2018

Stochastic Optimal Control of Grid-Level Storage

Yuhai Hu
Lehigh University

Stochastic Optimal Control of Grid-Level Storage

by

Yuhai Hu

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Industrial and Systems Engineering

Lehigh University

January 2018

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

_____
Dissertation Advisor

Committee Members:

_____
Boris Defourny, Committee Chair

_____
Lawrence V. Snyder

_____
Luis F. Zuluaga

_____
Alberto J. Lamadrid

Acknowledgements

This thesis would have been impossible without the support of the professors, my fellow graduate students and friends, the staff in the Industrial and System Engineering Department of Lehigh University as well as my family.

I would like to express my deepest gratitude, first and foremost, to my academic advisor Professor Boris Defourny. I feel highly fortune to have Professor Defourny as my advisor. His expertise and vision have guided me through my research. All the discussions and conversation we had together, from technical subjects to scientific topics in general, are just like lighthouses in my exploratory journey of research. In addition to his generous support on my research, I also received so much sincere and useful advice from him in many other aspects such as writing and speaking skills. Personally, he is also a true role model to me.

I would like to sincerely thank the remaining members of my thesis committee - Professor Lawrence V. Snyder, Professor Luis F. Zuluaga and Professor Alberto J. Lamadrid - for sharing their knowledge and experience with me over the past few years. I especially thank Professor Snyder who took me as a first-year PhD student. I learned so much from his attitudes to research, to friends and to family. Professor Zuluaga and Professor Lamadrid have provided me with many valuable inspirational suggestions on my research.

I would like to thank the other professors in the ISE department - Professor Tamás Terlaky, Professor Theodore K. Ralphs, Professor Katya Scheinberg, Professor Curtis E Frank, Professor Martin Takáč. - for sharing their expertise on operations research and showing passion on teaching students.

I would like to convey my appreciation to the great and dedicated staff member in the ISE department including Rita R. Frey, Kathy Rambo, Abby Barlok, Ana Quiroz and many

others. I sincerely thank these people for making the department as a home for me.

During my six years study at Lehigh University, I am fortunate to have met so many new friends. I had wonderful time and discussions with Wei Xia, Xi He, Xiaolong Kuang, Chenxin Ma, Ziyi Guo, Shu Tu, Dan Li, Shuyi Wang, Miao Bai, Zheng Han, Choat Inthawongse, Sertalp B Çay, Pelin Çay, Matt Menickelly and many others.

Finally, I would like to thank my loving parents, who are always being there for me whenever I need and give me a sweet home.

Contents

Acknowledgements **iv**

List of Tables **x**

List of Figures **xi**

Abstract **1**

1 Introduction **3**

1.1 Energy Storage Systems 3

1.2 Battery Energy Storage System 6

1.3 Outline of Contribution 7

2 p-Periodic Markov Decision Process **8**

2.1 Introduction 8

2.1.1 Motivation 9

2.1.2 Contributions and Related Work 9

2.1.3 Organization 10

2.2 Periodic Discounted Problems 11

2.2.1 Problem Formulation 11

2.2.2 Optimality Conditions 12

2.3 Near-optimality bounds for greedy policies 14

2.3.1 Error bound for difference between optimal and approximate value functions 15

2.3.2 Bounding ϵ_k . . . 19
2.4 Value iteration . . . 23
2.4.1 Iteration mechanism . . . 24
2.4.2 Convergence rate . . . 24
2.5 Application to grid-level storage operations . . . 26
2.5.1 Model Description . . . 26
2.5.2 Finite State Approximation . . . 29
2.5.3 Results . . . 31
2.6 Conclusion and future work . . . 33

3 Battery Operation with Aging **35**
3.1 Introduction . . . 35
3.1.1 Contributions and Related Work . . . 35
3.1.2 Organization . . . 38
3.2 Model Description . . . 38
3.3 Threshold Policy . . . 40
3.4 Analysis . . . 44
3.4.1 Threshold Policy Evaluation . . . 45
3.4.2 Threshold Policy Optimization . . . 47
3.5 Error Analysis . . . 53
3.6 Extensions of the Storage Device Model . . . 58
3.7 Computational Results . . . 60
3.7.1 Performance of the Proposed Algorithm . . . 60
3.7.2 Impact of Non-Ideal Battery Characteristics . . . 62
3.7.3 Economic Value of the Finite-Life Model . . . 64
3.8 Extension of the Price Model . . . 64
3.8.1 Threshold Policy Evaluation . . . 66
3.8.2 Threshold Policy Optimization . . . 67
3.8.3 Illustration . . . 73
3.9 Conclusion . . . 74

4 On the Price Impact of Distributed Energy Storage **76**

4.1 Introduction . . . 76

4.2 Technical Methods . . . 78

4.2.1 Markov Games . . . 78

4.2.2 Nash Equilibrium and Bimatrix Games . . . 79

4.2.3 Model Description and Assumptions . . . 81

4.2.4 Model Formulation . . . 82

4.2.5 Demand and Supplier Curve . . . 85

4.3 Algorithms . . . 85

4.4 Numerical Experiments . . . 87

4.4.1 Parameters setting . . . 87

4.4.2 Numerical Results . . . 87

4.4.3 Sub-optimality of Policy . . . 88

4.4.4 Extensions: Implicit Curve for Demand and Supply . . . 90

4.4.5 Extensions: Storage Devices with Non-Perfect Efficiency . . . 90

4.5 Incomplete Information Game . . . 91

4.5.1 Model Description and Assumption . . . 91

4.5.2 Estimation of the Other Player's Charge Level . . . 91

4.5.3 Impact of incomplete information . . . 94

4.5.4 Summary . . . 96

4.6 Impact of storage ownership on the price . . . 97

4.6.1 Parameter setting . . . 97

4.6.2 Numerical Results . . . 97

4.7 Conclusion . . . 99

5 Conclusion **101**

Bibliography **102**

A Threshold policy **115**

A.1 Concavity . . . 116

A.2 Optimal basestock targets 117
A.3 Price monotonicity 119

B Price process **124**

B.1 Periodic Price Model with i.i.d noise 125
B.2 Periodic autoregressive process of order 1 125
B.3 PAR(1) with spikes 126

C Linear and Quadratic Curve Case Study **127**

C.1 Analytic Solution for Equilibrium Price and Reward Functions: Linear and Quadratic Curve Case 127
C.2 Numerical Experiment: Linear and Quadratic Curve Case 128

Biography **130**

List of Tables

2.1 Hourly prices: Lognormal parameters 28

3.1 Algorithm for optimizing the thresholds. 52

3.2 Algorithm for optimizing the price thresholds under battery capacity deterioration and inefficiencies. 58

3.3 Price thresholds calculated by Value Iteration. 61

3.4 Price thresholds calculated by our algorithm. 62

3.5 Value of taking into account the finite life of the storage device. 65

4.1 fix supplier's policy 89

4.2 fix consumer's policy 90

4.3 Update Uncertainty Set Algorithm 92

4.4 Reward comparison between full and incomplete information game 96

List of Figures

2.1 Interpretation of assumption 2.11 16

2.2 Interpretation of 2 20

2.3 A typical discretized price state. 30

2.4 Near-optimal periodic policy, from hour 1 (midnight to 1am) to hour 24. X-axis: charge level state (indexed from 1 to 50), Y-axis: period-dependent price state (indexed from 1 to 20). White: Charge at maximal rate (buy), Black: Discharge at maximal rate (sell), Gray: intermediate actions. 31

2.5 Running time (seconds) for different problem sizes and number of cores. . . 31

2.6 The threshold policy of 24 time periods 32

2.7 The threshold policy of 13^{th} period 32

3.1 Comparison of price thresholds, as a function of n. Continuous line: base case corresponding to Table 3.4 ($\gamma = 0.999$). Dashed: With capacity deterioration. Dotted: With capacity deterioration and charging-discharging inefficiency. 63

3.2 Information on storage usage. Continuous line: base case ($\gamma = 0.999$). Dashed: With capacity deterioration. Dotted: With capacity deterioration and charging-discharging inefficiency. 63

3.3 Simulation of the optimal policy on the same sample path of a price process but starting from two different ages. Decisions: Charge (▲), Discharge (▼). 74

4.1 Bellman residual for both players as a function of iteration. *red*: consumer. *blue*: supplier. Inset graph: first 20 iterations. 87

4.2 Charging amount for both players (y-axis) as a function of the demand curve level (parameter b as x-axis). *red*: consumer. *blue*: supplier. 88
4.3 Next charge levels for both players (y-axis) as a function of the demand curve level (parameter b as x-axis). *red*: consumer. *blue*: supplier. 88
4.4 Distribution of time duration of incomplete information game 95
4.5 Evolution of uncertainty set 96
4.6 price volatility 98
4.7 Cumulated reward as a function of storage ownership 99

C.1 Charging amount for both players (y-axis) as a function of the demand curve level (parameter $b/100$ as x-axis). *red*: consumer. *blue*: supplier. 129
C.2 Next charge levels for both players (y-axis) as a function of the demand curve level (parameter $b/100$ as x-axis). *red*: consumer. *blue*: supplier. . . . 129

Abstract

The primary focus of this dissertation is the design, analysis and implementation of stochastic optimal control of grid-level storage. It provides stochastic, quantitative models to aid decision-makers with rigorous, analytical tools that capture high uncertainty of storage control problems. The first part of the dissertation presents a p-periodic Markov Decision Process (MDP) model, which is suitable for mitigating end-of-horizon effects. This is an extension of basic MDP, where the process follows the same pattern every p time periods. We establish improved near-optimality bounds for a class of greedy policies, and derive a corresponding value-iteration algorithm suitable for periodic problems. A parallel implementation of the algorithm is provided on a grid-level storage control problem that involves stochastic electricity prices following a daily cycle. Additional analysis shows that the optimal policy is threshold policy. The second part of the dissertation is concerned with grid-level battery storage operations, taking battery aging phenomenon (battery degradation) into consideration. We still model the storage control problem as a MDP with an extra state variable indicating the aging status of the battery. An algorithm that takes advantage of the problem structure and works directly on the continuous state space is developed to maximize the expected cumulated discounted rewards over the life of the battery. The algorithm determines an optimal policy by solving a sequence of quasiconvex problems indexed by a battery-life state. Computational results are presented to compare the proposed approach to a standard dynamic programming method, and to evaluate the impact of refinements in the battery model. Error bounds for the proposed algorithm are established to demonstrate its accuracy. A generalization of price model to a class of Markovian regime-switching processes is also provided. The last part of this dissertation is

concerned with how the ownership of energy storage make an impact on the price. Instead of one player in most storage control problems, we consider two players (consumer and supplier) in this market. Energy storage operations are modeled as an infinite-horizon Markov Game with random demand to maximize the expected discounted cumulated welfare of different players. A value iteration framework with bimatrix game embedded is provided to find equilibrium policies for players. Computational results show that the gap between optimal policies and obtained policies can be ignored. The assumption that storage levels are common knowledge is made without much loss of generality, because a learning algorithm is proposed that allows a player to ultimately identify the storage level of the other player. The expected value improvement from keeping the storage information private at the beginning of the game is then shown to be insignificant.

Chapter 1

Introduction

1.1 Energy Storage Systems

Over the past few years, research and development in the power system has increased in energy storage techniques. One of the reasons behind this phenomenon is that renewable energy plays a more important role for the electrical power grid. According to a report published by the Energy Information Administration [1], the percentage of energy generated from renewables has increased from 9.49% in 2006 to 13.35% in 2015. In addition, estimates of the potential market for energy storage in the United States are quite large [33]. According to a white paper from Electric Power Research Institute (EPRI) [80] published in 2010, there is more than 128 gigawatts (GW) of grid-level energy storage installed worldwide. Among all Energy Storage Systems (ESS), Pumped Hydroelectric Storage (PHS) has the percentage of installed capacity around 99%, while Compressed Air Energy Storage (CAES), flywheels and battery storage together constitute the remaining 1% [111, 80].

The energy storage system are very useful to deal with inherent intermittency of renewable energy generation, for example, wind, solar, wave energy [15, 5, 27, 23]. Benefit may be derived from the devices by charging them when the price is low and discharging it when the price is high [5, 30, 10]. Other benefits may also be derived from providing ancillary services such as regulation [23, 37, 26]. As a result, researchers have focused

on energy storage techniques, regarding them as a way to integrate renewable energy resources. With proper energy storage techniques, power system operators may be able to balance the difference between demand and supply, provide backup energy, in addition to the flexibility on the demand side provided by demand response [61, 5].

Researchers have been increasingly taking storage into consideration in Optimal Power Flow (OPF) and Unit Commitment (UC). In [16], an OPF model with storage is provided and shows how storage system optimize of power generation across multiple time periods. In [99], a stochastic electricity market model is established (which is similar to a Unit Commitment problem). The results show the effects of wind power generation on system operation as well as on economic value of investments in energy storage system (CAES in their case).

Although different kinds of energy storage techniques are developing rapidly in the recent years, most newer storage technologies are still in the development phase and could not be deployed yet. There are still lots of challenges to integrate storage [64]. In addition, storage devices like lithium batteries are too expensive to be deployed massively within the power grid. However, with a better control system, storage operators may be able to reduce running cost, increase revenue and provide reliable services. Due to those reasons, many researchers have sought to establish a more profitable and reliable storage system by providing practical and efficient optimal control methods.

During the last decade, the application of control theory to grid-level storage problems has increased sharply. Sioshansi et al. [109] establish a dynamic programming model to maximize the profit of four services by controlling the operation of multiple distributed energy storage resources. The services include: energy and ancillary services sales, backup energy supply and transformer loading relief. In [91], Secomandi studies the optimal trading policy for a capacitated storage asset. In this paper, the author uses gas as the commodity, while the charge and discharge rates are limited. Powell et al. [67] provides a stochastic, dynamic program to model hourly dispatch and energy allocation in a grid with storage. In their paper, variations come from wind, solar and demand, and they use hydroelectric storage to smooth energy production over different time scales. Van De Ven et al. [104]

formulate a dynamic program to study an optimal battery control policy, to let the end-user minimize the total discounted cost. The uncertainty comes from demand and price. In [28], optimal control policies for single and multiple batteries are studied. Single battery case has an optimal policy with threshold structure and can be obtained by dynamic programming. Multiple batteries case is extended by a method to map the optimal solution of single case to multiple batteries case. In Su et al. [97], the expected magnitude of residual power imbalance process is minimized. The authors model the power imbalance problem as an infinite horizon stochastic control problem and the optimal policy turns out to be a greedy policy. In addition, short time scale and long time scale approximations for power imbalance are also provided as well as the corresponding necessary storage capacity. Dicorato et al. [21], Teleke et al. [100] and Brekken et al. [11] provide models to integrate a battery storage system with a large wind farm to improve dispatchability. Cruise et al. [18] provide a nonlinear programming model and a Lagrangian-based algorithm to solve it. P Malysz et al. [62] provide a MIP formulation to solve energy storage operation in grid-connected electricity microgrid.

When we deal with stochastic control of energy storage problems, most often we need to establish a price model. As electricity prices have unique characteristics compared to other commodity prices, many different models have been proposed and studied in the literature. In [46], a mean-reverting price process is incorporated to an energy commitment problem where wind farms and storage devices exist. However, since the demand of energy varies during days, throughout months and across years, the impact from stochastic demand also results in seasonality effects on electricity prices. In [63], a mean-reverting jump diffusion stochastic process is adopted as the electricity price model that incorporates seasonality effects and price spikes. A diversity of models for electricity prices are discussed in detail in [48, 58, 29, 45]. In Appendix B, we describe several price models. We also present a p-periodic Markov Decision Process model in Chapter 2 which captures seasonality effects.

1.2 Battery Energy Storage System

As we mentioned in the previous section, in 2010, 99% of installed energy capacity is Pumped Hydroelectric Storage. However, according to the Department of Energy [17], there are around 1630 energy storage projects all over the world, and about 824 projects of them (around 50%) are using battery technology as the storage method, such as Lead Acid, Lithium Ion and Zinc Bromine flow. The main advantages of Battery Energy Storage Systems (BESS) are their rapid response time, and their high energy densities.

Storage batteries are rechargeable electrochemical systems for storing energy [34]. They deliver chemical energy in the form of electric energy. Different types of batteries and their properties are introduced in detail in [23, 42]. Lead-acid batteries are the oldest and mature type of rechargeable batteries which are used mostly in grid-level storage. However, the technology has a low cycle life and battery operational lifetime. Nickel-based batteries have longer lifetimes than lead-acid, but the cost is 10 times more and the energy efficiency is also lower than lead-acid batteries. Zinc bromine (ZnBr) battery belongs to a class called flow battery [15], which has non-self-discharge capacity. A main drawback of ZnBr system is that it need a third pump system to circulate bromine complexes, which introduce extra installation and operation cost. Lithium-based battery battery storage systems seem to be a very promising technology. They have high energy density, high efficiency, and low self-discharge rates.

The charge/discharge mechanism for batteries can be described as follows: during the discharge process, once a load is connected, chemical energy is generated by electrochemical reactions in a basic cell, between two electrodes plunged into an electrolyte. Electrons from one electrode move to the other through an external electric load, where the electric energy is delivered. The process is reversed during charging [34].

A rechargeable battery can be charged and discharged many times. However, any battery has its lifetime, which means there is a limit on the number of charge/discharge cycles (we call it life cycles in the subsequent chapters). Hence, there is a significant difference when we model a BESS and a PHS. In addition, the impact of aging not only affects the life of the battery storage, but also the capacity of the battery. According to

[12], temperature and state of charge (SOC) will affect the life cycles of the battery. In addition, for lead-acid battery, from Peukert's Law, the rate of charge will also have an influence on the available capacity of battery [24].

1.3 Outline of Contribution

In Chapter 2, we formulate and solve a p-Periodic Markov Decision Process model for a grid-level storage control problem. We present computational results. We establish a tighter bound for the near-optimal solution. And we discuss the structure of the optimal policy.

In Chapter 3, we incorporate the aging phenomenon into a grid-level battery operation problem. Instead of the periodic MDP model used in Chapter 2, we use an infinite horizon MDP model in this chapter. An efficient and accurate algorithm for solving the model is established. We provide structural results for the optimal policy. We report on related computational results.

In Chapter 4, we discuss the impact on electricity prices from storage decisions.

In Chapter 5, we summarize our work so far, and our current ongoing research.

Whereas the material of Chapters 4 is ready for submission, most material of Chapter 2 has been published in

Yuhai Hu and Boris Defourny. Near-optimality bounds for greedy periodic policies with application to grid-level storage. In *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014 IEEE Symposium on*, pages 1–8. IEEE, 2014

and the work of Chapter 3 has been published in

Yuhai Hu and Boris Defourny. Optimal price-threshold control for battery operation with aging phenomenon: a quasiconvex optimization approach. *Annals of Operations Research*, pages 1–28, 2017

Chapter 2

p-Periodic Markov Decision Process

Most work of this chapter has been published in paper

Y. Hu and B. Defourny. Near-optimality bounds for greedy periodic policies with application to grid-level storage. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL-2014)*, pages 1–8, December 2014.

2.1 Introduction

For stochastic dynamic programs with seasonality effects, such as inventory or storage problems with daily demand patterns, rolling-horizon look-ahead policies [75] often appear as a well-suited class of policies. A drawback of look-ahead policies, however, is that end-of-horizon effects can be detrimental to the optimality of the decisions. For instance, the case opposing JP Morgan Ventures Energy Corporation (JPMVEC) to the Federal Energy Regulatory Commission (FERC) exposed bidding strategies that were designed to exploit flaws in the market clearing algorithm of the California Independent System Operator (CAISO); one flaw was directly related to the truncation of the planning horizon [31].

2.1.1 Motivation

Two countermeasures are classically considered to mitigate end-of-horizon effects. The first countermeasure is the introduction of a terminal reward function. This essentially amounts to approximate the value function around the state at the terminal stage — and this is an art that requires domain knowledge. The second countermeasure is the extension of the horizon over which the look-ahead is performed. This amounts to assume that end-of-horizon effects die out by the time the backward optimization reaches the first stages. A challenge of this approach is the increased complexity of the look-ahead optimization problem, and in certain contexts, the unavailability of data relative to the extended horizon — for instance, longer-term forecasts may not be available; for a multistep bidding problem, market participants may not have been required to submit offers further in the future; etc.

This chapter is motivated by the synthesis of these two mitigation strategies. We consider policies that solve a discounted periodic dynamic program, over an infinite horizon, constructed by replicating the look-ahead problem or by appending a steady-state cycle to the look-ahead. The rationale is that the structure of a policy optimal for a finite-horizon Markov decision problem on p stages is the same as the one for an infinite-horizon, discounted p-periodic Markov decision problem. However, the cyclo-stationary extension could significantly improve the approximation of the future reward process. Early use of this strategy can be found in the water reservoir operations literature [98, 107].

2.1.2 Contributions and Related Work

Periodic dynamic programs have of course been considered earlier [82, 106, 103], as well as variations thereof [41]. In particular, it is known that an optimal p-periodic Markov policy can be derived using p value functions coupled by a Bellman-type recursion. Periodic dynamic programs can be viewed as dynamic programs with stationary reward and state-transition functions over a state space augmented with the position of time in the cycle, and therefore results from abstract dynamic programming are directly available [6].

The contribution of this chapter is twofold.

- We study the near-optimality of nonstationary policies greedy for periodic approxi-

mate value functions, and provide bounds that are tighter than the general bounds used with stationary value functions on an augmented state space or specialized bounds established for periodic Markov Decision Processes.

- We formulate a periodic Markov Decision Process model for a grid-level energy storage control problem where random electricity prices follow daily patterns. The idea is to recalibrate the model every day and then solve it for operations on the next day in a rolling-horizon fashion. A numerical example with a daily cycle of 24 periods is provided.

2.1.3 Organization

In this chapter, we provide a p-Periodic Markov Decision Process model without aging phenomenon considered. We are concerned with the subproblem to be solved by the proposed class of policies, and we develop effective methods to solve periodic dynamic programs. Most of work in this chapter has been published in [39]. In particular, in addition to the results published in [39], we discuss the threshold-policy structure of the optimal policy.

This chapter is organized as follows. Section 2.2 defines the periodic Markov Decision Problem, and recalls the optimality conditions. Section 2.3 establishes bounds useful to control the near-optimality of periodic policies based on periodic approximate value functions. Section 2.4 describes a value-iteration algorithm, based on the results of the previous section. Section 2.5 formulates a simple model for optimizing grid-level storage operations given day-ahead and historical electricity prices, based on the periodic Markov Decision Process framework. It also reports on numerical work carried out to evaluate the effectiveness of a parallel implementation of the value iteration algorithm. Section 2.6 concludes the present chapter.

2.2 Periodic Discounted Problems

In this section, we recall the mathematical formulation and optimality conditions of the $p-$periodic discounted Markov Decision Problem (p-MDP).

"Periodic" refers to the way the reward and state transition functions vary cyclically over time — this is distinct from the notion of "periodic state" in the theory of Markov chains.

In our formulation, the state space and the action space are periodic. This proves to be useful to adapt the states to the time-dependent characteristics of the problem.

2.2.1 Problem Formulation

For some integer $p \geq 1$ referred to as the cycle length, let

$$\{(\mathscr{S}_i, \mathscr{A}_i, P_i, R_i)\}_{i=0,\ldots,p-1}$$

define a p-periodic Markov Decision Process (p-MDP):

- $\{\mathscr{S}_i\}_{i=0,\ldots,p-1}$ form a base collection of finite state spaces, such that the state S_t at time $t \geq 0$ is in $\mathscr{S}_i$ where $i = \mod(t,p)$, or equivalently, $t = i + kp$ for some integer $k \geq 0$;
- $\{\mathscr{A}_i\}_{i=0,\ldots,p-1}$ form a base collection of finite action spaces, such that the action A_t at time $t \geq 0$ is in $\mathscr{A}_i$ where $i = \mod(t,p)$.
- $P_i : \mathscr{S}_i \times \mathscr{A}_i \times \mathscr{S}_{i+1} \mapsto [0,1]$ for $i = 0, \ldots, p-2$ and $P_{p-1} : \mathscr{S}_{p-1} \times \mathscr{A}_{p-1} \times \mathscr{S}_0 \mapsto [0,1]$, form a base collection of state transition probability functions, such that $\text{Prob}(S_{t+1} = s' \mid S_t = s, A_t = a) = P_i(s,a,s')$ where $i = \mod(t,p)$.
- $R_i : \mathscr{S}_i \times \mathscr{A}_i \mapsto \mathbb{R}$ for $i = 0, \ldots, p-1$ form a base collection of bounded reward functions, such that the reward at time t given $S_t = s$, $A_t = a$, is $r_t = R_i(s,a)$ where $i = \mod(t,p)$.

We then define $\iota(t) = \mod(t,p)$ and define $(\mathscr{S}_t, \mathscr{A}_t, P_t, R_t)$ for $t \geq p$, where $\mathscr{S}_t = \mathscr{S}_{\iota(t)}$, $\mathscr{A}_t = \mathscr{A}_{\iota(t)}$, $P_t = P_{\iota(t)}$, and $R_t = R_{\iota(t)}$.

When $p = 1$, the problem of course reduces to a stationary MDP $(\mathscr{S}, \mathscr{A}, P, R)$.

Consider the class Π of admissible nonstationary Markov policies

$$\pi = \{A_t^\pi\}_{t \geq 0}, \tag{2.1}$$

where $A_t^\pi : \mathscr{S}_t \mapsto \mathscr{A}_t$ is the decision rule at time t, that maps the current state s to an action $a = A_t^\pi(s)$ selected from a subset $\mathcal{A}_t(s) \subset \mathscr{A}_t$ that represents a set of admissible actions given s. To streamline the notation, we just write $a \in \mathcal{A}_t$ instead of $a \in \mathcal{A}_t(s)$ in the sequel.

Let $\gamma \in (0, 1)$ be a discount factor, and $s \in \mathscr{S}_0$ an initial state. We consider the p-periodic Markov Decision Problem consisting in maximizing the expected discounted total return by the choice of an admissible nonstationary policy π:

$$V_0^*(s) = \max_{\pi \in \Pi} \mathbb{E}^\pi [\textstyle\sum_{t=0}^{\infty} \gamma^t R_t(S_t, A_t^\pi(S_t)) | S_0 = s], \tag{2.2}$$

where $\mathbb{E}^\pi$ emphasizes that the probability distribution of S_t depends on π.

2.2.2 Optimality Conditions

For brevity, we use the short-hand $P_{ss'}^t(a) = P_t(s, a, s')$. For all t, and for a fixed nonstationary policy π, the expected discounted cumulated reward-to-go at time t when being in state s and following policy π is given by

$$V_t^\pi(s) = R_t(s, A_t^\pi(s)) + \gamma \sum_{s' \in \mathscr{S}_{t+1}} P_{ss'}^t(A_t^\pi(s)) V_{t+1}^\pi(s'). \tag{2.3}$$

$V_0^\pi(s)$ is the value of policy π when starting from state s.

Due to the periodic structure of the problem, Bellman's principle of optimality leads

to a system of equations involving p value functions only,

$$V_i(s) = \max_{a \in \mathcal{A}_i} [R_i(s,a) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} P^i_{ss'}(a) V_{i+1}(s')]$$
$$\text{for } i = 0, 1, \ldots, p-2,$$
$$V_{p-1}(s) = \max_{a \in \mathcal{A}_{p-1}} [R_{p-1}(s,a) + \gamma \sum_{s' \in \mathcal{S}_0} P^{p-1}_{ss'}(a) V_0(s')]. \tag{2.4}$$

These equations are written more compactly as

$$V_i = T_i \, V_{i+1} \qquad \text{for } i = 0, 1, \ldots, p-2,$$
$$V_i = T_i \, V_0 \qquad \text{for } i = p-1,$$

where the operators T_i are defined from (2.4). By induction,

$$V_0 = (T_0 T_1 \ldots T_{p-1}) \, V_0,$$
$$V_1 = (T_1 \ldots T_{p-1} T_0) \, V_1,$$
$$\ldots$$
$$V_{p-1} = (T_{p-1} T_0 \ldots \, T_{p-2}) V_{p-1},$$

showing that V_i is a fixed point of the operator

$$\mathcal{T}_i = (T_i T_{i+1} \ldots T_{p-1} T_0 \ldots T_{i-1}) \ . \tag{2.5}$$

The operator $\mathcal{T}_i$ inherits the contractive mapping property of the operators T_i (Section 2.4.2 provides more details), and therefore, the system (2.4) admits a unique solution

$$V^* = (V_0^*, V_1^*, \ldots V_{p-1}^*) \ , \tag{2.6}$$

which we refer to as the *optimal periodic value function.*

We say that a policy $\pi = \{A_t^\pi\}_{t \geq 0}$ is greedy for a periodic value function $V =$

$(V^0, \ldots, V^{p-1})$ when

$$A_t^\pi(s) \in \arg\max_{a \in \mathcal{A}_i} [R_i(s,a) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P_{ss'}^i(a) V_{i+1}(s')]$$
$$\text{for } t = i + kp \text{ with } i \in \{0, 1, \ldots, p-2\},$$
$$A_t^\pi(s) \in \arg\max_{a \in \mathcal{A}_i} [R_i(s,a) + \gamma \sum_{s' \in \mathscr{S}_0} P_{ss'}^i(a) V_0(s')]$$
$$\text{for } t = i + kp \text{ with } i = p-1. \tag{2.7}$$

Let π^* be a policy greedy for V^* as defined by (2.6). Then π^* is nonstationary but periodic with cycle length p, and by definition of T_i, it is optimal for the problem (2.2).

Without loss of optimality, the search over the class Π of nonstationary policies is thus reduced to a search over the class Π_p of p-periodic admissible policies π, such that $A_t^\pi = A_i^\pi$ with $i = \iota(t)$:

$$V_0^*(s) = \max_{\pi \in \Pi_p} \mathbb{E}^\pi [\textstyle\sum_{t=0}^\infty \gamma^t R_t(S_t, A_t^\pi(S_t)) | S_0 = s] \ . \tag{2.8}$$

A policy $\pi \in \Pi_p$ is uniquely defined by $(A_0^\pi, \ldots, A_{p-1}^\pi)$.

2.3 Near-optimality bounds for greedy policies

This section establishes upper error bounds for the difference between the optimal return and the return of a policy greedy with respect to a periodic approximate value function. This situation covers the case of a policy greedy with respect to a periodic approximate value function obtained by value iteration.

The error bounds can be evaluated numerically under the assumption that the periodic optimal value function is known. As this assumption is not met in practice, we establish upper bounds to be used when the optimal value function is unknown.

2.3.1 Error bound for difference between optimal and approximate value functions

Let $V^* = (V_0^*, \ldots, V_{p-1}^*)$ be the optimal periodic value function (2.6), and $\pi^* \in \Pi_p$ the optimal p-periodic policy greedy for V^*.

In many cases, the value functions V_i^* are difficult or even impossible to evaluate. In order to overcome such situations, it is common to use approximate value functions.

Let $\widetilde{V} = (\tilde{V}_0, \ldots, \tilde{V}_{p-1})$ denote a periodic approximate value function, and let $\tilde{\pi} \in \Pi_p$ be a p-periodic policy greedy for $\widetilde{V}$, that is,

$$\begin{aligned} \tilde{\pi}_i &\in \operatorname*{argmax}_{a \in \mathcal{A}_i} [R_i(s,a) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P_{ss'}^i(a) \widetilde{V}_{i+1}(s')] \\ &\text{for } i = 0, \ldots, p-2, \\ \tilde{\pi}_{p-1} &\in \operatorname*{argmax}_{a \in \mathcal{A}_{p-1}} [R_{p-1}(s,a) + \gamma \sum_{s' \in \mathscr{S}_0} P_{ss'}^{p-1}(a) \widetilde{V}_0(s')]. \end{aligned} \tag{2.9}$$

Let $V_i^{\pi^*}$ denote the "value" of policy π^* at time i, and let $V_i^{\tilde{\pi}}$ denote the "value" of policy $\tilde{\pi}$ at time i, as defined by (2.3), where "value" at time i means the expected cumulated reward-to-go obtained by following the policy from time i onwards. By definition of π^* and V^*, we have $V_i^{\pi^*} \equiv V_i^*$, but in the case of $\tilde{\pi}$, in general we have

$$V_i^{\tilde{\pi}} \not\equiv \widetilde{V}_i \ .$$

Definition 2.3.1. *Given an optimal policy π^* associated to V^* and a policy $\tilde{\pi}$ greedy for $\widetilde{V}$, the function $\widetilde{L}_i : \mathscr{S}_i \mapsto \mathbb{R}$ is defined as the difference between the expected reward-to-go at time i of those two policies: For all $s \in \mathscr{S}_i$,*

$$\widetilde{L}_i(s) = V_i^*(s) - V_i^{\tilde{\pi}}(s) \ . \tag{2.10}$$

In particular, $\widetilde{L}_0(s)$ quantifies the suboptimality of policy $\tilde{\pi}$ for the periodic MDP started from initial state s.

Assumption. In each time period i, the value function V_i^* is approximated by $\widetilde{V}_i$, and

for all $s \in \mathscr{S}_i$, the difference between those value functions is bounded by ϵ_i:

$$|V_i^*(s) - \widetilde{V}_i(s)| \leq \epsilon_i \ . \tag{2.11}$$

To visualize the assumption,

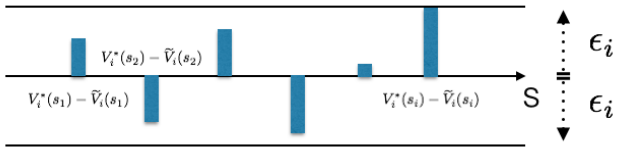


Figure 2.1: Interpretation of assumption 2.11

Under the assumption above, we provide the following proposition. The mechanism of the proof is based on [93].

