REINFORCEMENT LEARNING TECHNIQUES FOR CONTROLLING RESOURCES IN POWER NETWORKS

BY

ANUPAMA SUNIL KOWLI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Adjunct Professor Sean P. Meyn, Chair
Professor Peter W. Sauer
Assistant Professor Alejandro Dominguez-Garcia
Professor William H. Sanders

Abstract

As power grids transition towards increased reliance on renewable generation, energy storage and demand response resources, an effective control architecture is required to harness the full functionalities of these resources. Also needed are appropriate control mechanisms that help realize the value of storage and demand response resources as means of mitigating the impact of volatility from renewable energy. And there is a critical need for control techniques that recognize the unique characteristics of the different resources and exploit the flexibility afforded by them to provide ancillary services to the grid. The work presented in this dissertation addresses these needs.

The main contribution of this dissertation lies in the domain of control techniques. New stochastic control algorithms, capable of capturing the salient characteristics of the resources, are proposed for control synthesis. The principle advantage of these algorithms is that they can be devised in settings for which the precise distribution of the uncertainty and its temporal statistics are not known. The proposed algorithms are applied to power system control problems such as control of energy storage and demand response resources to extract ancillary services and coordination of dispatch of multiple resources in a power grid. Numerical studies demonstrate how these algorithms can be driven by real world data to successfully tune the control parameters to the underlying statistics of the system.

In addition to the development of new control techniques, this dissertation also provides an assessment of the impacts of variable and uncertain renewable generation and the flexibility provided by energy storage and demand response resources. Specifically, investigations quantifying the increased ancillary service needs of power grids with integrated renewable energy sources are reported. Also, control synthesis for provision of ancillary services by en-

ergy storage and demand response resources is discussed. Numerical studies presented in this context further the understanding of the impacts and the behavior of these new resources; this helps develop effective control strategies for them.

The technical contributions of this dissertation are three-fold. First, an architecture for analyzing approximate solutions of Markov decision processes (MDPs) is devised: the architecture allows us to examine the stability and performance of control policies derived from approximate MDP solutions. Second, two new algorithms based on parameterized Q-learning are presented for approximating solutions to MDPs. The first algorithm is based on reducing the error in the dynamic programming equation while the second algorithm is based on the linear programming approach to solve MDPs. Third, the connections between Q-learning and model predictive control (MPC) are established and leveraged to devise a new control approach. The new approach, referred to as Q-MPC, uses Q-learning to approximate the optimal terminal cost for MPC. The Q-MPC algorithm admits a stabilizing control policy under mild conditions and its computational efficiency is provided via numerical studies.

The control algorithms developed in this work have important practical impacts. In particular, we argue that the Q-MPC approach may be a better choice for economic dispatch of power grid resources as compared to the MPC implementation typically adopted for this problem. Numerical studies conducted on different test systems demonstrate how the computational complexity of the dispatch problem can be reduced via application of Q-MPC. The improvement in performance and greater adaptability of Q-MPC make it eminently suitable for large power grids with many resources and many sources of uncertainty.

The control techniques proposed in this dissertation impose minimal assumptions on the system model and allow the control to be "learned" based on actual dynamics of the system. These techniques provide a starting step towards the development of advanced control techniques that will be necessary for future power grids. They may also be used to improve operational decision-making tools, as demonstrated for the economic dispatch problem.

To my love, Ankur

Acknowledgments

As I was writing this dissertation, I realized that many people are responsible for helping me get to this stage. I want to take this opportunity to thank everyone who helped me become a doctor.

First and foremost, I want to thank Prof. Sean Meyn for being a patient teacher, an inspiring guide, an outstanding researcher and a brilliant scholar while also being a wonderful human being and a true friend. As a mentor, I could not have asked for more. He anchored me at times when I was drifting aimlessly and let me have a free reign when I wanted to explore. This dissertation would not have been possible without his expertise, insights and timely feedback.

I am also deeply indebted to Prof. Peter Sauer for always being there for me. I have been very fortunate to have had the opportunity to study under and work with him closely during my time at Illinois. He is a truly generous person, who offered his immense knowledge and unreserved help whenever I needed. And he always had my best interests at heart; for that, I thank him.

I am grateful to other members of my committee – Prof. William Sanders and Prof. Alejandro Dominguez-Garcia – for their insightful comments and constructive criticism. I am thankful for the financial support offered by the Trustworthy Cyber Infrastructure for Power Grid project. And, I would like to acknowledge Prof. George Gross, who introduced me to this wonderful area of electricity markets and power system planning and operations.

My sincere thanks go to the staff of both the Power and Energy Systems group at Everitt Lab and the Decision and Control group at Coordinated Sciences Lab/CSL: the Power staff, especially Karen Driscoll, made me feel at home when I first landed at Illinois, fresh off the plane from India. And when I had to leave this new home, Jana Lenz and Dan Jordan from CSL warmly

welcomed me and helped me find a new home at CSL.

I would like to thank Dr. Karanjit Kalsi and Dr. Marcelo Elizondo for the internship experience at Pacific Northwest National Laboratory. They gave me full support to facilitate my work there and helped me translate some of the theoretical advances made in the course of this research to more practical settings.

I have had the good fortune to meet some remarkable people at Illinois and forge wonderful friendships. I thank all my friends from Illinois for leaving me with many warm memories. In particular, I want to thank Gui and Matias for being with me through thick and thin; Abhishek, Ashish, Dayu, Kate, Nachiket, Renuka, Yun and Zeba for their support and encouragement; and Ehsan, Komal, Melanie and Rohith for the chats and the laughs.

These acknowledgments would be incomplete without mentioning my deepest gratitude to my parents for their guidance, unconditional love and unwavering support. My mother, especially, made a lot of effort to ensure I got my degree. Her efforts ranged from nagging me regarding my research progress to traveling across the oceans to be by my side whenever I needed her; I appreciate all that she has done for me and hope to make her proud some day. I also thank my sisters for always being there for me, for making me feel connected even across so many miles and for making me smile when things got tough. My in-laws also deserve a special mention, for their support and encouragement. And I am grateful to my cousin, Meenal, and her husband, Tejas, who have done more for me than any elder sibling would have.

Finally, I want to thank my husband, Ankur, for encouraging me to take the plunge and commit to a Ph.D. His sharp intellect, go-get-it attitude, jovial nature and cooking skills (or lack of thereof!) got me through my toughest times. Words fail me right now to convey the depth of my gratitude for him. It suffices to say that I would not be me without him.

Table of Contents

List of Figures

List of Abbreviations

ACOE	Average cost optimality equation
ACE	Area control error
ADP	Approximate dynamic programming
AHU	Air handling unit
BESS	Battery energy storage system
BPA	Bonneville Power Administration
DLC	Direct load control
DP	Dynamic programming
DR	Demand response
CAISO	California Independent System Operator
FERC	Federal Energy Regulatory Commission
HVAC	Heating, ventilation and air conditioning
ISO-NE	Independent System Operator New England
LQR	Linear quadratic regulator
MDP	Markov decision process
MPC	Model predictive control
NREL	National Renewable Energy Laboratory
PIA	Policy iteration algorithm
PJM	Pennsylvania-New Jersey-Maryland Interconnection
PNNL	Pacific Northwest National Laboratory
RL	Reinforcement learning
UC	Unit commitment
VFD	Variable frequency drive
VIA	Value iteration algorithm
VOLL	Value of lost load

Chapter 1

Introduction

This dissertation is concerned with the development of control-theoretic tools for management of resources in power grids. This chapter sets the stage for the work presented herein. It begins with a discussion on the motivation and the background for this research so as to enable the reader to better understand the nature of the problems of interest and the proposed solutions. A brief description of the state of the art, from an industry and academic point of view, is also provided to clarify which gaps this research intends to fill. The scope of this dissertation as well as the specific contributions are summarized. The chapter concludes with an outline of the remainder of this dissertation.

1.1 Background and Motivation

Electricity is considered the backbone of modern society. And the power grid – which transports electricity from its point of generation to its point of end-use consumption – is a fundamental part of the society's infrastructure. The electricity industry has significantly evolved over the past several decades and much effort has been expended to ensure provision of cheap and reliable electricity to consumers [1, 2]. Indeed, widespread electrification has been identified as one of the greatest engineering achievements of the 20^{th} century [3].

Maintaining system reliability has been and continues to be a primary concern in power system operations and planning [1,4]. In the operational domain, reliability concerns manifest themselves into four tasks [5]:

(OT1) balance supply and demand under normal and contingency conditions,
(OT2) control supply and demand to satisfy power flow constraints under

normal and contingency conditions,

(OT3) maintain voltages throughout the power system within prescribed limits under normal and contingency conditions, and,

(OT4) restart the power system after it collapses in case any of the above three tasks fail.

Economic considerations also play a vital role in driving operational and planning decisions in the sense that the goal is to maintain reliability at *lowest possible cost* [1, 2].

The responsibility of cost-effective management of available resources to complete the operational tasks **(OT1)**-**(OT4)** lies with a *system operator*. In a vertically integrated environment, the role of the system operator is portrayed by the single utility that owns and controls all generation, transmission and distribution assets. The operational and planning decisions of such a monopoly entity are subject to state and (possibly) federal regulation. In the restructured open-access environment, the system operations are managed by non-profit agencies such as independent system operators or regional transmission operators or transmission system operators; examples include Independent System Operator of New England (ISO-NE), California Independent System Operator (CAISO), Electric Reliability Council of Texas (ERCOT), Pennsylvania-New Jersey-Maryland (PJM) Interconnection and Bonneville Power Administration (BPA) in the United States and National Grid in the United Kingdom.

