    Initialize s arbitrarily.
    repeat for each step of episode
        Select a in s using policy derived from Q
        Take action a, observe reward r, and next state s'
        Q(s, a) ←
        Q(s, a) + α[ r + γ max_a Q(s', a) − Q(s, a) ]
        s ← s'
    until s is terminal
end
```

C. Definition of scenario

Rewards are the environment response to action taken by the agent. Available solar power output and consumer load are the two variables defining the dynamic environment of the consumer. The values of these two variables at time step t define the system state $s_t^i = \langle D_t, P_{sp} \rangle$. The actual states at time t and the future states at the next three time steps $t+1$ and $t+2$ and $t+3$ define the generic $Scenario_t^l = s_t^i; s_(t+1)^p; s_(t+2)^n; s_(t+3)^m$ where i, p,n,m are the indexes of the system states and l is the scenario index [10]. Furthermore, the level of battery charge R_t at time step t for $Scenario_t^l$ must also be considered. At time step t, the decision must be made about the actions to take at times t, $t+1$, $t+2$ and $t+3$. The 3-steps-ahead sequence of four actions decides actions decided at time t for battery scheduling in the scenario and is denoted as $A_t^j = [a_t, a_{t+1}, a_{t+2}, a_{t+3}]$. At each time step t, action a_t can assume values $a_t = a_o$ or $a_t = a_1$, where action a_o corresponds to covering part of the consumer electricity demand by discharging the battery, whereas action a_1 corresponds to purchasing all the electricity demanded by the consumer from the solar generator while charging the battery. In this view, 16 sequences of actions are possible for each scenario. Starting from battery charge level of R_5, it can reach a maximum level of R_9 and a minimum level of R_1. Battery scheduling is a continuous process, where the final level of battery charges R_{t+4} for one scenario becomes the initial level of battery charge R_t for the subsequent scenario and so on.

D. Reward function

We optimize the battery scheduling of the solar microgrid by reinforcement learning. This is a process of action-reward dynamics, driven by quantitative performance indicators which evaluate the action or sequence of actions undertaken and feedback the value to adjust future scheduling decisions. The optimization of the numerical reward is achieved through the choice of the actions a_0 and a_1 of battery scheduling. Since the battery cannot be charged and discharged at the same time, only one of these actions can be selected and performed at any time step t. The consumer aims at increasing its performance by selecting an optimal sequence of actions for a 3-step-ahead time period. The reward functions are the response we get from the environment for the actions taken. If it is charging (a_1) then the reward function is minimum of P_{sp} and $B_{difference}$ and if it is discharging (a_0) then the reward function is minimum of D_t and B_{level}. Here, $B_{difference}$ is the difference between maximum possible charge and the current battery level (B_{level}). The successful scheduling of the battery, and thus the increase of the microgrid performance with respect to the consumer goals, is done by considering 3-steps-ahead scenarios. The predicted values of the load and the

available solar power output for 3-steps-ahead is used to select the optimal sequence of actions A_t^j for the entire scenario.

E. Battery scheduling

In a 3-step-MDP, each action sequence involves four actions of the battery. The combinations of these four actions give 16 possible action sequences in a scenario. The action sequences, for example, may be charge-charge-charge-discharge (a_1, a_1, a_1, a_0) or discharge-charge-charge-discharge (a_0, a_1, a_1, a_0), etc. The best action for 3-step-ahead planning in a dynamic environment is chosen based on the 3-step Q learning algorithm, which takes the available solar power P_{sp}, battery level, and the load D_t, and gives the best possible action sequence to achieve consumer's long term objective, namely to reduce the power consumption from the grid. Keeping the initial battery level at R_5 and using tree diagram, 16 possible battery action sequences are identified. The actions a_1 and a_0 corresponds to charging and discharging respectively. When charged and discharged, the battery level increases and decreases respectively. At time t, the best possible sequence of action is chosen for the predicted solar power (P_{sp}) and the load (D_t) for $t+1$, $t+2$ and $t+3$. The actions are chosen not just considering the immediate environment but future as well. For the 16 different possible sequence of actions in the scenario, the Q values are calculated by using Q learning algorithm and the convergent value, Q^* is found for all the 16 sequence of actions. The sequence of actions corresponding to the maximum Q^* value (Q^*_{max}) is the best possible sequence of action for that scenario. The consumer agent takes this action sequence for optimal scheduling of the battery with the long term objectives.

F. Evaluation of actions of battery scheduling

Algorithm 2 is an outline of the 3-step-ahead Q learning algorithm. To obtain the optimal actions for battery scheduling, we adopt the Q-learning algorithm for which the total discounted reward for the sequence of action is calculated with the following equation

$$r(Scenario_t^l) = \gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) \quad (1)$$

where γ is the discount factor and $r(t)$ is the reward at time t based on the action performed (charging or discharging).