Proposition 1. *Let $V^* = (V_0^*, \ldots, V_{p-1}^*)$ be the optimal periodic value function (2.6), let π^* be the associated optimal policy, and let $\tilde{\pi}$ be a policy greedy for $\widetilde{V} = (\widetilde{V}_0, \ldots, \widetilde{V}_{p-1})$, where $\widetilde{V}$ satisfies the assumption (2.11) for $i = 0, \ldots, p-1$. Then for all states s, $\widetilde{L}_i(s) \leq \overline{L}_i$, where*

$$\overline{L}_i = \frac{\sum_{k=0}^{i} \gamma^{p+k-i}(2\epsilon_k) + \sum_{k=i+1}^{p-1} \gamma^{k-i}(2\epsilon_k)}{1-\gamma^p} \ . \tag{2.12}$$

Before establishing Proposition 1, we note that each $\overline{L}_i$ in (2.12) depends on all the ϵ_k, but the weighting differs among each i. In the stationary case ($p = 1$), the bound reduces to

$$\widetilde{L}_0 \leq 2\epsilon_0\gamma/(1-\gamma) \ .$$

With $\epsilon = \max_i \epsilon_i$, the bound also reduces to

$$\widetilde{L}_i \leq \frac{2\epsilon \sum_{k=1}^{p} \gamma^k}{1-\gamma^p} = \frac{2\epsilon}{1-\gamma^p}\frac{\gamma(1-\gamma^p)}{1-\gamma} = \frac{2\epsilon\gamma}{1-\gamma} \ ,$$

which is a bound known in the literature (see e.g. [6]). The bound (2.12) is tighter, since it does not replace ϵ_i by ϵ. In fact, $2\epsilon_i$ is the length of the interval where the difference $V_i^*(s) - \widetilde{V}_i(s)$ lies, according to the assumption which is equivalent to $-\epsilon_i \leq V_i^*(s) - \widetilde{V}_i(s) \leq$

ϵ_i. Now, if instead we assume that

$$\underline{\epsilon_i} \leq V_i^*(s) - \widetilde{V}_i(s) \leq \overline{\epsilon_i} \quad \text{for all } s \in \mathscr{S}_i \ ,$$

then Proposition 1 applies by formally setting

$$2\epsilon_k = \overline{\epsilon_k} - \underline{\epsilon_k} \ . \tag{2.13}$$

Proof of Proposition 1. For each period i, there exists a state, say z_i, that achieves the maximal loss $\widetilde{L}_i$ at this period:

$$\widetilde{L}_i(z_i) \geq \widetilde{L}_i(s) \text{ for all } s \in \mathscr{S}_i \ .$$

To this state z_i corresponds the optimal action

$$a = A_i^{\pi^*}(z_i),$$

and the action of the policy greedy for $\widetilde{V}$,

$$b = A_i^{\tilde{\pi}}(z_i).$$

Momentarily let us assume $i \leq p-2$. Since $\tilde{\pi}$ is greedy for $\widetilde{V}$, we have

$$\begin{aligned} & R_i(z_i, a) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P^i_{z_i s'}(a) \widetilde{V}_{i+1}(s') \\ \leq \ & R_i(z_i, b) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P^i_{z_i s'}(b) \widetilde{V}_{i+1}(s') \ . \end{aligned}$$

From the assumption $|V^*_{i+1}(s) - \widetilde{V}_{i+1}(s)| \leq \epsilon_{i+1}$ we have

$$\begin{aligned} & R_i(z_i, a) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P^i_{z_i s'}(a)(V^*_{i+1}(s') - \epsilon_{i+1}) \\ \leq\ & R_i(z_i, b) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P^i_{z_i s'}(b)(V^*_{i+1}(s') + \epsilon_{i+1}) \ , \end{aligned}$$

which is equivalent to

$$\begin{aligned} & R_i(z_i, a) - R_i(z_i, b) \leq 2\gamma\epsilon_{i+1} \\ & \quad + \gamma \sum_{s' \in \mathscr{S}_{i+1}} [P^i_{z_i s'}(b) V^*_{i+1}(s') - P^i_{z_i s'}(a) V^*_{i+1}(s')] \ . \end{aligned} \tag{2.14}$$

On the other hand, we have, by definition of $\widetilde{L}_i$,

$$\begin{aligned} & \widetilde{L}_i(z_i) = V^*_i(z_i) - V^{\tilde{\pi}}_i(z_i) = R_i(z_i, a) - R_i(z_i, b) \\ & \quad + \gamma \sum_{s' \in \mathscr{S}_{i+1}} [P^i_{z_i s'}(a) V^*_{i+1}(s') - P^i_{z_i s'}(b) V^{\tilde{\pi}}_{i+1}(s')] \ . \end{aligned} \tag{2.15}$$

Combining (2.14) and (2.15), we obtain (for $i = 0, \ldots, p-2$)

$$\begin{aligned} \widetilde{L}_i(z_i) &\leq 2\gamma\epsilon_{i+1} + \gamma \sum_{s'} P^i_{z_i s'}(b)[V^*_{i+1}(s') - V^{\tilde{\pi}}_{i+1}(s')] \\ &= 2\gamma\epsilon_{i+1} + \gamma \sum_{s'} P^i_{z_i y}(b) \widetilde{L}_{i+1}(s') \\ &\leq 2\gamma\epsilon_{i+1} + \gamma \sum_{s'} P^i_{z_i s'}(b) \widetilde{L}_{i+1}(z_{i+1}) \\ &= 2\gamma\epsilon_{i+1} + \gamma \widetilde{L}_{i+1}(z_{i+1}) \ , \end{aligned} \tag{2.16}$$

where the second inequality results from the definition of z_{i+1}.

Similarly, for $i = p-1$, we obtain

$$\widetilde{L}_{p-1}(z_{p-1}) \leq 2\gamma\epsilon_0 + \gamma \widetilde{L}_0(z_0) \ . \tag{2.17}$$

By induction,

$$\begin{aligned}\widetilde{L}_i(z_i) &\leq 2\gamma\epsilon_{i+1} + 2\gamma^2\epsilon_{i+2} + \cdots + 2\gamma^{p-1-i}\epsilon_{p-1} \\ &\quad + 2\gamma^{p-i}\epsilon_0 + \cdots + 2\gamma^{p-1}\epsilon_{i-1} + 2\gamma^p\epsilon_i + \gamma^p\widetilde{L}_i(z_i) \\ &= 2\sum_{k=i+1}^{p-1}\gamma^{k-i}\epsilon_k + 2\sum_{k=0}^{i}\gamma^{p-i+k}\epsilon_k + \gamma^p\widetilde{L}_i(z_i) \ ,\end{aligned}$$

and finally,

$$\widetilde{L}_i(s) \leq \widetilde{L}_i(z_i) \leq \frac{2\sum_{k=0}^{i}\gamma^{p-i+k}\epsilon_k + 2\sum_{k=i+1}^{p-1}\gamma^{k-i}\epsilon_k}{1-\gamma^p} \ .$$

□

2.3.2 Bounding ϵ_k

From the analysis above, we may obtain an upper bound for the suboptimality of a periodic policy greedy for the periodic approximate value function $\widetilde{V}$. However, in practice, the optimal periodic value function V^* is unknown, and therefore we cannot compute the ϵ_k's of the assumption. Fortunately, we may bound ϵ_k using quantities obtained in the course of one iteration of the value iteration algorithm. The mechanism of the proof is based on results from the theory of value iteration presented for instance in [6].

Definition 2.3.2. *Given $\widetilde{V}^\ell = (\widetilde{V}_0^\ell, \ldots, \widetilde{V}_{p-1}^\ell)$, let $\widetilde{V}^{\ell+1}$ be defined by one value-iteration performed as follows:*

$$\begin{aligned}\widetilde{V}_i^{\ell+1}(s) &= \max_{a\in\mathcal{A}_i}[R_i(s,a) + \gamma\sum_{s'\in\mathscr{S}_{i+1}} P_{ss'}^i(a)\widetilde{V}_{i+1}^\ell(s')] \\ &= (T_i\widetilde{V}_{i+1}^\ell)(s) \qquad\qquad \textit{for } i = 0,\ldots,p-2, \\ \widetilde{V}_i^{\ell+1}(s) &= \max_{a\in\mathcal{A}_i}[R_i(s,a) + \gamma\sum_{s'\in\mathscr{S}_0} P_{ss'}^i(a)\widetilde{V}_0^\ell(s')] \\ &= (T_i\widetilde{V}_0^\ell)(s) \qquad\qquad \textit{for } i = p-1.\end{aligned}$$

Definition 2.3.3. *Define δ_i as the maximal change of the value function relative to period i*

using the update described in Definition 2.3.2, over all states $s \in \mathscr{S}_i$:

$$\delta_i = \max_{s \in \mathscr{S}_i} \left| \widetilde{V}_i^{\ell+1}(s) - \widetilde{V}_i^{\ell}(s) \right|$$
$$= \begin{cases} \max_{s \in \mathscr{S}_i} |(T_i \widetilde{V}_{i+1}^{\ell})(s) - \widetilde{V}_i^{\ell}(s)| & \text{for } i = 0, \ldots, p-2 \\ \max_{s \in \mathscr{S}_0} |(T_i \widetilde{V}_0^{\ell})(s) - \widetilde{V}_i^{\ell}(s)| & \text{for } i = p-1 \ . \end{cases}$$

Proposition 2. *Let $\widetilde{V}^{\ell}$ be an approximation to the optimal periodic value function V^*, and let δ_i be defined as above. In this situation,*

$$\epsilon_i = \max_{s \in \mathscr{S}_i} |V_i^*(s) - \widetilde{V}_i^{\ell}(s)|$$

admits for $i = 0, \ldots, p-1$ the upper bound

$$\epsilon_i \leq \frac{\sum_{k=0}^{i-1} \gamma^{p+k-i} \delta_k + \sum_{k=i}^{p-1} \gamma^{k-i} \delta_k}{1 - \gamma^p} \ .$$

Based on the assumption 2.1, we can also interpret 2 as following:

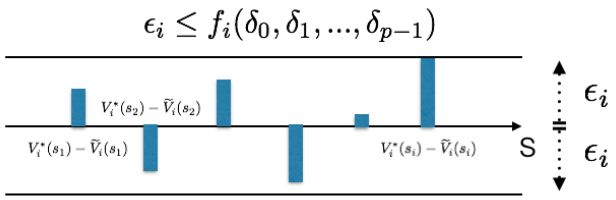


Figure 2.2: Interpretation of 2

Before establishing Proposition 2, we note that in the stationary case ($p = 1$), the expression reduces to

$$\epsilon_0 \leq \delta_0/(1-\gamma).$$

With $\delta = \max_i \delta_i$, the bound also reduces to

$$\epsilon_i \leq \frac{\delta}{1-\gamma^p} \sum_{k=0}^{p-1} \gamma^k = \frac{1}{1-\gamma^p} \frac{1-\gamma^p}{1-\gamma} = \frac{\delta}{1-\gamma} \ ,$$

which is a bound known in the literature. The bound in Proposition 2 is tighter, since it

does not replace δ_i by δ. In fact, if instead of δ_i from Definition 2.3.3, we define

$$\overline{\delta_i} = \max_{s \in \mathscr{S}_i} (\widetilde{V}_i^{\ell+1}(s) - \widetilde{V}_i^{\ell}(s)), \tag{2.18}$$

$$\underline{\delta_i} = \min_{s \in \mathscr{S}_i} (\widetilde{V}_i^{\ell+1}(s) - \widetilde{V}_i^{\ell}(s)), \tag{2.19}$$

and consider $\overline{\epsilon_i}$, $\underline{\epsilon_i}$ as defined in (2.13), then the result of Proposition 2 translates to

$$\overline{\epsilon_i} \leq \frac{\sum_{k=0}^{i-1} \gamma^{p+k-i} \overline{\delta_k} + \sum_{k=i}^{p-1} \gamma^{k-i} \overline{\delta_k}}{1-\gamma^p} \ ,$$
$$\underline{\epsilon_i} \geq \frac{\sum_{k=0}^{i-1} \gamma^{p+k-i} \underline{\delta_k} + \sum_{k=i}^{p-1} \gamma^{k-i} \underline{\delta_k}}{1-\gamma^p} \ .$$

Corollary 2.3.4. *The value for $2\epsilon_k$ in Proposition 1 can be set to*

$$2\epsilon_k = \frac{\sum_{k=0}^{i-1} \gamma^{p+k-i} (\overline{\delta_k} - \underline{\delta_k}) + \sum_{k=i}^{p-1} \gamma^{k-i} (\overline{\delta_k} - \underline{\delta_k})}{1-\gamma^p}. \tag{2.20}$$

As a preliminary to the proof of Proposition 2, we recall the following properties of the operators T_i .

1. Monotonicity: $V_{i+1} \succeq V'_{i+1}$ implies $T_i V_{i+1} \succeq T_i V'_{i+1}$, in the sense that $V_{i+1}(s) \geq V'_{i+1}(s)$ for all $s \in \mathscr{S}_{i+1}$ implies $(T_i V_{i+1})(s) \geq (T_i V'_{i+1})(s)$ for all $s \in \mathscr{S}_i$.

2. Uniform-shift: Let $\mathbf{1}_i : \mathscr{S}_i \mapsto \mathbb{R}$ denote the constant-valued function defined on $\mathscr{S}_i$ with value one. Then for any $c \in \mathbb{R}$, it holds that that

$$T_i(V_{i+1} + c\mathbf{1}_{i+1}) = T_i V_{i+1} + \gamma c \mathbf{1}_i,$$

since for all $s \in \mathscr{S}_i$,

$$\max_{a \in \mathcal{A}_i} [R_i(s,a) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P^i_{ss'}(s')(V_{i+1}(s) + c)]$$
$$= \max_{a \in \mathcal{A}_i} [R_i(s,a) + \gamma \sum_{s' \in \mathscr{S}_{i+1}} P^i_{ss'}(s')(V_{i+1}(s))] + \gamma c.$$

Proof of Proposition 2. By definition of δ_{i-1}, we have

$$\widetilde{V}_{i-1}^{\ell} + \delta_{i-1}\mathbf{1}_{i-1} \succeq T_{i-1}\widetilde{V}_{i}^{\ell}. \tag{2.21}$$

Applying T_{i-2} to both sides of the inequality, and using the uniform-shift and monotonicity properties of T_{i-2}, we get

$$T_{i-2}\widetilde{V}_{i-1}^{\ell} + \gamma\delta_{i-1}\mathbf{1}_{i-2} \succeq T_{i-2}T_{i-1}\widetilde{V}_{i}^{\ell} \ .$$

By definition of δ_{i-2}, we deduce

$$\widetilde{V}_{i-2}^{\ell} + \delta_{i-2}\mathbf{1}_{i-2} + \gamma\delta_{i-1}\mathbf{1}_{i-2} \succeq T_{i-2}T_{i-1}\widetilde{V}_{i}^{\ell} \ . \tag{2.22}$$

By repeating this process to cover a single cycle, we obtain the inequalities

$$\begin{aligned}\widetilde{V}_{i}^{\ell} + (\gamma^0\delta_i + \gamma^1\delta_{i+1} + \cdots + \gamma^{p-i-1}\delta_{p-1} + \\ \gamma^{p-i}\delta_0 + \cdots + \gamma^{p-1}\delta_{i-1})\mathbf{1}_i \\ \succeq T_iT_{i+1}\cdots T_{p-1}T_0\cdots T_{i-1}\widetilde{V}_{i}^{\ell} \ .\end{aligned}$$

By induction over an infinite number of cycles, we obtain

$$\begin{aligned}&\widetilde{V}_{i}^{\ell} + [(\gamma^0 + \gamma^p + \gamma^{2p} + \cdots)\delta_i + \\ &(\gamma^1 + \gamma^{p+1} + \gamma^{2p+1} + \cdots)\delta_{i+1} + \\ &\cdots + \\ &(\gamma^{p-1} + \gamma^{2p-1} + \gamma^{3p-1} + \cdots)\delta_{i-1}]\mathbf{1}_i \\ &= \widetilde{V}_{i}^{\ell} + \left(\frac{\gamma^0}{1-\gamma^p}\delta_i + \frac{\gamma^1}{1-\gamma^p}\delta_{i+1} + \cdots + \frac{\gamma^{p-i-1}}{1-\gamma^p}\delta_{p-1}\right. \\ &\left. + \frac{\gamma^{p-i}}{1-\gamma^p}\delta_0 + \cdots + \frac{\gamma^{p-1}}{1-\gamma^p}\delta_{i-1}\right)\mathbf{1}_i \\ &\succeq \lim_{N\to\infty}(T_iT_{i+1}\cdots T_{p-1}T_0\cdots T_{i-1})^N\widetilde{V}_{i}^{\ell} = \lim_{N\to\infty}\mathcal{T}_i^N\widetilde{V}_{i}^{\ell} \\ &= V_i^* \ ,\end{aligned}$$

where the last equality comes from the convergence of the value iteration algorithm for finding the fixed point V_i^* of the operator $\mathcal{T}_i$ as defined in (2.5).

Rearranging the terms of the inequality above, we obtain

$$\frac{1}{1-\gamma^p}\left(\sum_{k=0}^{i-1}\gamma^{p+k-i}\delta_k + \sum_{k=i}^{p-1}\gamma^{k-i}\delta_k\right)\mathbf{1}_i \succeq V_i^* - \widetilde{V}_i^\ell \ .$$

Over all states, this implies

$$\frac{\sum_{k=0}^{i-1}\gamma^{p+k-i}\delta_k + \sum_{k=i}^{p-1}\gamma^{k-i}\delta_k}{1-\gamma^p} \geq \max_{s\in\mathscr{S}_i}(V_i^*(s) - \widetilde{V}_i^\ell(s)) \ . \tag{2.23}$$

A similar reasoning starting from the inequality

$$\widetilde{V}_{i-1}^\ell - \delta_{i-1}\mathbf{1}_{i-1} \preceq T_{i-1}\widetilde{V}_i^\ell$$

leads to

$$-\frac{\sum_{k=0}^{i-1}\gamma^{p+k-i}\delta_k + \sum_{k=i}^{p-1}\gamma^{k-i}\delta_k}{1-\gamma^p} \leq \min_{s\in\mathscr{S}_i}(V_i^*(s) - \widetilde{V}_i^\ell(s))$$

which together with (2.23) implies

$$\begin{aligned}\frac{\sum_{k=0}^{i-1}\gamma^{p+k-i}\delta_k + \sum_{k=i}^{p-1}\gamma^{k-i}\delta_k}{1-\gamma^p} &\geq \max_{s\in\mathscr{S}_i}\ |V_i^*(s) - \widetilde{V}_i^\ell(s)| \\ &= \epsilon_i \ .\end{aligned}$$

□

2.4 Value iteration

In this section, we describe a value iteration algorithm suitable for periodic Markov Decision Processes in finite state-action spaces. The algorithm outputs a periodic value function $\widetilde{V}$ such that a periodic policy greedy for $\widetilde{V}$ is guaranteed to be (at least) η-optimal.

2.4.1 Iteration mechanism

The optimality condition $V_0 = (T_0 \dots T_{p-1})V_0$ suggests a value iteration algorithm for solving the p-periodic Markov Decision Problem.

Given $\eta > 0$, the algorithm returns a periodic value function $\widetilde{V} = (\widetilde{V}_0, \dots \widetilde{V}_{p-1})$ such that a policy greedy for $\widetilde{V}$ is guaranteed to be (at least) η-optimal.

1. Initialization: Guess an initial value function $\widetilde{V}_0^\ell$ for $\ell = 0$ (for instance, $\widetilde{V}_0^\ell \equiv 0$).

2. Value iteration: Compute successively

$$\begin{aligned}
\widetilde{V}_{p-1}^{\ell+1} &= T_{p-1}\widetilde{V}_0^\ell, \\
\widetilde{V}_{p-2}^{\ell+2} &= T_{p-2}\widetilde{V}_{p-1}^{\ell+1}, \\
&\dots \\
\widetilde{V}_0^{\ell+p} &= T_0\widetilde{V}_1^{\ell+p-1}.
\end{aligned}$$

   For $i = 0, \dots, p-1$, compute $\overline{\delta}_i, \underline{\delta}_i$ from (2.18),(2.19), and $2\epsilon_i$ from (2.20).
   Compute $\overline{L}_0$ from (2.12).

3. Set $\ell \leftarrow \ell + p$ and repeat Step 2 until the stopping criterion $\overline{L}_0 \le \eta$ is met.

The value-iteration algorithm utilizes the bounds of Section 2.3 which are adapted to each value function in the cycle.

As it can be seen from (2.10), $\tilde{L}_0(s) \le \eta$ indicates that a policy greedy with respect to the current value function $\tilde{V}^\ell$ is η-optimal. This does not imply that $\widetilde{V}_0^\ell$ represent the value of the policy, that is, that $\widetilde{V}_0^\ell$ has converged to V_0^*. To see this, consider the example of a policy greedy with respect to $V^* + c\mathbf{1}$ where $c\mathbf{1}$ is a constant-valued function: this policy is optimal and its value is $V_0^*(s)$ where s is the initial state.

2.4.2 Convergence rate

For a function $V_i : \mathscr{S}_i \mapsto \mathbb{R}$, we consider the sup-norm

$$||V_i||_{\infty,i} = \max_{s \in \mathscr{S}_i} |V_i(s)|,$$

where we write max instead of sup because $\mathscr{S}_i$ is finite.

Proposition 3. *The rate of convergence of the value iteration algorithm is governed by*

$$||\mathcal{T}_0^k V_0 - V_0^*||_{\infty,0} \leq \gamma^{kp} ||\widetilde{V}_0^0 - V_0^*||_{\infty,0} \ ,$$

computed using $\ell = kp$ iterations.

Proof. The mapping $\mathcal{T}_i = (T_i T_{i+1} \dots T_{p-1} T_0 \dots T_{i-1})$ is contractive with modulus γ^p, in the sense that for all functions V_i, V_i' from $\mathscr{S}_i$ to $\mathbb{R}$,

$$||\mathcal{T}_i V_i - \mathcal{T}_i V_i'||_{\infty,i} \leq \gamma^p ||V_i - V_i'||_{\infty,i} \ . \tag{2.24}$$

To see this, note first that the mappings T_i are contractions with modulus γ, in the sense that, for $i = 0, \dots, p-2$, the following property holds for all functions V_{i+1}, V_{i+1}' from $\mathscr{S}_{i+1}$ to $\mathbb{R}$:

$$||T_i V_{i+1} - T_i V_{i+1}'||_{\infty,i} \leq \gamma ||V_{i+1} - V_{i+1}'||_{\infty,i+1} \ ,$$

and for $i = p-1$, the following property holds for all functions V_0, V_0' from $\mathscr{S}_0$ to $\mathbb{R}$:

$$||T_{p-1} V_0 - T_{p-1} V_0'||_{\infty,p-1} \leq \gamma ||V_0 - V_0'||_{\infty,0} \ .$$

Then, by using the contractive property of the operators T_k successively for $k = i, i+1, \dots, p-1, 0, \dots, i-1$, one gets

$$\begin{aligned}
&||\mathcal{T}_i V_i - \mathcal{T}_i V_i'||_{\infty,i} \\
&= ||(T_i T_{i+1} \dots T_{i-1}) V_i - (T_i T_{i+1} \dots T_{i-1}) V_i'||_{\infty,i} \\
&\leq \gamma ||(T_{i+1} \dots T_{i-1}) V_i - (T_{i+1} \dots T_{i-1}) V_i'||_{\infty,i+1} \\
&\leq \ \cdots \leq \gamma^p ||V_i - V_i'||_{\infty,i} \ .
\end{aligned}$$

In particular, for $V_i' = V_i^*$, we have $\mathcal{T}_i V_i^* = V_i^*$ and therefore

$$||\mathcal{T}_i V_i - V_i^*||_{\infty,i} \leq \gamma^p ||V_i - V_i^*||_{\infty,i} \ .$$

It remains to set $i = 0$ and iterate k times the mapping $\mathcal{T}_0$ to get the result. □

2.5 Application to grid-level storage operations

We consider a grid-level storage control problem where the goal is to operate a battery to exchange electricity with the power grid at the current hourly spot price. The problem is formulated as p-periodic Markov Decision Process where the goal is to maximize the expected net proceeds from the purchase and selling of electricity over an infinite horizon.

A very appealing feature of the periodic Markov Decision Process model proposed in this chapter is that its computational tractability will not be affected by adopting shorter time periods, for instance periods of 5 minutes or less, making it suitable for various storage devices with different physical characteristics (power and energy capacities) or operating at different time scales.

In our numerical implementation, we use C as our programming language, with OpenMP for parallel computations.

2.5.1 Model Description

The state S_t is the current price S_t^{price} and the battery energy charge level S_t^{battery}. The decision A_t is the power at which we charge ($A_t < 0$) or discharge the battery ($A_t > 0$). The instantaneous reward R_t is the revenue over the time period: price $\times$ energy injected to the grid,

$$R_t(S_t, A_t) = S_t^{\text{price}} A_t \ \Delta_t$$

where Δ_t is the time period duration (1 hour).

The discount factor γ is set to 0.99. Hence the weight of the reward of tomorrow's

hour-1 is $\gamma^{24} = 0.78$, and the weight of the reward of next week's hour-1 is $\gamma^{168} = 0.18$.

The stochastic hourly price process is modeled as a cyclo-stationary process with a cycle length of $p = 24$ hours. The means are chosen to match the day-ahead prices posted by the independent system operator on the day preceding the exploitation of the policy. For the price volatility, we use the historical volatility of prices on similar days, although using implied volatility of options on forward contracts should yield better predictive distributions. Inter-hour correlations are neglected, but as the current price is in the state, specifying an order-1 Markov model for the price would simply change the state transition probabilities without increasing the complexity of the periodic MDP.

The hourly prices are assumed to follow a lognormal distribution $LN(\mu_i, \sigma_i^2)$. Other distributions could easily be accommodated. We formulate the problem on parameters estimated from PJM price data for one day of 2013, with historical volatilities estimated from prices of the corresponding month. The estimated parameters μ_i, σ_i are given in Table 2.1, along with the corresponding mean prices $\exp(\mu_i + \sigma_i^2/2)$.

For the battery, we use the parameters of a GM Chevy Volt battery pack repurposed for energy storage by ABB, having $C^{\text{battery}} = 10$ kWh of usable capacity [4]. We assume a power rating of $P^{\text{battery}} = 5$ kW, such that a full charge over the 10 kWh range can be done in 2 hours if desired. Intermediate charge and discharge rates are allowed, including the null injection $A_t = 0$ (pure storage).

The battery state transition function is given by

$$S_{t+1}^{\text{battery}} = S_t^{\text{battery}} - A_t \, \Delta_t,$$

with a state-dependent action space defined by the constraints

$$\begin{aligned} -P^{\text{battery}} \leq A_t \leq P^{\text{battery}} \qquad & \text{(power capacity)}, \\ S_t^{\text{battery}} - C^{\text{battery}} \leq A_t \, \Delta_t \leq S_t^{\text{battery}} \qquad & \text{(energy capacity)}. \end{aligned}$$

A more detailed battery model could easily be accommodated.

Table 2.1: Hourly prices: Lognormal parameters

Hour i	μ_i					
1–6	3.24	3.06	2.54	2.39	-1.13	-0.57
7–12	2.82	3.15	3.26	3.39	3.82	3.8
13–18	3.82	3.97	4.16	3.94	3.84	3.86
19–24	3.86	3.75	3.75	3.53	3.47	3.41

Hour i	σ_i					
1–6	0.14	0.25	0.59	0.78	1.21	1.07
7–12	0.25	0.16	0.2	0.3	0.24	0.23
13–18	0.28	0.4	0.33	0.35	0.27	0.19
19–24	0.21	0.26	0.25	0.12	0.12	0.11

Hour i	expected price $\exp(\mu_i + \sigma_i^2/2)$					[\$/MWh]
1–6	25.79	22.00	15.09	14.79	0.67	1.00
7–12	17.31	23.64	26.58	31.03	46.94	45.90
13–18	47.43	57.40	67.66	54.67	48.25	48.33
19–24	48.52	43.98	43.87	34.37	32.37	30.45

2.5.2 Finite State Approximation

For each hour i, the support of the price distribution is partitioned into $N = 20$ cells $[c_{i,j-1}, c_{i,j})$ of probability $1/N$, and discrete price levels $s_{i,j}^{\text{price}}$ are determined by computing the conditional expectation of the price given the cell:

$$\begin{aligned} c_{i,j} &= F_i^{-1}(j/N) = \exp(\mu_i + \sigma_i \Phi^{-1}(j/N)), \qquad (2.25) \\ &\qquad j = 0, \ldots, N, \\ s_{i,j}^{\text{price}} &= \mathbb{E}[S_i^{\text{price}} \mid c_{i,j-1} \leq S_i^{\text{price}} \leq c_{i,j}] \\ &= \int_{c_{i,j-1}}^{c_{i,j}} x f_i(x) dx \Big/ \int_{c_{i,j-1}}^{c_{i,j}} f_i(x) dx \\ &= \frac{e^{\mu_i + \sigma_i^2/2} \left(\Phi(\Phi^{-1}(\frac{j}{N}) - \sigma_i) - \Phi(\Phi^{-1}(\frac{j-1}{N}) - \sigma_i) \right)}{1/N}, \end{aligned}$$

where F_i^{-1} and f_i denote the inverse cumulative distribution function (inverse cdf) and probability density function (pdf) of the price, and Φ and Φ^{-1} denote the cdf and inverse cdf of the standard normal distribution, respectively.

By so doing, the rewards associated to the discrete prices will give the correct expected rewards for the original continuous distribution conditionally to being in the cell.

As for the battery state, we partition the operating range into a uniform grid of 50 discrete levels.

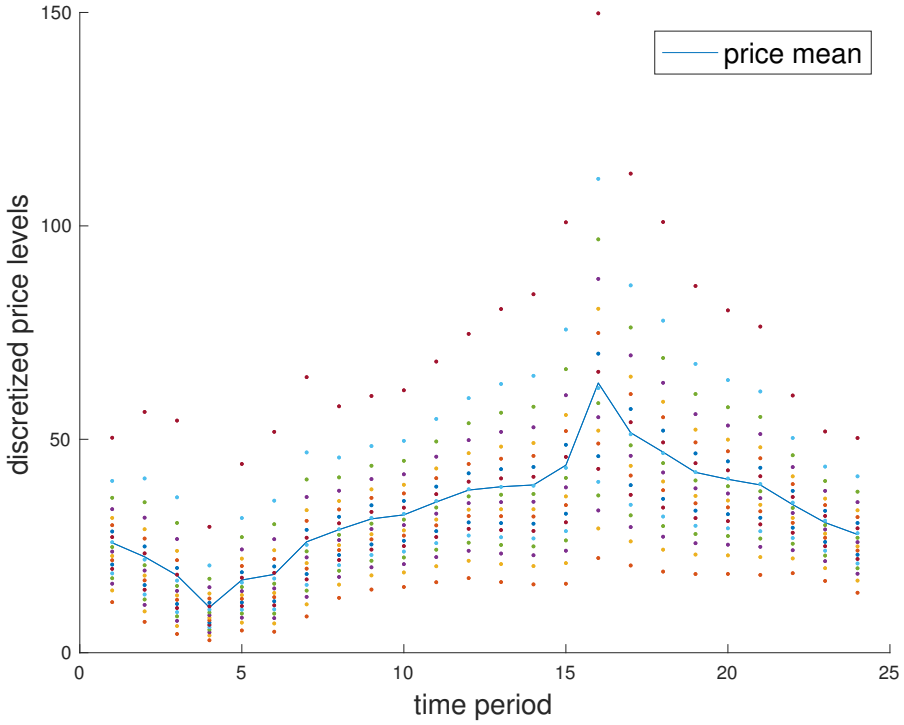


Figure 2.3: A typical discretized price state.

Figure 2.3 shows the discretized price levels who approximates the Lognormal distribution. Each time period we have 20 price levels and each price level has probability 0.05.

2.5.3 Results

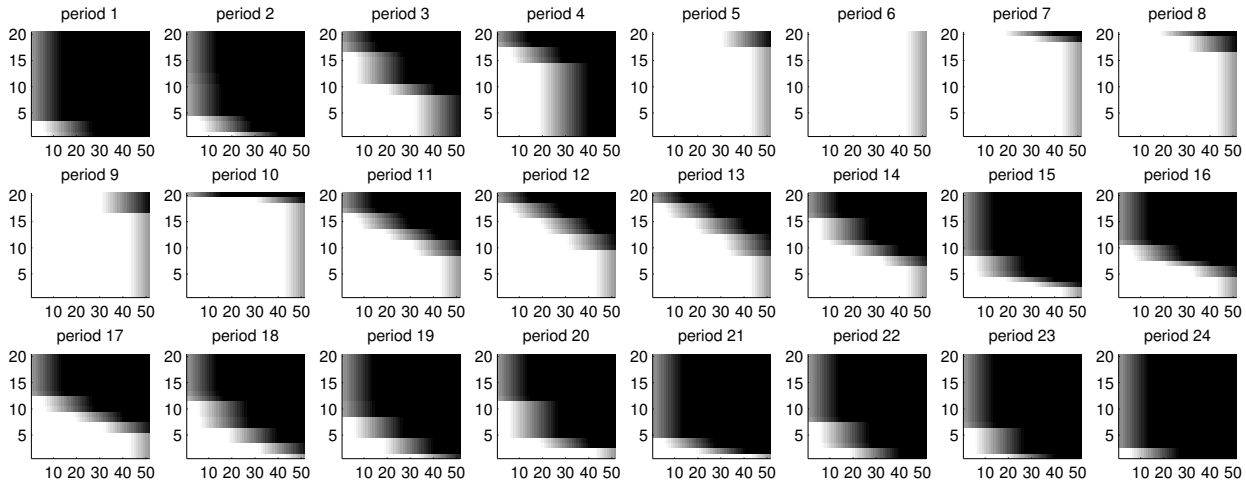


Figure 2.4: Near-optimal periodic policy, from hour 1 (midnight to 1am) to hour 24. X-axis: charge level state (indexed from 1 to 50), Y-axis: period-dependent price state (indexed from 1 to 20). White: Charge at maximal rate (buy), Black: Discharge at maximal rate (sell), Gray: intermediate actions.

Figure 2.4 shows the near-optimal periodic policy returned by the value iteration algorithm, corresponding to the problem data of Table 2.1. The periodic policy and the price cells $c_{i,j}$ should be loaded into the battery controller and recomputed periodically, typically every day. Actions in real-time would be selected according to the charge level and the price-cell index hit by the spot price.

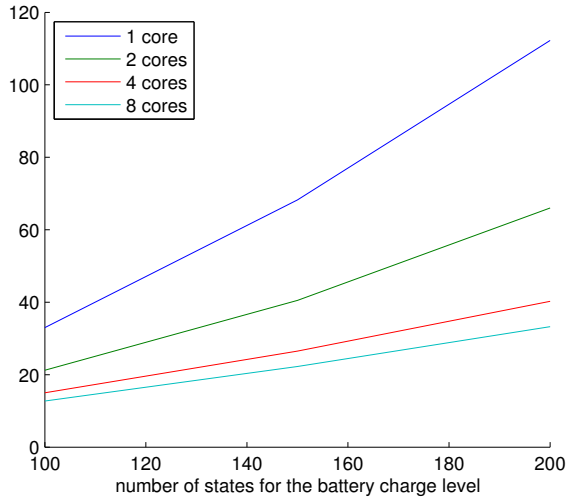


Figure 2.5: Running time (seconds) for different problem sizes and number of cores.

Figure 2.5 depict the running times for computing a near-optimal periodic policy, as a function of the number of cores used in our parallel implementation. The results of our experiments are consistent with parallel computing theory, which predicts that the more cores we use, the less efficiency gains we should get [83].