Ancillary services from generators and other resources provide the system operators with the much-needed flexibility to manage tasks **(OT1)**-**(OT4)** [5]. In the context of this dissertation, the term ancillary services is used to encompass all services that are needed for supporting the delivery of electricity from generators to consumers and the following classification is adopted for our research:

(AS1) active power services needed to manage the supply-demand balance at various time scales (many existing services fall under this category; a list is provided in Table 1.1),

(AS2) reactive power services needed to maintain voltages within prescribed limits throughout the grid, and,

(AS3) back-up reserve services needed to enable black-start of the system in

the event of a blackout.

The ancillary services are used directly for the previously enumerated operational tasks: the first two tasks **(OT1)** and **(OT2)** are managed by service **(AS1)** while the last two tasks, **(OT3)** and **(OT4)**, directly correspond to services **(AS2)** and **(AS3)** respectively.

Table 1.1: List of Ancillary Services in Category **(AS1)**

Condition	Response Times	Typical Services	Current and Potential Service Providers
normal operation	slow (several hours to several days)	scheduling (baseline)	nuclear, coal, hydro, load shifting
	moderate (several minutes to couple of hours)	load following, energy imbalances	hydro, gas turbines, chillers in air conditioners
	fast (seconds to couple of minutes)	primary response, regulation	governors, flywheels, batteries, commercial air conditioners
contingency opertation	moderate (several minutes to couple of hours)	replacement, spinning & non-spinning reserves	hydro, gas turbines, direct load control, flexible manufacturing
	fast (seconds to couple of minutes)	spinning reserves	gas turbines, under-frequency load shedding

With increased awareness of global warming and green house gas emissions, *environmental* impacts of electricity production have been included within the scope of power system planning and operations, in addition to the usual goals of *low cost* and *reliability*. Considerable investments have been made in wind and solar energy resources [6]. The deployment of renewable resources presents major challenges in system operations because of the variable and unpredictable nature of their outputs. In particular, renewable generators are *intermittent*, with their outputs frequently oscillating between zero and full capacity, as seen in Figure 1.1. Furthermore, the minute-to-minute generation outputs as well as the average daily generation exhibit high *volatility* – sharp

fluctuations and large deviations from the mean values – as also evidenced from Figure 1.1. And there is a significant amount of *uncertainty* regarding the realized outputs, especially when predictions are taken hours or days in advance.

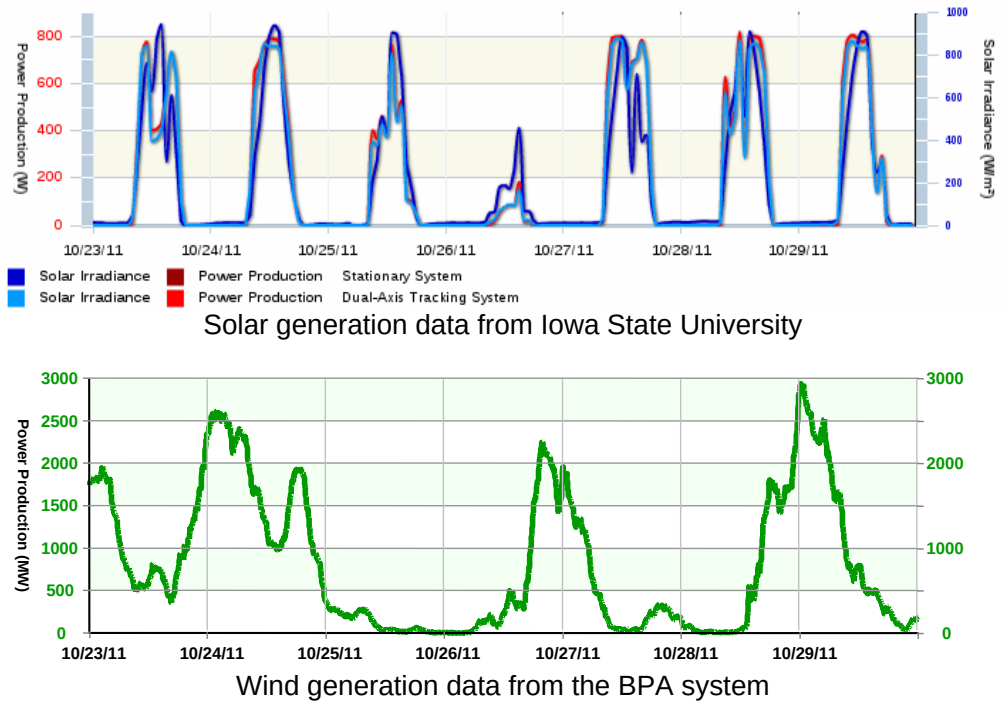


Figure 1.1: Plots of wind and solar generation outputs for a typical week to showcase the intermittent and volatile nature of renewables.

The large-scale deployment of renewable generation is expected to introduce higher variability, greater uncertainty and increased dynamics in the power grid [7–9]. The transition towards a sustainable future requires appropriate resources and technologies along with associated decision and control methodologies to mitigate such impacts. Specifically, there is a need to rethink operational practices for power systems with high penetration of renewable generation [10–13]. Also needed are resources providing various ancillary services that offer system operators flexibility on varying time scales to maintain system reliability and integrity [14–16].

The most accepted, and also the most widely sought after, source of flexibility today is gas turbine generators: these are often ramped up/down in

response to the needs of the power grid [17]. However, ancillary services procured from gas turbines are neither cheap nor environmentally friendly. Energy storage is a cleaner alternative and, perhaps, provides more flexibility as well: it can be used to smooth out the high frequency variations in renewable generation outputs as well as to store energy during periods of surplus generation and discharged to augment renewable supply during shortfalls [18–20], thus avoiding spilling of wind/solar energy as well as purchasing of expensive gas-turbine-based ancillary services. However, energy storage exists in limited form today and is not always economical.

The growth in supply-side ancillary service providers is not commensurate with the growth in renewable technologies. Hence, demand-side alternatives are being sought. Loads which possess inherent storage-like characteristics (for example, refrigeration systems, water heaters, air conditioners and electric space heaters) are a useful source for ancillary services. The power consumption of these loads can be manipulated around the nominal operating points without impacting the overall end-use [21–23]. Additionally, other demand-side resources that can ramp up or down their power consumption in response to needs of the grid are also an important source of flexibility; examples include agricultural farms and manufacturing plants [24, 25].

Significant efforts from academic research, industry innovation and utility pilot programs have been directed towards exploring the myriad sources of flexibility. For instance, research on energy storage technology has seen a boost [26, 27] along with a growth in the allied business sector [28]. Mechanisms to manipulate end-use power consumption have been created and deployed via so-called demand response (DR) programs[1] to exploit demand-side flexibility [30–32]. Many studies have been undertaken to determine flexibility afforded by the different resources and to investigate the interplay of such resources with variable renewable generation (see for example reports [16,20,33]; a survey is provided in Section 1.3).

Likewise, power system operational paradigms have experienced a gradual change. Some of the changes in the operational practices have been acceler-

[1]DR programs encourage demand-side participation in power system operations; such participation leads to load changes induced in response to the needs of the grid either in lieu of incentive payments or to exploit lower electricity prices [29].

ated by government policy. For instance, the regulatory push from renewable portfolio standards has caused system operators to attempt to maximize usage of volatile renewable energy in the grid. Consequently, many operators like BPA and CAISO have changed their scheduling and commitment policies, adopting intra-hour or near-real-time scheduling mechanisms to ensure reliability in spite of uncertain wind and solar energy usage [34,35]. Other changes have been necessary to leverage the new resources and technologies available to power system operators: an example includes the fast regulation service introduced in ERCOT to harness the fast response potential of flywheels and DR loads [36]. Operational and control practices are expected to undergo further modifications with the availability of advanced technologies such as phasor measurement units and smart meters.

Is changing operational practices and adding DR into the resource mix enough to facilitate deployment of renewable generation? Most certainly not. We argue that successful integration of renewable resources will require

- thorough accounting of ancillary service needed for power grids with a high penetration of renewable resources;
- comprehensive assessment of the quantity, quality and types of ancillary services that can be delivered by different resources; and
- control techniques that recognize the unique characteristics of different resources and effectively harness their flexibility potential to provide ancillary services to the grid.

Our research addresses the above needs, with primary focus on control techniques for future power grids with integrated renewable, storage and DR resources.

This dissertation quantifies the impacts of variable and uncertain renewable generation and determines the flexibility provided by energy storage and demand response resources. Specifically, preliminary investigations on the amount of ancillary services needed for reliable integration of renewable generation are reported. Also, provision of ancillary services by storage and DR resources is discussed and control algorithms for extracting ancillary services from such resources are developed. And, techniques for coordinating the dispatch of different resources in a power grid operations are proposed.

The main contributions of the reported research lie in the domain of control techniques. New stochastic control algorithms capable of capturing the salient characteristics of new resources and future power grids are proposed for control synthesis. The development of the algorithms is facilitated by recent advances in Markov decision theory, approximate dynamic programming (ADP) and reinforcement learning (RL). Based on numerical experiments, the proposed control algorithms are found to be remarkably effective for practical problems in power grids. The control algorithms developed in the course of this research provide a starting step towards the development of advanced control techniques that will be necessary for future power grids.

The next section discusses power grid operations from a control theory point of view and provides some insights on developing a control architecture for future power grids.

1.2 A Control Perspective on Grid Operations

As discussed in Section 1.1, the power system operations involve control of resources in order to balance supply and demand, satisfy network constraints, maintain an adequate voltage profile and ensure black-start capabilities. The system operator has many different resources at his disposal to complete these tasks. These resources include thermal generators, energy storage and flexible loads/DR.

Each resource – be it a generator or a storage unit or a responsive load – has unique operational characteristics and is subject to a range of physical constraints. For example, all generators operate under ramping and capacity constraints; however, their response times can be vastly different. Likewise, battery storage systems may be capable of fast response but their operation is subject to complex intertemporal constraints associated with their state of charge. Also, energy usage in a building system is flexible, but it is also subject to constraints that are not fully understood today. The complexity of the power grid with its diverse set of resources, each with its own dynamical properties and constraints, makes power system operations and control challenging.