The algorithm starts by setting all Q-values of possible sequences of actions equal to zero.

$$Q(\text{Scenario}_t^l, A_t^j)_m = Q(\text{Scenario}_t^l, A_t^j)_{m-1} + \alpha \left(r(\text{Scenario}_t^l, A_t^j)_m - Q(\text{Scenario}_t^l, A_t^j)_{m-1} \right) \quad (2)$$

When the Scenario_t^l occurs, a sequence of actions is selected randomly, actions are performed and the associated Q value is calculated. The convergent Q value (Q^*) is calculated using the equation 2 where m is the occurrence of scenario. The random selection of sequence of actions for this scenario continues until all the 16 sequence of action's Q-values converge to Q^* values. The duration of this convergence step, called *burn-in period* p of this scenario (Scenario_t^l) is measured by the number of occurrences of the same scenario to reach the convergent Q^* value. The continuous repetition of similar scenarios with random selection of sequence of actions allows to define the Q^* values of all sequences. Thus the Q^* values for all the 16 possible sequences of actions for this scenario are identified. The sequence of actions with highest Q^* value (Q^*_{max}) among 16 Q^* values is the best possible action with regard to the long term objectives and is performed at the future occurrences of this scenario. γ is the discount factor which determines the importance of future rewards and α is the learning rate which influence the speed of convergence. We have taken both the values to be 0.8.

Algorithm 2: 3-step-ahead Q learning

1. Initialize to 0 the Q-values of all possible action sequences for each scenario and set time $t = 0$.
   /* Find the best action sequence and perform actions on the battery. */
2. **loop**
3. For time t, let the forecast values of current P_{sp} and D_t be $P_{sp(t+1)}$,$P_{sp(t+2)}$,$P_{sp(t+3)}$ and D_{t+1},D_{t+2}, D_{t+3} for 3 steps ahead. The states $Scenario_t^l = [s_t^i$; s_{t+1}^p; s_{t+2}^n; $s_{t+3}^m]$ are the forecasts states of solar conditions for the local generation, simulated with the Markov chain model.
4. Based on identified $Scenario_t^l$ and battery charge R_t at time step t, define all possible actions sequences of battery scheduling for 3 steps ahead.
5. Apply the policy for selection of sequence of actions.
6. Perform the selected sequence under real system conditions, simulated using the Markov chain model for real solar conditions. Update the value of the sequence performed.
7. Move to time step $t + 4$.
8. **end**

IV. Case Study and Simulation Analysis

Two scenarios have been chosen with the goal of representing two opposite situations: Scenario 1 is characterized by low solar power output (1000W, 2000W, 1000W, 3000W) and medium and high values of load (4000W, 5000W, 3000W, 6000W), whereas Scenario 2 is characterized by high solar power output (6000W, 4000W, 5000W, 3000W) and low load (2000W, 3000W, 4000W, 2000W). The initial charge of the battery at time t is $3000W$ in both cases. Based on values of forecast solar power output (P_{sp}) and load (D_t) at time steps t, $t+1$, $t+2$ and $t+3$ and the initial battery charge, sixteen sequences of actions for battery scheduling are identified for every scenario. The total discounted reward is calculated by using the equation 1. Solar power and load data are supplied to the Q learning frame work and Q values are calculated for all possible sequence of action and then the convergent values Q^* are calculated with 16 similar scenarios as burn-in period using equation 2. Then out of the 16 Q^* values one with the highest

Q^* value (Q^*_{max}) gives the optimal sequence of actions. This optimal sequence of action is taken by the consumer agent to optimize the long term objective. The simulation program was implemented in Python language. The sequence actions corresponding to the maximum Q^* value obtained from the output of the program is compared with the theoretical value and found that the consumer agent is taking the best sequence of action for optimal scheduling of the battery to achieve the consumer's long term objective.