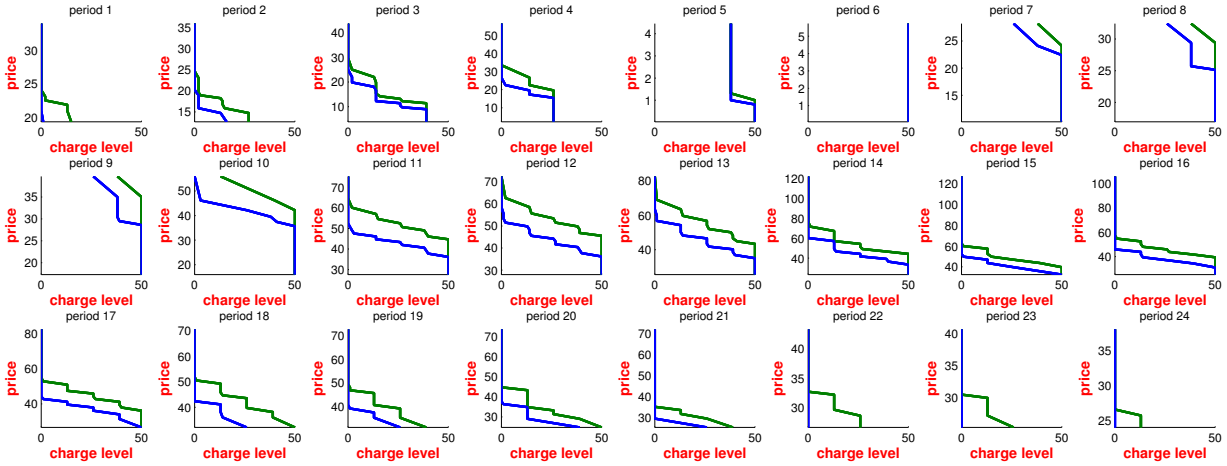


Figure 2.6: The threshold policy of 24 time periods

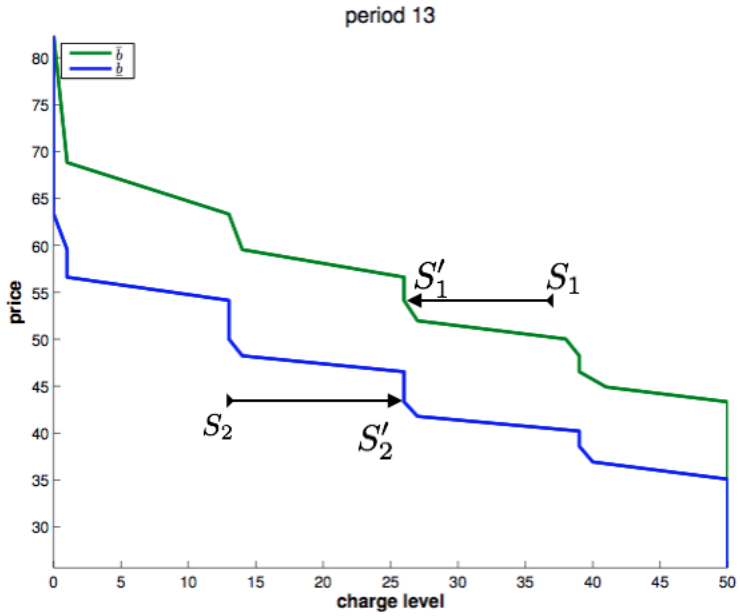


Figure 2.7: The threshold policy of 13^{th} period

Figure 2.6 shows the threshold policy of the whole 24 time periods and particularly,

Figure 2.7 shows the threshold policy of time period 13. The region under blue line is the region where we charge and the target charge level (which may not be reached due to power rate) is on the blue line along the horizontal direction, i.e., S_2' is the target level for state S_2. And the region above the green line is the discharging region where the target charge level is on the green line along the horizontal direction, i.e., S_1' is the target level for state S_1. Finally, the region between green and blue line is where we do nothing. The proof of the property of threshold policy is provided in the Appendix A in detail.

2.6 Conclusion and future work

In this chapter, we revisit the framework of periodic Markov Decision Processes, motivated by the use of cyclo-stationary models to approximate the expected return of reward processes in non-stationary environments subject to seasonal effects. We apply the approach to a grid-level storage control problem to obtain a near-optimal periodic policy, computed efficiently by combining various techniques proposed in the chapter.

Although the numerical example demonstrates the effectiveness of the approach for a cycle of 24 periods of 1 hour, cycles defined on a much larger number of periods can be accommodated without losing tractability, for instance 1440 periods of 5 minutes for a daily cycle. This favorable property of the approach proposed in this chapter comes from the choice of considering policies based on the greedy optimization of value functions.

The ability to accommodate short duration periods is especially important for battery storage control problems, for two related reasons. First, in contrast to hydro storage, the capacity of batteries is tiny. Profitable operations can thus come from increasing the frequency of profitable charge-discharge cycles during the day, in addition to providing regulation services for the system operator. Second, if wholesale electricity spot prices are updated every five minutes, it makes sense to have a control policy adapted to this time resolution. The spot price fed into the battery storage control problem can then be interpreted as the expected average spot price over the next 5 minutes. The charging action determined by the model should be implemented at a uniform rate over the next 5 minutes. As the 5-min spot price is much more volatile than the hourly spot price, operating the

battery at the 5-minute time scale is much more effective than operating it at the one-hour time scale.

The present work can be extended in several directions. The theoretical analysis could be extended to handle the case of an approximate evaluation of the Bellman iterations, and to handle approximate periodic value functions of a given approximation architecture. The value of a periodic policy used in a rolling-horizon fashion for appropriate classes of nonstationary problems could be studied in theory and numerically. The concept of using a cyclo-stationary reward process to approximate value functions at a terminal stage could be adapted to other approaches to stochastic optimization besides the Markov Decision Process framework.

A more realistic model for the battery should be implemented, to include the limited life cycles for batteries. We will discuss this model in the next chapter.

Chapter 3

Battery Operation with Aging

The material of this chapter has been published in paper

Yuhai Hu and Boris Defourny. Optimal price-threshold control for battery operation with aging phenomenon: a quasiconvex optimization approach. *Annals of Operations Research*, pages 1–28, 2017

3.1 Introduction

With renewable energy having an increasing impact on power grid operations, research and development has been very active in energy storage technology [27, 96]. Meanwhile, more attention has been given to improve the operation of energy storage devices [57, 63]. While the basic strategy for market-based energy storage operations is to buy at low prices and sell at high prices [104], with battery storage the charging-discharging cycles also count against the life of the storage device [71]. The battery-life effect may affect optimal storage operations much more significantly than with other storage technologies. Therefore, the present paper focuses on optimal control algorithms for storage devices where aging induced by operations is significant.

3.1.1 Contributions and Related Work

The structure of optimal policies for battery operation problems is well known and has been analyzed previously, see e.g. [79, 104, 38, 92, 53, 54]. However, since batteries are

expensive to replace — see e.g. [2] — the aging phenomenon is expected to have a strong influence on the overall return on investment. Batteries experience degradations in terms of capacity fading and increased resistance with time. The factors affecting degradation include operation temperature, depth of discharge, and state-of-charge [49]. [71] provide a model which takes capacity degradations into account, through a so-called degradation function. In their analysis, they assume that the lost capacity is replaced immediately and model this as an instantaneous penalty. Our approach differs in that with our model the stage of deterioration of the battery is being tracked through an additional state variable, and the goal is to maximize the expected profit from operations over the entire battery life.

The contributions of this chapter can be summarized as follows:

- We incorporate the aging phenomenon into a grid-level battery operation problem and formulate it as a Markov Decision Process with expected cumulated discounted rewards [84, 77, 44] where random electricity prices follow a given distribution. The optimal policy for this problem has the structure of a threshold policy, similarly to many problems admitting an optimal monotone policy [101, 36] that are often encountered in inventory theory.

- The value of modeling the end-of-life of the battery in the optimal control problem, versus neglecting it, is assessed in our computational work.

- We provide an algorithm for optimizing the set of state-dependent thresholds defining the policy. The algorithm is based on solving a sequence of quasiconvex optimization problems. The algorithm could be viewed as a policy-iteration scheme that utilizes problem structure to determine optimal parameters in the optimal order, owing to the decomposition of the global optimization problem into subproblems that can be solved to near-optimality. Under the assumption that the prices are independent and identically distributed (i.i.d.), the proposed algorithm works directly in the continuous state space, and finds the exact optimal thresholds within the tolerance of the quasiconvex optimization subroutine.

- Computational results confirm that the proposed algorithm is faster and more accurate than general-purpose dynamic programming algorithms such as value iteration. With the proposed algorithm, there are no details to tune (such as the details of the finite state space approximation for value iteration), and the complexity is not affected by the discount factor.

- We provide the error analysis of the propagation of suboptimal solutions caused by the finite tolerance of the optimization subroutine. If the absolute tolerance of the subroutine solving the quasiconvex optimization problems is $\epsilon \geq 0$, the error on the value function of the problem is bounded by $\epsilon/(1-\gamma)$, where $\gamma \in (0,1)$ is the discount factor.

- We later extend our threshold optimization approach to a more general type of price process, in particular, to a class of Markovian regime-switching processes. Our analysis establishes the global convergence and locally quadratic convergence of the sequence of iterates used to find the optimal thresholds.

Relating our results back to the existing algorithmic literature on dynamic programming, we note that the error bound for the proposed algorithm is the same as the error bound of approximate value iteration, which is known to be $\limsup_{k\to\infty} ||J_k - J^*|| \leq \epsilon/(1-\gamma)$, see e.g. [8], where in this expression J^* is the optimal value function, J_k is the approximate value function at iteration k assuming that $\sup_s |J_k(s) - [TJ_{k-1}](s)| \leq \epsilon$, and TJ_{k-1} denotes the update of the approximate value function J_{k-1} by exact Bellman iteration. There is a difference in those bounds, however, in that with approximate value iteration ϵ can be relatively large (it is the largest approximation error among all states) while with the proposed approach, ϵ is related to the user-controlled tolerance of the optimization subroutine.

Concerning the convergence rate result, we recall that under ideal conditions, policy iteration converges to an optimal policy at a quadratic rate [78]. However, as pointed out by [87], these conditions involve an exact optimization for each state. The algorithm proposed here operates over the parameters of a parametric policy, belonging to a class containing

an optimal policy. The resulting problem, being finite-dimensional, can be solved to an arbitrary precision, thereby circumventing the impossibility, in a continuous state space, of finding an optimal decision for each state individually.

3.1.2 Organization

The chapter is organized as follows. Section 3.2 formulates the Markov Decision Process model for the battery operation problem. Section 3.3 provides the results related to the structure of the optimal policy. Section 3.4 develops the procedure that finds a sequence of optimal thresholds describing the optimal policy by maximizing a sequence of quasiconcave objective functions. Section 3.5 provides our error analysis for the purpose of showing that the proposed scheme is robust with respect to the finite tolerance of the optimization subroutine. Section 3.6 refines the battery model to take into account capacity deterioration and charging inefficiencies. Section 3.7 compares the proposed algorithm to the classical value iteration algorithm, illustrates the effect of variants in the battery model, and assesses the economic value of having the battery controlled by a policy aware of finite-life effects. Section 3.8 extends the price process model to a Markovian regime-switching model and establishes convergence rate results. Finally, Section 3.9 concludes.

3.2 Model Description

This section describes the Markov Decision Process of the battery operation problem subject to battery aging. The time horizon for this problem is infinite, as battery aging is assumed not to occur when the battery is idle. The problem is described using the following notation.

- $\mathscr{S}$ is the state space. The state at time t, denoted S_t, has three components: the charge level $c_t \in \{0, 1\}$, the remaining life $n_t \in \{0, 1, \ldots, N\}$ interpreted as a number of remaining charging cycles, and the market price p_t, assumed to follow a given exogenous distribution. Thus, $S_t = (c_t, n_t, p_t)$.

- $\mathscr{A}$ is the decision space. The decision at time t, denoted A_t, is either -1 (discharge),

0 (idle), or +1 (charge). It is convenient to write $\mathscr{D}$ for the subset of $\mathscr{S} \times \mathscr{A}$ of feasible state-action pairs. The set of admissible decisions when being in state s is then denoted $D(s) = \{a \in \mathscr{A} : (s, a) \in \mathscr{D}\}$.

- P is the state transition probability function, that describes for each $(s, a) \in \mathscr{D}$ the probability of going to the next state S_{t+1} when being in state $S_t = s$ and choosing action $A_t = a$.

- $R : \mathscr{S} \times \mathscr{A} \mapsto \mathbb{R}$ is the reward function, such that the reward at time t given $S_t = s$, $A_t = a$, is $r_t = R(s, a)$.

The goal for this problem is to maximize the expected return over the infinite horizon, that is, to maximize the expected discounted sum of instantaneous rewards:

$$V(s) = \max_{\pi} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A^{\pi}(S_t)) \;\middle|\; S_0 = s \right] , \tag{3.1}$$

where $\gamma \in (0, 1)$ is the discount factor, and π is the policy that maps states to admissible decisions, as $a_t = A^{\pi}(s_t) \in D(s_t)$. Without loss of optimality we assume that the policy is stationary, see e.g. [77].

The transition function can be described as follows.

- The price follows a known distribution. The prices are independent and identically distributed (i.i.d) among time periods. (This is revisited in Section 3.8.)

- When $n_t \geq 1$, whenever the battery is discharged ($a_t = -1$) the remaining life cycle counter n_t is decremented of one unit. During charge or idle operations, the remaining life cycle counter does not change. Thus,

$$(c_{t+1}, n_{t+1}) = \begin{cases} (c_t + a_t, n_t) & \text{if } a_t \in \{0, 1\}, \\ (c_t + a_t, n_t - 1) & \text{if } a_t = -1. \end{cases} \tag{3.2}$$

- $n_t = 0$ is a terminal state (end of battery life). By the transition rule above, the charge level always becomes $c_t = 0$ when entering $n_t = 0$. The price p_t continues to

evolve but we may as well assume the full state process S_t is stopped when entering $n_t = 0$.

The reward function is described as

$$R(s_t, a_t) = p_t(-a_t) \tag{3.3}$$

under the convention that a_t is admissible in state s_t. In states with $c_t = 1$, $a_t \in \{-1, 0\}$. In states with $n_t \geq 1$ and $c_t = 0$, $a_t \in \{0, 1\}$. (The battery storage model is revisited in Section 3.6.)

3.3 Threshold Policy

In this section, we analyze the properties of the problem modeled in Section 3.2. Proposition 4 shows that there are price thresholds that determine the optimal decisions.

Proposition 4. *For each remaining life cycles $n \geq 1$,*
a) There exists a critical price level $\theta_n^{1,*}$, depending on n, such that an optimal action in each state $S_t = (c_t, n_t, p_t)$ for $c_t = 1$ is

$$A^\pi(1, n_t, p_t) = \begin{cases} -1 & \text{if } p_t \geq \theta_{n_t}^{1,*}, \\ 0 & \text{if } p_t < \theta_{n_t}^{1,*}. \end{cases}$$

b) There exists a critical price level $\theta_n^{0,*}$, depending on n, such that an optimal action in each state $S_t = (c_t, n_t, p_t)$ for $c_t = 0$ is

$$A^\pi(0, n_t, p_t) = \begin{cases} 1 & \text{if } p_t \leq \theta_{n_t}^{0,*}, \\ 0 & \text{if } p_t > \theta_{n_t}^{0,*}. \end{cases}$$

Proof of Proposition 4. (For convenience, in this proof we omit the subscript t in c_t, n_t, p_t and s_t, a_t. The next price is still written p_{t+1}.)

The instantaneous reward $R(s,a)$ defined in (3.3) only depends on the price state variable p and the decision a, so for simplicity we use the shorthand notation $R(p,a)$. We write the value function (3.1) as

$$V(c,n,p) = \max_{a \in \mathcal{A}(c,n)} Q(c,n,p,a)$$

where

$$Q(c,n,p,a) = R(p,a) + \gamma \mathbb{E}[V(c+a, n - 1_{\{n \geq 1, a=-1\}}, p_{t+1})].$$

Note that $V(c,0,p) = 0$ since there is no future reward when n attains 0.

Fix some arbitrary prices p, p' such that $p < p'$. Whenever $a = 0$ we have $Q(c,n,p,a) = Q(c,n,p',a) = \mathbb{E}[V(c,n,p_{t+1})]$ since $R(\cdot,0) = 0$ and the distribution of p_{t+1} does not depend on p.

If $c = 1$ then

$$\begin{aligned} Q(1,n,p,-1) &= p + \gamma \mathbb{E}[V(0,n-1,p_{t+1})] \\ &< p' + \gamma \mathbb{E}[V(0,n-1,p_{t+1})] = Q(1,n,p',-1). \end{aligned}$$

Thus if $c = 1$ and $a = -1$ is optimal at price p, meaning $Q(1,n,p,-1) \geq Q(1,n,p,0)$, we also have

$$Q(1,n,p',-1) > Q(1,n,p,-1) \geq Q(1,n,p,0) = Q(1,n,p',0),$$

that is, $a = -1$ is also optimal at any price $p' > p$. Hence there exists a critical threshold price $\theta_n^{1,*}$ such that $a = 0$ if $p < \theta_n^{1,*}$ and $a = 1$ if $p \geq \theta_n^{1,*}$.

If $c = 0$ then

$$\begin{aligned} Q(0,n,p,1) &= -p + \gamma \mathbb{E}[V(1,n,p_{t+1})] \\ &> -p' + \gamma \mathbb{E}[V(1,n,p_{t+1})] = Q(0,n,p',1). \end{aligned}$$

Thus if $c = 0$ and $a = 1$ is optimal at price p', meaning $Q(0,n,p',1) \geq Q(0,n,p',0)$, we

also have

$$Q(0,n,p,1) > Q(0,n,p',1) \geq Q(0,n,p',0) = Q(0,n,p,0),$$

that is, $a = 1$ is also optimal at any price $p < p'$. Hence there exists a critical threshold price $\theta_n^{0,*}$ such that $a = 0$ if $p > \theta_n^{0,*}$ and $a = 1$ if $p \leq \theta_n^{0,*}$. □

The optimal thresholds can be related to the optimal value function (3.1). This is done in Proposition 5, below. First, we state two simple properties of the value function.

Lemma 3.3.1. *The value function $V(c,n,p)$ is nondecreasing in n for each fixed (c,p).*

Proof. Fix c and p. If $n < n'$ then $V(c,n,p) \leq V(c,n',p)$ since it is always possible, starting from n', to pretend we are starting at $n'' = n$, use the corresponding optimal policy, and then remain forever idle ($a = 0$) when n'' reaches 0, even though $n' - n$ cycles remain in reality. □

Lemma 3.3.2. *Assuming the price is always nonnegative, the value function $V(c,n,p)$ is nondecreasing in c for each fixed (n,p). In fact, $0 \leq V(1,n,p) - V(0,n,p) \leq p$.*

Proof. Fix c, c' such that $0 = c < c' = 1$. We have $V(c',n,p) \geq V(c,n,p)$ since from c' it is always possible to pretend we are starting from c and use the corresponding policy, except that the first time the policy prescribes $a_t = 1$ assuming $c_t = 0$, which costs p_t, we use $a_t = 0$ at $c_t = 1$, which costs 0. In both cases we arrive at $c_{t+1} = 1$ and $n_{t+1} = n$. This establishes $V(1,n,p) - V(0,n,p) \geq 0$ assuming p_t is always nonnegative.

We also have

$$\begin{aligned}
p + V(0,n,p) &= p + \max\{Q(0,n,p,0), Q(0,n,p,1)\} \\
&= p + \max\{\gamma\mathbb{E}[V(0,n,p_{t+1})], -p + \gamma\mathbb{E}[V(1,n,p_{t+1})]\} \\
&= \max\{p + \gamma\mathbb{E}[V(0,n,p_{t+1})], \gamma\mathbb{E}[V(1,n,p_{t+1})]\} \\
&\geq \max\{p + \gamma\mathbb{E}[V(0,n-1,p_{t+1})], \gamma\mathbb{E}[V(1,n,p_{t+1})]\} \\
&= \max\{Q(1,n,p,-1), Q(1,n,p,0)\} = V(1,n,p),
\end{aligned}$$

where the inequality is due to Lemma 3.3.1. Thus $V(1,n,p) - V(0,n,p) \leq p$. □

Proposition 5. *There exist optimal price thresholds $\theta_n^{0,*}$, $\theta_n^{1,*}$ that can be expressed using the optimal value function* (3.1)*:*

$$\theta_n^{0,*} = \gamma(\mathbb{E}[V(1,n,p_{t+1})] - \mathbb{E}[V(0,n,p_{t+1})]), \tag{3.4}$$

$$\theta_n^{1,*} = \gamma(\mathbb{E}[V(1,n,p_{t+1})] - \mathbb{E}[V(0,n-1,p_{t+1})]). \tag{3.5}$$

In particular, $\theta_n^{1,*} - \theta_n^{0,*} = \gamma(\mathbb{E}[V(0,n,p_{t+1})] - \mathbb{E}[V(0,n-1,p_{t+1})]) \geq 0$.

Proof. Fix n. An optimal threshold $\theta_n^{0,*}$ at life n can be uniquely chosen as

$$\begin{aligned}
\theta_n^{0,*} &= \text{argmax}_p\{p : Q(0,n,p,1) \geq Q(0,n,p,0)\} \\
&= \text{argmax}_p\{p : -p + \gamma\mathbb{E}[V(1,n,p_{t+1})] \geq \gamma\mathbb{E}[V(0,n,p_{t+1})]\} \\
&= \text{argmax}_p\{p : -p + \gamma\mathbb{E}[V(1,n,p_{t+1})] = \gamma\mathbb{E}[V(0,n,p_{t+1})]\} \\
&= \gamma(\mathbb{E}[V(1,n,p_{t+1})] - \mathbb{E}[V(0,n,p_{t+1})]).
\end{aligned}$$

Similarly, an optimal threshold $\theta_n^{1,*}$ at life n can be uniquely chosen as

$$\begin{aligned}
\theta_n^{1,*} &= \text{argmin}_p\{p : Q(1,n,p,-1) \geq Q(1,n,p,0)\} \\
&= \text{argmin}_p\{p : p + \gamma\mathbb{E}[V(0,n-1,p_{t+1})] \geq \gamma\mathbb{E}[V(1,n,p_{t+1})]\} \\
&= \text{argmin}_p\{p : p + \gamma\mathbb{E}[V(0,n-1,p_{t+1})] = \gamma\mathbb{E}[V(1,n,p_{t+1})]\} \\
&= \gamma(\mathbb{E}[V(1,n,p_{t+1})] - \mathbb{E}[V(0,n-1,p_{t+1})]).
\end{aligned}$$

The sign of $(\theta_n^{1,*} - \theta_n^{0,*})$ is due to Lemma 3.3.1 applied to each $p = p_{t+1}$. □

This section concludes with a few remarks.

- Proposition 5 shows that the existence of two distinct thresholds is due to the finiteness of the battery life. Indeed, if $n \to \infty$, we have $\lim_{n\to\infty} V(0,n,p_{t+1}) = \lim_{n\to\infty} V(0,n-1,p_{t+1})$ and thus $\lim_{n\to\infty}(\theta_n^{1,*} - \theta_n^{0,*}) = 0$. Furthermore, if $n \to \infty$, the single inequality in the proof of Lemma 3.3.2 becomes an equality, leading to

$\lim_{n\to\infty}[V(0,n,p_{t+1}) - V(1,n,p_{t+1})] = p$. It then follows from (3.4) and (3.5) that

$$\lim_{n\to\infty} \theta_n^{1,*} = \lim_{n\to\infty} \theta_n^{0,*} = \gamma\mathbb{E}[p_{t+1}], \tag{3.6}$$

recovering a known result for battery operations without life limit constraints.

- Independently of the state of charge, the battery is necessarily idle when the price remains in the interval $(\theta_{n_t}^{0,*}, \theta_{n_t}^{1,*})$. This because the battery remains idle when $c_t = 0$ and the price is above $\theta_{n_t}^{0,*}$, and when $c_t = 1$ and the price is below $\theta_{n_t}^{1,*}$.

- Proposition 5 shows that $\theta_n^{0,*} \leq \theta_n^{1,*}$ for all n, and Equation (3.6) suggests that when a new storage device is put in service with n sufficiently large, the two thresholds coincide. This can suggest that as the battery ages ($n \to 0$), the spread $(\theta_n^{1,*} - \theta_n^{0,*})$ widens. The existence of an optimal policy for which the spread widens monotonically with n will be formally established by Proposition 10 in Section 3.4, and also illustrated in Section 3.7. Since the spread widens, the storage device is expected to spend more time sitting idle as it ages.

3.4 Analysis

Dynamic programming methods such as value iteration, policy iteration or linear optimization can solve the model described in Section 3.2, or at least a finite-state approximation thereof. However, we seek to avoid an explicit computation of the value function, due to the following shortages:

- If the price state variable is discretized, the determination of the optimal thresholds is inaccurate.

- Convergence may be slow, especially when the discount factor is close to 1.

- There is limited value in establishing the structure of an optimal policy if ultimately, numerical calculations are unable to exploit it.

In this section, we analyze the problem of directly finding the optimal thresholds without constructing a value function such as $V(c, n, p)$. This section establishes that we can calculate the optimal thresholds by maximizing a sequence of quasiconcave functions.

3.4.1 Threshold Policy Evaluation

In this section, we fix a threshold policy π described by some arbitrary thresholds θ_n^0 and θ_n^1 for $1 \leq n \leq N$, and we evaluate the expected value of this policy:

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} R(S_t, A^\pi(S_t)) \;\middle|\; S_0 = s \right]. \tag{3.7}$$

We assume that the prices p_t are independent and identically distributed (i.i.d.) (The price process model is extended in Section 3.8.) We assume that p_t admits a probability density function (pdf) denoted f. The corresponding cumulative distribution function (cdf) is denoted F.

At the beginning of period t, the price p_t is posted. The state (c_t, n_t, p_t) is known, and so is the action a_t since we fixed the policy. As we know a_t, at the beginning of period t we can also predict the next state of charge c_{t+1} and the next remaining life n_{t+1}. The only part of the next state that remain uncertain is the next price p_{t+1} (denoted p' in the sequel).

Let π_n be the probability that charging will occur at the next period, given that the remaining life will be n at the next period, the next state of charge will be $c = 0$, and given the price of the current period. Noting that the next price and next action are independent of the current price conditionally to knowing n and c, charging will occur if the next price p' is less than or equal to θ_n^0, thus

$$\pi_n = \text{Prob}(p' \leq \theta_n^0) = \textstyle\int_0^{\theta_n^0} f(p)dp = F(\theta_n^0)\,. \tag{3.8}$$

Let e_n^0 denote the expected price at the next period, if charging occurs at the next period, the remaining life at the next period is n, and knowing the current price. By the same

token we have

$$e_n^0 = \mathbb{E}[p' \mid p' \le \theta_n^0] = \int_0^{\theta_n^0} pf(p)dp \Big/ \pi_n. \tag{3.9}$$

Similarly, let ρ_n be the probability that discharging will occur at the next period, given that the remaining life will be n at the next period, the next state of charge will be $c = 1$, and knowing the current price. Discharging will occur if the next price p' is greater than or equal to θ_n^1, thus

$$\rho_n = \text{Prob}(p' \ge \theta_n^1) = \int_{\theta_n^1}^{\infty} f(p)dp = 1 - F(\theta_n^1). \tag{3.10}$$

Let e_n^1 denote the expected price at the next period, if discharging occurs at the next period, the remaining life at the next period is n, and knowing the current price. We have

$$e_n^1 = \mathbb{E}[p' \mid p' \ge \theta_n^1] = \int_{\theta_n^1}^{\infty} pf(p)dp \Big/ \rho_n. \tag{3.11}$$

Let $\bar{V}_n^{c,\pi}$ with $c \in \{0, 1\}$, $1 \le n \le N$, denote the expected value function at the *next* state, when the *next* state of charge is c, the *next* remaining-life state is n, the current price is known, and we follow policy π. By the same logic as before,

$$\bar{V}_n^{0,\pi} := \mathbb{E}_{p'}[V^\pi(0, n, p')], \tag{3.12}$$

$$\bar{V}_n^{1,\pi} := \mathbb{E}_{p'}[V^\pi(1, n, p')]. \tag{3.13}$$

We evaluate $\bar{V}_n^{c,\pi}$ by backward induction, as follows. If charging occurs at the next period, the expected reward is $-e_n^0$. If discharging occurs at the next period, the expected reward is e_n^1. If the battery remains idle, the reward is 0. For convenience, set $\bar{V}_0^{0,\pi} = 0$; then by backward induction, with $n = 1, \ldots, N$,

$$\bar{V}_n^{1,\pi} = (1 - \rho_n)[0 + \gamma \bar{V}_n^{1,\pi}] + \rho_n[e_n^1 + \gamma \bar{V}_{n-1}^{0,\pi}], \tag{3.14}$$

$$\bar{V}_n^{0,\pi} = (1 - \pi_n)[0 + \gamma \bar{V}_n^{0,\pi}] + \pi_n[-e_n^0 + \gamma \bar{V}_n^{1,\pi}]. \tag{3.15}$$

Equivalently,

$$\bar{V}_n^{0,\pi} = \frac{-\pi_n e_n^0}{1-\gamma(1-\pi_n)} + \frac{\gamma\pi_n \bar{V}_n^{1,\pi}}{1-\gamma(1-\pi_n)}, \tag{3.16}$$

$$\bar{V}_n^{1,\pi} = \frac{\rho_n e_n^1}{1-\gamma(1-\rho_n)} + \frac{\gamma\rho_n \bar{V}_{n-1}^{0,\pi}}{1-\gamma(1-\rho_n)}. \tag{3.17}$$

The value function (3.7) for policy π can then be obtained from the values $\bar{V}_n^{c,\pi}$, noting that the value of being in state $s = (c, n, p)$ and following policy π is

$$V^\pi(s) = \begin{cases} -p + \gamma \bar{V}_n^{1,\pi} & \text{if } c = 0 \text{ and } p \le \theta_n^0, \\ \gamma \bar{V}_n^0 & \text{if } c = 0 \text{ and } p > \theta_n^0, \\ p + \gamma \bar{V}_{n-1}^{0,\pi} & \text{if } c = 1 \text{ and } p \ge \theta_n^1, \\ \gamma \bar{V}_n^{1,\pi} & \text{if } c = 1 \text{ and } p < \theta_n^1. \end{cases} \tag{3.18}$$

We stress that the entities intervening in (3.8–3.18), namely π_n, ρ_n, e_n^0, e_n^1, $\bar{V}_n^{0,\pi}$, $\bar{V}_n^{1,\pi}$, all depend on the policy π under evaluation, which is parameterized by the thresholds $\theta = \{\theta_n^0, \theta_n^1\}_{1 \le n \le N}$. That is, these entities depend on θ.

3.4.2 Threshold Policy Optimization

Consider $\bar{V}_n^{1,\pi}$ in (3.17). Observe that ρ_n and e_n^1 only depend on θ_n^1, while $\bar{V}_{n-1}^{0,\pi}$ does not depend on θ_n^1, and actually only depends on θ_k^1 and θ_k^0 for $k = 1, \ldots, n-1$. Also observe that since $\gamma\rho_n/(1-\gamma(1-\rho_n))$ is nonnegative by virtue of $0 < \gamma < 1$ and $0 \le \rho_n \le 1$, we can maximize $\bar{V}_n^{1,\pi}$ by first having $\bar{V}_{n-1}^{0,\pi}$ maximized, and then maximizing $\bar{V}_n^{1,\pi}$ over θ_n^1 only.

The same reasoning applies to $\bar{V}_n^{0,\pi}$ in (3.16): we can maximize $\bar{V}_n^{0,\pi}$ by first having $\bar{V}_n^{1,\pi}$ maximized, and then maximizing $\bar{V}_n^{0,\pi}$ over θ_n^0 only, since π_n and e_n^0 only depend on θ_n^0, while $\bar{V}_n^{1,\pi}$ only depends on θ_n^1 and on θ_k^0 and θ_k^1 for $k = 1, \ldots, n-1$, and since $\gamma\pi_n/(1-\gamma(1-\pi_n))$ is nonnegative.

Therefore we essentially have $\bar{V}_n^{0,\pi} := \bar{V}^0(\theta_n^0, \bar{V}_n^{1,\pi})$, $\bar{V}_n^{1,\pi} := \bar{V}^1(\theta_n^1, \bar{V}_{n-1}^{0,\pi})$, and we can determine optimal threshold parameters $\theta_n^{0,*}, \theta_n^{1,*}$ by setting $\bar{V}_0^{0,*} = 0$ for convenience and

solving the nested scalar maximization problems for $n = 1, \ldots, N$,

$$\begin{aligned}
\bar{V}_n^{1,*} &= \max_{\theta_n^1} \bar{V}^1(\theta_n^1, \bar{V}_{n-1}^{0,*}) = \max_{\theta_n^1} \frac{\rho_n e_n^1 + \gamma \rho_n \bar{V}_{n-1}^{0,*}}{1 - \gamma(1 - \rho_n)} \\
&= \max_{\theta_n^1} \frac{\int_{\theta_n^1}^{\infty} p f(p) dp + \gamma(1 - F(\theta_n^1)) \bar{V}_{n-1}^{0,*}}{1 - \gamma F(\theta_n^1)}, \qquad (3.19) \\
\bar{V}_n^{0,*} &= \max_{\theta_n^0} \bar{V}^0(\theta_n^0, \bar{V}_n^{1,*}) = \max_{\theta_n^0} \frac{-\pi_n e_n^0 + \gamma \pi_n \bar{V}_n^{1,*}}{1 - \gamma(1 - \pi_n)} \\
&= \max_{\theta_n^0} \frac{-\int_0^{\theta_n^0} p f(p) dp + \gamma F(\theta_n^0) \bar{V}_n^{1,*}}{1 - \gamma(1 - F(\theta_n^0))}. \qquad (3.20)
\end{aligned}$$

Hence we have a chain of optimization problems to solve in the following order:

$$0 \overset{\theta_1^{1,*}}{\longmapsto} \bar{V}_1^{1,*} \overset{\theta_1^{0,*}}{\longmapsto} \bar{V}_1^{0,*} \overset{\theta_2^{1,*}}{\longmapsto} \bar{V}_2^{1,*} \overset{\theta_2^{0,*}}{\longmapsto} \bar{V}_2^{0,*} \ldots \overset{\theta_N^{1,*}}{\longmapsto} \bar{V}_N^{1,*} \overset{\theta_N^{0,*}}{\longmapsto} \bar{V}_N^{0,*}.$$

The following lemma is used in the proof of Proposition 6 below.

