POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Controlling Lithium-Ion Batteries Through Reinforcement Learning

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Oscar Francesco Pindaro

Advisor: Prof. Marcello Restelli

Co-advisors: Marco Mussi, Francesco Trovò

Academic year: 2020-2021

1. Introduction

A *Smart Grid* is an electric grid integrated with Information Technologies that allows to monitor, manage and repair the electric network, resulting in more efficient energy distribution and overall greater reliability and availability of the electric systems.
Historically, the electric grid has had a simple structure, and it can be divided into four parts:

- *Energy Producers*: generate electric power in chemical or nuclear plants of monumental dimensions.
- *Trasmission Grid*: transfers efficiently energy over long distances.
- *Distribution Grid*: distributes and delivers electricity to the final users.
- *End Users*: consume passively electric energy.

A Smart Grid enhances the functionalities of a traditional grid by leveraging information technologies. The main change of paradigm is given by the support of a *two-way flow* of electricity and information. This allows the Smart Grid to be composed of loosely coupled control subsystems that exchange information and can detect and react to events happening in the grid. *Two way flow electricity* allows end users to produce and sell energy for monetary compensation.

In this context, solar energy production becomes a valid and cheap alternative to traditional sources such as fossil fuels, and it can be produced at different scales, from domestic to industrial use cases. It has the advantage of being a source of inexhaustible free energy, and its availability is not controlled by market conditions or third party actors. However, solar energy production is characterized by a high variation in availability, and it is very challenging to predict. Solar energy is gathered through *photovoltaic panels*, and they can be used to reduce energy consumption from the grid or generate revenue by exploiting fluctuations in energy prices. Energy production and peak user demand are not aligned, and therefore, accumulation systems are used to store energy surpluses. In a domestic environment, a controller decides how to operate the accumulations system, and generate profit by meeting the domestic system energy demand and selling energy in excess to the Smart Grid.
These decisions need to take into account three main challenges. The first one is *Energy Arbitration*. An arbitration is the purchase and sale on a particular asset aimed at generating a profit from variations in the listed price of the asset. In order to perform arbitration, the controller needs to be able to make predictions on

future market prices, and understand which are the most profitable moments for selling energy. *Weather Forecasting* is fundamental when dealing with solar energy production. In fact, a controller should be able to predict energy availability. The last challenge is the *Degradation* that accumulation systems are subject to. They are mainly composed of Lithium-Ions battery packs, a very efficient and high-energy-density battery technology. This type of accumulation systems are affected by a degradation process that lowers their capacity and efficiency over time, caused by the natural aging that each battery incurs in, environmental impacts (such as storing conditions), and the dynamic loading. Battery degradation is a highly non-linear process, which progresses at different rates in different moments of the life of the battery.
The profit achieved by the controller depends on conflicting factors: it should be able to store energy for future uses, while avoiding too intensive battery cycling. A battery purchase is very burdensome, and it is crucial to find the best way to control the battery and generate as much profit as possible.

Original Contribution The novelty of this work is the design of a controller that is able to solve simultaneously the three challenges above mentioned, with a focus on battery degradation management. Long-term profit maximization is achieved by taking into account the revenue generated with energy arbitrage and the cost caused by battery degradation.
The proposed method is able to amortize the battery cost on an unknown time horizon, since battery life is heavily influenced by cycling conditions. The controller also performs weather forecasting by taking into account the daily and annual periodicity.
An interpretation of the behaviour of the controller is also discussed, allowing to understand better which are the relevant physical quantities that contribute in the generation of a higher profit.
These considerations allow the controller to generate up to 15% more in profit with respect to state of the art techniques.

2. Background

2.1. Lithium-Ion Batteries

Chemistry Lithium-Ion batteries are used to convert electricity into chemical energy and vice-versa. A battery cell is a stacking of three main components: the anode, the electrolyte, and the cathode. The *anode* is usually composed of a metallic element that releases ions and electrons when oxidized. These electric particle are consumed in the *cathode* with a reduction reaction that produces energy and consequently voltage. The *electrolyte* is used to separate anode from cathode so that reactions can be controlled. During a battery discharge, electrons leave the battery from the anode, and Lithium ions migrate from the cathode to the anode through the electrolyte.