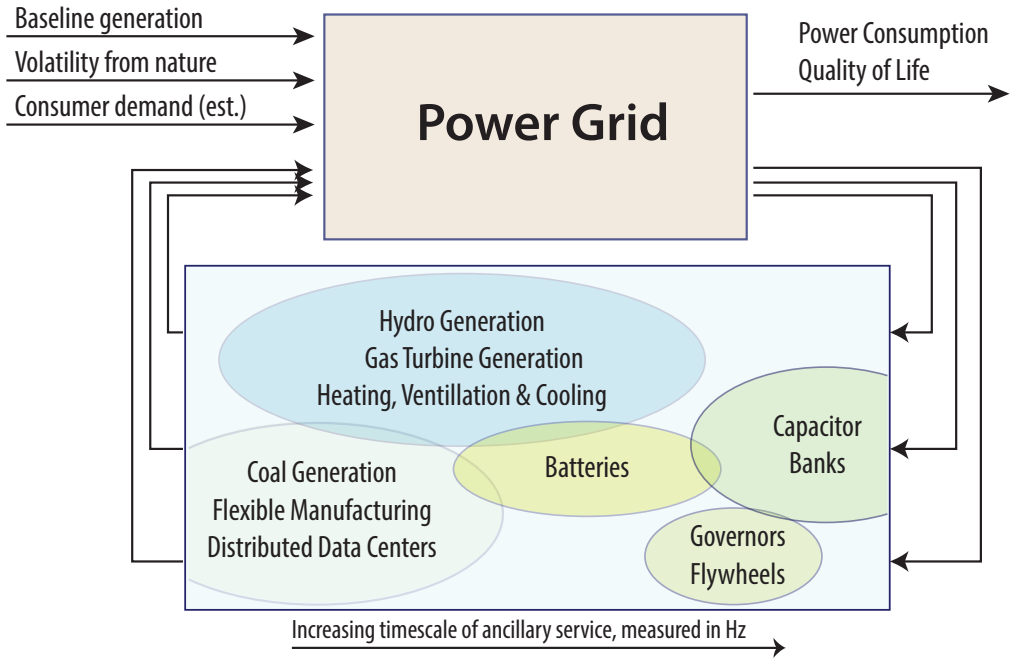


Figure 1.2: Power grid operations from a control theory point-of-view.

In essence, the power grid is a giant interconnection of many generators and loads connected via wires and many power electronic interfaces. From a control perspective, the grid can be viewed as a massive feedback control loop system whose architecture depends on various constituent components, as shown in Figure 1.2. The operation of these different resources such as coal generators, gas turbines, flexible loads and so on, depends on their own dynamics and constraints but can be manipulated to various degrees for supporting the needs of the power grid. The needs of the power grid are, in themselves, a manifestation of the needs of the producers and consumers, environment, policy mandates and other factors. The control architecture is responsible for translating these needs into operational commands for the various resources.

A key step towards creating the "right" architecture for future power grids – with renewable, storage and DR resources – is to devise operational and control policies that simultaneously optimize reliability, economic and environmental objectives while taking into account the dynamics, uncertainties and physical limitations of the grid and its constituent resources [10–13]. A successful architecture will require answers to the three issues raised in Section 1.1, which

in turn lead to a multitude of questions; some of these include:

(Q1) How to characterize the uncertainty and volatility associated with outputs of renewable generators?

(Q2) What are the impacts of increasing penetration of renewable generation on ancillary service needs of the grid and deployment of other resources?

(Q3) Which ancillary services can be tapped from storage and DR resources? What is the quality and quantity of these services and the associated costs?

(Q4) How can the flexibility afforded by different resources complement the volatility of renewable generation and aid in its deployment?

(Q5) How to control different resources while balancing the possibly competing goals of low cost, reliability and minimal environmental impact?

(Q6) What is the information structure required for controlling the grid and its constituent resources?

The objective of this research is to provide modeling and analytical tools to answer these questions. An allied goal is to provide insights for designing operational tools, guide policy development and steer long-term planning.

The following section summarizes research directed towards understanding the questions enumerated above. Specific gaps in the current state of the art are pointed out to emphasize the contributions of this dissertation.

1.3 Survey of the State of the Art

Renewable generators, flexible loads and energy storage are emerging as important resources in the electricity industry. Several industry developments, regulatory initiatives and academic research and reports are instrumental in furthering the participation of these resources in power system operations. This section provides a brief summary of these developments along with research pertaining to questions **(Q1)** to **(Q6)** listed in Section 1.2.

The move towards large-scale integration of renewable generation is motivated by environmental concerns and accelerated by government mandates

that have set aggressive targets in renewable portfolio standards and/or energy policies [37,38]. Consequently, many researchers have attempted to understand question **(Q1)** – the unique characteristics of energy derived from wind and solar resources – and a few studies are reviewed here. In an attempt to characterize variability inherent to renewable generation, reference [39] proposes the use of power spectral density of the generation outputs to determine the fill-in power that must be provided to compensate for the output fluctuations. A comprehensive assessment of the forecast uncertainty associated with the outputs is contained in [40,41]. An allied concern is the impact of geographical diversity on the variability and uncertainty of renewable generation. Studies indicate that spatial separation can smooth out the variability in the aggregate wind and solar power outputs, and also reduce the associated prediction error [41, 42].

Many investigations have focused on question **(Q2)**, the system-wide impacts of renewable generation deployment. A fairly well-accepted conclusion from a multitude of these studies is that as volatile renewable generation increasingly contributes towards the meeting the system load, more ancillary services – in the form of standby operational reserves as well as frequency regulation reserves – will be required to ensure system reliability (see [8, 9, 43, 44] and the references therein). It is argued in [45, 46] that faster responding resources are needed to minimize the impact of renewable energy deployment.

The past decade has also witnessed the wide adoption of DR programs, brought about by the confluence of many factors: absence of large storage alternatives, increased ancillary service needs synonymous with renewable deployments, quest for clean sources of ancillary services and regulatory push in the form of FERC orders 745[2] and 755.[3] The most significant success stories concerning DR involve bilateral contracts between utility companies and their large industrial and commercial consumers [49, 50].

The recognition of the flexibility potential of demand-side resources is not

[2]The 2011 FERC ruling 745 requires system operators to pay the same price for DR capacity as they would pay for generation capacity, thus leveling the playing field for supply- and demand-side resources [47].

[3]The 2011 FERC ruling 755 dictates system operators to compensate regulating resources based on their performance, thereby encouraging participation of fast-responding storage and DR resources in frequency regulation [48].

new. In fact, the newer practice of DR programs finds its roots in demand-side management practices employed by utilities during the 1980s and 1990s [51,52]. Seminal work by Schweppe *et al.* [53], which considers control of thermal loads for frequency regulation, provides the first theoretical foundation for answering question **(Q3)** concerned with characterization of DR-based ancillary services. Recent academic endeavors have been focused on aggregating residential loads such as refrigerators, air conditioners and water heaters for ancillary service provision (see [54–58] and the references therein). Likewise, DR potential of commercial loads has also been analyzed [59–61]. Many of these references focus on the quantity rather than quality of services that the DR resources can provide. In particular, little information is known of the level of uncertainty associated with DR. The time response characteristics of DR resources are not well understood either.

DR provides many different ancillary services to the power grid and offers a potential means to manage the impacts of renewable resources. Indeed, question **(Q4)** – dealing with the interplay between renewable generation and flexible loads – has been the focus of many studies. For instance, the use of price responsive demand to facilitate renewable integration has been proposed in many studies [62–65]. Likewise, coordinating the energy consumption of residential loads for ancillary service provision to mitigate impacts of renewable intermittency has also been extensively studied [54, 58].

Energy storage is another source of flexibility for the system operator. There is a vast body of literature concerned with tapping this flexibility for ancillary services in wind- and solar-integrated systems, thus making inroads towards answering questions **(Q3)** and **(Q4)**. For instance, references [18,19] provide a comprehensive assessment of the types of services that different storage technologies can provide. Likewise, the control of a generic storage resource to regulate wind farm outputs has been studied (see [66–68] and the references therein). Many investigations focus on deploying specific storage technologies in power grids with renewable resources. For example, use of storage potential of electric vehicles for supporting renewable generation deployment is studied in [69, 70]. Likewise, technical and economic aspects of employing advanced storage technologies such as compressed air energy storage and hydrogen fuel

cells in power grids have also been analyzed [71, 72].

As power grids transition towards increased use of renewable resources, the value of the flexibility afforded by different resources is expected to increase [73, 74]. These studies further the understanding of question **(Q4)** by proposing an quantitative framework to characterize operational flexibility. Qualitative evaluation of operational flexibility of generation, storage and DR resources is provided in [15], [20] and [54, 56] respectively. Modeling frameworks that capture such flexibility potential of a resource along with the associated dynamics and uncertainty have also been proposed [67, 75].

Question **(Q5)** pertaining to the coordination of different grid resources and the control architecture for the future grids has attracted considerable attention over the last decade. The reader is referred to reference [11] for a recent survey on the control implementations for the power grid and its component resources. Of particular interest is how to deal with the limited controllability of renewable generators like wind turbines and solar panels; some suggestions and requirements are outlined in [11]. Likewise, the importance of a new modeling framework which captures the impacts of dynamics and uncertainty for operational decision-making tools is stressed in [13]. Suggestions for practical implementations of such decision support tools are proposed in [12], and ramifications of multi-objective optimization to balance reliability, economic and environmental objectives are discussed in [76].

Recent advances in communication, computing and metering technologies may have profound impacts on the information architecture of future power grids. However, few studies focus on the related question **(Q6)**. The reader is referred to references [77] and [78] for a comparative analysis of centralized, decentralized and hierarchical architectures.