Lemma 3.4.1. *$\bar{V}_n^{0,\pi}$ is upper-bounded by*

$$\bar{V}_\infty^{0,*} = \frac{1}{1-\gamma} \left[\gamma\mu F(\gamma\mu) - \int_0^{\gamma\mu} p f(p) dp \right] \text{ where } \mu := \mathbb{E}[p_{t+1}].$$

In particular, this implies the inequality

$$(1 - \gamma)\bar{V}_n^{0,\pi} \leq \gamma\mu. \qquad (3.21)$$

Proof of Lemma 3.4.1. By definition, $\bar{V}_n^{0,\pi} \leq \bar{V}_n^{0,*}$. From Lemma 3.3.1, we know that $V(0, n, p) \leq V(0, n+1, p)$ for each p. In particular, $V(0, n, p) \leq \lim_{n' \to \infty} V(0, n', p)$. Therefore, by the monotone convergence theorem,

$$\bar{V}_n^{0,*} = \mathbb{E}[V(0, n, p_{t+1})] \leq \mathbb{E}[\lim_{n' \to \infty} [V(0, n', p_{t+1})]] = \lim_{n' \to \infty} \bar{V}_{n'}^{0,*} := \bar{V}_\infty^{0,*}.$$

By (3.6), $\bar{V}_\infty^{0,*}$ is attained at $\theta = \lim_{n \to \infty} \theta_n^0 = \lim_{n \to \infty} \theta_n^1 = \gamma\mu$, where $\mu = \mathbb{E}[p_{t+1}]$.

Define

$$\rho = \lim_{n\to\infty} \rho_n = 1 - F(\gamma\mu), \qquad \pi = \lim_{n\to\infty} \pi_n = F(\gamma\mu),$$
$$e^0 = \lim_{n\to\infty} e_n^0 = \int_0^{\gamma\mu} pf(p)dp/\pi, \qquad e^1 = \lim_{n\to\infty} e_n^1 = \int_{\gamma\mu}^{\infty} pf(p)dp/\rho.$$

Note the relations $\rho = 1-\pi$ and $\pi e^0 + \rho e^1 = \mu$, which do not hold for finite n. Substituting (3.17) into (3.16) with $n \to \infty$, we get

$$\bar{V}_\infty^{0,*} = \frac{-\pi e^0}{1-\gamma(1-\pi)} + \frac{\gamma\pi}{1-\gamma(1-\pi)}\left[\frac{\mu - \pi e^0}{1-\gamma\pi} + \frac{\gamma(1-\pi)\bar{V}_\infty^{0,*}}{1-\gamma\pi}\right].$$

After some simple manipulations the solution reduces to

$$\bar{V}_\infty^{0,*} = \frac{\pi(\gamma\mu - e^0)}{1-\gamma} = \frac{\gamma\mu F(\gamma\mu) - \int_0^{\gamma\mu} pf(p)fp}{1-\gamma}.$$

(3.21) is due to $\bar{V}_n^{0,*} \le \bar{V}_\infty^{0,*}$, $\gamma\mu \ge 0$, $F(\gamma\mu) \le 1$, and $\int_0^{\gamma\mu} pf(p)fp \ge 0$. □

Proposition 6. *The function $\bar{V}^0(\theta_n^0, \bar{V}_n^{1,*})$ is quasiconcave in the threshold θ_n^0. Furthermore, if the price distribution is supported on $(0, +\infty)$, then $\bar{V}^0(\theta_n^0, \bar{V}_n^{1,*})$ is strictly quasiconcave in θ_n^0.*

Similarly, the function $\bar{V}^1(\theta_n^1, \bar{V}_{n-1}^{0,*})$ is quasiconcave in θ_n^1, and strictly quasiconcave in θ_n^1 if the price distribution is supported on $(0, +\infty)$.

Proof. The partial derivative of $\bar{V}^0$ with respect to θ_n^0 is

$$\frac{\partial \bar{V}^0}{\partial \theta_n^0} = \frac{f(\theta_n^0) g(\theta_n^0, \bar{V}_n^{1,*})}{[1-\gamma(1-F(\theta_n^0))]^2},$$

where we have defined

$$g(\theta_n^0, \bar{V}_n^{1,*}) := \gamma(1-\gamma)\bar{V}_n^{1,*} - (1-\gamma)\theta_n^0 - \gamma\theta_n^0 F(\theta_n^0) + \gamma\int_0^{\theta_n^0} pf(p)dp. \tag{3.22}$$

If we can show that there is a point θ_n^0 such that $\frac{\partial \bar{V}^0}{\partial \theta_n^0} \ge 0$ on $(0, \theta_n^0)$ and $\frac{\partial \bar{V}^0}{\partial \theta_n^0} \le 0$ on (θ_n^0, ∞), then we will have showed that the function $\bar{V}^0$ is quasiconcave in θ_n^0, see e.g. [9].

If the inequalities are strict, we will have showed that the function is strictly quasiconcave.

Clearly, the denominator is positive: $[1-\gamma(1-F(\theta_n^0))]^2 > 0$, since $0 < \gamma < 1$. Clearly, $f(\theta_n^0) \geq 0$, and if the price is supported on $(0, +\infty)$, then $f(\theta_n^0) > 0$ on that interval. Therefore we are left with studying the sign of g.

We have

$$g(0, \bar{V}_n^{1,*}) = \gamma(1-\gamma)\bar{V}_n^{1,*} \geq 0,$$

assuming $\bar{V}_n^{1,*} \geq 0$ since the always-idle policy attains the zero expected value.

The partial derivative of g with respect to θ_n^0 is

$$\begin{aligned}\frac{\partial g}{\partial \theta_n^0} &= -(1-\gamma) - \gamma F(\theta_n^0) - \gamma\theta_n^0 f(\theta_n^0) + \gamma\theta_n^0 f(\theta_n^0) \\ &= -1 + \gamma(1 - F(\theta_n^0)) < 0.\end{aligned}$$

Therefore $g(\theta_n^0, \bar{V}_n^{1,*})$ is decreasing in θ_n^0, and considering the sign of g at 0, we can conclude that there exist a critical θ_n^0 such that $g(\theta_n^0, \bar{V}_n^{1,*}) > 0$ on $(0, \theta_n^0)$ and $g(\theta_n^0) < 0$ on (θ_n^0, ∞). It follows that $\bar{V}^0$ is quasiconcave in θ_n^0, and strictly quasiconcave provided $f(\theta_n^0) > 0$ on $(0, +\infty)$.

The proof is similar for $\bar{V}^1$, except that it requires Lemma 3.4.1. The proof of the quasiconcavity $\bar{V}^1$ in θ_n^1 relies on the following expressions:

$$\begin{aligned}&\frac{\partial \bar{V}^1}{\partial \theta_n^1} = \frac{f(\theta_n^1) h(\theta_n^1, \bar{V}_{n-1}^{0,*})}{(1-\gamma F(\theta_n^1))^2}, \\ \text{where} \quad &h(\theta_n^1, \bar{V}_{n-1}^{0,*}) := -\theta_n^1 - \gamma(1-\gamma)\bar{V}_{n-1}^{0,*} + \theta_n^1\gamma F(\theta_n^1) + \gamma\int_{\theta_n^1}^{\infty} p f(p) dp, \qquad (3.23) \\ &h(0, \bar{V}_{n-1}^{0,*}) = -\gamma(1-\gamma)\bar{V}_{n-1}^{0,*} + \gamma\mathbb{E}[p_{t+1}], \qquad \frac{\partial h}{\partial \theta_n^1} = -1 + \gamma F(\theta_n^1).\end{aligned}$$

Equation (3.21) in Lemma 3.4.1 is used to assert that $h(0, \bar{V}_{n-1}^{0,*}) \geq 0$. Therefore, since $\partial h / \partial \theta_n^1 < -(1-\gamma) < 0$, there exists a critical θ_n^1 such that $h > 0$ on $(0, \theta_n^1)$ and $h < 0$ on (θ_n^1, ∞). □

First-order optimality conditions for quasiconvex optimization problems are well known but rely on a generalization of the notion of subdifferential [32, 47]. If the problem is strictly

quasiconvex, then the usual subdifferential or gradient is sufficient to express first-order optimality conditions.

The proof of Proposition 6 shows that the objective is not strictly quasiconcave due to the possibility of having regions where the probability density function of the price is zero. This leads to the following result.

Proposition 7. *Sufficient optimality conditions for the thresholds θ_n^0, θ_n^1 are given by the implicit equations*

$$g(\theta_n^{0,*}, \bar{V}_n^{1,*}) = 0, \tag{3.24}$$

$$h(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) = 0, \tag{3.25}$$

with the functions g and h defined by (3.22) *and* (3.23)*. The solutions to* (3.24) *and* (3.25) *are unique.*

Proof. Without loss of optimality, we can always restrict the thresholds to lie on the set where the density of the price is nonzero. Therefore, for maximizing $\bar{V}^0$ and $\bar{V}^1$, we can replace the necessary conditions $\partial \bar{V}^0 / \partial \theta_n^0 = 0$ and $\partial \bar{V}^1 / \partial \theta_n^1 = 0$ by the sufficient conditions (3.24), (3.25).

The uniqueness of the solution to each equation follows from the implicit function theorem [25] applied separately to each equation, where

$$\frac{\partial g}{\partial \theta_n^0} = -1 + \gamma(1 - F(\theta_n^0)) < 0, \qquad \frac{\partial h}{\partial \theta_n^1} = -1 + \gamma F(\theta_n^1) < 0,$$

showing that the partial derivatives are invertible. □

Proposition 8. *Let the thresholds $\theta_n^{0,*}, \theta_n^{1,*}$ be determined by* (3.24) *and* (3.25)*. Then we have*

$$\bar{V}_n^{1,*} = \bar{V}_{n-1}^{0,*} + \theta_n^{1,*}/\gamma, \qquad \bar{V}_n^{0,*} = \bar{V}_n^{1,*} - \theta_n^{0,*}/\gamma.$$

Table 3.1: Algorithm for optimizing the thresholds.

Let $\bar{V}_0^{0,*} = 0$. *Then for* $n = 1, \ldots, N$:

1. *Solve* $h(\theta_n^1, \bar{V}_{n-1}^{0,*}) = 0$ *given by* (3.23) *to find* $\theta_n^{1,*}$.

2. *Compute* $\bar{V}_n^{1,*} = \bar{V}_{n-1}^{0,*} + \theta_n^{1,*}/\gamma$.

3. *Solve* $g(\theta_n^0, \bar{V}_n^{1,*}) = 0$ *given by* (3.22) *to find* $\theta_n^{0,*}$.

4. *Compute* $\bar{V}_n^{0,*} = \bar{V}_n^{1,*} - \theta_n^{0,*}/\gamma$.

Proof. From (3.23), $h(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) = 0$ implies in particular

$$\int_{\theta_n^{1,*}}^{\infty} pf(p)dp = \theta_n^{1,*}/\gamma + (1-\gamma)\bar{V}_{n-1}^{0,*} - \theta_n^{1,*}F(\theta_n^{1,*}).$$

This expression is then substituted into the objective in (3.19).

From (3.22), $g(\theta_n^{0,*}, \bar{V}_n^{1,*}) = 0$ implies in particular

$$\int_0^{\theta_n^{0,*}} pf(p)dp = (1-\gamma)(\theta_n^{0,*}/\gamma - \bar{V}_n^{1,*}) + \theta_n^{0,*}F(\theta_n^{0,*}).$$

This expression is then substituted into the objective in (3.20).

Alternatively, we use (3.4) and (3.5) to directly show the result. □

Our algorithm for determining optimal thresholds is based on Propositions 7 and 8. It is summarized in Table 3.1.

Proposition 9. *Suppose* $g(\theta_n^{0,*}, \bar{V}_n^{1,*}) = 0$ *and* $h(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) = 0$. *Then the solution* θ^0 *to* $g(\theta^0, \tilde{V}^1) = 0$ *for* $\tilde{V}^1$ *in a neighborhood of* $\bar{V}_n^{1,*}$, *and the solution* θ^1 *to* $h(\theta^1, \tilde{V}^0) = 0$ *for* $\tilde{V}^0$ *in a neighborhood of* $\bar{V}_{n-1}^{0,*}$, *admit first-order expansions described as*

$$\theta^0 = \theta_n^{0,*} + \frac{(1-\gamma)\gamma}{1-\gamma(1-F(\theta_n^{0,*}))}(\tilde{V}^1 - \bar{V}_n^{1,*}) + o(|\tilde{V}^1 - \bar{V}_n^{1,*}|),$$
$$\theta^1 = \theta_n^{1,*} - \frac{(1-\gamma)\gamma}{1-\gamma F(\theta_n^{1,*})}(\tilde{V}^0 - \bar{V}_{n-1}^{0,*}) + o(|\tilde{V}^0 - \bar{V}_{n-1}^{0,*}|).$$

Proof of Proposition 9. This follows from implicit differentiation applied to the implicit equation $g(\theta_n^0, \bar{V}_n^1) = 0$ in a neighborhood of the solution $(\theta_n^{0,*}, \bar{V}_n^{1,*})$ and from implicit differentiation of $h(\theta_n^1, \bar{V}_{n-1}^0) = 0$ in a neighborhood of the solution $(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*})$, noting that $\partial g/\partial\theta_n^0 < 0$, $\partial h/\partial\theta_n^1 < 0$, and $\partial g/\partial\bar{V}_n^1 = \gamma(1-\gamma) = -\partial h/\partial\bar{V}_{n-1}^0$. □

Proposition 10. *The calculated optimal thresholds are such that $\theta_n^{0,*}$ is nondecreasing in n and $\theta_n^{1,*}$ is nonincreasing in n.*

Proof of Proposition 10. From Lemma 3.3.1 at each $p = p_{t+1}$ we have $\bar{V}_n^{1,*} - \bar{V}_{n-1}^{1,*} = \mathbb{E}[V(1,n,p_{t+1}) - V(1,n-1,p_{t+1})] \geq 0$ and $\bar{V}_n^{0,*} - \bar{V}_{n-1}^{0,*} = \mathbb{E}[V(0,n,p_{t+1}) - V(0,n-1,p_{t+1})] \geq 0$, that is, $\bar{V}_n^{1,*} - \bar{V}_{n-1}^{1,*} \geq 0$ and $\bar{V}_n^{0,*} - \bar{V}_{n-1}^{0,*} \geq 0$.

The threshold $\theta_n^{0,*}$ is defined by the implicit equation $g(\theta_n^{0,*}, \bar{V}_n^{1,*}) = 0$. From the proof of Proposition 6, g is decreasing in θ_n^0. It can be seen from (3.22) that g is increasing in $\bar{V}_n^{1,*}$. Therefore, since the threshold $\theta_{n-1}^{0,*}$ is described by the implicit equation $g(\theta_{n-1}^{0,*}, \bar{V}_{n-1}^{1,*}) = 0$, it follows that $\bar{V}_{n-1}^{1,*} \leq \bar{V}_n^{1,*}$ implies $\theta_{n-1}^{0,*} \leq \theta_n^{0,*}$.

Similarly, $\theta_n^{1,*}$ is defined by $h(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) = 0$. From the proof of Proposition 6, h is decreasing in θ_n^1. It can be seen from (3.23) that h is decreasing in $\bar{V}_{n-1}^{0,*}$. Therefore, since the threshold $\theta_{n+1}^{1,*}$ is defined by $g(\theta_{n+1}^{1,*}, \bar{V}_n^{0,*}) = 0$, it follows that $\bar{V}_n^{0,*} \geq \bar{V}_{n-1}^{0,*}$ implies $\theta_{n+1}^{1,*} \leq \theta_n^{1,*}$. □

Recall from Proposition 5 that $\theta_n^{1,*} - \theta_n^{0,*} \geq 0$, and from (3.6) that $\lim_{n\to\infty}(\theta_n^{1,*} - \theta_n^{0,*}) = 0$. Proposition 10 allows us to conclude that as n increases, $\theta_n^{1,*} - \theta_n^{0,*}$ monotonically decreases to 0.

It also follows from Proposition 10 that $\bar{V}_{n-1}^{1,*} - \bar{V}_{n-1}^{0,*} \leq \bar{V}_n^{1,*} - \bar{V}_n^{0,*}$, given (3.4) combined with $\theta_{n-1}^{0,*} \leq \theta_n^{0,*}$.

3.5 Error Analysis

This section evaluates the propagation of errors when the steps of the algorithm summarized in Table 3.1 are not carried out exactly.

Given $\tilde{V}_{n-1}^0$, let

$$\tilde{V}_n^{1,*} = \max_{\theta_n^1} \bar{V}_n^1(\theta_n^1, \tilde{V}_{n-1}^0), \qquad \tilde{\theta}_n^{1,*} = \operatorname*{argmax}_{\theta_n^1} \bar{V}_n^1(\theta_n^1, \tilde{V}_{n-1}^0)$$

be the exact optimal value to the perturbed optimization problem obtained by replacing $\bar{V}_{n-1}^{0,*}$ by $\tilde{V}_{n-1}^0$ in (3.19), and the corresponding optimal solution uniquely defined by $g(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) = 0$. In particular we have $\bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) = \tilde{V}_n^{1,*}$.

It is assumed that a near-optimal value $\tilde{V}_n^1$ can be found for the perturbed optimization problem, in the sense that

$$\tilde{V}_n^1 + \epsilon \geq \tilde{V}_n^{1,*} \quad \text{for some } \epsilon > 0. \tag{3.26}$$

We denote by $\tilde{\theta}_n^1$ the near-optimal solution that attains $\tilde{V}_n^1 = \bar{V}^1(\tilde{\theta}_n^1, \tilde{V}_{n-1}^0)$.

Similarly, let

$$\tilde{V}_n^{0,*} = \max_{\theta_n^0} \bar{V}_n^0(\theta_n^0, \tilde{V}_n^1), \qquad \tilde{\theta}_n^{0,*} = \operatorname*{argmax}_{\theta_n^0} \bar{V}_n^0(\theta_n^0, \tilde{V}_n^1)$$

be the exact optimal value to the perturbed optimization problem obtained by replacing $\bar{V}_n^{1,*}$ by $\tilde{V}_n^1$ in (3.20), and the corresponding optimal solution uniquely defined by $h(\tilde{\theta}_n^{0,*}, \tilde{V}_n^1) = 0$. In particular we have $\bar{V}^0(\tilde{\theta}_n^{0,*}, \tilde{V}_n^1) = \tilde{V}_n^{0,*}$.

It is assumed that a near-optimal value $\tilde{V}_n^0$ can be found for the perturbed optimization problem, in the sense that

$$\tilde{V}_n^0 + \epsilon \geq \tilde{V}_n^{0,*} \quad \text{for some } \epsilon > 0. \tag{3.27}$$

We denote by $\tilde{\theta}_n^0$ the near-optimal solution that attains $\tilde{V}_n^0 = \bar{V}^0(\tilde{\theta}_n^0, \tilde{V}_n^1)$.

Lemma 3.5.1. *The function* $c_\gamma(x) := \frac{x}{1-\gamma(1-x)}$ *defined over* $x \in [0,1]$ *satisfies* $0 \leq c_\gamma(x) \leq 1$.

Proof. Assume $x \in [0,1]$ and $0 < \gamma < 1$. The function c_γ is increasing since $dc_\gamma/dx = (1-\gamma)/[1-\gamma(1-x)]^2 > 0$, thus its minimum and maximum are attained at $x = 0$ and

$x = 1$ respectively. $\square$

Lemma 3.5.2. *Suppose* $|\tilde{V}_{n-1}^0 - \bar{V}_{n-1}^{0,*}| \leq \delta_n$. *Then we have* $|\tilde{V}_n^1 - \bar{V}_n^{1,*}| \leq \gamma\delta_n + \epsilon$.

Proof. By definition of $\tilde{V}_n^1$ and $\tilde{V}_n^{1,*}$, we have

$$0 \leq \tilde{V}_n^{1,*} - \tilde{V}_n^1 = \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) - \bar{V}^1(\tilde{\theta}_n^1, \tilde{V}_{n-1}^0) \leq \epsilon. \tag{3.28}$$

By definition of $\tilde{V}_n^{1,*}$, for any θ_n^1 we have

$$\tilde{V}_n^{1,*} = \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) \geq \bar{V}^1(\theta_n^1, \tilde{V}_{n-1}^0). \tag{3.29}$$

First, consider the case where $\bar{V}_{n-1}^{0,*} \geq \tilde{V}_{n-1}^0$ and thus $\bar{V}_{n-1}^{0,*} - \tilde{V}_{n-1}^0 \leq \delta_n$. We will establish that $0 \leq \bar{V}_n^{1,*} - \tilde{V}_n^1 \leq \gamma\delta_n + \epsilon$.

From (3.16), for any fixed θ_n^1 we have

$$\bar{V}^1(\theta_n^1, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\theta_n^1, \tilde{V}_{n-1}^0) = \frac{\gamma\rho_n}{1 - \gamma(1 - \rho_n)}(\bar{V}_{n-1}^{0,*} - \tilde{V}_{n-1}^0).$$

Lemma 3.5.1 with $x = \rho_n \in [0, 1]$ then implies

$$0 \leq \bar{V}^1(\theta_n^1, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\theta_n^1, \tilde{V}_{n-1}^0) \leq \gamma(\bar{V}_{n-1}^{0,*} - \tilde{V}_{n-1}^0). \tag{3.30}$$

We then proceed with evaluating

$$\begin{aligned}
&\bar{V}_n^{1,*} - \tilde{V}_n^1 \\
&= \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\tilde{\theta}_n^1, \tilde{V}_{n-1}^0) \\
&\leq \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) + \epsilon \quad \text{by rightmost inequality of (3.28)} \\
&\leq \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\theta_n^{1,*}, \tilde{V}_{n-1}^0) + \epsilon \quad \text{by (3.29) with } \theta_n^1 = \theta_n^{1,*} \\
&\leq \gamma(\bar{V}_{n-1}^{0,*} - \tilde{V}_{n-1}^0) + \epsilon \quad \text{by rightmost inequality of (3.30) with } \theta_n^1 = \theta_n^{1,*} \\
&\leq \gamma\delta_n + \epsilon.
\end{aligned}$$

We also have

$$\begin{aligned}
\bar{V}_n^{1,*} - \tilde{V}_n^1 &= \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\tilde{\theta}_n^1, \tilde{V}_{n-1}^0) \\
&\geq \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) \quad \text{by leftmost inequality of (3.28)} \\
&\geq \bar{V}^1(\tilde{\theta}_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) \quad \text{by definition of } \bar{V}_n^{1,*} \\
&\geq 0 \quad \text{by (3.30) with } \theta_n^1 = \tilde{\theta}_n^{1,*}.
\end{aligned}$$

Second, consider the case where $\bar{V}_{n-1}^{0,*} \leq \tilde{V}_{n-1}^0$ and thus $\tilde{V}_{n-1}^0 - \bar{V}_{n-1}^{0,*} \leq \delta_n$. We will establish that $\epsilon \leq \tilde{V}_n^1 - \bar{V}_n^{1,*} \leq \gamma\delta_n$.

From (3.16) and Lemma 3.5.1 we have, for any θ_n^1,

$$0 \leq \bar{V}^1(\theta_n^1, \tilde{V}_{n-1}^0) - \bar{V}^1(\theta_n^1, \bar{V}_{n-1}^{0,*}) \leq \gamma(\tilde{V}_{n-1}^0 - \bar{V}_{n-1}^{0,*}). \tag{3.31}$$

We then proceed with evaluating

$$\begin{aligned}
\tilde{V}_n^1 - \bar{V}_n^{1,*} &= \bar{V}^1(\tilde{\theta}_n^1, \tilde{V}_{n-1}^0) - \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) \\
&\leq \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) - \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) \quad \text{by (3.29)} \\
&\leq \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) - \bar{V}^1(\tilde{\theta}_n^{1,*}, \bar{V}_{n-1}^{0,*}) \quad \text{by definition of } \bar{V}_n^{1,*} \\
&\leq \gamma(\tilde{V}_{n-1}^0 - \bar{V}_{n-1}^{0,*}) \quad \text{by rightmost inequality of (3.31) with } \theta_n^1 = \tilde{\theta}_n^{1,*} \\
&\leq \gamma\delta_n.
\end{aligned}$$

We also have

$$\begin{aligned}
\tilde{V}_n^1 - \bar{V}_n^{1,*} &= \bar{V}^1(\tilde{\theta}_n^1, \tilde{V}_{n-1}^0) - \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) \\
&\geq \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) - \epsilon - \bar{V}^1(\theta_n^{1,*}, \bar{V}_{n-1}^{0,*}) \quad \text{by (3.28)} \\
&\geq \bar{V}^1(\tilde{\theta}_n^{1,*}, \tilde{V}_{n-1}^0) - \bar{V}^1(\tilde{\theta}_n^{1,*}, \bar{V}_{n-1}^{0,*}) - \epsilon \quad \text{by definition of } \bar{V}_n^{1,*} \\
&\geq -\epsilon \quad \text{by leftmost inequality of (3.31) with } \theta_n^1 = \tilde{\theta}_n^{1,*}.
\end{aligned}$$

From the two cases, it follows that the inequality of the proposition is verified. □

Lemma 3.5.3. *Suppose* $|\tilde{V}_n^1 - \bar{V}_n^{1,*}| \leq \delta'_n$. *Then we have* $|\tilde{V}_n^0 - \bar{V}_n^{0,*}| \leq \gamma\delta'_n + \epsilon$.

The proof of Lemma 3.5.3 is almost identical to the proof of Lemma 3.5.2 and is thus omitted.

The results of Lemmas 3.5.2 and 3.5.3 remain valid under the assumption that $|\tilde{V}_n^1 - \tilde{V}_n^{1,*}| \leq \epsilon$ and $|\tilde{V}_n^0 - \tilde{V}_n^{0,*}| \leq \epsilon$. Those inequalities are slightly weaker than (3.26) and (3.27) and correspond to the case where the optimal value is not estimated precisely. The corresponding proof works by replacing 0 by $-\epsilon$ in the lefthand side of (3.28).

Proposition 11. *Let the algorithm in Table 3.1 be implemented, except that* $\bar{V}_{n-1}^{0,*}$ *is replaced by* $\tilde{V}_{n-1}^0$ *following* (3.27)*, and* $\bar{V}_n^{1,*}$ *is replaced by* $\tilde{V}_n^1$ *following* (3.26)*. We initialize with* $\tilde{V}_0^0 = 0$*. Then, the propagation of errors among iterations is such that*

$$|\bar{V}_n^{1,*} - \tilde{V}_n^1| \leq \frac{1-\gamma^{2n-1}}{1-\gamma}\epsilon \quad \text{and} \quad |\bar{V}_n^{0,*} - \tilde{V}_n^0| \leq \frac{1-\gamma^{2n}}{1-\gamma}\epsilon,$$

guaranteeing that $|\bar{V}_n^{c,*} - \tilde{V}_n^c| \leq \epsilon/(1-\gamma)$ *for all* n *and for* $c = 0, 1$.

Proof. Using Lemma 3.5.2 and Lemma 3.5.3 alternatively, starting from $V_0^{0,*} = \tilde{V}_0^0 = 0$, we obtain

$$\begin{aligned}
|\bar{V}_1^{1,*} - \tilde{V}_1^1| &\leq \gamma \cdot 0 + \epsilon = \epsilon, \\
|\bar{V}_1^{0,*} - \tilde{V}_1^0| &\leq \gamma\epsilon + \epsilon, \\
|\bar{V}_2^{1,*} - \tilde{V}_2^1| &\leq \gamma(\gamma\epsilon + \epsilon) + \epsilon = \gamma^2\epsilon + \gamma\epsilon + \epsilon, \\
|\bar{V}_2^{0,*} - \tilde{V}_2^0| &\leq \gamma^3\epsilon + \gamma^2\epsilon + \gamma\epsilon + \epsilon,
\end{aligned}$$

and thus in general for $n \geq 1$,

$$|\bar{V}_n^{1,*} - \tilde{V}_n^1| \leq \sum_{k=0}^{2n-2} \gamma^k\epsilon = \frac{1-\gamma^{2n-1}}{1-\gamma}\epsilon, \qquad |\bar{V}_n^{0,*} - \tilde{V}_n^0| \leq \sum_{k=0}^{2n-1} \gamma^k\epsilon = \frac{1-\gamma^{2n}}{1-\gamma}\epsilon.$$

□

Table 3.2: Algorithm for optimizing the price thresholds under battery capacity deterioration and inefficiencies.

Set $(C_0/C_1)\bar{V}_0^{0,*} = 0$ *for convenience. Then for* $n = 1, \ldots, N$*:*

1. *Solve* $h(\theta_n^{1,*}, (\eta_n^{dis})^{-1}(C_{n-1}/C_n)\bar{V}_{n-1}^{0,*}) = 0$ *with* h *given by* (3.23) *to find* $\theta_n^{1,*}$.

2. *Compute* $\bar{V}_n^{1,*} = (C_{n-1}/C_n)\bar{V}_{n-1}^{0,*} + (\eta_n^{dis}/\gamma)\theta_n^{1,*}$.

3. *Solve* $g(\theta_n^{0,*}, \eta_n^{ch}\bar{V}_n^{1,*}) = 0$ *with* g *given by* (3.22) *to find* $\theta_n^{0,*}$.

4. *Compute* $\bar{V}_n^{0,*} = \bar{V}_n^{1,*} - \theta_n^{0,*}/(\gamma\eta_n^{ch})$.

3.6 Extensions of the Storage Device Model

This section considers modifications of the storage device model, so as to represent other non-ideal characteristics of battery storage. It shows how the results of the previous sections can be extended to solve the modified problem.

First, we now recognize that charging cycles adversely affect storage capacity. The battery model is completed by a *capacity function* C that describes the storage capacity as a function of the remaining cycles:

$$C_n = C(n). \tag{3.32}$$

We assume C is nondecreasing in n, and nonnegative.

Second, we now recognize that energy losses occur during the charging and discharging operations. We suppose that the cost of charging a unit-capacity battery is $(\eta_n^{\mathrm{ch}})^{-1}p_t > p_t$, and the reward of discharging a unit-capacity battery is $\eta_n^{\mathrm{dis}}p_t < p_t$, where $\eta_n^{\mathrm{ch}}, \eta_n^{\mathrm{dis}} \in (0, 1]$ are the charging and discharging efficiency coefficients, possibly dependent on the remaining life n.

Proposition 12. *Suppose the storage device has non-ideal characteristics: a deteriorating storage capacity* $C(n)$*, and charging and discharging efficiencies* $\eta_n^{ch}, \eta_n^{dis} \in (0, 1]$*. Then, optimal thresholds can be computed by the algorithm in Table 3.2.*

Proof. Observe that if the capacity of the battery is a constant β, the reward function defined for a unit capacity battery is scaled by β, and so are the value functions. Furthermore, the threshold policy for the unit-capacity device is still optimal for the β-capacity device.

Suppose we keep the convention that $\bar{V}_n^1$ and $\bar{V}_n^0$ are defined for unit-capacity storage devices. Then, (3.16) and (3.17) become

$$C_n \bar{V}_n^{0,\pi} = \frac{-(\eta_n^{\text{ch}})^{-1} C_n \pi_n e_n^0}{1-\gamma(1-\pi_n)} + \frac{\gamma \pi_n C_n \bar{V}_n^{1,\pi}}{1-\gamma(1-\pi_n)},$$
$$C_n \bar{V}_n^{1,\pi} = \frac{\eta_n^{\text{dis}} C_n \rho_n e_n^1}{1-\gamma(1-\rho_n)} + \frac{\gamma \rho_n C_{n-1} \bar{V}_{n-1}^{0,\pi}}{1-\gamma(1-\rho_n)}.$$

We get the optimal threshold $\theta_n^{1,*}$ by maximizing the scaled objective

$$\bar{V}_n^1/\eta_n^{\text{dis}} = \frac{\rho_n e_n^1}{1-\gamma(1-\rho_n)} + \frac{\gamma \rho_n (\eta_n^{\text{dis}})^{-1} (C_{n-1}/C_n) \bar{V}_{n-1}^{0,*}}{1-\gamma(1-\rho_n)},$$

that is, by solving $h(\theta_n^{1,*}, (\eta_n^{\text{dis}})^{-1}(C_{n-1}/C_n)\bar{V}_{n-1}^{0,*}) = 0$. The optimal $\bar{V}_n^{1,*}$ is then described as

$$\bar{V}_n^{1,*} = (C_{n-1}/C_n)\bar{V}_{n-1}^{0,*} + (\eta_n^{\text{dis}}/\gamma)\theta_n^{1,*}.$$

Similarly, we get the optimal threshold $\theta_n^{0,*}$ by maximizing the scaled objective

$$\eta_n^{\text{ch}} \bar{V}_n^0 = \frac{-\pi_n e_n^0}{1-\gamma(1-\pi_n)} + \frac{\gamma \pi_n \eta_n^{\text{ch}} \bar{V}_n^{1,*}}{1-\gamma(1-\pi_n)},$$

that is, by solving $g(\theta_n^{0,*}, \eta_n^{\text{ch}} \bar{V}_n^{1,*}) = 0$. The optimal $\bar{V}_n^{0,*}$ is then described as

$$\bar{V}_n^{0,*} = \bar{V}_n^{1,*} - (\gamma \eta_n^{\text{ch}})^{-1} \theta_n^{0,*}.$$

□

3.7 Computational Results

Our computational results are presented in this section.

We consider a problem where the discount factor is $\gamma = 0.999$, and the battery has a life of $N = 2000$ cycles. We assume the price follows a lognormal distribution $LN(\mu_0, \sigma_0^2)$ with $\mu_0 = 4$ and $\sigma_0 = 0.50$. Thus, the optimal price threshold for the infinite-life battery is $\theta^*_\infty = \gamma \mathbb{E}[p_{t+1}] = \gamma \exp(\mu_0 + \sigma_0^2/2) = 61.8059$. We also consider a variant of the problem where $\gamma = 0.9999$ and thus $\theta^*_\infty = 61.8616$.

Our codes are written in Matlab and are run on a pc equipped with a 2.80GHz Intel Xeon processor.

3.7.1 Performance of the Proposed Algorithm

We compare the proposed approach to the value iteration algorithm from the literature [8], in terms of accuracy and computational speed. Both algorithms have access to the density f and cumulative distribution function F of the price.

The value iteration (VI) algorithm approximates the continuous price state into discrete price states for the purpose of assigning a decision to each price level. Thus, in value iteration there is a tradeoff between accuracy and complexity. We use a uniform grid of price states $p^m = 0.00, 0.01, \ldots, 500.00$, noting that $\text{Prob}(p_{t+1} > 500) < 10^{-5}$. The price p^m is assigned the probability $w^m = F((p^{m+1}+p^m)/2) - F((p^m+p^{m-1})/2)$. By convention, for $p^m = 0$ we have $p^{m-1} = 0$, and for $p^m = 500$ we have $p^{m+1} = \infty$. We also take advantage of the analysis of the paper, by maximizing the discrete-price approximation V_{VI} of $V(c, n, p)$ sequentially: for a fixed n, we use the following iteration over k until convergence,

$$
\begin{aligned}
V_{\text{VI}}^{k+1}(1, n, p^m) = \max\{ &+ p^m + \gamma \textstyle\sum_i w^i V_{\text{VI}}^*(0, n-1, p^i), \\
&0 + \textstyle\sum_i w^i V_{\text{VI}}^k(1, n, p^i)\} \quad \text{for all } m, \\
V_{\text{VI}}^{k+1}(0, n, p^m) = \max\{ &- p^m + \gamma \textstyle\sum_i w^i V_{\text{VI}}^{k+1}(1, n, p^i), \\
&0 + \textstyle\sum_i w^i V_{\text{VI}}^k(0, n, p^i)\} \quad \text{for all } m.
\end{aligned}
$$

Table 3.3: Price thresholds calculated by Value Iteration.