Characterization The main physical quantities that characterize a battery are State of Charge (SoC), Voltage (V), Current (I) and State of Health (SoH). *SoC* $\sigma_t \in [0, 1]$ is the amount of battery capacity currently available with respect to the maximum capacity. It expresses the capabilities of the battery of interacting with external components. The distance between the maximum and the minimum values of SoC during a charge or discharge is called Depth of Discharge (DoD). The *Voltage (V)* of a battery is determined by its chemical characteristics and it is influenced by environmental conditions. It can be modeled as a polynomial and exponential function of SoC. *Current (I)* is defined as the number of charges passing through a point in a unit of time. When a current is applied to a battery, its SoC varies, since charges are being introduced or taken out from the battery. Batteries are subject to a degradation process that lowers their overall capacity overtime, and the real capacity quickly moves away from the nominal value. *State of Health (SoH)* indicates the remaining capacity of a battery and is defined as:

$$SoH_t = \frac{C_{t,max}}{C_{0,max}} \tag{1}$$

where $C_{t,max}$ is the capacity at time t and $C_{0,max}$ is the nominal capacity of the battery. SoH evolution is a highly non-linear process that is caused by a variety of factors. Most of the degradation is concentrated at the beginning and end

of the battery life, with a heavy slow down in the battery degradation rate during its mid-life. Degradation is caused by irreversible reactions between the anode and the electrolyte. Capacity fade is a consequence of the irreversible consumption of lithium ions that cause the creation of the Solid Electrolyte Interphase (SEI), a layer of non-reactive compounds that limits the amount of Lithium ions that can be exchanged between anode and cathode. Degradation $D_t = 1 - SoH_t$ can also be used to describe the battery health.

Degradation Model A degradation model allows to simulate the dynamics that causes a battery to loose its capacity over time. This work uses the model proposed by Xu et al. [1]. Battery degradation is a non-linear process that depends on factors such as charging and discharging current, time and temperature, but also its current state of life. These factors can be grouped into two stress functions: calendar and cycling ageing. The *calendar ageing* f_{cal} is the degradation stress that a battery suffers independently from its use. The *cycling ageing* f_{cyc} is caused by the direct use of the battery. Every cycle is modeled as a single stress event, **independent** from the others. The overall battery degradation D can be expressed as:

$$D = 1 - \alpha_{sei} e^{-\beta_{sei} f_d} - (1 - \alpha_{sei}) e^{f_d} \tag{2a}$$

$$f_d = f_{cal}(t, \bar{\sigma}, \bar{T}) + f_{cyc}(\sigma_{0:t}, T_{0:t}) \tag{2b}$$

where t is the operational life of the battery, $\bar{\sigma}$ is the mean SoC, $\bar{T}$ is the mean temperature, and $\sigma_{0:t}$ and $T_{0:t}$ are respectively the SoC profile and temperature profile that the battery was subject to. Eq. (2b) is a stress function that combines calendar and cycling ageing. Eq. (2a) can be divided into two components: one that takes into account the capacity loss cause by the SEI formation, and another that considers capacity fading at a rate proportional to the battery life.

2.2. Reinforcement Learning

Markov Decision Process The problem can be formalized as a Markov Decision Process (MDP). An MDP is a mathematical framework used to model control problems. Such a framework is composed of several components. Formally, an MDP $\mathcal{M}$ is defined as a tuple:

$$\mathcal{M} := (\mathcal{S}, \mathcal{A}, P(s'|s,a), R(s,a), \gamma, \mu_0),$$

where $\mathcal{S}$ is the set of states, which contain all the possible states characterizing the system under analysis, $\mathcal{A}$ is the set of actions the controller is allowed to perform in each state $s \in \mathcal{S}$, $P(s'|s,a)$ is the state transition probability matrix, specifying the probability to go to state s' for each generic state/action (s,a) pair ($P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$), $R(s,a)$ is the reward function, defining for each state/action pair (s,a) the expected immediate reward obtained ($R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$), $\gamma \in [0,1]$ is the discount factor, which define how much the controller is evaluating future rewards ($\gamma \approx 0$ for greedy controller and $\gamma \approx 1$ for foresighted ones), and $\mu_0 \in [0,\ 1]^{|\mathcal{S}|}$ is the initial distribution of the states. The state in MDPs should be defined s.t. it contains all the information that allows predicting the future, independently from the past (a.k.a. Markovian Property).