The investigations reviewed in this section provide insights on the questions raised in Section 1.2; these insights can be applied towards developing a control architecture for the future grid. However, one roadblock in devising such an architecture is the limited understanding of question **(Q1)** pertaining to the uncertainty and the dynamics that come into play as far as renewable generation is concerned. This limits the development of appropriate control strategies for these resources. The research reported in this dissertation offers

possible solutions; the specific contributions are outlined in the next section.

1.4 Scope and Contributions

The objective of this dissertation is to provide modeling and analytical tools to answer the questions enumerated in Section 1.2. The answers to these questions provide insights for designing operational tools, developing new policies and long-term planning.

The principal contribution of our research is the new control techniques for management of resources in power grids. Tools from stochastic control are used to shed light on the questions **(Q2)** to **(Q5)** and provide a work-around for the limited knowledge regarding question **(Q1)**. The reported research can be divided into two separate sets of investigations. One set of investigations answers questions **(Q2)** and **(Q3)** by analyzing the impacts of renewable and DR resources. The other set develops new control techniques for investigating questions **(Q4)** and **(Q5)**, in spite of the limited understanding of the uncertainty associated with renewable generation and DR.

The scope of these investigations is limited to active power control and the associated ancillary services; that is, services falling under category **(AS1)**. Extensions to reactive power control entail modeling changes; in particular, AC power flow models and voltage dynamics need to be explicitly considered. These extensions are left as an exercise for some future work.

1.4.1 Impact of Renewable and DR Resources

Investigations on questions **(Q2)** and **(Q3)** are presented in Chapter 2. These investigations provide a preliminary assessment of the amount of ancillary services needed for reliable deployment of renewable generation. Also, demonstrations of how DR-based ancillary services can aid renewable integration and participate in power system operations are provided.

The first study focuses on the operational impacts of wind generation deployment and the availability of DR-based load following and load shifting services. A unit commitment framework is adopted for analysis: the day-ahead scheduling problem for generation and DR resources is cast as an allocation

of resources under uncertainty and modeled as a two-stage stochastic control problem. It is assumed that the distribution of the uncertainty is known and that DR is available for provision of balancing as well as load shifting ancillary services. A proof of concept of the effectiveness of both DR services in facilitating wind generation deployment is provided via numerical simulations. Also, the amount of ancillary services needed is characterized based on the statistics of the renewable generation outputs.

The second study investigates how frequency regulation services can be extracted from heating, ventilation and air conditioning (HVAC) loads of commercial buildings. A feedforward architecture is proposed to modify the HVAC power consumption to track the regulation signal. The performance of the controller is tested via simulation experiments. An important observation from this study is that to avoid conflict with the existing temperature control system of the building, the low frequency variations of the regulation signal should be filtered out. Simulations demonstrate how the proposed controller has minimal impact on the building environment and can successfully track a high frequency regulation signal.

1.4.2 Learning-based Control in Power Grids

The work presented in Chapter 2 assumes that the distribution of the uncertainty is known. In the chapters 3, 5 and 6, this requirement is relaxed and, instead, the uncertainty is *learned* as a part of the solution. Such control synthesis is made possible by use of ADP- and RL-based stochastic control techniques. The main advantage of these algorithms is that they require very little knowledge of the underlying stochastic processes and can accommodate real-world data. Due to this feature, the techniques are eminently suitable for dealing with renewable generation and DR resources. Chapters 3, 5 and 6 include several examples to illustrate the application of these techniques to practical problems. The proposed control techniques provide tools to answer questions **(Q4)** and **(Q5)** raised in Section 1.2 in spite of the limited understanding of question **(Q1)**.

In Chapter 3, two RL techniques – SARSA and TD learning – are used to develop control strategies for energy storage and DR resources to enable

provision of load following and frequency regulation services. As an example, a control scheme for an energy storage unit operating in conjunction with a volatile wind generator is devised so that the combined resource can meet steady as well as time-varying demand. The control problem is cast as a Markov decision process (MDP) and solved using ADP and RL techniques. The learning technique is employed to size storage units for real wind farm locations so that the aforementioned objectives are met. Next, the control schemes for heating and cooling loads to provide frequency regulation services are examined. The resulting control problem is structurally similar to the control problem of combined wind-storage resource operation. The numerical examples presented here illustrate how ADP and RL can be successfully applied to devise controllers for newer generation ancillary service providers such as storage and DR resources.

In Chapter 5, two new techniques for Q-learning are devised and applied to an economic dispatch problem. Simulation experiments indicate that the learning algorithms can be successfully tuned to the underlying statistics of the system, thus avoiding the need to impose restrictive assumptions on the uncertainties in the system model. Furthermore, the Q-learning techniques are used to enhance the performance of model predictive control (MPC). We argue that the combination of the two approaches, referred to as *Q-MPC*, is an effective mechanism to address control problems arising in constrained power grids with higher levels of uncertainty.

The usual implementation of the economic dispatch problem can be cast in the MPC framework and its computational complexity can be reduced via use of the Q-MPC approach. The numerical experiments reported in chapters 5 and 6 showcase the capabilities of the Q-MPC approach to handle large-scale control problems associated with dispatch of resources in constrained power networks. For the three different text systems studied in our experiments, the Q-MPC approach expends the least computational effort as compared to standard MPC implementations, in the sense that good performance is obtained even for small prediction horizons. The improvement in performance and greater adaptability of Q-MPC is of particular importance in large grids with many resources and many sources of uncertainty.

1.4.3 Contributions to RL and ADP

In addition to demonstrating applications of ADP- and RL-based control techniques to power system control problems, our research also makes significant contributions to the area of stochastic control. The technical contributions of this dissertation are three-fold: an analytical architecture to examine approximate MDP solutions, two new Q-learning techniques and a proposed improvement on the standard MPC framework.

In Chapter 4, an architecture is developed for analyzing approximate MDP solutions that may be obtained via ADP, RL or any other technique. The central component of this architecture is the error in the approximation: bounds on the error are used to provide sufficient conditions for stability as well as for establishing performance bounds. Additionally, closed form solutions for relaxations to the MDP model are obtained; these solutions provide a starting point for constructing an approximate solution to the MDP.

In Chapter 5, two new parameterized Q-learning algorithms are presented. These algorithms provide a significant improvement over the standard Q-learning technique proposed by Watkins and Dayan in [79]. Watkins and Dayan's algorithm requires a parameterization of *all* Q-functions and, hence, fails for large state/action spaces. Q-learning based on a finite-dimensional parameterization has been considered only in very special cases, such as optimal stopping [80], a particular class of queuing models [81], and a class of deterministic models in [82].

Although MPC is a popular technique and finds many applications in power systems, its speed and complexity leave scope for improvement. In Chapter 6, the new Q-learning algorithms devised in the preceding chapter are used to improve the MPC technique. Specifically, the Q-learning algorithms are used to approximate the *optimal* terminal cost for the MPC implementation. The resulting controller admits a stabilizing policy under mild conditions. Also, guidelines for approximating the MPC terminal cost are provided.

1.5 Dissertation Outline

The dissertation contains six additional chapters. Chapter 2 starts with a review of the power system operations and discusses the impacts of wind and DR deployments from an operational standpoint. A unit commitment framework is introduced for scheduling resources in a power grid with wind energy and DR sources. Numerical studies using this framework provide an estimate on the increased ancillary service needs associated with wind generation deployments as well as a proof of concept on how these needs can be met with DR-based ancillary services. Additionally, a feedforward control architecture is presented for manipulating power consumption in HVAC loads of commercial building for the purposes of frequency regulation. Numerical experiments indicate that, as long as the frequency content of the regulation signal is suitably constrained, the regulation service provision has minimal impacts on the indoor environment of the building.

Chapter 3 introduces the reader to tools from Markov decision theory, ADP and RL. It also presents a modeling framework for the power grid. These models and tools are applied to develop control policies for energy storage and DR resources. In particular, the control of storage resources to smoothen the variability of wind generation is studied. Likewise, the problem of controlling thermal loads for frequency regulation is also studied. The control problems presented here correspond to control synthesis on medium to fast time scales.

In Chapter 4, an analytical framework to judge the quality of approximate solutions to MDPs is presented. An error criterion is defined in the context of average cost optimization for MDPs. Relaxations to the MDP models, called fluid models, are introduced and closed form solutions for the fluid value function are derived to form the basis for approximating the MDP solutions. Bounds on the error in the approximation are used to define sufficient conditions for stability of the control policy derived from the approximate MDP solutions. Performance bounds for the approximate policy are also provided.

In Chapter 5, two new techniques for approximating solutions to MDPs are devised. Specifically, two parameterized Q-learning algorithms based on Bellman error reduction and linear programming approach to solving MDPs are proposed. The implementation of the algorithms for power system control

applications is discussed via numerical examples. A specific case of economic dispatch is considered. Our results show how Q-learning can be used to improve the fluid model-based approximations for MDPs.

Chapter 6 describes the marriage of Q-learning with MPC resulting in the new Q-MPC approach to control both stochastic and deterministic systems. The stability of the controller is established under mild conditions. The approach is applied to the economic dispatch problem. Numerical experiments demonstrate how Q-MPC provides close-to-optimal solutions for small prediction horizons. The dissertation concludes with a summary of the research and discussion of the avenues for future work in Chapter 7.

Chapter 2

Operational Impacts of Wind and DR Integration

This chapter provides an overview of the operational decision-making process for a power system. It also includes investigations on understanding the role of renewable generation and DR loads in power system operations.

The first investigation is concerned with the interplay between variable and uncertain renewable generation and ancillary services from DR resources. As an example, wind-integrated systems are considered and naive models for the uncertain wind energy outputs are assumed. Two types of ancillary services are considered: balancing service and load shifting. Numerical experiments quantify the increased ancillary service needs associated with use of wind generation. They also serve to demonstrate the effectiveness of DR-based ancillary services in meeting these needs and facilitating wind generation deployment.