		Life state n					
		10	50	100	500	1000	2000
$\gamma = 0.999$	$\theta_n^{1,*}$	131.61	95.76	83.15	64.37	62.11	61.82
	$\theta_n^{0,*}$	33.78	44.46	49.80	60.12	61.60	61.80
$\gamma = 0.9999$	$\theta_n^{1,*}$	194.31	148.74	131.19	95.67	83.13	72.92
	$\theta_n^{0,*}$	23.76	30.43	34.06	44.61	49.91	55.11

Note that only $\sum_i w^i V_{\mathrm{VI}}^*(0, n-1, p^i)$ is transferred to the next fixed-point problem relative to the next n. Intermediate variables (not shown) store sums calculated once. In summary, we solve small fixed-point problems to converge to $V_{\mathrm{VI}}^*(\cdot, n, \cdot)$, for $n = 1, \ldots, N$, instead of maximizing V_{VI} via a single large fixed-point problem.

The results are reported in Tables 3.3 and 3.4. It can be seen from the tables that the threshold values are very close. In terms of computational times however, for the case $\gamma = 0.999$ it takes 9.6 seconds to solve the problem with value iteration, compared to 0.75 seconds to solve the problem exactly with the proposed threshold optimization algorithm. Thus, the proposed algorithm is significantly faster.

If we reduce the number of price states to accelerate value iteration, the accuracy starts deteriorating, relatively to the exact results of Table 3.4. For instance, if we truncate the price distribution at 350 instead of 500, noting that $\mathbb{P}(p_{t+1} > 350) \simeq 10^{-4}$, we obtain $\theta_{10}^1 = 131.55$ instead of the exact value 131.6191, while the computational time is reduced to 6.7 seconds.

It is well known that the complexity of value iteration or policy iteration is greatly affected by the proximity to 1 of the discount factor, see [88]. This is confirmed here, where we observe that value iteration now takes 40 seconds if the discount factor is set to $\gamma = 0.9999$. Unlike value iteration, the complexity of the proposed algorithm is independent of the discount factor, and it takes a similarly short time of 0.78 seconds to compute the thresholds with $\gamma = 0.9999$.

Table 3.4: Price thresholds calculated by our algorithm.

		Life state n					
		10	50	100	500	1000	2000
$\gamma = 0.999$	$\theta_n^{1,*}$	131.6191	95.7515	83.1412	64.3610	62.1062	61.8106
	$\theta_n^{0,*}$	33.7848	44.4674	49.8020	60.1277	61.6049	61.8028
$\gamma = 0.9999$	$\theta_n^{1,*}$	194.4449	148.7545	131.1973	95.6708	83.1240	72.9190
	$\theta_n^{0,*}$	23.7513	30.4317	34.0608	44.6154	49.9148	55.1110

3.7.2 Impact of Non-Ideal Battery Characteristics

We illustrate the impact of the deteriorating battery capacity and of inefficiencies on the optimal price thresholds. Using $\gamma = 0.999$ and the same lognormal distribution for the price, the capacity deterioration is described by specifying the capacity when n cycles remain,

$$C_n = n/(100 + n) \quad \text{for } n = 1, \ldots, N.$$

In particular, $\lim_{n\to\infty} C_n = 1$. The efficiencies are $\eta_n^{\text{ch}} = \eta_n^{\text{dis}} = 0.9$. (Hence the round-trip efficiency is $\eta_n^{\text{ch}}\eta_n^{\text{dis}} = 0.81$.) The impact of the model modifications are presented in Figure 3.1. Since the threshold computation algorithm works by rescaling the value functions that parameterize the equations $h = 0$ and $g = 0$, and since the impact of altering the value function is well understood (proof of Proposition 10), we expect to see a greater value for the difference $\theta_n^{1,*} - \theta_n^{0,*}$, which is indeed an effect visible on the figure.

We also report some indicators on the usage of the battery device. We report the probability $\pi_n = F(\theta_n^{0,*})$ that a discharged battery buys energy, and the probability $\rho_n = 1 - F(\theta_n^{1,*})$ that a charged battery sells energy, as a function of n. This is done in Figure 3.2a for the three battery models described above. As the number of remaining cycles decreases, the probability of battery operation (charging or discharging) decreases.

It is useful to relate the probabilities π_n, ρ_n to the expected number of periods the battery occupies state n, i.e. to the expected n-th cycle time. This cycle time is equal to

$$\tau_n := \textstyle\sum_{i=0}^{\infty}(i+1)\pi_n^i(1-\pi_n) + \sum_{j=0}^{\infty}(j+1)\rho_n^j(1-\rho_n) = 1/\pi_n + 1/\rho_n,$$

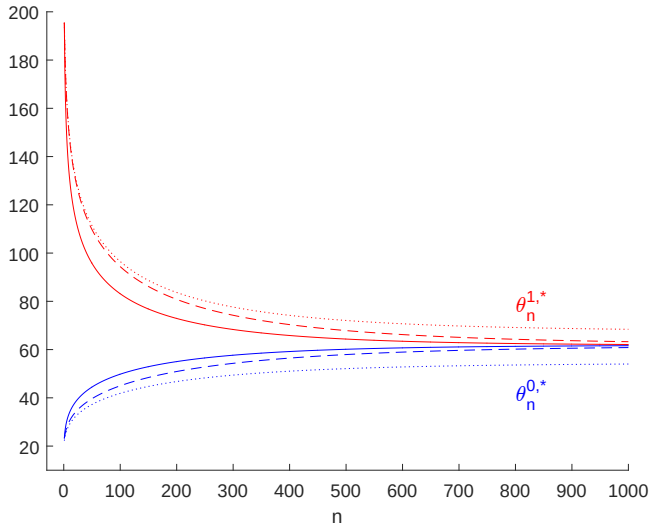
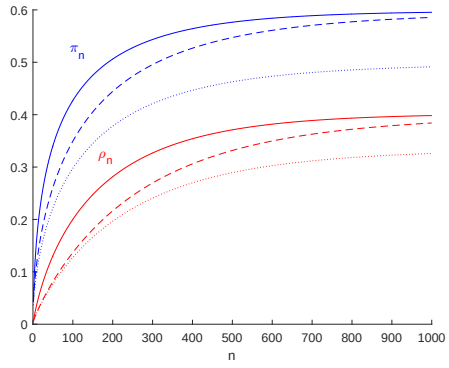
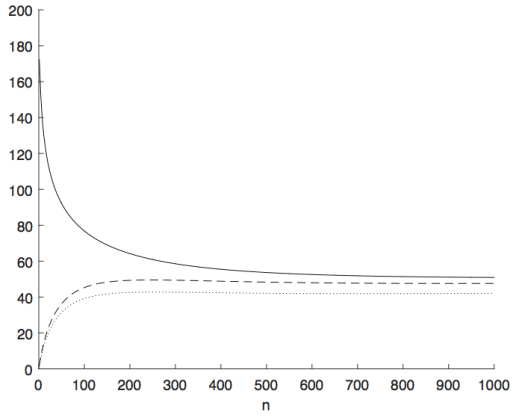


Figure 3.1: Comparison of price thresholds, as a function of n. Continuous line: base case corresponding to Table 3.4 ($\gamma = 0.999$). Dashed: With capacity deterioration. Dotted: With capacity deterioration and charging-discharging inefficiency.



(a) Probabilities π_n (charging) and ρ_n (discharging) under the optimal policy.



(b) Expected discounted profit per charging cycle under the optimal policy.

Figure 3.2: Information on storage usage. Continuous line: base case ($\gamma = 0.999$). Dashed: With capacity deterioration. Dotted: With capacity deterioration and charging-discharging inefficiency.

where i and j correspond to the time spent idle before charging and discharging respectively.

We also show, on Figure 3.2b, the expected discounted profit of the n-th charge-discharge cycle, assuming the time is set to 0 when entering state n. This profit, denoted $\bar{v}_n$, is defined in three steps, with i and j interpreted as the idle times,

$$\begin{aligned}
\bar{v}_n^1 &= \textstyle\sum_{j=0}^{\infty}(1-\rho_n)^j\rho_n\gamma^j(\eta^{\text{dis}}e_n^1) = \rho_n(\eta^{\text{dis}}e_n^1)/[1-\gamma(1-\rho_n)] \\
\bar{v}_n^0 &= \textstyle\sum_{i=0}^{\infty}(1-\pi_n)^i\pi_n\gamma^i(-(e_n^0/\eta^{\text{ch}}) + \gamma\bar{v}_n^1) = \dfrac{\pi_n(-e_n^0/\eta^{\text{ch}} + \gamma\bar{v}_n^1)}{1-\gamma(1-\pi_n)}, \\
\bar{v}_n &= C_n \cdot \bar{v}_n^0.
\end{aligned}$$

3.7.3 Economic Value of the Finite-Life Model

Finally, we comment on the value of taking into account the finiteness of the battery life in formulating the battery control problem, and on the impact of the discount factor γ.

To do this, we adopt the thresholds that are optimal for an infinite-life battery ($N \to \infty$), and compute the value of the objective (3.7) under this policy. For the perfectly efficient battery with infinite life, the thresholds are given by (3.6). Table 3.5 compares those values to the optimal objective values attained by the policy aware of the finiteness of battery life. The difference in expected values underscores the importance of implementing the control law optimized with the correct assumptions on battery life. When $\gamma = 0.999$, we have $\gamma^N = 0.1352$ if $N = 2000$, indicating that the objective will not weight much the return obtained at the end of the battery life. When $\gamma = 0.9999$, we have $\gamma^N = 0.8187$ if $N = 2000$, which explains that in this case, the value of taking into account the finiteness of the battery life is now clearly visible.

3.8 Extension of the Price Model

This section considers an extension of the price process to a regime-switching price model, in order to be able to capture more realistic price processes.

We assume that the price p_t, conditionally to being in state $m_t = m$, has density f_m (cdf: F_m), where $m_t \in \{1, ..., M\}$ is a finite-state Markov chain. The state m_t is used to

Table 3.5: Value of taking into account the finite life of the storage device.

	Battery life N					
	10	50	100	500	1000	2000
$\gamma = 0.999$						
$\bar{V}_N^{0,\pi} \mid \theta_n^c = \theta_\infty^c$	496	2287	4144	10655	11986	12174
$\bar{V}_N^{0,*} \mid \theta_n^c = \theta_n^{c,*}$	1230	3936	5985	11191	12057	12175
Improvement	147.97%	72.12%	44.44%	5.03%	0.59%	0.01%
$\gamma = 0.9999$						
$\bar{V}_N^{0,\pi} \mid \theta_n^c = \theta_\infty^c$	644	3185	6284	28218	49625	78187
$\bar{V}_N^{0,*} \mid \theta_n^c = \theta_n^{c,*}$	1990	7460	12773	39862	60335	84689
Improvement	208.99%	134.22%	103.28%	41.27%	21.58%	8.31%

indicate the regime for the price process. The one-step state transition probability matrix for m_t is denoted $\boldsymbol{T} \in \mathbb{R}^{M \times M}$. Thus:

$$\text{Prob}(m_{t+1} = j \mid m_t = i) = T_{ij}, \qquad \text{Prob}(p_t \leq p \mid m_t = m) = F_m(p). \tag{3.33}$$

The regime state can be used as a device to distinguish periods of low prices versus price spikes, periods of low versus high price volatility, the time index of a seasonal process, etc. See [65] and [108] for discussions.

In the limit case where the densities f_m degenerate to a single point, m_t becomes a sufficient statistic for p_t, and the price process degenerates to a finite-state Markov chain.

From the conditional independence assumption in (3.33), an optimal policy can still be described using a finite number of thresholds, now indexed by the state m. Thus, we can restrict the search to policies such that

$$a_t = A^\pi(c_t, n_t, p_t, m_t) = \begin{cases} 1 & \text{if } c_t = 0 \text{ and } p_t \leq \theta^0_{n_t,m_t}, \\ -1 & \text{if } c_t = 1 \text{ and } p_t \geq \theta^1_{n_t,m_t}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.34}$$

where $\theta^0_{n,m}$, $\theta^1_{n,m}$ are the threshold parameters describing a policy π.

3.8.1 Threshold Policy Evaluation

Generalizing previous definitions, let $\bar{V}^{c,\pi}_{n,m}$ be the expected value function at the successor state when the *current* regime state is m, the *next* state of charge is $c \in \{0,1\}$, the *next* remaining-life state is n, and the policy π is followed. Thus, the expectation is over the next regime state and price. It is easy to verify on the two-regime case ($M = 2$) that by backward induction,

$$\begin{bmatrix} \bar{V}^{0,\pi}_{n,1} \\ \bar{V}^{0,\pi}_{n,2} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} (1-\pi_{n,1})[0+\gamma\bar{V}^{0,\pi}_{n,1}] + \pi_{n,1}[-e^0_{n,1}+\gamma\bar{V}^{1,\pi}_{n,1}] \\ (1-\pi_{n,2})[0+\gamma\bar{V}^{0,\pi}_{n,2}] + \pi_{n,2}[-e^0_{n,2}+\gamma\bar{V}^{1,\pi}_{n,2}] \end{bmatrix},$$

$$\begin{bmatrix} \bar{V}^{1,\pi}_{n,1} \\ \bar{V}^{1,\pi}_{n,2} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} (1-\rho_{n,1})[0+\gamma\bar{V}^{1,\pi}_{n,1}] + \rho_{n,1}[e^1_{n,1}+\gamma\bar{V}^{0,\pi}_{n-1,1}] \\ (1-\rho_{n,2})[0+\gamma\bar{V}^{1,\pi}_{n,2}] + \rho_{n,2}[e^1_{n,2}+\gamma\bar{V}^{0,\pi}_{n-1,2}] \end{bmatrix},$$

where

$$\pi_{n,m} = F_m(\theta^0_{n,m}), \qquad e^0_{n,m} = \textstyle\int_0^{\theta^0_{n,m}} p f_m(p)dp \Big/ \pi_{n,m}, \qquad \text{(3.8'-3.9')}$$

$$\rho_{n,m} = 1 - F_m(\theta^1_{n,m}), \qquad e^1_{n,m} = \textstyle\int_{\theta^1_{n,m}}^{\infty} p f_m(p)dp \Big/ \rho_{n,m}. \qquad \text{(3.10'-3.11')}$$

By convention, $\pi_{n,m}e^0_{n,m} := 0$ if $\pi_{n,m} = 0$, and $\rho_{n,m}e^1_{n,m} := 0$ if $\rho_{n,m} = 0$. This arises if one respectively never charges or never discharges when being in regime m.

We can write this in vector form, as follows (the extension to $M > 2$ is then immediate). Let $\text{Diag}(\boldsymbol{x})$ denote the diagonal matrix with diagonal $\boldsymbol{x}$. Let $\boldsymbol{I}$ be the identity matrix. Define

$$\bar{\boldsymbol{V}}^{c,\pi}_n = \begin{bmatrix} \bar{V}^{c,\pi}_{n,1} \\ \bar{V}^{c,\pi}_{n,2} \end{bmatrix}, \qquad \boldsymbol{e}^c_n = \begin{bmatrix} e^c_{n,1} \\ e^c_{n,2} \end{bmatrix}, \qquad \boldsymbol{\pi}_n = \begin{bmatrix} \pi_{n,1} \\ \pi_{n,2} \end{bmatrix}, \qquad \boldsymbol{\rho}_n = \begin{bmatrix} \rho_{n,1} \\ \rho_{n,2} \end{bmatrix},$$

$$\boldsymbol{D}_{\pi_n} = \text{Diag}(\boldsymbol{\pi}_n), \qquad \boldsymbol{D}_{\rho_n} = \text{Diag}(\boldsymbol{\rho}_n).$$

From the expression $\bar{\boldsymbol{V}}^{0,\pi}_n = \boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\pi_n})\gamma\bar{\boldsymbol{V}}^{0,\pi}_n + \boldsymbol{T}\boldsymbol{D}_{\pi_n}(-\boldsymbol{e}^0_n + \gamma\bar{\boldsymbol{V}}^{1,\pi}_n)$ and similarly from

the expression $\bar{\boldsymbol{V}}_n^{1,\pi} = \boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\rho_n})\gamma\bar{\boldsymbol{V}}_n^{1,\pi} + \boldsymbol{T}\boldsymbol{D}_{\rho_n}(\boldsymbol{e}_n^1 + \gamma\bar{\boldsymbol{V}}_{n-1}^{0,\pi})$ we obtain

$$\bar{\boldsymbol{V}}_n^{0,\pi} = \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\boldsymbol{D}_{\pi_n}(-\boldsymbol{e}_n^0 + \gamma\bar{\boldsymbol{V}}_n^{1,\pi}), \qquad \boldsymbol{A}_{\pi_n} = \boldsymbol{I} - \gamma\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\pi_n}), \tag{3.16'}$$

$$\bar{\boldsymbol{V}}_n^{1,\pi} = \boldsymbol{A}_{\rho_n}^{-1}\boldsymbol{T}\boldsymbol{D}_{\rho_n}(\boldsymbol{e}_n^1 + \gamma\bar{\boldsymbol{V}}_{n-1}^{0,\pi}), \qquad \boldsymbol{A}_{\rho_n} = \boldsymbol{I} - \gamma\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\rho_n}). \tag{3.17'}$$

We stress that (3.8'-3.9'), (3.10'-3.11'), (3.16'), (3.17') are valid for a policy described by any, not necessarily optimal, $\theta = \{\boldsymbol{\theta}_n^0, \boldsymbol{\theta}_n^1\}_{1\le n\le N}$ (where $\boldsymbol{\theta}_n^c = \begin{bmatrix} \theta_{n,1}^c \\ \theta_{n,2}^c \end{bmatrix}$ if $M = 2$.)

3.8.2 Threshold Policy Optimization

Let the notation $\bar{\boldsymbol{V}}_n^{0,\pi} = \bar{\boldsymbol{V}}^0(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^1)$ stress that there is a functional relation between $(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^1)$ and $\bar{\boldsymbol{V}}_n^{0,\pi}$, given by (3.16'). Let $\bar{\boldsymbol{V}}_n^{0,*} = \bar{\boldsymbol{V}}^0(\boldsymbol{\theta}_n^{0,*}, \bar{\boldsymbol{V}}_n^{1,*})$ denote the optimal expected value function attained by a policy with optimal parameters $\boldsymbol{\theta}_n^{0,*}$. Similarly let $\bar{\boldsymbol{V}}_n^{1,*} = \bar{\boldsymbol{V}}^1(\boldsymbol{\theta}_n^{1,*}, \bar{\boldsymbol{V}}_{n-1}^{0,*})$ with optimal parameters $\boldsymbol{\theta}_n^{1,*}$. By backward induction, if $\bar{\boldsymbol{V}}_n^{1,*}$ is optimal, then each $\bar{V}_{n,m}^{0,\pi}$ is maximized by optimizing over $\boldsymbol{\theta}_n^0$.

For convenience, in the sequel we write $\bar{\boldsymbol{V}}_n^0$ instead of $\bar{\boldsymbol{V}}^0$ even though the index n does not alter the function. We adopt the approach of seeking to optimize $\bar{V}_{n,m}^0$ for all m at the same time. To do this, we evaluate the derivative of $\bar{V}_{n,i}^0$ with respect to $\theta_{n,j}^0$ for each i, j, that is, we evaluate the Jacobian matrix $\boldsymbol{J}_n^0$ with elements $J_{n,ij}^0$.

Let $\boldsymbol{f}_n^0$ denote the vector with m-th element $f_{n,m}^0 = f_m(\theta_{n,m}^0)$.

Proposition 13. *The Jacobian matrix $\boldsymbol{J}_n^0$ with elements $J_{n,ij}^0 = \partial\bar{V}_{n,i}^0/\partial\theta_{n,j}^0$, evaluated at $(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*})$, is described by*

$$\boldsymbol{J}_n^0 = \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\,\mathrm{Diag}(\boldsymbol{f}_n^0)\,\mathrm{Diag}(\gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^0) - \boldsymbol{\theta}_n^0). \tag{3.35}$$

Equivalently,

$$\boldsymbol{J}_n^0 = \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\,\mathrm{Diag}(\boldsymbol{f}_n^0)\,\mathrm{Diag}(\boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{g}(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*})), \tag{3.36}$$

where

$$\boldsymbol{g}(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*}) = \gamma(\boldsymbol{I} - \gamma\boldsymbol{T})\bar{\boldsymbol{V}}_n^{1,*} - (\boldsymbol{I} - \gamma\boldsymbol{T})\boldsymbol{\theta}_n^0 - \gamma\boldsymbol{T}\boldsymbol{D}_{\pi_n}\boldsymbol{\theta}_n^0 + \gamma\boldsymbol{T}\boldsymbol{D}_{\pi_n}\boldsymbol{e}_n^0. \tag{3.22'}$$

Proof. Calculus is done using the differential operator d, in order to identify the Jacobian matrix $\boldsymbol{J}_n^0$ via $d\bar{\boldsymbol{V}}_n^0 = \boldsymbol{J}_n^0 d\boldsymbol{\theta}_n^0$, see [60] Chapter 9. We define $\boldsymbol{b}_n^0 = -\boldsymbol{e}_n^0 + \gamma\bar{\boldsymbol{V}}_n^{1,*}$, thus from (3.16'), $\bar{\boldsymbol{V}}_n^0 = \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\boldsymbol{D}_{\pi_n}\boldsymbol{b}_n^0$. Then

$$\begin{aligned} d\bar{\boldsymbol{V}}_n^0 &= d(\boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\boldsymbol{D}_{\pi_n}\boldsymbol{b}_n^0) \\ &= d(\boldsymbol{A}_{\pi_n}^{-1})\boldsymbol{T}\boldsymbol{D}_{\pi_n}\boldsymbol{b}_n^0 + \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}d(\boldsymbol{D}_{\pi_n})\boldsymbol{b}_n^0 + \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\boldsymbol{D}_{\pi_n}d\boldsymbol{b}_n^0. \end{aligned} \tag{3.37}$$

We have $\partial\pi_{n,m}/\partial\theta_{n,m}^0 = f_{n,m}^0$. From $\partial e_{n,m}^0/\partial\theta_{n,m}^0 = (\theta_{n,m}^0 - e_{n,m}^0)f_{n,m}^0/\pi_{n,m}$ and $\partial e_{n,m}^0/\partial\theta_{n,j}^0 = 0$ for $j \neq m$ we obtain

$$d\boldsymbol{b}_n^0 = -d\boldsymbol{e}_n^0 = \boldsymbol{D}_{\pi_n}^{-1}\operatorname{Diag}(\boldsymbol{e}_n^0 - \boldsymbol{\theta}_n^0)\operatorname{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0.$$

We calculate, for any fixed vector $\boldsymbol{x} \in \mathbb{R}^M$,

$$\begin{aligned} d(\boldsymbol{D}_{\pi_n})\boldsymbol{x} &= \operatorname{Diag}(\boldsymbol{f}_n^0)\operatorname{Diag}(d\boldsymbol{\theta}_n^0)\boldsymbol{x} = \operatorname{Diag}(\boldsymbol{x})\operatorname{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0, \\ d(\boldsymbol{A}_{\pi_n})\boldsymbol{x} &= \gamma\boldsymbol{T}d(\boldsymbol{D}_{\pi_n})\boldsymbol{x} = \gamma\boldsymbol{T}\operatorname{Diag}(\boldsymbol{x})\operatorname{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0, \\ d(\boldsymbol{A}_{\pi_n}^{-1})\boldsymbol{x} &= -\boldsymbol{A}_{\pi_n}^{-1}d(\boldsymbol{A}_{\pi_n})\boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{x} = -\boldsymbol{A}_{\pi_n}^{-1}\gamma\boldsymbol{T}\operatorname{Diag}(\boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{x})\operatorname{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0. \end{aligned}$$

Therefore overall, (3.37) becomes

$$\begin{aligned} d\bar{\boldsymbol{V}}_n^0 &= -\boldsymbol{A}_{\pi_n}^{-1}\gamma\boldsymbol{T}\operatorname{Diag}(\boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\boldsymbol{D}_{\pi_n}\boldsymbol{b}_n^0)\operatorname{Diag}(\boldsymbol{f}_0^n)d\boldsymbol{\theta}_n^0 \\ &\quad + \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\operatorname{Diag}(\boldsymbol{b}_n^0)\operatorname{Diag}(\boldsymbol{f}_0^n)d\boldsymbol{\theta}_n^0 \\ &\quad + \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\operatorname{Diag}(\boldsymbol{e}_n^0 - \boldsymbol{\theta}_n^0)\operatorname{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0 \\ &= \boldsymbol{A}_{\pi_n}^{-1}\boldsymbol{T}\operatorname{Diag}(\boldsymbol{y}_n^0)\operatorname{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0 \qquad (\boldsymbol{y}_n^0 \text{ defined below}) \end{aligned}$$

where we have defined

$$\begin{aligned}\boldsymbol{y}_n^0 &:= -\gamma \boldsymbol{A}_{\pi_n}^{-1} \boldsymbol{T} \boldsymbol{D}_{\pi_n} \boldsymbol{b}_n^0 + \boldsymbol{b}_n^0 + \boldsymbol{e}_n^0 - \boldsymbol{\theta}_n^0 \\ &= -\gamma \bar{\boldsymbol{V}}_n^0 + (-\boldsymbol{e}_n^0 + \gamma \bar{\boldsymbol{V}}_n^{1,*}) + \boldsymbol{e}_n^0 - \boldsymbol{\theta}_n^0 \\ &= \gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^0) - \boldsymbol{\theta}_n^0. \end{aligned} \tag{3.38}$$

By identification via $d\bar{\boldsymbol{V}}_n^0 = \boldsymbol{J}_n^0 d\boldsymbol{\theta}_n^0$, we get (3.35).

To get (3.36) and (3.22'), we use $\boldsymbol{y}_n^0 = \boldsymbol{A}_{\pi_n}^{-1} \boldsymbol{A}_{\pi_n} \boldsymbol{y}_n^0$ and

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*}) &:= \boldsymbol{A}_{\pi_n} \boldsymbol{y}_n^0 \\ &= \gamma \boldsymbol{A}_{\pi_n} \bar{\boldsymbol{V}}_n^{1,*} - \gamma \boldsymbol{T} \boldsymbol{D}_{\pi_n} (-\boldsymbol{e}_n^0 + \gamma \bar{\boldsymbol{V}}_n^{1,*}) - \boldsymbol{A}_{\pi_n} \boldsymbol{\theta}_n^0 \\ &= \gamma(\boldsymbol{A}_{\pi_n} - \gamma \boldsymbol{T} \boldsymbol{D}_{\pi_n}) \bar{\boldsymbol{V}}_n^{1,*} + \gamma \boldsymbol{T} \boldsymbol{D}_{\pi_n} \boldsymbol{e}_n^0 - \boldsymbol{A}_{\pi_n} \boldsymbol{\theta}_n^0 \\ &= \gamma(\boldsymbol{I} - \gamma \boldsymbol{T}) \bar{\boldsymbol{V}}_n^{1,*} + \gamma \boldsymbol{T} \boldsymbol{D}_{\pi_n} \boldsymbol{e}_n^0 - (\boldsymbol{I} - \gamma \boldsymbol{T} + \gamma \boldsymbol{T} \boldsymbol{D}_{\pi_n}) \boldsymbol{\theta}_n^0. \end{aligned}$$

□

From Proposition 13 the following remarks are in order:

- At optimality, the Jacobian matrix must be identically zero. To see this, note that at any optimal $\bar{V}_{n,i}^0$, the threshold policy needs to satisfy $\partial \bar{V}_{n,i}^0 / \partial \theta_{n,j}^0 = 0$ for all j to be optimal.

- From (3.35), the Jacobian matrix is identically zero if $\gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^0) - \boldsymbol{\theta}_n^0 = 0$. Thus a threshold vector $\boldsymbol{\theta}_n^{0,*}$ satisfying

$$\boldsymbol{\theta}_n^{0,*} = \gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^{0,*}) \tag{3.4'}$$

  is optimal.

- From (3.4'), Step 4 in Table 3.1 generalizes to $\bar{\boldsymbol{V}}_n^{0,*} = \bar{\boldsymbol{V}}_n^{1,*} - \boldsymbol{\theta}_n^{0,*}/\gamma$.

- It is easy to check that $\boldsymbol{A}_{\pi_n}$ is strictly row diagonally dominant, with positive diagonal elements. Hence $\boldsymbol{A}_{\pi_n}$ is invertible, and furthermore, $\boldsymbol{A}_{\pi_n}^{-1}$ is nonnegative, see e.g. [73].

- From (3.36), the Jacobian matrix is identically zero if

$$\boldsymbol{g}(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*}) = \mathbf{0}. \tag{3.24'}$$

- Equation (3.36) eliminates $\bar{\boldsymbol{V}}_n^0$ from (3.35), to eliminate a recursion that would otherwise appear in the second-order differentiation (used in Prop. 14 below).

- It is easy to see that (3.22') reduces to (3.22) in the case $M = 1$, where necessarily $\boldsymbol{T} = 1$, $\boldsymbol{D}_{\pi_n} = \pi_n = F(\theta_n^0)$, and $\boldsymbol{A}_{\pi_n} = 1 - \gamma(1 - \pi_n)$.

The following proposition is the counterpart of Proposition 7. Let $\boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^0}$ denote the Jacobian matrix of $\boldsymbol{g}$ with respect to $\boldsymbol{\theta}_n^0$, evaluated at $\boldsymbol{\theta}_n^0$.

Proposition 14. *A sufficient optimality condition for the thresholds $\boldsymbol{\theta}_n^0$ is given by the implicit equation* (3.24'). *At any $\boldsymbol{\theta}_n^0$ we have $\boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^0} = -\boldsymbol{A}_{\pi_n}$. The solution $\boldsymbol{\theta}_n^{0,*}$ is unique, assuming the thresholds lie on the support of the price distributions.*

Proof. The Jacobian matrix of $\boldsymbol{g}$ with respect to $\boldsymbol{\theta}_n^0$ can be identified by calculating

$$\begin{aligned}
d\boldsymbol{g} &= -(\boldsymbol{I} - \gamma\boldsymbol{T})d\boldsymbol{\theta}_n^0 + \gamma\boldsymbol{T}d(\boldsymbol{D}_{\pi_n})(-\boldsymbol{\theta}_n^0 + \boldsymbol{e}_n^0) - \gamma\boldsymbol{T}\boldsymbol{D}_{\pi_n}d\boldsymbol{\theta}_n^0 + \gamma\boldsymbol{T}\boldsymbol{D}_{\pi_n}d\boldsymbol{e}_n^0 \\
&= (-\boldsymbol{I} + \gamma\boldsymbol{T} - \gamma\boldsymbol{T}\boldsymbol{D}_{\pi_n})d\boldsymbol{\theta}_n^0 + \gamma\boldsymbol{T}\,\mathrm{Diag}(-\boldsymbol{\theta}_n^0 + \boldsymbol{e}_n^0)\,\mathrm{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0 \\
&\quad + \gamma\boldsymbol{T}\boldsymbol{D}_{\pi_n}[\boldsymbol{D}_{\pi_n}^{-1}\,\mathrm{Diag}(\boldsymbol{\theta}_n^0 - \boldsymbol{e}_n^0)\,\mathrm{Diag}(\boldsymbol{f}_n^0)d\boldsymbol{\theta}_n^0] \qquad \text{(cf. Proof of Prop. 13)} \\
&= -(\boldsymbol{I} - \gamma\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\pi_n}))d\boldsymbol{\theta}_n^0 \\
&= -\boldsymbol{A}_{\pi_n}d\boldsymbol{\theta}_n^0,
\end{aligned}$$

thus the Jacobian matrix of $\boldsymbol{g}$ is $\boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^0} = -\boldsymbol{A}_{\pi_n}$, which is strictly row diagonally dominant and thus invertible. The uniqueness of the solution to $\boldsymbol{g}(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*}) = \mathbf{0}$ then follows from the implicit function theorem. □

Several methods can be used to solve the implicit equation (3.24'). The following proposition furnishes a simple iterative procedure.

Proposition 15. *To find the solution* $\boldsymbol{\theta}_n^{0,*}$ *to* $\boldsymbol{g}(\boldsymbol{\theta}_n^{0,*}, \bar{\boldsymbol{V}}_n^{1,*}) = \mathbf{0}$ *for a fixed* $\bar{\boldsymbol{V}}_n^{1,*}$*, let the sequence of iterates* $(\boldsymbol{\theta}_n^{0,k}, \bar{\boldsymbol{V}}_n^{0,k})$*,* $k = 1, 2, \ldots$*, be defined by*

$$\boldsymbol{\theta}_n^{0,k+1} = \gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^{0,k}), \tag{3.39}$$

where $\bar{\boldsymbol{V}}_n^{0,k}$ *is the value relative to* $\boldsymbol{\theta}_n^{0,k}$ *given by* (3.16') *with* $\bar{\boldsymbol{V}}_n^{1,\pi}$ *set to* $\bar{\boldsymbol{V}}_n^{1,*}$*, and the other entities computed for* $\boldsymbol{\theta}_n^{0,k}$*.*

Then, it holds that the sequence of iterates $\boldsymbol{\theta}_n^{0,k}$ *converges to* $\boldsymbol{\theta}_n^{0,*}$*. Additionally, if* $\boldsymbol{\theta}_n^{0,k}$ *is sufficiently close to* $\boldsymbol{\theta}_n^{0,*}$*, the rate of convergence is quadratic.*

Proof. Newton's method is used to solve the nonlinear system $\boldsymbol{g}(\boldsymbol{\theta}_n^0, \bar{\boldsymbol{V}}_n^{1,*}) = \mathbf{0}$. Assuming a full Newton step, the update is

$$\begin{aligned}
\boldsymbol{\theta}_n^{0,k+1} &= \boldsymbol{\theta}_n^{0,k} - \boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^{0,k}}^{-1} \boldsymbol{g}(\boldsymbol{\theta}_n^{0,k}, \bar{\boldsymbol{V}}_n^{1,*}) \\
&= \boldsymbol{\theta}_n^{0,k} - [-\boldsymbol{A}_{\pi_n^k}^{-1}][\boldsymbol{A}_{\pi_n^k} \boldsymbol{y}_n^{0,k}] \qquad \text{with } \boldsymbol{y}_n^{0,k} \text{ defined as in (3.38)} \\
&= \boldsymbol{\theta}_n^{0,k} + \gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^{0,k}) - \boldsymbol{\theta}_n^{0,k} \\
&= \gamma(\bar{\boldsymbol{V}}_n^{1,*} - \bar{\boldsymbol{V}}_n^{0,k}).
\end{aligned}$$

Theorem 11.2 in [69] ensures that for a starting point $\boldsymbol{\theta}_n^{0,1}$ sufficiently close to a solution $\boldsymbol{\theta}_n^{0,*}$ with $\boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^{0,*}}$ nonsingular, the sequence of iterates converges to $\boldsymbol{\theta}_n^{0,*}$. Now, since $\boldsymbol{J}_{\boldsymbol{g},\theta_n^0} = -\boldsymbol{A}_{\pi_n}$ is strictly diagonally dominant for any $\boldsymbol{\theta}_n^0$, we have $\|\boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^0}^{-1}\|_\infty < 1/\min_i\{|A_{\pi_n,ii}| - \sum_{j\neq i} |A_{\pi_n,ij}|\}$, see [105]. Each row of $\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\pi_n})$ has the sum of its elements between 0 and 1, implying that each row of $\boldsymbol{A}_{\pi_n}$ has the sum of its elements between $1-\gamma$ and 1. Therefore, for any $\boldsymbol{\theta}_n^0$ arbitrarily far from $\boldsymbol{\theta}_n^{0,*}$,

$$\|\boldsymbol{J}_{\boldsymbol{g},\boldsymbol{\theta}_n^0}^{-1}\|_\infty < 1/(1-\gamma), \tag{3.40}$$

which implies $\boldsymbol{\theta}_n^{0,k+1} - \boldsymbol{\theta}_n^{0,*} = o(\|\boldsymbol{\theta}_n^{0,k} - \boldsymbol{\theta}_n^{0,*}\|)$ (through the proof of Theorem 11.2). Moreover, since $\boldsymbol{g}$ is differentiable and thus in particular Lipschitz continuous, by the cited Theorem 11.2 it holds that for $\boldsymbol{\theta}_n^{0,k}$ sufficiently close to $\boldsymbol{\theta}_n^{0,*}$ we have $\boldsymbol{\theta}_n^{0,k+1} - \boldsymbol{\theta}_n^{0,*} = \mathcal{O}(\|\boldsymbol{\theta}_n^{0,k} - \boldsymbol{\theta}_n^{0,*}\|^2)$, indicating convergence of the quadratic kind. □

We conclude this section with the following remarks.