The goal of a controller is, given the elements described above, to define a policy $\pi(a|s)$ to select, for each state $s \in \mathcal{S}$, the action $a \in \mathcal{A}$ which maximizes the discounted sum of rewards V:

$$V = \sum_{t=1}^{\mathcal{T}} \gamma^{t-1} r_t, \tag{3}$$

where r_t is the reward collected by the policy $\pi(a|s)$, and $\mathcal{T} \in \mathbb{N}$ is the time horizon.

Learning in such environments requires either the availability of a dataset composed by interactions of the form $\{(s_t, a_t, r_t)\}_t = 1^N$, where s_t is the state, a_t the action performed, and r_t the instantaneous reward at time t, or the capability to interact with the environment, which generates sequence of interactions to be used for learning. Basing our learning on such information, one might approach the policy learning task in several ways.

One of the most popular and effective learning methodology is the model-free one, in which the estimate of the policy is provided without estimating explicitly the environment, i.e., without estimating the transition probability distribution $P(s'|s,a)$. Instead, the estimation of the value function $Q(s,a)$, which characterize the cumulative discounted reward for each state/action pair, and choosing as a policy in

each state s the action a^* maximizing it, formally $\hat{\pi}(a|s) := \arg\max_a Q(s,a)$. This estimation was performed by using the Fitted Q-Iteration algorithm.

Fitted Q-Iteration Fitted Q-Iteration (FQI) [2] is used to estimate the action-value function and derive a control policy from a batch of transitions sampled from the environment. The transitions are sampled with a policy that tries to visit most of the state space.
A transition is a four-tuples $\langle s_t, a_t, r_t, s_{t+1}\rangle$, where s_t is the starting state of the transition, a_t is the action drawn from the exploratory policy, r_t is the reward obtained by the agent after performing the action a_t in the state s_t and, finally, s_{t+1} is the ending state reached after performing the action a_t in the state s_t. Also, let Q_i be the action-value function computed on a time of horizon of i steps.
The idea behind FQI is to build at every training step i an approximation of Q_i from the transition dataset with a supervised learning method. Once the approximator has been trained, the control policy can be retrieved by maximizing the approximated Q-function with respect to a. The approximator used in this work is a tree-boosting algorithm called XGBoost.

3. Related Works

PV power generation, battery degradation, and energy arbitrage are complex problems often studied individually.
Sui et al. [3] study the problem of scheduling charge, discharge, and resting periods while using multiple batteries. The proposed scheduler has to keep the SoC of every battery over a given level while minimizing the degradation caused by high temperatures. It models two different characteristics of a Lithium-Ion battery: *rate capacity effect* and *recovery effect*. Due to the former, a battery shows a smaller overall capacity when discharged at high currents, while the latter influences the battery voltage recovery after a continuous discharge process. The scheduler takes advantage of these two effects and extends the battery life. This work considers fixed charge/discharge currents. While this approach simplifies the control problem, it does not allow the scheduler to choose between different charging or discharging profiles that could achieve the same performances with lower effects on the degradation. Another shortcoming of this work is that SoH modeling is influenced only by temperature, and other important factors such as DoD, SoC, and current rate are not considered. Moreover, no economic considerations are done w.r.t. SoH, and the objective of the scheduler is just to use for as long as possible a battery while avoiding cycles that generate short-term high degradation.