The second investigation concerns the use of HVAC loads to provide frequency regulation service to the power grid. A feedforward control approach for manipulating the HVAC power consumption in commercial buildings is proposed and tested on a detailed model of a real building. Numerical experiments provide insights on how to enable effective regulation signal tracking for the controller without compromising in the occupants' comfort.

The two detailed examples of DR-based ancillary services presented here correspond to two different time scales. The first investigation is concerned with moderate time scales (minutes) while the second investigation focuses on fast time scales (seconds).

2.1 Overview of Power System Operations

Operational decisions in power systems are concerned with controlling the grid resources in the most *economical* manner without violating the various constraints on the individual resources as well as the system as a whole. The time scales for these decisions range from a few seconds to several days. This section describes the operations corresponding to these time scales for active power control of resources. The associated decision-making processes can viewed in three domains: commitment, dispatch and regulation. Figure 2.1 depicts these processes and the interactions between them.

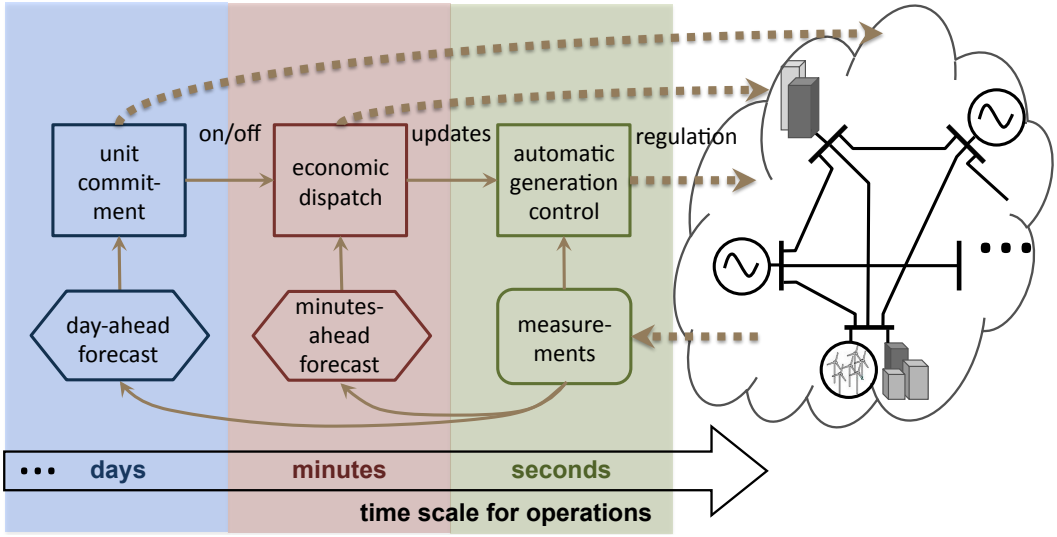


Figure 2.1: The key stages in power system operations.

The system operator schedules resources one day prior to the actual production and delivery of energy in such a way that physical constraints are met and supply-demand balance is maintained. As the supply and demand are not perfectly predictable in the day-ahead decision-making process, the resource outputs are updated during the dispatch process based on revised estimates of real-time conditions. These updates are computed every 5 to 10 minutes. The faster fluctuations in supply and demand are managed by regulating the resource outputs in response to deviations in the system frequency and scheduled tie-line flows.

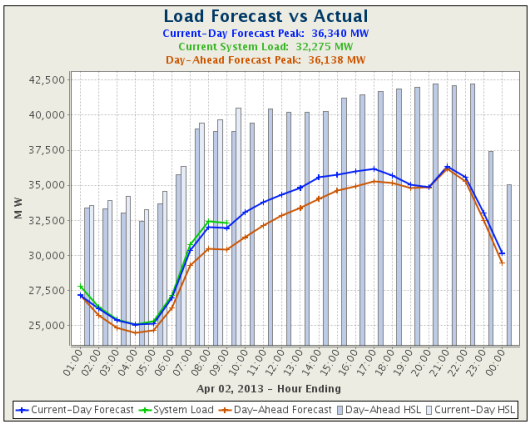
The scheduling decisions are concerned with turning ON or OFF genera-

tion units and are determined via the unit commitment (UC) process. The objective of UC is to determine the least cost resource schedule to meet the predicted load demand for a specified time period while satisfying the limitations imposed by the physical resources as well as those imposed by the various operating policies. The UC process determines the energy and reserve commitments for the system resources a day ahead of real time and is usually performed once per day to account for the predictable, larger, slow changes in demand. The scheduling process is performed for a day at a time, with a rolling look-ahead horizon of up to 3 days with a half-hourly or hourly time resolution.

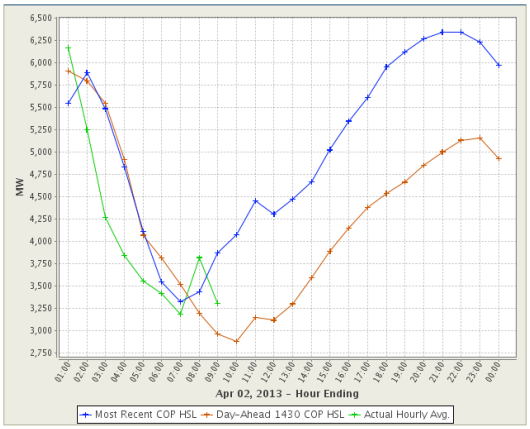
The economic dispatch uses a revised load and generation forecast – minutes-ahead or hour-ahead of real time – to fine-tune the resource schedule determined by the UC process. The dispatch process allocates the total load on the system among the committed resources so as to minimize operational costs. Scheduled interchanges, reserve availability and various operational and physical constraints are considered in the dispatch process. The dispatch process is performed once every few minutes and has a look-ahead horizon of a few hours.

The system states are continuously changing due to fluctuations in the load and generation. The fast fluctuations manifest themselves in terms of deviations in the system frequency as well as tie-line flows. For stability and reliability reasons, the impact of these deviations should be minimized within prescribed limits; this is achieved by manipulating the operating points of a certain set of resources through the regulation process, also referred to as automatic generation control or load frequency control. The objective of the regulation process is to minimize frequency deviations and regulate tie-line interchanges. Accordingly, an appropriate error signal, known as the area control error, is computed and used to construct regulation commands that dictate how the regulating resources should change their operating points to minimize the consequences of the supply-demand deviations.

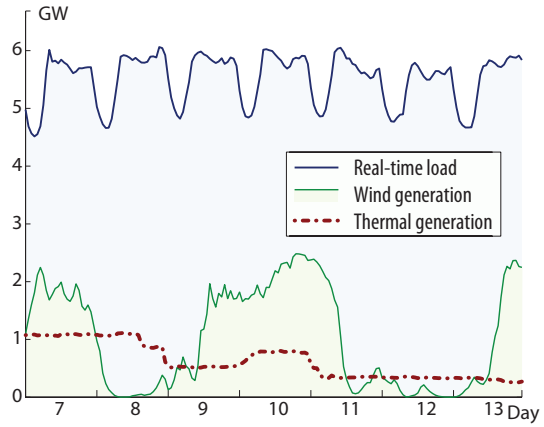
The operational practices outlined here have been traditionally concerned with control of generation resources, with variations in the load and unplanned equipment outages being the main sources of uncertainty. With increased



(a) ERCOT load: prediction versus actual



(b) ERCOT wind generation: prediction versus actual



(c) BPA load and generation for a typical week

Figure 2.2: Comparing wind generation and load demand patterns.

reliance on variable and uncertain renewable generation, these operational practices may no longer apply *as is* and may require modifications.

As such, managing variability and uncertainty is not an uncommon occurrence for power system operators. The current operational practices are designed to be robust, to some extent, to the uncertainty and variability in thermal generation as well as electric loads. This is made possible, in part, by demand exhibiting fairly well understood patterns with low prediction errors, as seen in Figures 2.2(a) and 2.2(c). However, wind generation patterns (see Figures 2.2(b) and 2.2(c)) are not as well understood — *it is unclear whether there exists any pattern at all.* Also, with limited control on their outputs, the role of renewable power plants in operations is ambiguous. In fact, wind generation deployments often impact thermal generation in the sense that it needs to be ramped up or down to compensate for lower or higher wind generation, as evidenced in Figure 2.2(c).

As more renewable resources replace conventional generation, ancillary services from alternative sources such as DR will need to be procured. This may necessitate changes to operating paradigms. In what follows, the interplay between wind energy and DR is investigated.

2.2 Supporting Wind Generation with Demand Response

A major concern in the day-ahead decision making process for systems with high penetration of renewable resources is how to deal with the uncertainty and variability introduced by these resources. An ad-hoc approach that has been adopted by many system operators is to manage the volatility in renewable generation outputs through procurement and deployment of supply-side reserves. This section is concerned with how DR can be effectively utilized to compensate the need for stand-by reserves in systems with deep penetration of renewable resources (as an example, wind generation is considered).

The determination of optimal reserve levels in a stochastic power system model can be cast as a variant of the newsboy's problem for optimal inventory modeling [9, 83, 84]. The economic efficiency of this optimal outcome has been

established in [9,83,84] and the references therein. The analysis in [9] concludes that as penetration of wind resources increases, more reserves will be required to ensure system reliability. Similar results have also been established in unit commitment-based analysis which explicitly consider uncertainty due to wind forecasts [43, 85]. Thus, there is the need for fast responding reserves from both the supply- and the demand-side. While the nature and costs of supply-side reserves are well understood, the demand-side options have not been fully explored. With the implementation of Smart Grid technology, it is envisioned that the demand-side will play a key role in facilitating reliable integration of the wind resources [29, 62, 86–89].