- The results for optimizing $\bar{\boldsymbol{V}}_n^1$ in (3.17') over $\boldsymbol{\theta}_n^1$ given $\bar{\boldsymbol{V}}_{n-1}^{0,*}$ are established similarly. Let $\boldsymbol{f}_n^1$ denote the vector with m-th element $f_m(\theta_{n,m}^1)$. Let $\boldsymbol{J}_n^1$ denote the Jacobian matrix of $\bar{\boldsymbol{V}}_n^1$ with respect to $\boldsymbol{\theta}_n^1$. Let $\boldsymbol{J}_{\boldsymbol{h},\boldsymbol{\theta}_n^1}$ be the Jacobian matrix of the function $\boldsymbol{h}$ (defined below) with respect to $\boldsymbol{\theta}_n^1$. Then we have

$$\begin{aligned}
\boldsymbol{J}_n^1 &= \boldsymbol{A}_{\rho_n}^{-1}\boldsymbol{T}\,\mathrm{Diag}(\boldsymbol{f}_n^1)\,\mathrm{Diag}(\gamma(\bar{\boldsymbol{V}}_n^1 - \bar{\boldsymbol{V}}_{n-1}^{0,*}) - \boldsymbol{\theta}_n^1) \\
&= \boldsymbol{A}_{\rho_n}^{-1}\boldsymbol{T}\,\mathrm{Diag}(\boldsymbol{f}_n^1)\,\mathrm{Diag}(\boldsymbol{A}_{\rho_n}^{-1}\boldsymbol{h}(\boldsymbol{\theta}_n^1, \bar{\boldsymbol{V}}_{n-1}^{0,*})), \\
\boldsymbol{h}(\boldsymbol{\theta}_n^1, \bar{\boldsymbol{V}}_{n-1}^{0,*}) &= -\boldsymbol{\theta}_n^1 - \gamma(\boldsymbol{I} - \gamma\boldsymbol{T})\bar{\boldsymbol{V}}_{n-1}^{0,*} + \gamma\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{D}_{\rho_n})\boldsymbol{\theta}_n^1 + \gamma\boldsymbol{T}\boldsymbol{D}_{\rho_n}\boldsymbol{e}_n^1. \qquad (3.23')
\end{aligned}$$

  A sufficient optimality condition for the thresholds $\boldsymbol{\theta}_n^1$ is given by the implicit equation

$$\boldsymbol{h}(\boldsymbol{\theta}_n^1, \bar{\boldsymbol{V}}_{n-1}^{0,*}) = 0. \qquad (3.25')$$

  We have $\boldsymbol{J}_{\boldsymbol{h},\boldsymbol{\theta}_n^1} = -\boldsymbol{A}_{\rho_n}$, which is always invertible, so the solution $\boldsymbol{\theta}_n^{1,*}$ is unique, assuming the thresholds lie on the support of the price distributions.

- The solution $\boldsymbol{\theta}_n^{1,*}$ to $\boldsymbol{h}(\boldsymbol{\theta}_n^{1,*}, \bar{\boldsymbol{V}}_{n-1}^{0,*}) = \boldsymbol{0}$ for a fixed $\bar{\boldsymbol{V}}_{n-1}^{0,*}$ can be obtained via the sequence of iterates $(\boldsymbol{\theta}_n^{1,k}, \bar{\boldsymbol{V}}_n^{1,k})$, $k = 1, 2, \ldots$, defined by

$$\boldsymbol{\theta}_n^{1,k+1} = \gamma(\bar{\boldsymbol{V}}_n^{1,k} - \bar{\boldsymbol{V}}_{n-1}^{0,*}), \qquad (3.41)$$

  where $\bar{\boldsymbol{V}}_n^{1,k}$ is the value relative to $\bar{\boldsymbol{\theta}}_n^{1,k}$ obtained via (3.17') with $\bar{\boldsymbol{V}}_{n-1}^{0,\pi}$ set to $\bar{\boldsymbol{V}}_{n-1}^{0,*}$.

- The overall algorithm for determining all the optimal thresholds is described by Table 3.1 of Section 3.4, with all the entities in Table 3.1 now referring to the entities in bold letters defined in the present section. If the implicit equations are solved by the iterative method described in Proposition 15, then Step 2 is a byproduct of Step 1, and Step 4 is a byproduct of Step 3. Since the convergence is quadratic, only a few iterations are typically needed to complete Step 1 and Step 3. What is more, $\boldsymbol{\theta}_{n+1}^{c,*}$ tends to be close to $\boldsymbol{\theta}_n^{c,*}$, making $\boldsymbol{\theta}_{n+1}^{c,1} = \boldsymbol{\theta}_n^{c,*}$ a natural starting point of the

iteration for $\boldsymbol{\theta}_{n+1}^{c,*}$.

- Non-ideal battery characteristics are handled exactly as in Section 3.6, so it is easy to combine the results obtained so far. The starting point is the policy evaluation step to calculate the expected value functions per unit of capacity,

$$C_n \bar{\boldsymbol{V}}_n^{0,\pi} = \boldsymbol{A}_{\pi_n}^{-1} \boldsymbol{T} \boldsymbol{D}_{\pi_n} (-C_n \boldsymbol{e}_n^0 / \eta_n^{\text{ch}} + \gamma C_n \bar{\boldsymbol{V}}_n^{1,\pi}),$$
$$C_n \bar{\boldsymbol{V}}_n^{1,\pi} = \boldsymbol{A}_{\rho_n}^{-1} \boldsymbol{T} \boldsymbol{D}_{\rho_n} (C_n \boldsymbol{e}_n^1 \eta_n^{\text{dis}} + \gamma C_{n-1} \bar{\boldsymbol{V}}_{n-1}^{0,\pi}). \tag{3.42}$$

  The overall optimization algorithm for determining the optimal thresholds under non-ideal battery characteristics is described by Table 3.2 of Section 3.6, with all the entities in the table now referring to the entities in bold letters defined in the present section.

3.8.3 Illustration

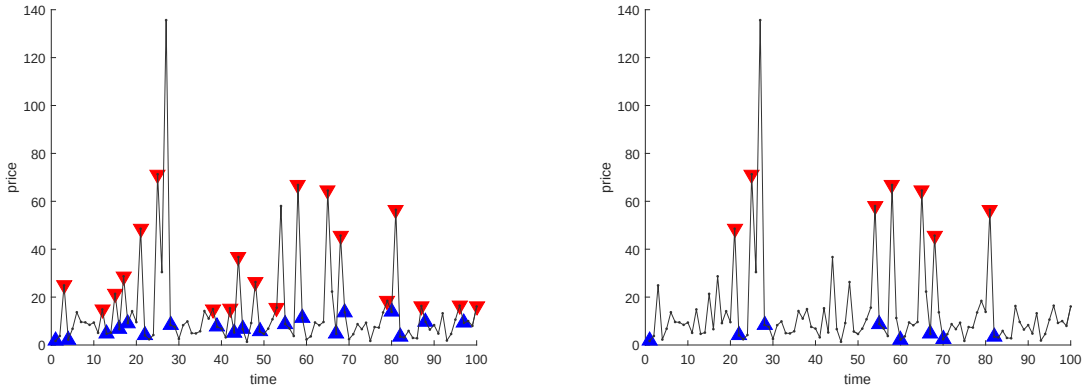
We illustrate the behavior of the optimal policy on a simple regime-switching price process with two regimes.

As in Section 3.7.2, the battery has efficiency parameters $\eta_n^{\text{ch}} = 0.9$, $\eta_n^{\text{dis}} = 0.9$. The capacity deteriorates following $C_n = n/(100 + n)$.

The Markov chain governing the regime m_t has $\boldsymbol{T} = \begin{bmatrix} 0.90 & 0.10 \\ 0.95 & 0.05 \end{bmatrix}$. The price p_t follows $LN(\mu_{m_t}, \sigma_{m_t}^2)$ with $\mu_1 = 2$, $\sigma_1 = 0.7$ for regime 1 and $\mu_2 = 4$, $\sigma_2 = 0.5$ for regime 2. The discount factor is $\gamma = 0.999$.

Figure 3.3 depicts the behavior of the optimal policy on a sample path of the price process over $t = 1, \ldots, 100$. The charging and discharging decisions are marked with triangles on the sample path. On Figure 3.3a, the battery starts from $n_t = 1000$ at $t = 1$ and attains $n_t = 981$ at $t = 100$. On Figure 3.3b, the battery starts from $n_1 = 50$ and attains $n_{100} = 43$. Thus the battery completes 19 cycles in case (a) but only 7 cycles in case (b).

If we were to operate the battery using the optimal thresholds but assuming 1000 cycles remain while actually 50 remain, the policy, being more active, would exhaust the battery



(a) Starting from $n = 1000$ remaining cycles.

(b) Starting from $n = 50$ remaining cycles.

Figure 3.3: Simulation of the optimal policy on the same sample path of a price process but starting from two different ages. Decisions: Charge (▲), Discharge (▼).

sooner and attain $C_{50}\bar{V}_{50,1}^{0,\pi|n_1=1000} = 265$ calculated using the recursion (3.42). This is to be compared to the optimal value with the correct assumption of 50 cycles remaining, equal to $C_{50}\bar{V}_{50,1}^{0,\pi|n_1=50} = 485$. Thus, there is high value in considering battery aging when optimizing operations.

3.9 Conclusion

This chapter considers the market-based battery operation problem with aging phenomenon. The problem is formulated as an infinite-horizon Markov Decision Process with a continuous price state and a discrete remaining life state. An efficient optimization algorithm is proposed which exploits the structure of the optimal policy and is based on solving a sequence of optimization problems. An error analysis and a computational study demonstrate the performance of the algorithm in terms of accuracy and efficiency. Since the algorithm is fast, it would be practical to embed it into a rollout scheme used as a policy for approximately solving more complex problems.

While the stochastic price models used in this chapter are simple, the solution to the stochastic optimal control problem is already sufficient to offer managerial insights. In particular, the awareness of the finiteness of the battery life dramatically increases the optimal price spread required for storage operations. The spread widens when the battery

approaches its end-of-life. Capacity deterioration compounds the widening effect. This behavior contrasts with the optimal price spread caused by the need for compensating for inefficiencies during a single storage cycle, which is present as soon as the storage device is put into service. This suggests that the contribution of storage devices to market-based operations could sharply decrease when devices near their end-of-life, with the typical price spreads of the market becoming insufficient to justify participation. Decreased participation could thus occur much sooner than expected from a count of remaining cycles.

Chapter 4

On the Price Impact of Distributed Energy Storage

4.1 Introduction

Given installed electricity storage devices for consumers and suppliers, we are concerned with the problem of managing the inflows and outflows of electricity between the power grid and the storage devices, to maximize the expected discounted cumulated welfare of consumers and suppliers in the market over an infinite horizon.

In this chapter, two key aspects that are taken into consideration are (i) the sensitivity of the supply and demand to electricity prices, that are to be determined endogenously to maintain the power balance, and (ii) the sensitivity of the demand curve to exogenous stochastic factors.

Furthermore, since prices affect in opposite directions the utility of the demand and the utility of the supply, and since storage actions influence prices, an issue that arises for determining optimal battery operations and equilibrium prices is the role of ownership of battery capacity. The repartition between the demand side and the supply side is therefore expected to play a role, inasmuch as market participants are expected to behave strategically.

Our problem relates closely to storage management problems found in the commodity

storage literature [72, 85], for which rational expectation (RE) models have been proposed. [110] provides a stochastic valuation of energy storage framework and an energy storage optimization model. However, this deterministic model does not consider price uncertainty. [95] considers the impact of large storage devices on the electricity price, thus on the welfare on generators and consumers separately. Welfare effects from storage in different market are studies [94]. To capture the recursive relationship between decisions and future expected prices, the RE model is formulated as an optimal control problem [56], which can in theory be solved by dynamic programming, possibly by exploiting a favorable structure [35]. Otherwise, approximate dynamic programming (ADP) techniques are needed to solve the dynamic program heuristically [76, 7]. [43] describes approximate solution techniques that are applied to solve RE models.

In the context of managing energy storage resources, several stochastic models have been proposed and investigated, often using an exogenous stochastic processes for the price, [112, 14]. In these models, the value of energy storage devices comes from buying low price energy and selling high price energy. However, in reality, electricity prices are affected by the supply and demand of energy. Hence, if storage devices are connected to the grid, they also balance supply and demand and have an influence on the electricity prices. While the influence on prices may be negligible for a single device acting in isolation, it would be imprudent to ignore the impact on prices of large ensembles of devices working in a coordinated fashion, e.g. to correct for imbalances.

To address a variety of questions related to the presence of energy storage, several market equilibrium models have been proposed [89, 109, 59, 26, 51, 50, 20, 3]. In these models, the prices are produced as a byproduct of balancing supply and demand, following the logic of the spot pricing of electricity [90]. However, a limitation of existing equilibrium models, sometimes apparent only when it comes to the numerical work, lies in the use of net demand curves whose evolution is deterministic, in contrast to the stochastic dynamic programming framework adopted here where the net demand curve is kept stochastic within the numerical solution algorithm. Keeping uncertainty at each stage of the multistage storage problem is critical to justify energy storage economically, otherwise conventional

generation can be planned to be started up with any lead time to adapt production and reserves to net demand variations at minimal cost.

The chapter is organized as follows. Section 4.2 provides an introduction on Nash equilibrium and bimatrix games. Section 4.2.3 formulates the Markov Game model. Section 4.3 provides the value iteration framework to obtain equilibrium policies for both players. Section 4.4 shows the numerical experiments and generalizes the basic model to implicit curve and non-perfect efficiency cases. Section 4.5 discusses the effect of incomplete information on charge level. Section 4.6 presents the impact from the ownership of energy storage. Section 4.7 concludes.

4.2 Technical Methods

We formulate our model as a Markov game in Section 4.2.4. During each time period in the Markov game, we deal with bimatrix games and find Nash equilibrium. In this section, we introduce related concepts and algorithms in the literature.

4.2.1 Markov Games

A typical Markov game $\Gamma = [S, N, \mathbf{A}, T, R, \gamma]$ includes state space S, a set of players $N = \{1, 2, ..., n\}$, a set of actions for each player in each state $\{A_{i,s}\}_{i\in N, s\in S}$, a transition function $T : S \times \mathbf{A} \times S \mapsto [0, 1]$ giving transition probabilities, a reward function $R : S \times \mathbf{A} \mapsto \mathbf{R}^n$ and a discount factor γ.

A stationary policy π_i for player i is a mapping $\pi_i : S\times\mathbf{A} \mapsto [0, 1]$ assigning probabilities to state-action pairs in the sense that in state s the action a is chosen with probability $\pi_i(s, a)$.

The value of the set of policies $\pi = \{\pi_i\}_{i\in N}$ for player i can be described by

$$V^i_\pi(s) = \sum_{a\in\mathbf{A}} \pi(s, a)Q^i_\pi(s, a) \ ,$$
$$Q^i_\pi(s, a) = R^i(s, a) + \gamma \sum_{s'\in S} T(s, a, s')V^i_\pi(s') \ .$$

A Nash equilibrium set of policies π^* will be such that

$$V_i^{\{\pi_1^*,\ldots,\pi_{i-1}^*,\pi_i^*,\pi_{i+1}^*,\ldots,\pi_n^*\}}(s) \geq V_i^{\{\pi_1^*,\ldots,\pi_{i-1}^*,\pi_i,\pi_{i+1}^*,\ldots,\pi_n^*\}}(s)$$

for all $s \in S$ and each $i \in N$.

Several algorithms have been proposed in the literature to solve Markov games. [70] discuss challenges to solve Markov games in multi-agent systems. [55] describes several Value Iteration based reinforcement-learning algorithms to find a Nash equilibrium in multiagent Markov games. Although some theoretical guarantees under certain assumptions are provided, convergence results in the general Markov game case remain unknown.

[113] show that value iteration may not be sufficient to learn an equilibrium policy in general Markov games. In 2012, [19] provide a polynomial-time algorithm called FolkEgal to find a Nash equilibrium for repeated two-players stochastic games.

4.2.2 Nash Equilibrium and Bimatrix Games

In this section, we discuss several algorithms that have been introduced to find Nash equilibria in finite bimatrix games.

A bimatrix game is a simultaneous game for two players where each player has a finite number of possible actions. There are two payoff matrices A, B for the two players. A, B are $m \times n$ matrices where m, n are cardinalities of action space of row player and column player. If the row player selects the i-th action and the column player selects the j-th action, the payoff to the row player is $A[i, j]$ and that to the column player is $B[i, j]$. The players can also play mixed strategies. A mixed strategy for row player is a probability vector x with length m such that $\sum_{i=1}^{m} x_i = 1$, and that for column player is a probability vector y with length n such that $\sum_{i=1}^{n} y_i = 1$. The expected payoff of the row player is $x^T Ay$ and the expected payoff of the column player is $x^T By$.

Assume we maximize the revenue for both players. If there is a pair (x^*, y^*) such that

$$(x^*)^T A y^* \geq x^T A y^*, \forall\ x\ ,$$
$$(x^*)^T B y^* \geq (x^*)^T B y, \forall\ y\ ,$$

then (x^*, y^*) is a Nash equilibrium.

A bimatrix game is called zero-sum game if $A + B = 0$. Otherwise, it is a non-zero-sum game.

In 1950, Nash proved that every finite game has a mixed strategy Nash equilibrium [68]. A variety of methods for finding a Nash equilibrium have been provided since then. [52] provide an algorithm to find a Nash equilibrium for a bimatrix games. This algorithm is a pivoting algorithm and referred to as the Lemke-Howson algorithm nowadays. The main idea is to formulate the problem as a special case of linear complementarity problem (LCP) and use the Lemke-Howson to solve the LCP. The LCP formulation is described as follows:

$$w = Mz + q,\ w \geq 0,\ z \geq 0,\ w^T z = 0\ ,$$

where

$$M = -\begin{bmatrix} 0 & A \\ B^T & 0 \end{bmatrix},\ q = \begin{bmatrix} e_m \\ e_n \end{bmatrix},\ w = \begin{bmatrix} e_m - Ay \\ e_n - B^T x \end{bmatrix},\ z = \begin{bmatrix} x \\ y \end{bmatrix}\ .$$

The vectors $e_m,\ e_n$ refer to the vectors of all $\mathbf{1}'$s of size m and n respectively. More details can be found in [66].

[74] provide a search method, referred to as Porter, Nudelman and Shoham (PNS) for computing a Nash equilibrium. The idea is to eliminate conditionally dominated actions and use heuristics to determine as quickly as possible the support of the mixed strategies, i.e. the set of decisions with a nonzero probability to be chosen. This algorithm can be generalized to $n-$players games.

[86] present a mixed integer programming (MIP) formulation to find a Nash equilibrium. Four different formulations are provided. Their first formulation makes it possible to specify a supplementary objective that can be used to select an equilibrium among the set of mixed equilibria, such as maximizing the payoff of a given player, or minimizing the payoff difference between the two players.

4.2.3 Model Description and Assumptions

In our model, there are two players in the market, the consumer and the supplier, both of whom control storage devices. There could be multiple storage devices, which are distributed. We aggregated them by owner type, i.e. consumer or supplier. Storage device owners influence the electricity price through operating their own storage devices until an equilibrium price is attained that balances supply and demand of power including the net power injection from storage.

Our basic setting corresponds to a full information stochastic game, which can be generalized to incomplete information game, see Section 4.5.

The full information game is played as follows. The initial state X_0 is given. At stage $t = 0, 1, ...,$ the current state X_t is revealed to both players. The state consists of state variables described in Section 4.2.4. Then both players make their decisions at the same time. Based on the current state and decisions from both players, the next state X_{t+1} is drawn according to a transition function and corresponding probability and . Rewards for both players at stage t are computed and then the game proceeds to stage $t + 1$.

First, we describe the supply curve, demand curve as follows:

$$S_t = f_s(P_t), \quad D_t = f_d(P_t),$$

where P_t is the price at time t. The quantity of energy[1] change for the storage is

$$H_t = h(P_t) = S_t - D_t \ . \tag{4.1}$$

Note that if $H_t > 0$, the storage device is being charged while $H_t < 0$ means that is being discharged. Since we have two players in the market, the value of energy change in equation (4.1) is aggregated, consisting of two parts: energy from supplier's storage H_t^s and that from consumer's storage H_t^c,

$$H_t = H_t^s + H_t^c \ .$$

We assume that the supply curve s is nondecreasing in the price and the demand curve d is nonincreasing in the price. This assumption is natural since in general, if price increases, there is less demand and more supply. Based on the equation (4.1), the function h giving the net energy being charged is nondecreasing in the price. We strengthen these conditions by requiring that h is continuous and increasing in price P_t. It then follows that h has an inverse h^{-1} which is continuous and increasing in H_t. The increasing property means that higher prices are obtained with more energy withdrawn to charge the storage devices. The continuity assumption implies that any target price in a price interval can be obtained by adjusting the quantity of energy withdrawn for charging storage devices.

The demand curve s is assumed to depend on exogenous random variables. One example is provided in Section 4.2.5.

4.2.4 Model Formulation

In this section, we formulate the mathematical model. The problem is described using the following notation.

- $\mathscr{X}$ is the state space. The state at time t, denoted X_t, has three components: the

[1]We think of supply and demand as power. Then H_t is the total net power that is withdrawn from the grid by the storage devices. We need exact balance at all times, hence (4.1). The model is simplified by assuming the prices and power injections or withdrawals are held constant during a period of time, then one can view S_t and D_t as power integrated over one time period, that is, energy.

stored energy level for consumer X_t^c and the stored energy level for supplier X_t^s and a demand curve state W_t. Thus, $X_t = (X_t^c, X_t^s, W_t)$. As mentioned earlier, it's a full information game where each player can observe the full state.

- $\mathscr{H} = \mathscr{H}^c \times \mathscr{H}^s$ is the decision space. The decision at time t, denoted H_t, has two components: the storage decision from consumer H_t^c and supplier H_t^s. From the discussion later in section (4.3), we may have mixed strategies (probability vectors for possible actions), denoted as (u^c, u^s). Let $U^c : \mathscr{X} \times \mathscr{H}^c \mapsto (0, 1]$ be the decision probability function of consumer c, such that $p = U^c(x, H^c)$ is the probability that consumer c choose action H^c while being in state x. Let $U^s : \mathscr{X} \times \mathscr{H}^s \mapsto (0, 1]$ be the decision probability function of supplier s, such that $p = U^s(x, H^s)$ is the probability that supplier s choose action H^s while being in state x. Assuming the decision space $\mathscr{H}$ is discrete and consumer's action $H^c \in \{H_1^c, H_2^c, ..., H_m^c\}$ (m possible actions in total), then consumer's probability vector u^c can be computed as $u_i^c = U^c(x, H_i^c), i = 1, 2..., m$. Similar results can be obtained for supplier, $u_i^s = U^s(x, H_i^s), i = 1, 2, ..., n$.

- $P : \mathscr{X} \times \mathscr{H} \times \mathscr{X} \mapsto [0, 1]$ is the state transition probability function, such that $\text{Prob}(X_{t+1} = x' \mid X_t = x, H_t = a) = P(x, a, x')$.

- $R^c : \mathscr{X} \times \mathscr{H} \mapsto \mathbb{R}$ is the reward function for consumer such that the reward at time t given $X_t = x$, $H_t^c = a^c$, $H_t^s = a^s$ is $r_t^c = R^c(x, a^c, a^s)$.

  $R^s : \mathscr{X} \times \mathscr{H} \mapsto \mathbb{R}$ is the reward function for supplier such that the reward at time t given $X_t = x$, $H_t^s = a^c$, $H_t^s = a^s$ is $r_t^s = R^s(x, a^c, a^s)$.

  The detail of definition of reward function R^c, R^s is described later in this section.

The charge levels of consumer and supplier are kept track of through X_t^c, X_t^s, which represents the quantities of stored energy at the beginning of each time period t. Then, the different between two consecutive time periods is

$$X_{t+1}^c - X_t^c = g^c(H_t^c) \ ,$$
$$X_{t+1}^s - X_t^s = g^s(H_t^s) \ .$$

In particular, if the storage devices have perfect efficiency, we have $g^c(H_t^c) = H_t^c$ and $g^s(H_t^s) = H_t^s$. The generalized inefficiency model is discussed further in Section 4.4.5.

With storage capacities K^c and K^s for consumer's and supplier's storage devices and since charge levels must be nonnegative, it is straightforward that $0 \leq X_t^c \leq K^c$ and $0 \leq X_t^s \leq K^s$. We scaled the energy units by X_t^c, X_t^s, K^c, K^s so that one unit of power during one time period corresponds to one unit of energy.

Both players want to maximize their expected reward along the infinite horizon

$$V^c_{\pi_c,\pi_s} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma R_t^c \mid X_0\right],$$
$$V^s_{\pi_c,\pi_s} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma R_t^s \mid X_0\right].$$

We are looking for policies π_c^*, π_s^* such that

$$V^c_{\pi_c^*,\pi_s^*} \geq V^c_{\pi_c,\pi_s^*} \qquad \forall \pi_c \,,$$
$$V^s_{\pi_c^*,\pi_s^*} \geq V^c_{\pi_c^*,\pi_s} \qquad \forall \pi_s.$$

This is a stochastic dynamic program over an infinite horizon where $\gamma \in (0,1)$ is the discount factor and X_0 is the initial state.

Define demand reward R_t^d, consumer's storage reward R_t^{cs}, generation reward R_t^g and supplier's storage reward R_t^{ss} as

$$R_t^d = \textstyle\int_0^{D_t} [{f_d}^{-1}(q) - P_t] dq \,,$$
$$R_t^g = \textstyle\int_0^{S_t} [P_t - {f_s}^{-1}(q)] dq \,,$$
$$R_t^{cs} = -P_t H_t^c \,,$$
$$R_t^{ss} = -P_t H_t^s \,,$$

where ${f_d}^{-1}, {f_s}^{-1}$ are inverse function of f_d, f_s defined in Section 4.2.3.

Then consumer's reward function R_t^c is the sum of demand reward R_t^d and consumer's storage reward R_t^{cs} while supplier's reward function R_t^s is the sum of generation reward R_t^g

and supplier's storage reward R_t^{ss}

$$R_t^c = R_t^d + R_t^{cs} \ , \tag{4.2}$$
$$R_t^s = R_t^g + R_t^{ss} \ .$$

4.2.5 Demand and Supplier Curve

In theory, the curve functions related to demand and supply can be either closed-form functions or implicit function. In our basic model, we assume a closed-form expression for the equilibrium price as a function of the demand and supply. In particular, the inverse function of supply curve is quadratic and that of demand curve is linear, as follows:

$$f_s{}^{-1}(q) = cq + dq^2, \quad f_d{}^{-1}(q) = b - aq \ ,$$

where $a, c, d > 0$ are fixed parameters and b is an random variable with an i.i.d Normal distribution. The detail of analytic solutions for equilibrium price and reward functions can be found in Appendix C.1.

More general, if the curves are represented in implicit functions, we can still find the equilibrium price numerically, which is described in Section 4.4.4.

4.3 Algorithms

In the example mentioned in 4.2.5, we assume the actions for both players are known. In this section, we show how these actions can be obtained by finding a Nash Equilibrium. As we mentioned in section 4.2, a mixed Nash Equilibrium strategy is guaranteed to exist in finite games.

For most one-objective function MDP problems, we enumerate all the possible actions for each state and pick the most profitable action. Since we have two players in our problem, we find the optimal actions for both players through finding the Nash equilibrium. We discretize the state space, including the charge levels for both storage devices as well as the distribution parameter for demand curve. With discretized charge levels, we obtain

discretized action space for both players.

We provide the Value Iteration framework to solve the Markov game. However, value iteration schemes are not guaranteed to find an equilibrium for non-zero sum Markov games [113]. In Section 4.4.3, we show numerically the gap between the equilibrium policy and the sub-optimal policy we get is very small.

1. Initialization: Guess initial value functions $V_0^c(x), V_0^s(x)$ for all possible states (for instance, $V_0^c(x) \equiv 0, V_0^s(x) \equiv 0$ for $\forall\ x$). Set iteration $k = 1$.

2. For all state x and actions H_i^c, H_j^s (i is the index for consumer's action and j is that for supplier), compute the pure-strategy cumulated cost-to-go values

$$\begin{aligned} A_{k,ij}(x) &= R^c(x, H_i^c, H_j^s) + \gamma \mathbb{E}[V_{k-1}^c(x')|x, H_i^c, H_j^s] \ , \\ B_{k,ij}(x) &= R^s(x, H_i^c, H_j^s) + \gamma \mathbb{E}[V_{k-1}^s(x')|x, H_i^c, H_j^s] \ . \end{aligned} \tag{4.3}$$

3. For each state x, find the optimal strategies (probability vectors) $u^c(x), u^s(x)$ for both players by solving the bimatrix game with payoff matrices (A_k, B_k), see Section 4.2.

Then the value functions can be updated based on

$$\begin{aligned} V_k^c(x) &= (u^c(x))^\top A_k(x) u^s(x) \ , \\ V_k^s(x) &= (u^c(x))^\top B_k(x) u^s(x) \ . \end{aligned} \tag{4.4}$$

4. Set $k \leftarrow k + 1$ and repeat Step 2 until a maximum iteration number is reached. We obtain the policy U^c, U^s.

We use the PATH Solver [22] to find Nash equilibrium in our Value Iteration framework. PATH is provided by Ferris et al., which solves linear complementarity problem. This solver is efficient but it stops whenever it finds a Nash equilibrium (Multiple Nash equilibriums are possible for some A, B) and due to the LCP formulation, there is no easy way of specifying a selection process when several equilibria exist.

4.4 Numerical Experiments

In this section, we show numerical results on a linear/piecewise linear curves case and discuss the sub-optimality of policies obtained. Another simpler case study is shown in C.2

4.4.1 Parameters setting

We use a linear demand curve and piecewise linear supply curve,

$$d^{-1}(q) = b - 60q$$

where b has a Normal Distribution $\mathcal{N}(800, 60^2)$, which is approximated by a discrete distribution with 61 different states.

Capacities for both storages are $K_c = K_s = 0.4$, discretized with step length 0.05.

4.4.2 Numerical Results

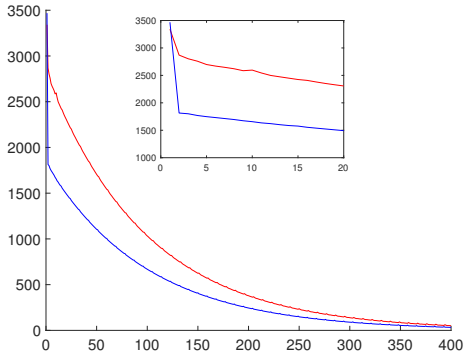


Figure 4.1: Bellman residual for both players as a function of iteration. *red*: consumer. *blue*: supplier. Inset graph: first 20 iterations.

Figure (4.1) shows the difference between two consecutive stages' value functions for both players. In this particular case, the policy stabilizes after first 20 iterations.

Figure (4.2) shows charging amount and (4.3) show the next charge levels for both players when the value iteration algorithm stops. When mixed strategy occurs, the weighted mean of decision is shown. These policies are highly dependent on the curvature of demand

and supplier curves. It instructs the decision makers how they should behave in different situations.

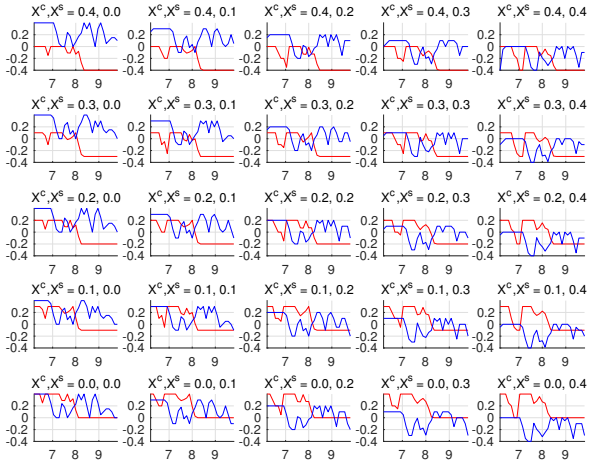


Figure 4.2: Charging amount for both players (y-axis) as a function of the demand curve level (parameter b as x-axis). *red*: consumer. *blue*: supplier.