4. Algorithm

This work considers a domestic environment in which an energy producer (e.g., a photovoltaic plant) and an energy consumer (i.e., the house) are interacting. This plant produces energy in a periodic way, alternating periods in which the domestic environment has energy surplus and others with energy shortage. In the case of an energy shortage (surplus) in the domestic environment, a controller must define the amount of energy to take (store) from the battery and, consequently, the energy to get (introduce) in the public network. The solution presented below provides a high-level controller which manages the energy flow. Such a controller defines the amount of energy taken/given to the battery, given a set of battery capacity constraints, and the amount taken/given to the public energy network. The key idea is to provide a balance between the economic loss given by the difference between the buy/sell process provided by the public network and the economic loss given by the battery degradation.
As shown in Fig. 1, the controller is able to measure the net power $P_{h,t}$ coming from the domestic system. Part of this power is assigned to the battery ($P_{b,t}$), while the remaining power $P_{g,t}$ is sold (bought) from the grid at the market place price c_g (p_g). This work considers stationary energy prices.

4.1. Reinforcement Learning Model

The above-mentioned process is formalized as a *Markov Decision Process* $\mathcal{M}$. In what follows, the elements defining the MDP for the analysed problem are described.

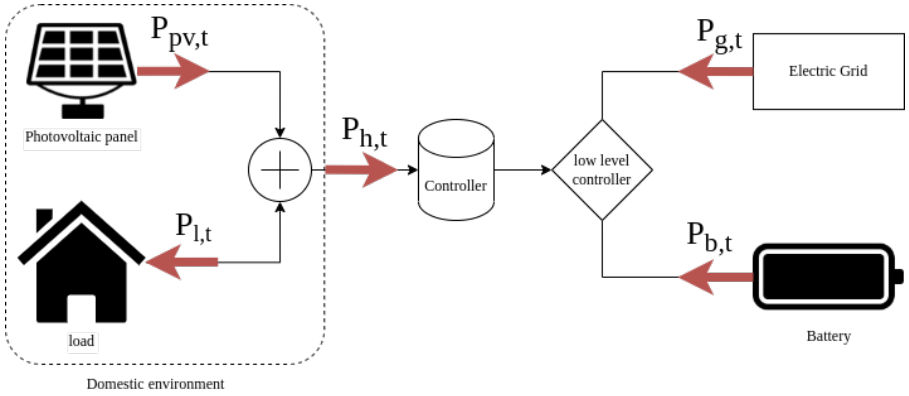

Figure 1: The domestic environment is composed by the residence of an user and a photo-voltaic panel. The controller can read the net power consumption $P_{h,t}$ and decide how much power will be routed to the battery and to the electric grid.

State The state vector $s \in \mathcal{S}$ is defined as follows:

$$s = (\sigma_t, T_{b,t}, \delta_t, I_t^{req}, P_{PV,t}, \cos(\varphi_{d,t}), \sin(\varphi_{d,t}), \cos(\varphi_{y,t}), \sin(\varphi_{y,t})) \quad (4)$$

where t is the current time step. σ_t, $T_{b,t}$ and δ_t are respectively the SoC, the battery temperature and the DoD, and are used to keep track of the state of the battery. I_t^{req} is the maximum C-rate that the battery would be subjected to if all the net power $P_{h,t}$ would be directed to the battery. The last four components are a trigonometric encoding used to represent the daily and annual periodicity.

Action The agent can select eleven discrete actions $\alpha_t \in \{0.0, 0.1, \ldots, 1.0\}$ at every control step t. At every time step, the controller has to decide the percentage α_t of the net power $P_{h,t}$ that will be directed to the battery. The remaining power is sent to the electric grid. The low level controller checks that the action respect the physical constraints of the battery.

Reward The rationale behind the definition of the MDP reward $R(s, a)$ consists in the reduction the amount of money paid to maintain the system over time:

$$r_t = \frac{f_{d,t} - f_{d,t-1}}{f_d^{max}} c_b + p_g E_{g,t} \mathbb{1}_{\mathbb{R}^-}(P_{g,t}) - c_g E_{g,t} \mathbb{1}_{\mathbb{R}^+}(P_{g,t}) \quad (5)$$

where $f_{d,t}$ and f_d^{max} are respectively the linear degradation at time step t and the maximum linear degradation corresponding to maximum admissible real degradation $D_{t,max}$ and $E_{g,t}$ is the energy exchanged with the electric grid. Eq. (5) takes into account the profit made by the agent by exchanging energy with the electric grid and the amortization of the battery value during the operational period. The battery value is amortized by considering the variation in *linear degradation* f_d. The γ factor used in this problem has been set to 1.