In what follows, the impacts of deploying volatile wind generation and harnessing the flexibility in demand-side consumption are investigated. The focus is on provision of demand-side reserve capacity and leveling of the load profile. The analysis is based on a UC model that explicitly represents the uncertainty in the day-ahead scheduling decisions using a two-stage stochastic control problem. Simulation results characterize the increased ancillary service needs in terms of the variability of the wind generation. Also, a proof of concept that DR-based reserve capacity provides an effective mechanism to manage the volatility and uncertainty of wind resources is provided. These results also demonstrate how load leveling relaxes the constraints imposed on the unit commitment solution and, thus, helps accommodate renewable generation.

The analysis described here *does not* extend to the complex feedback loop found in a system in which prices to consumers vary according to the current environment, as in many DR programs such as real-time pricing mechanisms. In fact, this work is aligned with the viewpoint propagated by Callaway and Hiskens in [90]:

... in order to achieve full responsiveness, direct load control (as opposed to price response) is required to enable fast time scale, predictable control opportunities, especially for the provision of ancillary services such as regulation and contingency reserves.

2.2.1 Scheduling of Generation and DR Resources

From the discussion in Section 2.1, it is clear that the system operator procures energy and reserve commitments from resources the day before real time based on predictions of the supply and demand. The resources are dispatched closer to real time based on revised estimates of system conditions; the reserves are deployed to manage the deviations from the promised supply and demand energy commitments. In this section, a stochastic framework which mimics this decision process of the operator is described.

The Setup

It is assumed that the power grid under consideration has wind resources. The "use all wind" policy is adopted: that is, the controllable generators serve the net-load imposed on the system after all available wind generation is absorbed into the system. The following DR scenarios are considered:

- The availability of DR capacity as reserves from the implementation of incentive-based DR scheme such as emergency DR program, and,
- The leveling of the demand profile brought about by the consumer response to time-of-use prices that are known in advance, on a day-ahead basis (or longer).

The consumer loads providing DR reserve capacity are modeled analogous to very reliable, fast-start generators that only contribute towards system reserves. The amount of reserves that such consumers can provide is constrained by the amount of load they are consuming – they can only reduce as much load as they consume. The response of the consumers enrolled in the time-of-use pricing scheme is modeled by modifying the demand profile. It is assumed that implementation of time-use-rates will cause consumers to shift from high-price to low-price hours, thereby resulting in load leveling. Furthermore, pricing structure is assumed to induce responses in the consumption patterns of the enrolled consumers that are perfectly predictable in the day-ahead.[1] Thus,

[1]While such an assumption would most likely not hold for the real world, the analysis without this assumption would be complicated by the representation of rationality in consumer behavior which is far beyond the scope of this work.

the day-ahead demand predictions are considered to be representative of the impacts of pricing-based DR program.

Stochastic Unit Commitment Problem

The UC problem is cast as a *two-stage stochastic program* with *recourse actions* – a special case of the multi-stage stochastic optimal control problem. Such programs are often applied to inventory management problems [91]. The multi-stage decision framework mimics the decision-making process of the system operator wherein the resources are scheduled and reserves are procured in the day-ahead while the deviations from the scheduled supply and demand in real time are managed using the procured reserves.

In the proposed UC formulation, the determination of the day-ahead commitment schedule is modeled as the first stage of the stochastic control problem. The real-time balancing operations depend on both the day-ahead commitment schedule as well as the real-time conditions, which are not known in the day-ahead. The real-time supply-demand balance is maintained by dispatching reserves procured in the day-ahead and/or shedding load[2] subject to the constraints on both the reserves dispatched and the load shed. The system operator's real-time decision-making process is modeled as a recourse action – the second stage of the stochastic control problem.

A compact formulation of the stochastic UC problem for a power system with I generators and J DR reserve providers for a scheduling period of T time steps is as follows:

$$\min_{\substack{\underline{\hat{\boldsymbol{u}}}_{\mathsf{G}},\underline{\hat{\boldsymbol{\varepsilon}}}_{\mathsf{G}},\underline{\hat{\boldsymbol{r}}}_{\mathsf{G}} \\ \underline{\hat{\boldsymbol{u}}}_{\mathsf{D}},\underline{\hat{\boldsymbol{r}}}_{\mathsf{D}}}} \quad \sum_{t} \left\{ \sum_{i} f_{\mathsf{G}i}\left(\hat{u}_{\mathsf{G}i}(t), \hat{\varepsilon}_{\mathsf{G}i}(t), \hat{r}_{\mathsf{G}i}(t)\right) + \sum_{j} f_{\mathsf{D}j}\left(\hat{u}_{\mathsf{D}j}(t), \hat{r}_{\mathsf{D}j}(t)\right) + \mathsf{E}\left[\varphi^{\star}\left(\underline{\hat{x}}(t)\right)\right] \right\}$$

$$s.t. \qquad \sum_{i} \hat{\varepsilon}_{\mathsf{G}i}(t) + \hat{\varepsilon}_{\mathsf{W}}^{\mathsf{tot}}(t) \;=\; \hat{\varepsilon}_{\mathsf{D}}^{\mathsf{tot}}(t) \qquad \forall\, t \tag{2.1a}$$

$$(\hat{\boldsymbol{u}}_{\mathsf{G}i}, \hat{\boldsymbol{\varepsilon}}_{\mathsf{G}i}, \hat{\boldsymbol{r}}_{\mathsf{G}i}) \in \mathbf{X}_{\mathsf{G}i} \qquad \forall\, i \tag{2.1b}$$

$$(\hat{\boldsymbol{u}}_{\mathsf{D}j}, \hat{\boldsymbol{r}}_{\mathsf{D}j}) \in \mathbf{X}_{\mathsf{D}j} \qquad \forall\, j \tag{2.1c}$$

$$(\underline{\hat{\boldsymbol{u}}}_{\mathsf{G}}, \underline{\hat{\boldsymbol{\varepsilon}}}_{\mathsf{G}}, \underline{\hat{\boldsymbol{r}}}_{\mathsf{G}}, \underline{\hat{\boldsymbol{u}}}_{\mathsf{D}}, \underline{\hat{\boldsymbol{r}}}_{\mathsf{D}}) \in \mathbf{X}_{\mathsf{T}}\ . \tag{2.1d}$$

[2]Such load shedding induced by the operator is not the same as the voluntary load curtailments by the consumers enrolled in DR programs. It is typically invoked only under extreme conditions.

In (2.1), $\hat{u}_{\mathsf{G}i}(t) \in \{0,1\}$ is the commitment status of generator i at time t; $\hat{\varepsilon}_{\mathsf{G}i}(t) \in \{0\} \cup \left[\varepsilon_{\mathsf{G}i}^{\mathsf{min}}, \varepsilon_{\mathsf{G}i}^{\mathsf{max}}\right]$ is the energy output of i with $\varepsilon_{\mathsf{G}i}^{\mathsf{min}}$ and $\varepsilon_{\mathsf{G}i}^{\mathsf{max}}$ as its capacity limits; and $\hat{r}_{\mathsf{G}i}(t)$ is its spinning reserve commitment. Similarly, $\hat{u}_{\mathsf{D}j}(t) \in \{0,1\}$ is the commitment status of DR reserve provider j at time t and $\hat{r}_{\mathsf{D}j}(t)$ is its spinning reserve commitment. The boldface notations $\hat{\boldsymbol{u}}_{\mathsf{G}i}$, $\hat{\boldsymbol{\varepsilon}}_{\mathsf{G}i}$ and $\hat{\boldsymbol{r}}_{\mathsf{G}i}$ denote the trajectories of the corresponding variables. For instance, $\hat{\boldsymbol{u}}_{\mathsf{G}i} := \{\hat{u}_{\mathsf{G}i}(1), \ldots, \hat{u}_{\mathsf{G}i}(T)\}$. The underlined boldface notations represent the collection of trajectories for all units. For instance, $\underline{\hat{\boldsymbol{u}}}_{\mathsf{G}} := \{\hat{\boldsymbol{u}}_{\mathsf{G}i} : i = 1, \ldots, I\}$. Constraint (2.1b) represents the supply-demand balance constraint for the predicted values of total demand $\hat{\varepsilon}_{\mathsf{D}}^{\mathsf{tot}}(t)$ and total wind generation $\hat{\varepsilon}_{\mathsf{W}}^{\mathsf{tot}}(t)$. The set $\mathbf{X}_{\mathsf{G}i}$ represents the physical constraints – such as the minimum up/down time constraints, ramping limits and capacity limits – on a generator i. Similarly, $\mathbf{X}_{\mathsf{D}j}$ represents the constraints on DR resource j. The set $\mathbf{X}_{\mathsf{T}}$ represents the operational constraints imposed by the transmission network. The function $f_{\mathsf{G}i}(\cdot)$ denotes the offer function of generator i: it implicitly incorporates the fuel charges, start-up/shut-down costs, capacity costs for reserve provision and other operating costs. Similarly, function $f_{\mathsf{D}j}(\cdot)$ represents the offer function of DR resource j. If the expectation term $\mathsf{E}[\cdot]$ in (2.1a) is ignored, this problem formulation is similar to the conventional unit commitment problem.