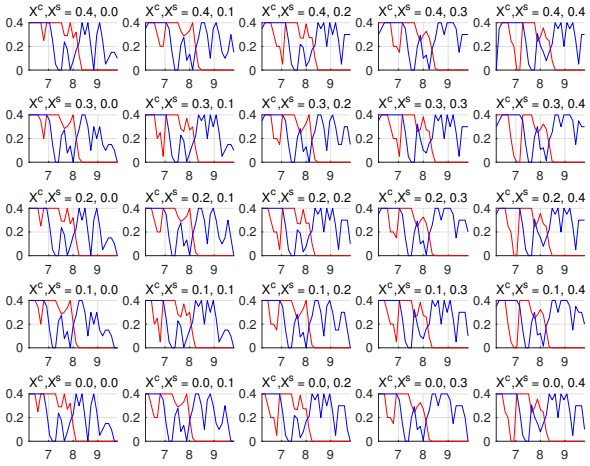


Figure 4.3: Next charge levels for both players (y-axis) as a function of the demand curve level (parameter b as x-axis). *red*: consumer. *blue*: supplier.

4.4.3 Sub-optimality of Policy

After the value iteration stops, we obtain policy U^c, U^s for both players. With these fixed policies, we can compute the value of the policy for consumer and supplier as $V^c(U^c, U^s)$

and $V^s(U^c, U^s)$. By the definition of Nash equilibrium policy, if U^c and U^s are equilibrium policy, we should have

$$V^c(U^c, U^s) \geq V^c(\hat{U}^c, U^s) \ ,$$
$$V^s(U^c, U^s) \geq V^s(U^c, \hat{U}^s) \ ,$$

where $\hat{U}^c$ and $\hat{U}^s$ are some policies other than U^c and U^s.

When we fix the policy for supplier U^s, we can compute the optimal policy for consumer. This is equivalent to a traditional one-player game where optimal policy is guaranteed through value iteration. We denote the optimal policy as $\hat{U}^c$ and the new value of policy is $V^c(\hat{U}^c, U^s)$.

Then we compute the difference

$$V^c(\hat{U}^c, U^s) - V^c(U^c, U^s) \ .$$

Table 4.1: fix supplier's policy

	old policy	new policy	difference	percent
consumer	279142.73	279142.78	+0.05	0%
supplier	180706.58	180671.11	-35.47	-0.02%

In Table 4.1, The gain for consumer is very small while the loss for supplier is around 0.02%. This means both players have little intention to change their current policies.

Similarly, if we fixed consumer's policy U^c, we compute the optimal policy for supplier, denoted as $\hat{U}^s$. Then we compute the difference

$$V^s(U^c, \hat{U}^s) - V^s(U^c, U^s) \ .$$

Table 4.2: fix consumer's policy

	old policy	new policy	difference	percent
consumer	279142.73	279132.41	-10.32	-0.004%
supplier	180706.58	180706.64	+0.06	0%

Hence, we can conclude that both players are very unlikely to deviate from the current policy and the policies from our framework is very close to the Nash equilibrium policies.

4.4.4 Extensions: Implicit Curve for Demand and Supply

In Section 4.2.5, we assume the demand and response curve are given explicitly. So we can compute the price and rewards in closed form. In general, we may find the equilibrium price P_t by solving equation (4.1)

$$H_t = f_s(P_t) - f_d(P_t)$$

numerically, e.g. by zero-finding using Brents method. A data driven example is provided in Section 4.6.

4.4.5 Extensions: Storage Devices with Non-Perfect Efficiency

In the basic model in Section 4.2.4, we assume the storage devices have perfect efficiency. We can generalize this model into inefficiency case easily.

$$g^c(y) = \begin{cases} \eta^c_{ch} y & y \geq 0 \\ \frac{1}{\eta^c_{dis}} y & y < 0 \end{cases}, \; g^s(y) = \begin{cases} \eta^s_{ch} y & y \geq 0 \\ \frac{1}{\eta^s_{dis}} y & y < 0 \end{cases}. \tag{4.5}$$

$\eta^c_{ch}, \eta^c_{dis} \in (0, 1]$ are consumer's efficiency parameters for charging and discharging respectively while $\eta^s_{ch}, \eta^s_{dis} \in (0, 1]$ are those of supplier's. Thus, different storage technologies can be modeled with different efficiency parameters.

4.5 Incomplete Information Game

The Markov game described in Section 4.2.4 is a full information game, where both players know the supply curve, the demand curve and charge levels of both storages. In this section, we investigate the consequences of not sharing the charge level information.

4.5.1 Model Description and Assumption

This section is based on Section 4.2.4. The storage capacity K^i are still common knowledge. The main difference is that the charge level is not shared to the other player, which makes the problem no longer a full information game. The game is played as follows.

1. As the game begins, each player starts with his own charge level $c^i_{t=1}$ (where i is the index for players). For player i, he also maintains an uncertainty set $[lb_1^{-i}, ub_1^{-i}] = [0, K^{-i}]$ of the other player $-i$.

2. At time period t, the demand curve and the supply curve information are revealed. Both players make their decisions at the same time, then they are able to observe the price (which depends on their actions).

3. Based on the new information (and older observation), they may update their uncertainty set $[lb^i_t, ub^i_t]$.

4. Based on transition functions, charge levels are updated. In addition, the rewards for both players are computed and the game proceeds to the next time period $t+1$.

4.5.2 Estimation of the Other Player's Charge Level

At time period t, after the players make their decisions H^c_t, H^s_t, they observe the equilibrium price P_t . Based on equation (4.1)

$$H^c_t + H^s_t = H_t = S_t - D_t = f_s(P_t) - f_d(P_t) \ ,$$

where f_s and f_d are known supply and demand curve, they can compute the other player's action (in closed form or numerically). Note that the other player's charge level is still

unknown.

However, based on the uncertainty charge level set and actions of the other player, it is possible for each player to estimate the other player's charge level. For simplicity, in the following context, we assume we plays as the consumer and the following algorithm shows how one can get a good estimation of the supplier's charge level.

Table 4.3: Update Uncertainty Set Algorithm

Algorithm: Updated Uncertainty Set
Data: Capacity K^s, Uncertainty Set $[lb_t, ub_t]$,
Action Quantity H_t^s
Update:
$lb_{t+1} = \max\{0, \min\{lb_t + H_t^s, K^s\}\}$
$ub_{t+1} = \max\{0, \min\{ub_t + H_t^s, K^s\}\}$

With the Update Uncertainty Set algorithm above, we have following propositions:

Proposition 16. *The uncertainty charge level set is non-expansive, i.e.,*

$$ub_{t+1} - lb_{t+1} \leq ub_t - lb_t \ . \tag{4.6}$$

Proof. We discuss the possible H_t^s case by case

- If $H_t^s = 0$, then $ub_{t+1} = ub_t$ and $lb_{t+1} = lb_t$. Inequality (4.6) holds as equality.
- If $H_t^s > 0$, then $lb_{t+1} = lb_t + H_t^s$. Since $ub_{t+1} = \min\{ub_t + H_t^s, K^s\}$
  - if $ub_{t+1} = ub_t + H_t^s$, inequality (4.6) holds as equality.
  - if $ub_{t+1} = K^s < ub_t + H_t^s$, clearly, $ub_{t+1} - lb_{t+1} < ub_t - lb_t$.
- If $H_t^s < 0$, then $ub_{t+1} = ub_t + H_t^s$. Since $lb_{t+1} = \max\{lb_t + H_t^s, 0\}$
  - if $lb_{t+1} = lb_t + H_t^s$, inequality (4.6) holds as equality.
  - if $lb_{t+1} = 0 > lb_t + H_t^s$, clearly, $ub_{t+1} - lb_{t+1} < ub_t - lb_t$.

Hence, we finish our proof. □

Note that we don't have to store all the historical actions of the other player if we maintain our uncertainty set every time period.

Proposition 17. *For some finite time period T, if*

$$|\sum_{t=1}^{T} H_t^s| \geq \epsilon$$

for some $\epsilon > 0$, then we can guarantee that there exist $\delta = K^s - \epsilon$ such that

$$ub_T - lb_T \leq \delta \ ,$$

we call this property almost sure contraction.

In particular, if $\epsilon = K^s$, which indicates that $\delta = 0$, $ub_T = lb_T$, the charge level of the other player is known exactly.

Proof. Assume at time period $t = 1$, we have uncertainty set $[lb, ub]$ where $lb \geq 0, ub \leq K^s$, then at time period T, we have uncertainty set

$$\Omega = [lb + \sum_{t=1}^{T} H_t^s, ub + \sum_{t=1}^{T} H_t^s] \ .$$

- If $\sum_{t=1}^{T} H_t^s > 0$, then the lower bound $lb_T = lb + \sum_{t=1}^{T} H_t^s \geq 0 + \epsilon$ while upper bound $ub_T \leq K^s$. Hence, the range of the uncertainty set $ub_T - lb_T \leq K^s - \epsilon = \delta$.
- If $\sum_{t=1}^{T} H_t^s < 0$, then the lower bound $lb_T \geq 0$ while upper bound $ub_T = ub + \sum_{t=1}^{T} H_t^s \leq ub - \epsilon \leq K^s - \epsilon$. Hence, the range of the uncertainty set $ub_T - lb_T \leq K^s - \epsilon = \delta$.

The $\delta = 0$ case means that during the T time periods, if the cumulated action $\sum_{t=1}^{T} H_t^s$ of the supplier is K^s (charged an amount corresponding to his capacity K^s) or $-K^s$ (discharged an amount corresponding to his capacity K^s), then we can conclude immediately his charge level is K^s (if charge) or 0 (if discharge). □

4.5.3 Impact of incomplete information

In last section, we discuss that as a player (consumer), how we can estimate the other player's (supplier's) charge level as the game proceed. This raises a question whether knowing this information will affect the overall rewards for each player and the total welfare.

Pessimistic Decision

From proposition 17, we can see that as the game proceeds long enough, both players are most probably certain about the other player's charge level, making the problem as a full information game, which has been studied in Section 4.2.4. Here, we are interested in the time periods where information is incomplete. We assume the two players adopt a pessimistic approach based on their uncertainty set when they are not sure about the other player's charge level.

Assume we've computed the pay-off matrices $A(s), B(s)$ for each state $s = (x^c, x^s, w)$, we show how we select our pessimistic action as a consumer.

- Our current charge level x_c is known as well as the demand curve information w. We have the uncertainty set of supplier's charge level $\hat{X}^s = [lb^s, ub^s]$.
- For each $x^s \in \hat{X}^s$, we obtain the pay-off matrix $A(x^s)$, which depends on x^s. For each fixed x^s, the pay-off matrix $A(x^s)$ has dimension $m \times n$, where m, n are numbers of possible actions for consumer and supplier respectively, $m = |\mathscr{H}^c|, n = |\mathscr{H}^s|$. Hence, the matrix of $A(\hat{X}^s)$ has dimension $m \times n \times k$, where $k = |\hat{X}^s|$, the cardinality of the uncertainty set of supplier's charge level.
- The pessimistic action for consumer is

$$H^{c*} = \arg \max_{H^c \in \mathscr{H}^c} \min_{H^s \in \mathscr{H}^s, x^s \in \hat{X}^s} A(\hat{X}^s) \ .$$

Similarly, we can obtain the pessimistic action for supplier

$$H^{s*} = \arg \max_{H^s \in \mathscr{H}^s} \min_{H^c \in \mathscr{H}^c, x^c \in \hat{X}^c} B(\hat{X}^c) \ .$$

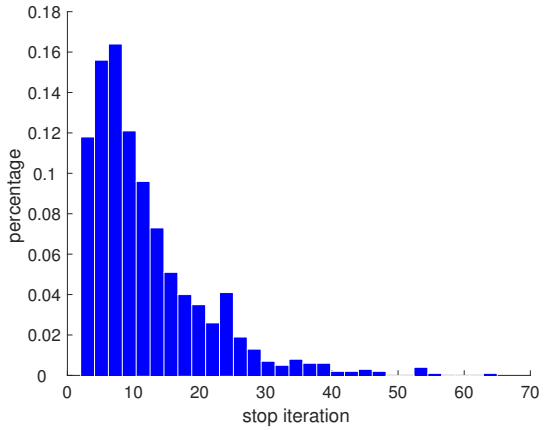


Figure 4.4: Distribution of time duration of incomplete information game

If the consumer takes the pessimistic action, he is guaranteed to get pay-off at least A_p no matter which charge level the supplier is at and what he chooses, where $A_p = \max_{H^c \in \mathscr{H}^c} \min_{H^s \in \mathscr{H}^s, x^s \in \hat{X}^s} A(\hat{X}^s)$.

Simulation and Numerical Results

We use similar parameter setting from Section 4.4.1. The initial uncertainty set is the interval $[0, K^i]$. The simulation is described as follows:

1. At each time period t, the random parameter b is drawn according to the discretized normal distribution approximation.

2. For uncertainty case, players adopt pessimistic actions described in section 4.5.3, uncertainty set are updated.

3. For certainty case, players adopt actions based on the optimal policy.

4. If both uncertainty set have length 1, simulation stops (game enters full information phase).

Figure 4.4 shows the distribution of time duration of incomplete information game. We can see the two players basically know each other's charge level after 40 time periods.

Table 4.4 shows the comparison of rewards for both players between certainty case (full information) and uncertainty case (incomplete information). In particular, this simulation

stops at iteration 10.

Table 4.4: Reward comparison between full and incomplete information game

	Full	Incomplete
consumer	24.86	24.88
supplier	14.82	14.81

There are no big difference (less than 0.1%) in the cumulated rewards for certainty case and uncertainty case for both players.

Figure 4.5 shows an example of the evolution of uncertainty set. We stopped the simulation when both players know the other player charge level exactly.

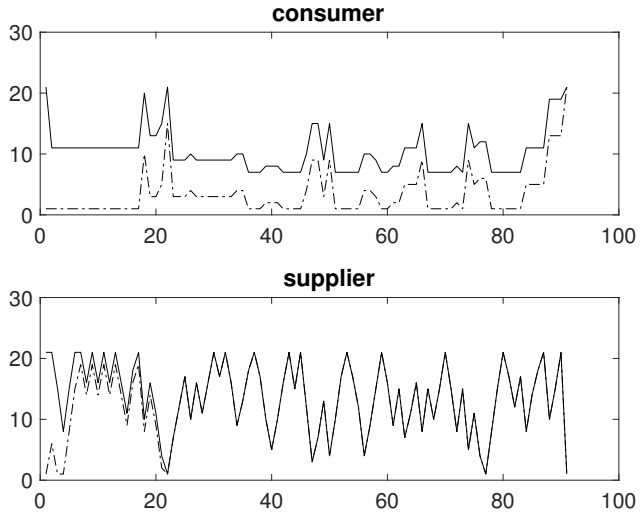


Figure 4.5: Evolution of uncertainty set

4.5.4 Summary

Based on the numerical results, we conclude that

- The cumulated rewards of each player doesn't have much difference even if they don't know the other player's charge level at the beginning.
- This difference will be diminished even more as the game enters the full information phase.

- Overall, keeping the charge level as long as possible may benefit a little, but not much (as we can see in Table 4.4), it is not recommended to keep this information private.

4.6 Impact of storage ownership on the price

In this section, we discuss the impact of the storage ownership on the price. For the extreme cases, where only one of them controls energy storage device, the device owner holds the whole power to affect the price by operating the device. However, when both of them control some storage devices, they try to affect the price to achieve their best interests.

4.6.1 Parameter setting

In our numerical work, we assume there are 8 units of storage in total. The parameters are the same as those in Section 4.4.1. We discuss 9 different cases, where consumer-supplier ownership are $[.8, 0], [.7, .1], [.6, .2], \ldots, [0, .8]$.

4.6.2 Numerical Results

All the following rewards are approximated on simulation with time period length $T = 200$,

$$\bar{R}_t = \frac{1}{1-\gamma^T}\mathbb{E}\sum_{t=0}^{T-1}\gamma^t R_t \ .$$

With different distribution of ownership, the price volatility is shown in Fig 4.6, with the increases of supplier's ownership, the volatility of price increases.

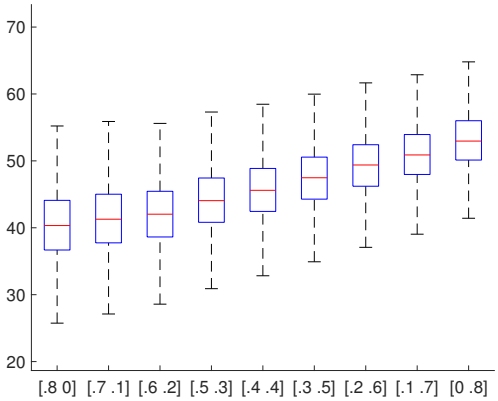
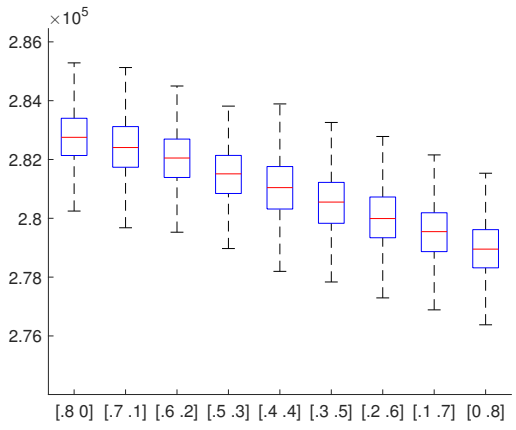
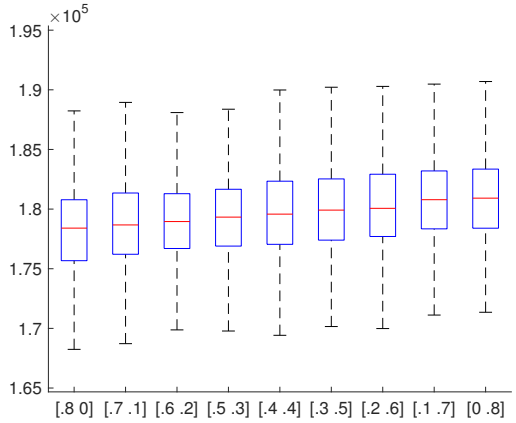


Figure 4.6: price volatility

The total cumulated rewards are shown in Fig. 4.7. With more storage ownership over his competitor, the player increases his reward. Compared to consumer, when supplier controls more storage, he intends to increase the variance of price. We believe this is due to the very steep curvature for the supplier curve. Instead, in this simulation, consumer favors more stable price.



consumer



supplier

Figure 4.7: Cumulated reward as a function of storage ownership

4.7 Conclusion

This chapter considers the impact of distributed energy storage ownership under uncertainty on the price. The problem is formulated as an infinite-horizon Markov Game with a random demand curve. A value iteration framework is provided to find the sub-optimal equilibrium policies for both players. The actions at each iteration are obtained by solving bimatrix games. A computational study demonstrates the gap between the policies obtained from our approach and the Nash equilibrium policies is very small. Another computational study shows that the benefit from hiding charge level information is not significant. Thus, it is suggested to share the information with each other. The impact

of storage ownership study offers insights about how much storage a player will need if he wants to make a real impact on the price and his profits.

Chapter 5

Conclusion

In this dissertation, we present models for stochastic optimal control of grid-level storages, especially battery storages. In the basic storage operation problem, storage device owner is the decision maker. The owner makes profit by buying and selling energy from the grid, given the price varies during different time periods. The stochasticity comes from the uncertainty of price or the uncertainty of demand and supplier. A popular way to model a storage operation problem is Markov Decision Process (MDP). An infinite horizon MDP with discount factor can be formulated naturally and the objective function is to maximize the expected discounted total profit for the device owner. Classical algorithms for solving MDP problems include value iteration and policy iteration.

We propose several MDP-based models in different market settings and provide related algorithms and analysis to find the optimal or sub-optimal policies. Corresponding numerical experiments perform well and give us more insights about the model.

- In Chapter 2, we extend basic MDP model into p-periodic MDP model, which is suitable for mitigating end-of-horizon effects. The idea is to recalibrate the model every p time periods in a rolling-horizon fashion. We provide a tighter bound with stationary value functions on an augmented state space than general bounds. An implementation on grid-level storage operation problem where price follows daily patterns is studied.

- In Chapter 3, battery degradation (aging phenomenon) is incorporated in the MDP

model. We introduce an extra state variable remaining life cycle to denote the aging status of the battery. Electricity price is assumed to follow a given independent and identically distributed distribution or a Markovian regime-switching process. Compared to value iteration, by utilizing the problem structure, we provide a faster and more accurate algorithm to solve the problem. The algorithm returns the optimal policies by solving a sequence of quasiconvex optimization problems.

- In Chapter 4, instead of the single player in previous two chapters, we consider two-players problems. Electricity price now not only depends on the relationship between demand and supplier, which is stochastic, but also on the decisions of two storage device owners. The problem is formulated as an infinite-horizon Markov game and a value iteration framework is proposed to find the sub-optimal policies for both players. In each iteration, on contrast to return the optimal action by maximizing the Q function in MDP model, we solve a bimatrix game in Markov game. In addition, we also study a related incomplete information game and the impact of repartition of the energy storage.

Bibliography

[1] U.S. Energy Information Administration. Total electric power industry & renewable sources database. `http://www.eia.gov/electricity/monthly/`. Accessed: 2017-01-18.

[2] A.A. Akhil, G. Huff, A.B. Currier, B.C. Kaun, D.M. Rastler, S.B. Chen, A.L. Cotter, D.T. Bradshaw, and W.D. Gauntlett. DOE/EPRI electricity storage handbook in collaboration with NRECA. Technical Report SAND2015-1002, Sandia National Laboratories, January 2015.

[3] Ahmed SA Awad, J David Fuller, Tarek HM El-Fouly, and Magdy MA Salama. Impact of energy storage systems on electricity market equilibrium. *IEEE Transactions on Sustainable Energy*, 5(3):875–885, 2014.

[4] Sandeep Bala. Energy storage systems in the distribution grids of the future. http://www.lehigh.edu/engineering/news/events/ine/pdf/Bala.pdf. Accessed: 2014-06-01.

[5] John P Barton and David G Infield. Energy storage and its use with intermittent renewable energy. *Energy Conversion, IEEE Transactions on*, 19(2):441–448, 2004.

[6] D. P. Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, Nashua, NH, 2013.

[7] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 2007.

[8] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, third edition, 2005.

[9] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[10] Kyle Bradbury, Lincoln Pratson, and Dalia Patiño-Echeverri. Economic viability of energy storage systems based on price arbitrage potential in real-time us electricity markets. *Applied Energy*, 114:512–519, 2014.

[11] Ted KA Brekken, Alex Yokochi, Annette Von Jouanne, Zuan Z Yen, Hannes Max Hapke, and Douglas A Halamay. Optimal energy storage sizing and control for wind power applications. *Sustainable Energy, IEEE Transactions on*, 2(1):69–77, 2011.

[12] M Broussely, Ph Biensan, F Bonhomme, Ph Blanchard, S Herreyre, K Nechev, and RJ Staniewicz. Main aging mechanisms in li ion batteries. *Journal of Power Sources*, 146(1):90–96, 2005.

[13] Markus Burger, Bernhard Klar, Alfred M ller, and Gero Schindlmayr. A spot market model for pricing derivatives in electricity markets. *Quantitative Finance*, 4(1):109–122, 2004.

[14] René Carmona and Michael Ludkovski. Pricing asset scheduling flexibility using optimal switching. *Applied Mathematical Finance*, 15(5-6):405–447, 2008.

[15] Juan Manuel Carrasco, Leopoldo Garcia Franquelo, Jan T Bialasiewicz, Eduardo Galván, RC Portillo Guisado, Ma AM Prats, José Ignacio León, and Narciso Moreno-Alfonso. Power-electronic systems for the grid integration of renewable energy sources: A survey. *Industrial Electronics, IEEE Transactions on*, 53(4):1002–1016, 2006.

[16] K Mani Chandy, Steven H Low, Ufuk Topcu, and Huan Xu. A simple optimal power flow model with energy storage. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 1051–1057. IEEE, 2010.

[17] Sandia Corporation. Doe international energy storage database. http://www.energystorageexchange.org/projects. Accessed: 2017-01-21.

[18] James Cruise, Lisa Flatley, Richard Gibbens, and Stan Zachary. Optimal control of storage incorporating market impact and with energy applications. *arXiv preprint arXiv:1406.3653*, 2014.

[19] Enrique Munoz De Cote and Michael L Littman. A polynomial-time nash equilibrium algorithm for repeated stochastic games. *arXiv preprint arXiv:1206.3277*, 2012.

[20] Pedro Crespo Del Granado, Stein W Wallace, and Zhan Pang. The value of electricity storage in domestic homes: a smart grid perspective. *Energy Systems*, 5(2):211–232, 2014.

[21] Maria Dicorato, Giuseppe Forte, Mariagiovanna Pisani, and Michele Trovato. Planning and operating combined wind-storage system in electricity market. *Sustainable Energy, IEEE Transactions on*, 3(2):209–217, 2012.

[22] Steven Dirkse, Michael Ferris, and Todd Munson. The path solver. http://pages.cs.wisc.edu/ ferris/path.html.

[23] KC Divya and Jacob Østergaard. Battery energy storage technology for power systemsan overview. *Electric Power Systems Research*, 79(4):511–520, 2009.

[24] Dennis Doerffel and Suleiman Abu Sharkh. A critical review of using the peukert equation for determining the remaining capacity of lead-acid and lithium-ion batteries. *Journal of Power Sources*, 155(2):395–400, 2006.

[25] A.L. Dontchev and R.T. Rockafellar. *Implicit Functions and Solution Mappings*. Springer, New York, NY, 2009.

[26] Easan Drury, Paul Denholm, and Ramteen Sioshansi. The value of compressed air energy storage in energy and reserve markets. *Energy*, 36(8):4959–4973, 2011.

[27] B. Dunn, H. Kamath, and J. M. Tarascon. Electrical energy storage for the grid: A battery of choices. *Science*, 334(6058):928–935, 2011.

[28] Tomaso Erseghe, Andrea Zanella, and Claudio G Codemo. Optimal and compact control policies for energy storage units with single and multiple batteries. *IEEE Transactions on Smart Grid*, 5(3):1308–1317, 2014.

[29] Alvaro Escribano, J Ignacio Peña, and Pablo Villaplana. Modelling electricity prices: International evidence. *Oxford bulletin of economics and statistics*, 73(5):622–650, 2011.

[30] Jim Eyer and Garth Corey. Energy storage for the electricity grid: Benefits and market potential assessment guide. *Sandia National Laboratories*, 20(10):5, 2010.

[31] Federal Energy Regulatory Commission. *Order Approving Stipulation and Consent Agreement, In Re Make-Whole Payments and Related Bidding Strategies*, July 2013. Docket Nos. IN11-8-000, IN13-5-000.

[32] H.J. Greenberg and W.P. Pierskalla. Quasi-conjugate functions and surrogate duality. *Cahiers du Centre d'Etudes de Recherche Opérationnelle*, 15:437–448, 1973.

[33] Imre Gyuk, P Kulkarni, JH Sayer, JD Boyes, GP Corey, and GH Peek. The united states of storage [electric energy storage]. *Power and Energy Magazine, IEEE*, 3(2):31–39, 2005.

[34] Ioannis Hadjipaschalis, Andreas Poullikkas, and Venizelos Efthimiou. Overview of current and future energy storage technologies for electric power applications. *Renewable and Sustainable Energy Reviews*, 13(6):1513–1522, 2009.

[35] Lars Peter Hansen and Thomas J Sargent. *Recursive models of dynamic linear economies*. Princeton University Press, 2013.

[36] D.P. Heyman and M.J. Sobel. *Stochastic Models in Operations Research*, volume II: Stochastic Optimization. Dover, Mineola, NY, 2003.

[37] Mike Hoffman, Michael Kintner-Meyer, John DeSteese, and Artyom Sadovsky. Analysis tools for sizing and placement of energy storage in grid applications. In *ASME*

2011 5th International Conference on Energy Sustainability, pages 1565–1573. American Society of Mechanical Engineers, 2011.

[38] Y. Hu and B. Defourny. Near-optimality bounds for greedy periodic policies with application to grid-level storage. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL-2014)*, pages 1–8, December 2014.

[39] Yuhai Hu and Boris Defourny. Near-optimality bounds for greedy periodic policies with application to grid-level storage. In *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014 IEEE Symposium on*, pages 1–8. IEEE, 2014.

[40] Yuhai Hu and Boris Defourny. Optimal price-threshold control for battery operation with aging phenomenon: a quasiconvex optimization approach. *Annals of Operations Research*, pages 1–28, 2017.

[41] M. Jacobson, N. Shimkin, and A. Shwartz. Markov Decision Processes with slow scale periodic decisions. *Mathematics of Operations Research*, 28(4):777–800, 2003.

[42] Ami Joseph and Mohammad Shahidehpour. Battery storage systems in electric power systems. In *Power Engineering Society General Meeting, 2006. IEEE*, pages 8–pp. IEEE.

[43] Kenneth L Judd. *Numerical methods in economics*. MIT press, 1998.

[44] L.C.M. Kallenberg. Finite state and action MDPs. In E.A. Feinberg and A. Schwartz, editors, *Handbook of Markov Decision Processes*, pages 21–87. Kluwer, Boston, 2002.

[45] Janina C Ketterer. The impact of wind power generation on the electricity price in germany. *Energy Economics*, 44:270–280, 2014.

[46] Jae Ho Kim and Warren B Powell. Optimal energy commitments with storage and intermittent supply. *Operations research*, 59(6):1347–1360, 2011.

[47] K.C. Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, 90(1):1–25, 2001.

[48] Christopher R Knittel and Michael R Roberts. An empirical examination of restructured electricity prices. *Energy Economics*, 27(5):791–817, 2005.

[49] Michael Koller, Theodor Borsche, Andreas Ulbig, and Göran Andersson. Defining a degradation cost function for optimal control of a battery energy storage system. In *2013 IEEE PowerTech*, pages 1–6, 2013.

[50] Matt Kraning, Eric Chu, Javad Lavaei, Stephen Boyd, et al. Dynamic network energy management via proximal message passing. *Foundations and Trends® in Optimization*, 1(2):73–126, 2014.

[51] Alan D Lamont. Assessing the economic value and optimal structure of large-scale electricity storage. *IEEE Transactions on Power Systems*, 28(2):911–921, 2013.

[52] Carlton E Lemke and Joseph T Howson, Jr. Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics*, 12(2):413–423, 1964.

[53] D Lifshitz and G Weiss. Optimal energy management for grid-connected storage systems. *Optimal Control Applications and Methods*, 36(4):447–462, 2015.

[54] Doron Lifshitz and George Weiss. Optimal control of a capacitor-type energy storage system. *IEEE Transactions on Automatic Control*, 60(1):216–220, 2015.

[55] Michael L Littman. Value-function reinforcement learning in markov games. *Cognitive Systems Research*, 2(1):55–66, 2001.

[56] Lars Ljungqvist and Thomas J Sargent. *Recursive macroeconomic theory*. MIT press, 2012.

[57] Nils Löhndorf and Stefan Minner. Optimal day-ahead trading and storage of renewable energies — An approximate dynamic programming approach. *Energy Systems*, 1(1):61–77, 2010.

[58] Julio J Lucia and Eduardo S Schwartz. Electricity prices and power derivatives: Evidence from the nordic power exchange. *Review of Derivatives Research*, 5(1):5–50, 2002.

[59] Zhongjing Ma, Duncan Callaway, and Ian Hiskens. Decentralized charging control for large populations of plug-in electric vehicles. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 206–212. IEEE, 2010.

[60] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus*. Wiley, revised edition, 1999.

[61] Yuri V Makarov, Pengwei Du, Michael CW Kintner-Meyer, Chunlian Jin, and Howard F Illian. Sizing energy storage to accommodate high penetration of variable energy resources. *IEEE Transactions on sustainable Energy*, 3(1):34–40, 2012.

[62] Pawel Malysz, Shahin Sirouspour, and Ali Emadi. An optimal energy storage control strategy for grid-connected microgrids. *IEEE Transactions on Smart Grid*, 5(4):1785–1796, 2014.

[63] S. Moazeni, W.B. Powell, and A.H. Hajimiragha. Mean-conditional value-at-risk optimal energy storage operation in the presence of transaction costs. *IEEE Trans. Power Systems*, 30(3):1222–1232, 2015.

[64] Alaa Mohd, Egon Ortjohann, Andreas Schmelter, Nedzad Hamsic, and Danny Morton. Challenges in integrating distributed energy storage systems into future smart grid. In *Industrial Electronics, 2008. ISIE 2008. IEEE International Symposium on*, pages 1627–1632. IEEE, 2008.

[65] T.D. Mount, Y. Ning, and X. Cai. Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics*, 28(1):62–80, 2006.

[66] Katta G Murty and Feng-Tien Yu. *Linear complementarity, linear and nonlinear programming*. Citeseer, 1988.

[67] Juliana M Nascimento and Warren B Powell. An optimal approximate dynamic programming algorithm for the energy dispatch problem with grid-level storage. *Mathematics of Operations Research*, 34(1):210–237, 2009.

[68] John F Nash et al. Equilibrium points in n-person games. 1950.

[69] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

[70] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*, pages 441–470. Springer, 2012.

[71] Marek Petrik and Xiaojian Wu. Optimal threshold control for energy arbitrage with degradable battery storage. In *Uncertainty in Artificial Intelligence (UAI)*, 2015.

[72] Craig Pirrong. *Commodity price dynamics: A structural approach*. Cambridge University Press, 2011.

[73] R.J. Plemmons. M-matrix characterizations. I – Nonsingular M-matrices. *Linear Algebra and its Applications*, 18(2):175–188, 1977.

[74] Ryan Porter, Eugene Nudelman, and Yoav Shoham. Simple search methods for finding a nash equilibrium. *Games and Economic Behavior*, 63(2):642–662, 2008.

[75] W. B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, second edition, 2011.

[76] Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

[77] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, Hoboken, NJ, 1994.

[78] M.L. Puterman and S.L. Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research*, 4(1):60–69, 1979.

[79] Junjie Qin, Raffi Sevlian, David Varodayan, and Ram Rajagopal. Optimal electric energy storage operation. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–6. IEEE, 2012.

[80] DM Rastler. *Electricity energy storage technology options: a white paper primer on applications, costs and benefits*. Electric Power Research Institute, 2010.

[81] Eran Raviv, Kees E Bouwman, and Dick Van Dijk. Forecasting day-ahead electricity prices: Utilizing hourly prices. Technical report, Tinbergen Institute, 2013.

[82] J. O. Riis. Discounted Markov programming in a periodic process. *Operations Research*, 13(6):920–929, 1965.

[83] Seyed H Roosta. *Parallel processing and parallel algorithms: theory and computation*. Springer, 2000.

[84] S. M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, NY, 1983.

[85] Xiongwen Rui and Mario J Miranda. Solving nonlinear dynamic games via orthogonal collocation: An application to international commodity markets. *Annals of Operations Research*, 68(1):89–108, 1996.

[86] Tuomas Sandholm, Andrew Gilpin, and Vincent Conitzer. Mixed-integer programming methods for finding nash equilibria. 2005.