V. Evaluating Microgrid Performance

The forecast value of solar irradiance (G) is found from the National Renewable Energy Laboratory (NREL) for a 10 KW solar PV system in our campus located in Chennai, India, for the year 2014 and temperature (T) is taken from meteorological department for the same year. These two data are fed into already trained neural network to predict the solar power P_{sp} for the whole year. Then the solar power distribution for the whole year in monthly basis is drawn as shown in the Fig. 3. Various load patterns for residential, commercial and industry are studied and the load pattern of electrical engineering department in our campus is considered. The electrical load D_t of the department is predicted every

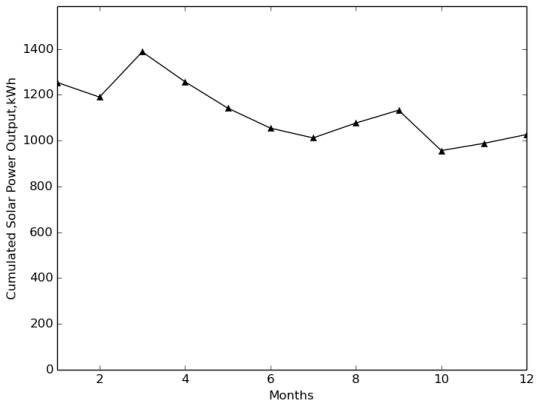

Fig. 3. Solar monthly power in a year

hour for the year 2014. The available solar power in hourly basis for a 10 KW unit is predicted and drawn as shown in Fig. 4. The hourly basis readings of P_{sp} and D_t datas are fed to 3-step-ahead Q learning algorithm for optimal scheduling of the batterries to acheive the long term objectives of the consumer.

Three indicators B_0, S_0 and P_g show the improvement in the performance of the microgrid by using the reinforcement learning. The increase in the utilization of electricity from the battery is estimated by B_0 which is defined as the ratio of the cumulative power used from the battery to the yearly cumulative load.

$$B_0 = \frac{\sum \text{battery to load}}{\sum \text{load}(D_t)}$$

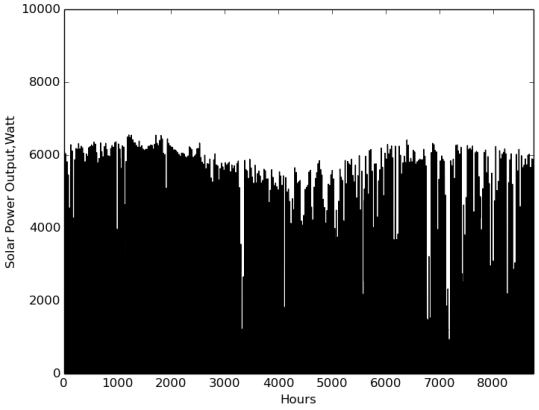

Fig. 4. Solar power in hourly basis

The increase of the utilization rate of the solar PV system evaluated by S_0 is defined as the ratio of the yearly cumulative power used from the solar PV system to the yearly cumulative available solar power.

$$S_0 = \frac{\sum \text{solar to battery}}{\sum \text{solar power}(P_{sp})}$$

Finally, parameter P_g indicates the cumulative annual power

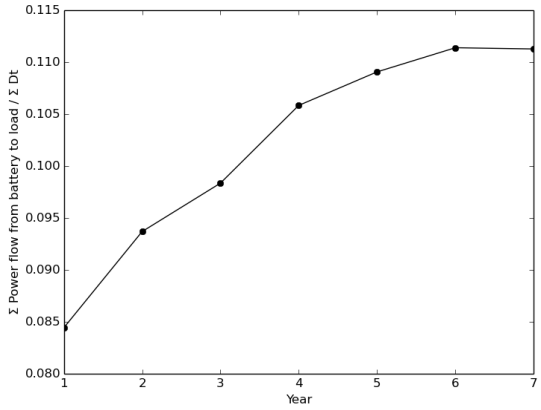

Fig. 5. Average utilization of battery (B_0)

received from the external grid.

$$P_g = \sum \text{Grid} = \sum \text{load}(D_t) - \sum \text{battery to load}$$

Yearly 50 simulations are done for all these three parameters and the average is calculated. The average values of the performance of parameters per year are simulated for seven years and the graphs are drawn. Fig. 5 and Fig. 6 show the average increase in the utilities of the battery and solar PV system; Fig. 7 shows the average reduction in the power consumption from the grid over the years. Because of the increase in the utilities of battery and solar power, the power from grid and hence the cost of power consumed is reduced

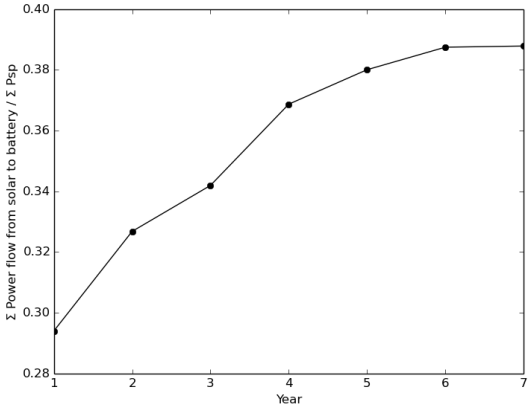

Fig. 6. Average utilization of solar power (S_0)

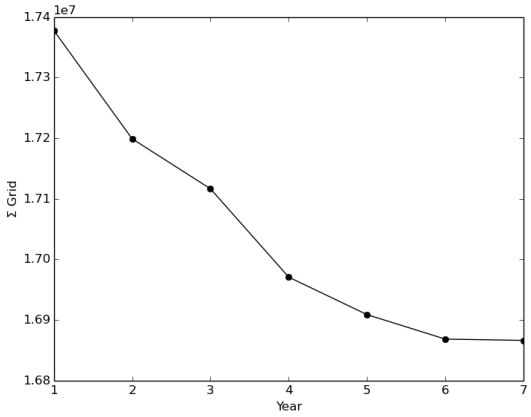

Fig. 7. Average power from Grid (P_g)

when the consumer agent is empowered with reinforcement learning capabilities. The upward trend in Fig. 5 and Fig. 6 shows the increase in the utility of the solar power and the downward trend in Fig. 7 shows the decrease of power consumption from grid. Thus the continuous improvement towards consumer objectives due to Q learning is observed.