In the stochastic UC formulation, variables $\hat{u}_{\mathsf{G}i}(t)$, $\hat{\varepsilon}_{\mathsf{G}i}(t)$, $\hat{r}_{\mathsf{G}i}(t)$, $\hat{u}_{\mathsf{D}j}(t)$ and $\hat{r}_{\mathsf{D}j}(t)$ are the first stage decision variables. The variable $\underline{\hat{x}}(t)$ is a shorthand notation for the time-t first stage variables; that is,

$$\underline{\hat{x}}(t) := \left[\{\hat{u}_{\mathsf{G}i}(t), \hat{\varepsilon}_{\mathsf{G}i}(t), \hat{r}_{\mathsf{G}i}(t) \;\forall\; i\} \cup \{\hat{u}_{\mathsf{D}j}(t), \hat{r}_{\mathsf{D}j}(t) \;\forall\; j\}\right].$$

Clearly, the real-time balancing costs at time t depend on $\underline{\hat{x}}(t)$. The function $\varphi^{\star}(\cdot)$ in (2.1a) represents the optimal costs of the second-stage (wherein the real-time balancing actions are modeled) as a function of the first stage

decisions $\underline{\hat{x}}(t)$. It can be computed as follows:

$$\varphi^{\star}\left(\underline{\hat{x}}(t)\right) := \min_{\underline{R}_{\mathsf{G}}(t),\underline{R}_{\mathsf{D}}(t),L(t)} \sum_{i} h_{\mathsf{G}i}\left(R_{\mathsf{G}i}(t)\right) \quad + \quad \sum_{j} h_{\mathsf{D}j}\left(R_{\mathsf{D}j}(t)\right) + \upsilon L(t) \tag{2.2a}$$

$$s.t. \sum_{i}\left[E_{\mathsf{G}i}(t)+R_{\mathsf{G}i}(t)\right]+ E_{\mathsf{W}}^{\mathsf{tot}}(t) \quad = E_{\mathsf{D}}^{\mathsf{tot}}(t) - \sum_{j} R_{\mathsf{D}j}(t) - L(t) \tag{2.2b}$$

$$0 \leq |R_{\mathsf{G}i}(t)| \quad \leq \quad |\hat{r}_{\mathsf{G}i}(t)| \qquad \forall\, i \tag{2.2c}$$

$$0 \leq |R_{\mathsf{D}j}(t)| \quad \leq \quad |\hat{r}_{\mathsf{D}j}(t)| \qquad \forall\, j \tag{2.2d}$$

$$0 \leq L(t) \quad \leq \quad \ell^{\mathsf{max}}(t) \tag{2.2e}$$

$$(\underline{E}_{\mathsf{G}}(t), \underline{R}_{\mathsf{G}}(t), \underline{R}_{\mathsf{D}}(t), L(t)) \quad \in \quad \mathbf{X}_{\mathsf{T}}|_{t}\ . \tag{2.2f}$$

In (2.2), the capital letters denote random variables: these capture the uncertainty associated with real-time conditions. Variables $R_{\mathsf{G}i}(t)$ and $R_{\mathsf{D}j}(t)$ are the dispatched reserves of generator i and DR resource j, respectively, while $L(t)$ represents the unserved load at time t. Note that $\underline{R}_{\mathsf{G}}(t) := \{R_{\mathsf{G}1}(t), \ldots, R_{\mathsf{G}I}(t)\}$; $\underline{R}_{\mathsf{D}}(t)$ is analogously defined. The variables $E_{\mathsf{G}i}(t)$, $E_{\mathsf{W}}^{\mathsf{tot}}(t)$ and $E_{\mathsf{D}}^{\mathsf{tot}}(t)$ denote the uncertain real-time realizations of the generation from conventional and wind resources and demand respectively. Functions $h_{\mathsf{G}i}\left(\cdot\right)$ and $h_{\mathsf{D}j}\left(\cdot\right)$ represent the real-time dispatch costs of the corresponding resources while υ is the value of lost load (VOLL) associated with involuntary load shedding. Constraint (2.2b) is the real-time supply-demand balance constraint while constraints (2.2c)-(2.2e) represent the limits on the corrective balancing actions followed by the network constraints in (2.2f).

The constraints (2.2c) and (2.2d) and the expectation term in (2.1a) explicitly represent the coupling between the first and second stage decision variables. The solution of (2.1)-(2.2), denoted by $(\underline{\hat{\boldsymbol{u}}}_{\mathsf{G}}^{\star}, \underline{\hat{\boldsymbol{\varepsilon}}}_{\mathsf{G}}^{\star}, \underline{\hat{\boldsymbol{r}}}_{\mathsf{G}}^{\star}, \underline{\hat{\boldsymbol{u}}}_{\mathsf{D}}^{\star}, \underline{\hat{\boldsymbol{r}}}_{\mathsf{D}}^{\star})$, is the least-cost day-ahead commitment schedule which meets the predicted demand requirements, satisfies the physical constraints, and minimizes the expected cost of real-time balancing operations.

2.2.2 Numerical Results

The stochastic UC problem formulation proposed in Section 2.2.1 provides a testing platform for the capability of DR resources in facilitating deployment

of volatile wind generation (WG). Several simulation studies are performed to investigate the interplay between wind generation and demand response.

The VOLL used in these studies is analogous to the cost of blackout considered in [92, 93]. Similarly, the DR reserves are similar to the DR-based load shedding described in these references.

As such, the integer constraints and stochastic nature of the problem impose a huge burden on the computing resources. Therefore, certain allowances are made in the simulations studies to ensure computational tractability. In particular,

- transmission constraints are ignored;
- generators and DR resources are assumed to be 100% available; and
- the expectation term in (2.1a) is approximated by a sample average, where the samples are obtained through Latin hypercube sampling.

Demand and wind generation forecast errors are the main sources of uncertainty modeled in the following simulation studies. The real-time demand and wind generation are modeled as

$$\begin{aligned} E_{\mathsf{D}}^{\mathsf{tot}}(t) &= (1+N_{\mathsf{D}})\,\hat{\varepsilon}_{\mathsf{D}}^{\mathsf{tot}}(t) \qquad \text{with} \quad N_{\mathsf{D}\mathsf{sim}}\mathcal{N}\left(0,\sigma_{\mathsf{D}}^2\right) \\ E_{\mathsf{W}}^{\mathsf{tot}}(t) &= (1+N_{\mathsf{W}})\,\hat{\varepsilon}_{\mathsf{W}}^{\mathsf{tot}}(t) \qquad \text{with} \quad N_{\mathsf{W}\mathsf{sim}}\mathcal{N}\left(0,\sigma_{\mathsf{W}}^2\right), \end{aligned}$$

respectively, with σ_{W} typically greater than σ_{D}. Since the focus of this exercise is to investigate how DR can help manage WG, the assumptions imposed above are not altogether unreasonable.

The investigations are performed on three- and ten-unit systems of [94] and [95]. Both test systems are modified to suit the analysis: namely, availability of wind generation and demand response reserves is assumed and suitable parameters are adopted. For simplicity, a scheduling horizon of 8 periods is considered. The load and wind forecast patterns are generated using data is obtained from ISO-NE [96] and NREL [97]. The modifications are detailed in [44].

Example A: Three Unit System

In Figure 2.3, the number of hours units 2 and 3 are committed is plotted for the following five cases:

- base case with a peak-valley load forecast profile (no wind generation, no DR)
- WG case which simulates the base case system with a specified wind generation forecast
- DR case A which simulates the WG case with DR reserves
- DR case B which simulates the WG case with a levelized load forecast profile
- DR case which combines A & B scenarios – WG with DR reserves and a levelized load forecast profile

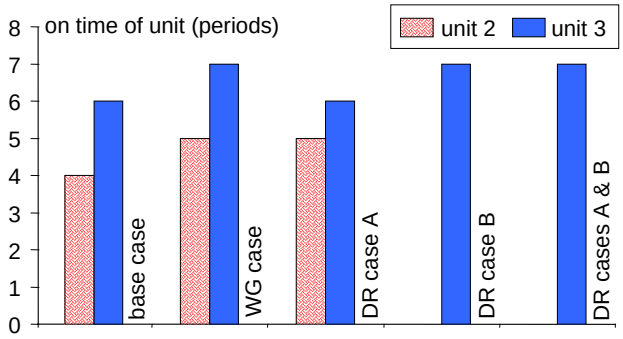


Figure 2.3: Operation of units 2 and 3 under different case scenarios for the 3-unit system.

Unit 1 is a base-load unit and is committed for all hours in all five cases. Notice that with a levelized load profile, the operator can do away with committing unit 2, the most expensive unit in the system. This has significant impact on the costs because when wind generation is deployed, unit 2 is kept ON only to provide reserves.

An added benefit of load leveling is increase in the load during the night hours when prices are typically low. This allows absorption of nighttime WG which may otherwise need to be curtailed to manage physical limitations on base load generators.

The economic impacts of WG and DR deployment are captured using the following cost metrics: cost of generation based on day-ahead commitment, start-up costs, capacity costs for generation reserves, capacity costs for DR reserves, and finally, the expected costs of real-time operations. In Figure 2.4,

these cost metrics for the system are presented for the five cases described above.

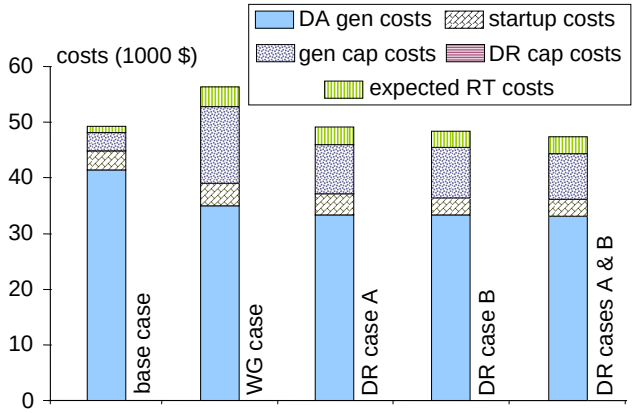


Figure 2.4: Cost metrics for different case scenarios for the 3-unit system.

With the deployment of zero-fuel-cost wind generation, the cost of generation reduces, but such cost reductions are offset by the increase in reserve capacity costs. Furthermore, injection of all available wind generation imparts variability to the net load, thus increasing start-up costs. In general, the real-time balancing costs are higher when wind generation is introduced. Combining wind deployments with DR – either as reserves or as levelized load profile – decreases the real-time costs, making wind resources viable for use.

Example B: Ten Unit System

When operational and economic impacts of DR resources are investigated for the ten unit system, results similar to those on the 3 unit system are obtained. Further studies are conducted on the ten unit system for sensitivity analysis.