[87] M.S. Santos and J. Rust. Convergence properties of policy iteration. *SIAM Journal on Control and Optimization*, 42(6):2094–2115, 2004.

[88] B. Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. In *Advances in Neural Information Processing Systems*, pages 386–394, 2013.

[89] Wolf-Peter Schill and Claudia Kemfert. Modeling strategic electricity storage: the case of pumped hydro storage in germany. *The Energy Journal*, pages 59–87, 2011.

[90] FC Schweppe, RD Tabors, MC Caraminis, and RE Bohn. Spot pricing of electricity. 1988.

[91] Nicola Secomandi. Optimal commodity trading with a capacitated storage asset. *Management Science*, 56(3):449–467, 2010.

[92] Zhen Shu and Panida Jirutitijaroen. Optimal operation strategy of energy storage system for grid-connected wind power plants. *IEEE Transactions on Sustainable Energy*, 5(1):190–199, 2014.

[93] Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

[94] Ramteen Sioshansi. When energy storage reduces social welfare. *Energy Economics*, 41:106–116, 2014.

[95] Ramteen Sioshansi, Paul Denholm, Thomas Jenkin, and Jurgen Weiss. Estimating the value of electricity storage in pjm: Arbitrage and some welfare effects. *Energy economics*, 31(2):269–277, 2009.

[96] A.R. Sparacino, G.F. Reed, R.J. Kerestes, B.M. Grainger, and Z.T. Smith. Survey of battery energy storage systems and modeling techniques. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–8, July 2012.

[97] Han-I Su and Abbas El Gamal. Modeling and analysis of the role of energy storage for renewable integration: power balancing. *Power Systems, IEEE Transactions on*, 28(4):4109–4117, 2013.

[98] S. Y. Su and R. A. Deininger. Modeling the regulation of Lake Superior under uncertainty of future water supplies. *Water Resources Research*, 10(1):11–25, 1974.

[99] Derk J Swider. Compressed air energy storage in an electricity system with significant wind power generation. *Energy Conversion, IEEE Transactions on*, 22(1):95–102, 2007.

[100] Sercan Teleke, Mesut E Baran, Subhashish Bhattacharya, and Alex Q Huang. Optimal control of battery energy storage for wind farm dispatching. *Energy Conversion, IEEE Transactions on*, 25(3):787–794, 2010.

[101] D. Topkis. *Supermodularity and Complementarity*. Princeton University Press, 1998.

[102] Donald M Topkis. *Supermodularity and complementarity*. Princeton University Press, 1998.

[103] T. Tüfekçi and R. Güllü. An iterative approximation scheme for repetitive Markov Processes. *Journal of Applied Probability*, 36(3):654–667, 1999.

[104] Peter M van de Ven, Nidhi Hegde, Laurent Massoulié, and Theodoros Salonidis. Optimal control of end-user energy storage. *Smart Grid, IEEE Transactions on*, 4(2):789–797, 2013.

[105] J.M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11:3–5, 1975.

[106] L. M. M. Veugen, J. van der Wal, and J. Wessels. The numerical exploitation of periodicity in Markov Decision Processes. *OR Spektrum*, 5:97–103, 1983.

[107] D. Wang and B. J. Adams. Optimization of real-time reservoir operations with Markov Decision Processes. *Water Resources Research*, 22(3):345–352, 1986.

[108] R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014.

[109] Xiaomin Xi and Ramteen Sioshansi. A dynamic programming model of energy storage and transformer deployments to relieve distribution constraints. *Computational Management Science*, 13(1):119–146, 2016.

[110] Nanpeng Yu and Brandon Foggo. Stochastic valuation of energy storage in wholesale power markets. *Energy Economics*, 64:177–185, 2017.

[111] Behnam Zakeri and Sanna Syri. Electrical energy storage systems: A comparative life cycle cost analysis. *Renewable and Sustainable Energy Reviews*, 42:569–596, 2015.

[112] Yangfang Zhou, Alan Scheller-Wolf, Nicola Secomandi, and Stephen Smith. Electricity trading and negative prices: storage vs. disposal. *Management Science*, 62(3):880–898, 2015.

[113] Martin Zinkevich, Amy Greenwald, and Michael L Littman. Cyclic equilibria in markov games. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 1641–1648. MIT Press, 2005.

Appendix A

Threshold policy

In Appendix A, we provide some nice properties of the model and some analysis on the optimal policy. The proof are based on [91].

The state variables are $s_i = (l_i, p_i)$ and the reward function is (let $\frac{\Delta}{k} = 1$)

$$R_{t+1} = \begin{cases} -p_t \frac{1}{\eta^{charge} a_t}, & a_t > 0 \\ 0 & \\ -p_t \eta^{discharge} a_t, & a_t < 0 \end{cases} .$$

For simplicity, let $\eta^c = \eta^{charge} < 1, \eta^d = \eta^{discharge} < 1$.

Correspondingly, the value function becomes

$$V_t(l_t, p_t, a_t) = \max_{a \in \mathcal{A}} [R_t(l_t, p_t, a_t) + \gamma \mathbb{E}_{\tilde{p}_{j+1}} [V_{t+1}(l_t + a_t, \tilde{p}_{j+1})]] .$$

And $V_T(l_T, p_T) = \max_{a \in \mathcal{A}} \ -p_T \eta^c a_T$, i.e., discharge or do nothing.

Let

$$\begin{aligned} U_T(l, p) &= 0 , \\ U_t(l, p) &= \gamma \mathbb{E}[V_{t+1}(l, \tilde{p}_{t+1}) | \tilde{p}_t = p] \quad t \in \mathcal{T} \backslash \{T\} . \end{aligned}$$

A.1 Concavity

In time period $t \in \mathcal{T}$. the functions $U_t(l,p)$ and $V_t(l,p)$ are concave in $l \in \mathcal{L}$ for each given p.

Proof: By induction.

For time period T: Clearly hold. Discharge as much as possible.

Assume the property holds for $t+1$, i.e., $V_{t+1}(l, p_{t+1})$ is concave in x, given p_{t+1},

$$V_{t+1}(l^\phi, p_{t+1}) \geq \phi V_{t+1}(l^1, p_{t+1}) + (1-\phi) V_{t+1}(l^2, p_{t+1}) \ .$$

Consider time period t. Pick $\phi \in [0,1]$ and l^1, l^2. Let $l^\phi = \phi l^1 + (1-\phi) l^2$, which clearly is in $\mathcal{L}$. Based on the assumption that price is finite, we know that $U_t(l,p)$ should be bounded, which implies that it is real value in $l \in \mathcal{L}$ for each given p. Discounting and taking expectations on both sides

$$U_t(l^\phi, p) \geq \phi U_t(l^1, p) + (1-\phi) U_t(l^2, p) \ , \qquad (*)$$

which implies that $U_t(l,p)$ is concave in $l \in \mathcal{L}$ for given p.

Let a^i be a feasible action at storage level $l^i, i = 1, 2$, and define $a^\phi = \phi a^1 + (1-\phi) a^2$. The convexity of the storage action set $\mathcal{C}$ implies that $(l^\phi, a^\phi) \in \mathcal{C}$, then

$$U_t(l^\phi + a^\phi, p) \geq \phi U_t(l^1 + a^1, p) + (1-\phi) U_t(l^2 + a^2, p) \ .$$

The reward function $R_t(a,p)$ is piecewise linear and concave in a given p. Combine with (*), $R_t(l_t, p_t, a_t) + \gamma \mathbb{E}_{\tilde{p}_{j+1}}[V_{t+1}(l_t + a_t, \tilde{p}_{j+1})$ is jointly concave in $(l,a) \in \mathcal{C}$ for given p. By proposition B-4 in Heyman, $V_t(l,p)$ is concave in $l \in \mathcal{L}$ for given p. Thus, the property holds in time period t. We finish our proof.

A.2 Optimal basestock targets

In each time period, there exist critical storage level $\underline{b}_t(p), \bar{b}_t(p) \in \mathcal{L}$ with $\underline{b}_t(p) < \bar{b}_t(p)$, which depend on price p such that an optimal action in each state (l, p) is

$$a_t^*(l,p) = \begin{cases} (\underline{b}_t(p) - l) \wedge \bar{a} & if\ l \in [\underline{l}, \underline{b}_t(p)) \\ 0 & if\ l \in [\underline{b}_t(p), \bar{b}_t(p)] \\ (\bar{b}_t(p) - l) \wedge \underline{a} & if\ l \in (\bar{b}_t(p), \bar{l}] \end{cases} .$$

where $\bar{a}, \underline{a}$ are maximum charge rate and discharge rate and $\bar{l}, \underline{l}$ are maximum and minimum capacity of the device.

Proof. Consider any time period t and pick state (l, p). Relax the ramping constraint $(\underline{a} \leq a \leq \bar{a})$ First. Let $y = l + a$ be the decision variable, then the optimization problem without ramping constraint become

$$\max_y \ R_t(y - l, p) + U_t(y, p)$$

Depending on whether $y \geq l$ or $y \leq l$, the respective objective function become

$$U_t(y,p) - \eta^c p y + \eta^c p l \ , \tag{A.1}$$

$$U_t(y,p) - \eta^d p y + \eta^d p l \ . \tag{A.2}$$

Then, the original problem can be approached by finding optimal solutions to the problems

$$\max_{y \in [l, \bar{l}]} U_t(y,p) - \eta^c p y + \eta^c p l \ , \tag{A.3}$$

$$\max_{y \in [\underline{l}, l]} U_t(y,p) - \eta^d p y + \eta^d p l \ , \tag{A.4}$$

and taking the optimal solution to the original problem to be the one with the highest objective function value.

In particular, at $l = \underline{l}$ and $l = \bar{l}$, the original problem become

$$\max_{y \in [\underline{l}, \bar{l}]} U_t(y,p) - \eta^c p y + \eta^c p l \ , \tag{A.5}$$

$$\max_{y \in [\underline{l}, \bar{l}]} U_t(y,p) - \eta^d p y + \eta^d p l \ , \tag{A.6}$$

which can be simplified to

$$\max_{y \in [\underline{l}, \bar{l}]} U_t(y,p) - \eta^c p y \ , \tag{A.7}$$

$$\max_{y \in [\underline{l}, \bar{l}]} U_t(y,p) - \eta^d p y \ . \tag{A.8}$$

Let $\underline{b}_t(p)$ and $\bar{b}_t(p)$ be optimal solutions to (7),(8) respectively.

Since $U_t(y,p)$ is concave in y given p and so is $U_t(y,p) - \eta^c p y$ and $U_t(y,p) - \eta^d p y$. In addition, it holds that $\eta^c \geq \eta^d$. Hence, the optimal solution to (7) $\underline{b}_t(p)$ is never greater than an optimal solution $\bar{b}_t(p)$ to (8). The optimal solution for (7) is $\frac{\partial U_t(y,p)}{\partial y} - \eta^c p$ while the optimal solution for (8) should be $\frac{\partial U_t(y,p)}{\partial y} - \eta^d p$.

Consider the original problem for any $l \in [\underline{l}, \underline{b}_t(p))$. It is clear that $\underline{b}_t(p)$ is an optimal (3). It also hold that l is an optimal solution to (4). The reason is that $\bar{b}_j(p)$ maximizes (2) when y can take any value in the whole range. But in (4), $y \leq l$. We know that $\bar{b}_j(p) \geq \underline{b}_j(p) \geq l$ is the optimal solution for (8), which implies that in the range $[\underline{l}, l]$, (4) is increasing in y, given p. Hence, l is the optimal solution to (4).

Moreover, l is a feasible solution to (3), so that

$$U_t(\underline{b}_t(l), p) - \eta^c p \underline{b}_t + \eta^c p l \geq U_t(l,p) \ .$$

Thus, we conclude that $\underline{b}_j(p)$ optimize the original problem and $a_j^*(l,p) = (\underline{b}_j(p) - x) \wedge \bar{C}$. Consider the range $l \in [\underline{b}_j(p), \bar{b}_j(p)]$, storage level l optimizes both (3) and (4) because $\underline{b}_j(p)$ and $\bar{b}_j(p)$ maximize (8) and (9) on $[\underline{l}, \bar{l}]$ and $\underline{l} \leq l \leq \bar{l}$. It follows that $a_j^*(l,p) = 0$.

The case where $l \in (\bar{b}_j(p), \bar{l}]$ can be dealt with similarly to the first case $l \in [\underline{l}, \underline{b}_t(p))$.

A.3 Price monotonicity

Assumption 2: (Spot-price process) For every $t \in \mathcal{T} \backslash T$:

(a) The distribution function of random variable $\tilde{p}_{t+1}$ conditional on the spot price is time period t stochastically increase in $p \in P_t$.

(b) The function $\delta_t \mathbb{E}_t[\frac{1}{\eta^c}\tilde{p}_{t+1}|\tilde{p}_t = p] - \eta^d p$ decreases in $p \in P_t$

Proposition 2 If assumption 2 holds, then every time period $j \in J$ the optimal basestock target function $\underline{b}_t(p)$ and $\bar{b}_t(p)$ decrease in the spot price $p \in P$.

Proof: By induction. And there are three statements to prove to hold at every iterations.

- the optimal basestock target function $\underline{b}_t(p)$ and $\bar{b}_t(p)$ decrease in the spot price $p \in P$
- $U_t'(l, p)$ increase in $p \in P_t$, given l.
- functions $U_t'(l, p) - \frac{1}{\eta^c}p$ and $U_t'(l, p) - \eta^d p$ decrease in $p \in P_t$ for each given l.

For stage T:

- The value function is

$$V_T(l_T, p_T) = \max_{a \in \mathcal{A}} \ -p_T \eta^c a_T \ .$$

  It is easy to verify that if $p_T < 0$ $\underline{b}_T(p) = \underline{l}$ and $\bar{b}_T(p) = \bar{l}$, otherwise, $\underline{b}_T(p) = \bar{b}_T(p) = \underline{l}$.

- $U_T'(l, p) = 0$. Trivial.

- Functions

$$U_T'(l, p) - \frac{1}{\eta^c}p = -\frac{1}{\eta^c}p$$
$$U_T'(l, p) - \eta^d p = -\eta^d p$$

  decrease in p, given l.

Consider j. By proposition 1, the objective functions (7) and (8) are concave in the decision variable y and $\underline{b}_t(p)$, $\bar{b}_t(p)$ are maximal solutions are these two functions. Hence, our goal is to show that the partial derivative respect to y,

$$U'_T(y,p) - \frac{1}{\eta^c}p$$

$$U'_T(y,p) - \eta^d p$$

both decrease in p given y.

In order to achieve this, pick $(l,p) \in \mathcal{L}\times\mathcal{P}$ and consider function $U_t(l,p) = \delta_t \mathbb{E}_t[V_{t+1}(l,\bar{p}_{t+1})|p_t = p]$. Focus on the function $V_{t+1}(l,z)$ in feasible state (x,z) in stage $t+1$. Consider the optimal action. There are five mutually exclusive cases need to be considered.

- Discharge is optimal but $\bar{b}_{t+1}(z)$ cannot be reached from l; that is, only $l + \underline{a}$ can be reached from l
- Discharge is optimal and $\bar{b}_{t+1}(z)$ can be reached from l
- Do nothing is optimal
- Charge is optimal and $\underline{b}_{t+1}(z)$ can be reached from l
- Charge is optimal and $\underline{b}_{t+1}(z)$ cannot be reached from l ;that is, only $l + \bar{a}$ can be reached from l

Accordingly, define the following mutually exclusive events:

- $A^1_{t+1}(l,z) := \{l + \underline{a} > \bar{b}_{t+1}(z)\}$
- $A^2_{t+1}(l,z) := \{l + \underline{a} < \bar{b}_{t+1}(z), l > \bar{b}_{t+1}(z)\}$
- $A^3_{t+1}(l,z) := \{\underline{b}_{t+1}(z) \leq l \leq \bar{b}_{t+1}(z)\}$
- $A^4_{t+1}(l,z) := \{l + \bar{a} > \bar{b}_{t+1}(z), l < \underline{b}_{t+1}(z)\}$
- $A^5_{t+1}(l,z) := \{l + \bar{a} < \underline{b}_{t+1}(z)\}$

Let $\mathbf{1}\{\}$ equals 1 if its argument is true and 0 otherwise. Then $V_{t+1}(l,z)$ can be written as

$$\begin{aligned}
V_{t+1}(l,z) =&[-\eta^d z\underline{a} + U_{t+1}(l+\underline{a},z)]\mathbf{1}\{A^1_{t+1}(l,z)\} \\
&+ \{-\eta^d z[\bar{b}_{t+1}(z) - l] + U_{t+1}(\bar{b}_{t+1}(z),z)\}\mathbf{1}\{A^2_{t+1}(l,z)\} \\
&+ U_{t+1}(l,z)\mathbf{1}\{A^3_{t+1}(l,z)\} \\
&+ \{-\frac{1}{\eta^c} z[\underline{b}_{t+1}(z) - l] + U_{t+1}(\underline{b}_{t+1}(z),z)\}\mathbf{1}\{A^4_{t+1}(l,z)\} \\
&+ [-\frac{1}{\eta^c} z\bar{a} + U_{t+1}(l+\bar{a},z)]\mathbf{1}\{A^5_{t+1}(l,z)\} \ .
\end{aligned}$$

Consider the function

$$\begin{aligned}
V'_{t+1}(l,z) =&U'_{t+1}(l+\underline{a},z)\mathbf{1}\{A^1_{t+1}(l,z)\} + (\eta^d z)\mathbf{1}\{A^2_{t+1}(l,z)\} \\
&+ U'_{t+1}(l,z)\{A^3_{t+1}(l,z)\} \\
&+ (\frac{1}{\eta^c})\mathbf{1}\{A^4_{t+1}(l,z)\} + U'_{t+1}(l+\bar{a},z)\mathbf{1}\{A^5_{t+1}(l,z)\} \ .
\end{aligned}$$

Arrange it and define $f^1_{t+1}(l,z)$ and $f^2_{t+1}(l,z)$

$$\begin{aligned}
V'_{t+1}(l,z) = & \\
&\left.\begin{aligned}
&[U'_{t+1}(l+\underline{a},z) - \eta^d z]\mathbf{1}\{A^1_{t+1}(l,z)\} \\
&+[\eta^d z - \eta^d z]\mathbf{1}\{A^2_{t+1}(l,z)\} \\
&+[\tfrac{1}{\eta^c} - \tfrac{1}{\eta^c}]\mathbf{1}\{A^4_{t+1}(l,z)\} \\
&+[U'_{t+1}(l+\bar{a},z) - \tfrac{1}{\eta^c} z]\mathbf{1}\{A^5_{t+1}(l,z)\}
\end{aligned}\right\} = f^1_{t+1}(l,z) \\
&\left.\begin{aligned}
&+(\eta^d)[\mathbf{1}\{A^1_{t+1}(l,z)\} + \mathbf{1}\{A^2_{t+1}(l,z)\}] \\
&+U'_{t+1}(l,z)\mathbf{1}\{A^3_{t+1}(l,z)\} \\
&+(\tfrac{1}{\eta^c})[\mathbf{1}\{A^4_{t+1}(l,z)\} + \mathbf{1}\{A^5_{t+1}(l,z)\}]
\end{aligned}\right\} = f^2_{t+1}(l,z)
\end{aligned}$$

We need to study the behavior of the functions $f^1_{t+1}(l,z)$ and $f^2_{t+1}(l,z)$ in z given l. Consider the determination of an optimal action in state (l,z) in period $t+1$ as p varies in $\mathcal{P}_{t+1}$. By the first induction hypothesis, there exist no more than four ordered prices that depend

on l, denoted, with a slightly abuse of notation, by $p^1_{t+1}(l), p^2_{t+1}(l), p^3_{t+1}(l), p^4_{t+1}(l)$ with $p^1_{t+1}(l) < p^2_{t+1}(l) < p^3_{t+1}(l) < p^4_{t+1}(l)$, that can be used to partition set $\mathcal{P}_{t+1}$ into mutually exclusive and exhaustive sets

- $\mathcal{P}^1_{t+1}(l) := (p^4_{t+1}(l), \infty) \cap \mathcal{P}_{t+1}$
- $\mathcal{P}^2_{t+1}(l) := (p^3_{t+1}(l), p^4_{t+1}(l)] \cap \mathcal{P}_{t+1}$
- $\mathcal{P}^3_{t+1}(l) := [p^2_{t+1}(l), p^3_{t+1}(l)] \cap \mathcal{P}_{t+1}$
- $\mathcal{P}^4_{t+1}(l) := [p^1_{t+1}(l), p^2_{t+1}(l)) \cap \mathcal{P}_{t+1}$
- $\mathcal{P}^5_{t+1}(l) := [0, p^1_{t+1}(l)) \cap \mathcal{P}_{t+1}$

which satisfy the property that $z \in \mathcal{P}^k_{t+1}$ if and only if $\mathbf{1}\{A^k_{t+1}(l,z)\} = 1$ for all $k \in \{1,2,3,4,5\}$. Consider different case, we obtain those inequalities as following:

$$
\begin{aligned}
U'_{t+1}(l+\underline{a}, z) - \eta^d z \leq 0, \forall\, z \in \mathcal{P}^1_{t+1} \\
U'_{t+1}(l, z) \leq \eta^d z, \forall\, z \in \mathcal{P}^2_{t+1} \\
U'_{t+1}(l+\underline{a}, z) - \eta^d z \geq 0, \forall\, z \in \mathcal{P}^2_{t+1} \\
U'_{t+1}(l, z) \geq \eta^c z, \forall\, z \in \mathcal{P}^4_{t+1} \\
U'_{t+1}(l+\bar{a}, z) - \frac{1}{\eta^c} z \leq 0, \forall\, z \in \mathcal{P}^4_{t+1} \\
U'_{t+1}(l+\bar{a}, z) - \frac{1}{\eta^c} z \geq 0, \forall\, z \in \mathcal{P}^5_{t+1} \\
\eta^d z \leq U'_{t+1}(l, z) \leq \frac{1}{\eta^c} z, \forall\, z \in \mathcal{P}^3_{t+1}\,.
\end{aligned}
$$

Consider $f^1_{t+1}(l,z)$, given l, this function is positive for $z \in \mathcal{P}^5_{t+1}$, zero for $z \in \mathcal{P}^2_{t+1} \cup \mathcal{P}^3_{t+1} \cup \mathcal{P}^4_{t+1}$, negative for $z \in \mathcal{P}^1_{t+1}$. Moreover, the third induction hypothesis implies that this function decreases in $z \in \mathcal{P}_{t+1}$. Assumption 2 and Corollary in [102] imply that

$$\delta_t \mathbb{E}[f^1_{t+1}(l, \bar{p}_{t+1}) | \bar{p}_t = p]$$

decreases in $p \in \mathcal{P}_t$.

Consider $f^2_{t+1}(l,z)$, given l, combined with the second induction hypothesis, we have $f^2_{t+1}(l,z)$ increases in $z \in \mathcal{P}_{t+1}$ and $f^2_{t+1}(l,z) \leq \frac{1}{\eta^c}, \forall\, z \in \mathcal{P}_{t+1}$.

Hence, we have

$$\delta_t \mathbb{E}[f^2_{t+1}(l,\bar{p}_{t+1})|\bar{p}_t = p] - \frac{1}{\eta^c} \text{ decrease in } p \in \mathcal{P}_t \ ,$$
$$\delta_t \mathbb{E}[f^2_{t+1}(l,\bar{p}_{t+1})|\bar{p}_t = p] - \eta^d \text{ decrease in } p \in \mathcal{P}_t \ .$$

And from Lemma 7, we have

$$U'_{t+1}(l,p) - \frac{1}{\eta^c} = \delta_t \mathbb{E}[f^1_{t+1}(\bar{p}_{t+1})|\bar{p} = p] + \delta_t \mathbb{E}[f^2_{t+1}|\bar{p}_t = p] - \frac{1}{\eta^c} \ ,$$
$$U'_{t+1}(l,p) - \eta^c = \delta_t \mathbb{E}[f^1_{t+1}(\bar{p}_{t+1})|\bar{p} = p] + \delta_t \mathbb{E}[f^2_{t+1}|\bar{p}_t = p] - \eta^c \ .$$

Then both $U'_{t+1}(l,p) - \frac{1}{\eta^c}$ and $U'_{t+1}(l,p) - \eta^c$ decrease in $p \in \mathcal{P}_t$ for given l.

Then we show that $U'_t l, p$ increases in $p \in \mathcal{P}_t$ for given l. We know that

$$U'_{t+1}(l + \bar{a}, z)\mathbb{1}\{z \in \mathcal{S}^5_{t+1}(l)\} + \frac{1}{\eta^c}\mathbb{1}\{z \in \mathcal{P}^4_{t+1}(l)\}$$

increases in $z \in \mathcal{P}^4_{t+1}(l) \cup \mathcal{P}^5_{t+1}(l)$.

And

$$\frac{1}{\eta^c}\mathbb{1}\{z \in \mathcal{P}^4_{t+1}(l)\} + U'_{t+1}(l,z)\mathbb{1}\{z \in \mathcal{S}^3_{t+1}(l)\} + (\eta^d)\mathbb{1}\{z \in \mathcal{P}^2_{t+1}(l)\}$$

increases in $z \in \mathcal{S}^4_{t+1}(l)\} \cup \mathcal{S}^3_{t+1}(l)\} \cup \mathcal{S}^2_{t+1}(l)\}$.

And

$$U'_{t+1}(l + \underline{a}, z)\mathbb{1}\{z \in \mathcal{S}^2_{t+1}(l)\} + (\eta^d)\mathbb{1}\{z \in \mathcal{P}^1_{t+1}(l)\}$$

increases in $z \in \mathcal{S}^2_{t+1}(l)\} \cup \mathcal{S}^1_{t+1}(l)\}$. Thus $V'_{t+1}(l,z)$ increases in $z \in \mathcal{P}_{t+1}$. And by Corollary 3.9.1(a) in [102], we have $U_t(l,p_t) = \delta_t \mathbb{E}[V'_{t+1}|\bar{p}_t = s]$ increases in $p \in \mathcal{P}_t$

Appendix B

Price process

The power industry has become an open, competitive environment and the uncertainty of price is one of the key components of this environment. In the related literature, many different models have been studied [81]. The complexity of a model depends on the data that is available. At the time these models were proposed, data on the deregulated electricity markets was scarce.

According to [58], in order to develop a price process model, the following properties should be taken into consideration: mean reversion, time of day effects, weekend/weekday effects, seasonal effects, time-varying volatility and extreme values/price spikes. Several price models are provided in [58], including mean-reverting process, time-varying mean, jump-diffusion process, time-dependent jump intensity, ARMAX.

In [13], Burger et al. present a general model called Spot Market Price simulation (SMaPS-model) that simultaneously takes into account seasonal patterns, price spikes, mean reversion, price dependent volatilities and long-term non-stationarity.

In my research, in order to incorporate the price process into the battery operation model, sometimes a relatively simple model has been chosen although a more complicated model can replace the simpler model. We give 3 examples of stochastic processes for the price.

B.1 Periodic Price Model with i.i.d noise

Assuming the price varies periodically with period T subject to i.i.d. noise, we have the following price process

$$p_{t+1} = \mu_j + \sigma_j \epsilon_{t+1}$$

where

$j =$ $\mathrm{mod}(t+1, T)$ the index of $t+1$ in the cycle of period T

$\mu_j =$ the expected price at time index j in the cycle

$\sigma_j =$ the standard deviation of random noise at time index j in the cycle

$\epsilon_{t+1} \sim$ $N(0,1)$ a standard normal noise

This process may be the simplest price process we can find. The price at certain time period t only depends on its own mean and variance. However, if we evaluate the logarithm of all the data before we use this price model, this model can be extend to so-called lognormal distribution, which we used in Chapter 2.

B.2 Periodic autoregressive process of order 1

Consider another model where the price is a periodic autoregressive model of order 1 (PAR(1) process) which exhibits mean-reversion to the periodic mean of period T. We can describe this as follows:

$$p_{t+1} - p_t = (\mu_j - \mu_i) + \kappa_i(\mu_i - p_t)\Delta_t + \sigma_j\sqrt{\Delta_t}\epsilon_{t+1}$$

where

$i =$ $\mathrm{mod}(t, T)$ the index of t in the cycle of period T

$j =$ $\mathrm{mod}(t+1, T)$ the index of $t+1$ in the cycle of period T

$\mu_j =$ the expected price at time index j in the cycle

$\kappa_i =$ the mean-reversion parameter at time index j in the cycle

$\sigma_j =$ the standard deviation of random noise at time index j in the cycle

$\epsilon_{t+1} \sim$ $N(0,1)$ a standard normal noise

Compared with Model 1, prices in Model 2 depend on its mean, variance and its precedent price. The parameter κ controls the rate at which the price reverts back to the nominal value μ_i.

B.3 PAR(1) with spikes

The third model we provide here incorporates a probability of random price spike. To describe this model, we first define

$w^b_{t+1} \sim$ $B(1, b)$: A Bernoulli random variable equal to 1 with probability b to indicate a spiking reg

$w^\epsilon_{t+1} \sim$ $N(0, 1)$ a standard Normal noise.

$f =$ A function to replace a base price by a spike price

Then the model can be described as follows:

$$p_{t+1} = w^b_{t+1}\xi_{t+1} + (1 - w^b_{t+1})f(\xi_{t+1})$$

$$(\xi_{t+1} - \xi_t) = (\mu_j - \mu_i) + \kappa(\mu_i - \xi_t)\Delta_t + \sigma_j\sqrt{(}\Delta_t)w^\epsilon_{t+1}$$

The numerical work for this model has not been finished yet, but hopefully, we can obtain a better prediction of price in a short term.

Appendix C

Linear and Quadratic Curve Case Study

C.1 Analytic Solution for Equilibrium Price and Reward Functions: Linear and Quadratic Curve Case

With fixed a, we obtain

$$\begin{aligned} S_t = f_s(P_t) &= \frac{-c + \sqrt{c^2 + 4dP_t}}{2d} \,, \\ D_t = f_d(P_t) &= \frac{b - P_t}{a} \,. \end{aligned} \tag{C.1}$$

If we know both players' actions H_t^c, H_t^s, then we know the quantity of aggregated energy change $H_t = H_t^c + H_t^s$. By solving equation (4.1), we have the price P_t given known actions. Plugging P_t back into equations (C.1), we obtain the demand and supply given known actions as well,

$$\begin{aligned} P_t(H_t) &= \frac{1}{2}\Big(2b + \frac{a^2}{d} + \frac{ac}{d} + 2aH_t - \frac{a\sqrt{a^2 + 2ac + c^2 + 4bd + 4adH_t}}{d}\Big) \,, \\ D_t(H_t) &= -\frac{a + c + 2dH_t - \sqrt{(a+c)^2 + 4bd + 4adH_t}}{2d} \,, \\ S_t(H_t) &= \frac{1}{2d}\Big(-c + \sqrt{\Phi(H_t) + \Psi(H_t)}\Big) \,. \end{aligned}$$

where

$$\Phi(H_t) = 2a^2 + c^2 + 4bd \ ,$$
$$\Psi(H_t) = 2a(c + 2dH_t - \sqrt{(a+c)^2 + 4bd + 4adH_t}) \ .$$

In addition, we compute the demand reward R_t^d and generation reward R_t^g

$$R_t^d = \frac{a}{8d^2}\Big(a + c + 2dH_t - \sqrt{(a+c)^2 + 4bd + 4adH_t}\Big)^2 \ ,$$
$$R_t^g = \frac{1}{24d^2}(c - \Phi(H_t))^2(c + \Phi(H_t)) \ .$$

The storage reward for consumer and supplier are

$$R_t^{cs} = -H_t^c P_t(H_t) \ ,$$
$$R_t^{ss} = -H_t^s P_t(H_t) \ .$$

With equation (4.2), we obtain the reward functions for both players given known actions H_t^c, H_t^s.

C.2 Numerical Experiment: Linear and Quadratic Curve Case

We use similar setting with Section 4.4.1, instead of the piecewise for supply curve, we use a simple quadratic curve

$$s^{-1}(q) = 2q + 0.5q^2$$

where b has a Normal Distribution $\mathcal{N}(800, 60^2)$, which is approximated by a discrete distribution with 61 different states.

Capacities for both storages are $K_c = K_s = 2$, discretized with step length 0.1.

In this linear/quadratic case, consumer has tendency to discharge with higher demand

level b while supplier tries to charge.

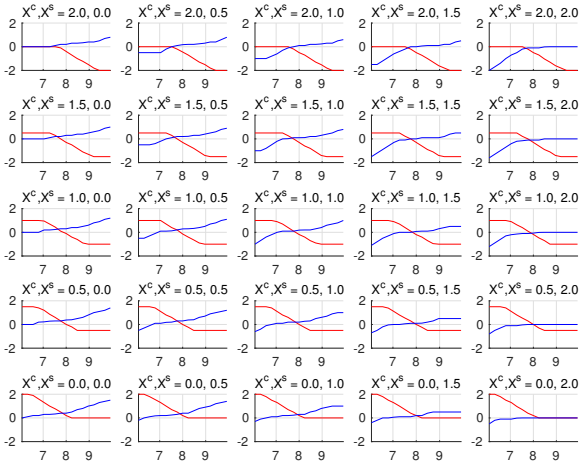


Figure C.1: Charging amount for both players (y-axis) as a function of the demand curve level (parameter $b/100$ as x-axis). *red*: consumer. *blue*: supplier.

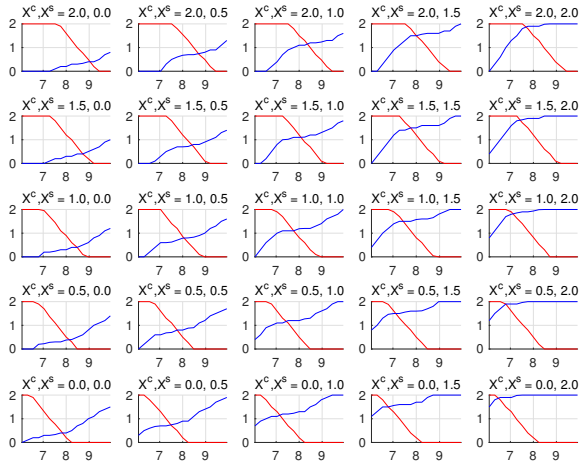


Figure C.2: Next charge levels for both players (y-axis) as a function of the demand curve level (parameter $b/100$ as x-axis). *red*: consumer. *blue*: supplier.

Biography

Yuhai Hu earned his Bachelor of Science degree in Industrial Engineering from Huazhong University of Science and Technology in 2012. In 2012, he joined the doctoral program in Industrial Engineering at Lehigh University. He is keenly interested in almost all areas of mathematical optimization and his research focus is on stochastic optimal control of grid-level storage. He spent the summer of 2015 working as Data Mining Intern at Bosch RTC in Palo Alto, CA and the fall of 2016 working part-time as Computational Modeling Intern at AirProducts in Allentown, PA.