5. Experimental Results

In order to test the solution proposed in Section 4, an online simulator which implements the OpenAI Gym standard framework is implemented. A core part of the simulation is the power signal generated by the PV $P_{PV,0:\mathcal{T}}$ and the auto-consumption profile $P_{l,0:\mathcal{T}}$, both generated from real data. The agent was trained with the FQI algorithm for 200 iterations using an XGBoost regressor of 1100 trees of maximum depth of 8. A total of 7 millions state transitions were sampled over a span of 100 episodes. An episode is run for 8 years or when the battery degradation exceeds the maximum allowed D_{max}. The agent was tested against 4 different Key Performance Indicators (KPI):

- *Profit*: the value of the objective function.
- *Battery Cost*: the first component of the objective function, expresses how much value of the battery was lost while cycling.
- *Energy Profit*: the second component of the objective function, it is the profit made by exchanging energy with the electric grid.
- *Degradation*: degradation D_t that the bat-

	Profit	Battery Cost	Energy Profit	Degradation
Agent	-2295.32	-1680.54	-614.78	0.1245
SoC20-80	-2454.99	-2144.69	-310.30	0.1589
OnlyBattery	-2365.93	-2141.18	-224.74	0.1586
OnlyGrid	-2354.15	-1531.66	-822.49	0.1135

Table 1: Average KPI values after 8 years

tery was subject to.

The performance of the agent is compared with 3 different baselines:

- *OnlyGrid*: The actions of this baseline are always set to 0. No power falls on the battery, and therefore only the calendar ageing impacts on the degradation.
- *OnlyBattery*: This baseline always uses the battery. Energy exchanges with the grid happen only when the battery is completely empty or full.
- *Soc20-80*: This baseline keeps the SoC between 0.2 and 0.8. This is the state-of-the-art control policy.

Table 1 report the performance of the agent and the three different baselines with respect to each KPI. All policies generated a negative profit, due to the stationary prices hypothesis that leaves little to no room for optimization. The agent is able to beat the state of the art techniques, achieving almost 150€ more in profit. It is important to note that the agent was able to reach this result without excelling in both battery cost or energy profit. The agent is able to save the most on battery cost, while being able to exchange energy with the electric grid. The agent is able to achieve these results by avoiding high stress cycling conditions, such as high SoC and DoD values. It tries to cycle at low SoC and charge the battery with low powered charges. Once the domestic environment enters in an energy deficit period, the battery is discharged as fast as possible, while meeting the house energy demand.

6. Conclusions

Photovoltaic panels are used in residential environments to produce cheap and clean energy, lowering electricity costs and increasing energy independence. The main difficulties in managing such systems are caused by the unpredictable nature of solar energy production and by the asynchronicity between energy production and consumption. To alleviate these limitations, an accumulation system is used, where energy in excess can be stored for later use. However, these accumulations systems are characterised by a process degradation influenced both by environmental factors and dynamic loading.

This work design a RL controller trained with Fitted Q-Iteration (FQI) and the ensemble tree regressor XGBoost, by considering a degradation model that allows to compute instantaneous SoH loss. The objective is to maximize the long term profit while exchanging energy with the electric grid and by amortizing the battery cost on the whole period accordingly to its use. The algorithm proposed outperforms the state-of-the-art techniques by up to 15%, with a control policy that keeps SoC values as low as possibles with slow low-powered charges and fast discharges. The controller achieves great generalization, since it is able to operate different battery capacities without a dedicated training and it has been designed on multiple energy consumption routines.

References

[1] Bolun Xu, Alexandre Oudalov, Andreas Ulbig, Göran Andersson, and Daniel S Kirschen. Modeling of lithium-ion battery degradation for cell life assessment. *IEEE Transactions on Smart Grid*, 9(2):1131–1140, 2016.

[2] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[3] Yu Sui and Shiming Song. A multi-agent reinforcement learning framework for lithium-ion battery scheduling problems. *Energies*, 13(8):1982, 2020.