VI. Conclusion

The optimal battery scheduling is done with Q learning, a model-free reinforcement learning algorithm. A simulation model was developed to represent the dynamic interactions between the consumer agent and its environment for autonomous optimization of battery scheduling to increase the utility of the battery, solar power, thereby reducing the power consumption from grid in the long run. Uncertainties in the solar power generator due to the stochastic nature of the irradiance and temperature are accounted. The proposed framework gives to the intelligent consumer the ability to explore and understand the stochastic environment and reuse this experience for selecting the optimal energy management actions to reduce dependency on the grid. Future work will focus on the extension to multiple agents integrating diverse renewable generators (solar and wind) and several intelligent consumers with conflicting requirements for electricity supply. Also, the work can be extended to include other parameters which influence the total optimisation of solar microgrid, such as operation of grid switch, operation of both solar and grid sharing the load, battery charging with grid and having a number of batteries or number of solar panels so that charging and discharging can be done simultaneously.

References

[1] Ross Guttromson, Steve Glover, "The advanced microgrid integration and interoperability," Sandia National Laboratories, Sandia report march 2014.

[2] Hatziargyriou ND, European transactions on electrical power, Special issue: "Microgrid and energy management," pp. 1139-141, December 2010.

[3] Reddy P P, Veloso M M, "Strategy learning for autonomous agents in smart grid markets," In Twenty-second international joint conference on artificial intelligence, pp. 1446-51, 2005.

[4] Chen C, Duan S, Cai T, Liu B, Hu G, "Smart energy management system for optimal microgrid economic operation," Renewable Power Generation, IET 5(3), pp.258-67, 2011.

[5] Mohamed F A, Koivo H N, "System modelling and online optimal management of MicroGrid with battery storage," International Journal on Electrical Power and Energy Systems, 32(5), pp.398-407, 2010.

[6] Colson C M, Nehrir M H, Pourmousavi S A, "Towards real-time microgrid power management using computational intelligence methods." IEEE, pp.1-8, 2010.

[7] Abdirahman M Abdilahi, M W Mustafa, G Aliyu, J Usman, "Autonomous Integrated Microgrid (AIMG) System," International Journal of Education and Research, Vol.2, no.1, pp.77-82, January 2014.

[8] Jun Z, Junfeng L, Jie W, and Ng H, "A multi-agent solution to energy management in hybrid renewable energy generation system," Renewable Energy, vol. 36, no. 5, pp.1352-63, 2011.

[9] Aymen Chaouachi, Rashad M, Kamel M, Ridha Andoulsi, and Ken Nagasaka, "Multiobjective Intelligent Energy Management for a Microgrid," IEEE Transactions, Industrial Electronics, 60(4), pp.1688-99, 2013.

[10] Kuznetsova E, Li Y F, Ruiz C, Zio E, Ault G, and Bel L K, "Reinforcement learning for microgrid energy management," Energy Journal, 59, pp.133-46, 2013.

[11] Sutton R S, Barto A G, "Reinforcement Learning: An introduction". London, England: The MIT Press, pp.1-398, 1998.

[12] Engin Karatepe, Mutlu Boztepe, Metin Cola, "Neural network based solar cell model," Energy Conversion and Management, 47, pp.1159-78, 2006.

[13] Hiyama T, Kitabayashi K, "Neural network based estimation of maximum power generation from PV module using environmental information," IEEE Transactions, Energy Conversion; 12 (3), pp.241-52, 1997.

[14] Yujin Lim, Hak-Man Kim, "Strategic bidding using reinforcement learning for load shedding in microgrid," Computers and Electrical Engineering, Elsevier, 2014.

[15] Busoniu L, Babuska R, De Schutter B, and Ernst D, "Reinforcement Learning and Dynamic Programming Using Function Approximators," CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2010.