First, a business case for DR reserves is posed. The UC problem is solved for a specified load and wind generation forecast for different VOLLs v. The expected real-time balancing costs, cost of procuring generation reserves and the expected load shed in real-time are plotted in Figure 2.5 for different v. The figure emphasizes the relationship between the VOLL and the procurement of expensive generation reserve capacity: As VOLL falls, it is more economical to shed load than to procure generation reserves. This exercise helps make a case for flexible loads such as commercial refrigeration systems to be used

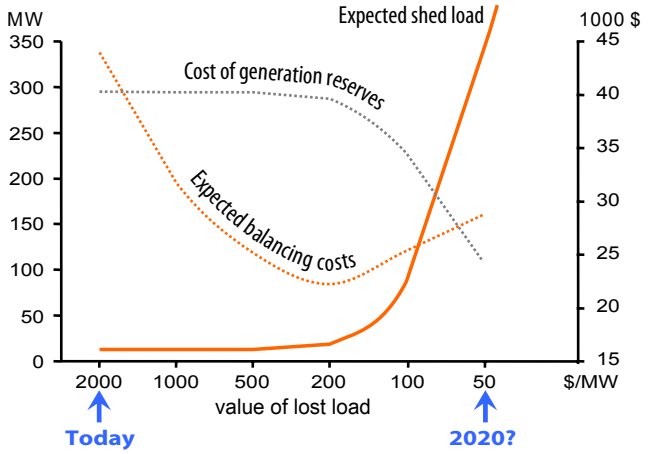


Figure 2.5: Optimal procurement of generation reserves as function of VOLL.

as reserves because for such loads, a temporary loss of load is typically not detrimental and hence the VOLL for such loads is lower than that of a critical load such as a hospital. Indeed, loads which participate in provision of reserves or direct load control (DLC) programs have low VOLL.

Next, impacts of increasingly uncertain wind forecasts are simulated by varying the parameter σ_W. The reserve generation capacity procured for the peak load period and the expected real-time balancing costs for different values of σ_W are plotted in Figure 2.6. From the figures, it is clear that while

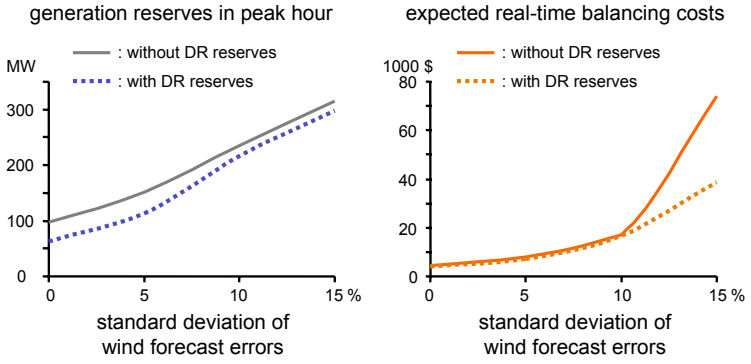


Figure 2.6: Impact of wind generation uncertainty on generation reserves and real-time costs.

increasingly uncertain wind generation can result in procurement of a large amount of reserve generation capacity and lead to higher real-time balancing costs, such impacts can be mitigated by DR reserves procured from responsive

loads. In this way, DR reserves provide an effective approach to managing wind uncertainty.

Remark. The analysis described in this section is limited by two assumptions. One, that the quantity of load that can be shed is known in advance for scheduling purposes. Second, that a consumer committed to provide DR reserve in the day-ahead will indeed curtail load when asked to do so. The first assumption has been partially addressed in [98] via proposal of establishing appropriate baseline load consumption models and quantifying the flexibility available in responsive demand resources. The second assumption can be justified by development of appropriate standards and/or market mechanisms for demand response.

2.3 Frequency Regulation from Commercial Building HVACs

The proper functioning of a power grid requires continuous matching of supply and demand, in spite of the randomness of electric loads and the uncertainty of power generation. A direct consequence of supply-demand mismatch is a deviation in the system frequency, which is closely monitored and controlled as discussed in Section 2.1. An important ancillary service used in managing the system frequency is the regulation service: it is deployed on the fastest time scale (seconds to minute) to correct the short-term power imbalance in load and generation to maintain system frequency within the prescribed limits. This service has been traditionally provided by generators by tracking a regulation signal sent by the grid operator that dictates changes in the generators' outputs. In this section, it is argued that (a) commercial buildings can be tapped for ancillary services, (b) HVAC systems can be manipulated for regulation service on faster time scales more effectively than generators, and (c) commercial buildings can provide this service at a very low cost.

Buildings account for 75% of total electricity consumption in the U.S., with roughly equal share between commercial and residential buildings [99]. With growing interest in procuring ancillary services from fast responding DR resources [22,100], buildings are a natural source of demand-side flexibility. The

choice of commercial buildings is motivated by several important factors. First, a commercial building can provide more flexibility (compared to a residential building) due to its much larger thermal inertia. Second, approximately one third of the commercial building floor space is equipped with variable frequency drive (VFD) that operates the HVAC equipment. It can be commanded to vary their speed and power consumption quickly and continuously, instead of in an on/off manner. This is a crucial advantage for providing regulation service, since the regulation signal to be tracked changes in the order of seconds. Third, a large fraction of commercial buildings in the United States are equipped with building automation systems (BAS). These systems can receive regulation signals from grid operators and manipulate the control variables needed for providing regulation service, without requiring additional equipment such as smart meters. Ancillary services can thus be provided at virtually no cost; these are obtained as a simple add-on to the current HVAC control system.

Many load control mechanisms explored for commercial buildings in the current literature are primarily concerned with low frequency changes in demand, i.e., the changes occur over a minutes/hours time scale. Here, the focus is on high frequency load changes in commercial buildings to provide regulation service to the grid. This section contains preliminary results showcasing the feasibility of extracting regulation service from commercial buildings. The power consumption of the fans in the building's HVAC system is the only source of flexibility considered. A feedforward control architecture is proposed to manipulate fan power consumption as needed for regulation purposes. While a simplified thermal model of a building is used for control design, the performance of the controller is tested via simulations on a high fidelity non-linear model. The simulations indicate that the controller performs on the complex model as predicted by the simplified model, thus justifying the adequacy of a simplified model for control synthesis.

Numerical experiments outlined here are performed on a model derived from a commercial building on the University of Florida campus (Pugh Hall). Results indicate that it is feasible to use up to 15% of the total fan power for regulation service to the grid, without noticeably impacting the building's

indoor environment and occupants' comfort, provided the bandwidth of regulation service is suitably constrained. To ensure the comfort of occupants, and to manage stress on HVAC equipment, both upper and lower bounds on bandwidth are necessary. Based on simulation experiments, this bandwidth is estimated to be $[1/\tau_0, 1/\tau_1]$, where $\tau_0 \approx 10$ minutes, and $\tau_1 \approx 8$ seconds.

2.3.1 Control Architecture for Regulation from HVAC loads

The regulation signal sent by the grid operator is typically a sequence of pulses at 4 second intervals, where the magnitude of the pulse indicates the amount of regulation required. In case of loads, the pulse's magnitude signals the amount of deviation in their power consumption asked by the grid operator. In what follows, a *regulation controller* is designed, as an add-on to the existing BAS, to enable provision of regulation service. It is a feedforward controller which is configured to change the power consumption of the building so that the change tracks the regulation signal sent by the grid operator. A high-level overview of the controller design is discussed here (for details, see [101, 102]).

BAS control architecture

A building is typically divided into several zones. At each zone, local controllers manipulate certain quantities such as zone temperature while a central controller maintains the air supply to the building at appropriate temperature and indoor air quality within prescribed limits. The building environment is maintained through complex interactions between the decentralized *zonal* controllers and central controller. They are briefly summarized here.

The main component of the central control system is the air handing unit (AHU). The AHU recirculates the return air from each zone and mixes it with fresh outside air. The mixed air is drawn through the cooling coil in the AHU by a supply fan, which cools the air and reduces its humidity. In cold/dry climates it may also reheat and humidify the air.

The air supplied by the AHU is distributed to each zone through ducts and controlled by actuating dampers and reheat coil in variable air volume box of

that zone. The zonal controller manipulates the mass flow rate of air going into the zone so that the zone temperature is maintained at a desired value.

As the zonal controllers change the damper positions in response to local disturbances (heat gains from solar radiation, occupants and so on), the differential pressure across the AHU fan changes. The AHU fan controller senses this change and activates the VFD of the supply fan to change the fan speed command $u(t)$ correspondingly. This causes the fan speed $v(t)$ to change in a way such that the differential pressure is maintained at a predetermined setpoint.

Modified BAS to Provide Regulation Service

To simplify the analysis, it is assumed that the power consumed by the furnace supplying hot water to the VAV boxes for reheating and the chiller/cooling tower providing chilled water to the cooling coil of the AHU are independent of the power consumed by the fan. The first decoupling assumption holds true since furnaces in typical HVAC systems consume natural gas instead of electricity. The second decoupling assumption is justified if the fan power deviations are of a *high frequency* and low magnitude, due to the large mechanical inertia of the chiller/cooling tower equipment. In addition, if the chilled water is supplied from a water storage tank, the decoupling assumption holds naturally.

Suppose the building is required to provide $r(t)$ (in kW) amount of regulation service at time t. The regulation controller perturbs the fan speed command so that the fan's power consumption is changed in a way such that the deviation in consumption tracks $r(t)$. The architecture of the control system is shown in Fig. 2.7.

The regulation signal $r(t)$ is transformed to a perturbation $u^r(t)$ to the fan speed command by the regulation controller. This command is then added to the nominal fan speed command $u^b(t)$ produced by the building's fan controller, which is purely dictated by the thermal dynamics of the building. If $p^b(t)$ is the nominal power consumption of the fan due to the thermal load on the building and $p^{b+r}(t)$ is the fan power consumption with the adjustments imposed by the regulation controller, then the deviation in power consumed

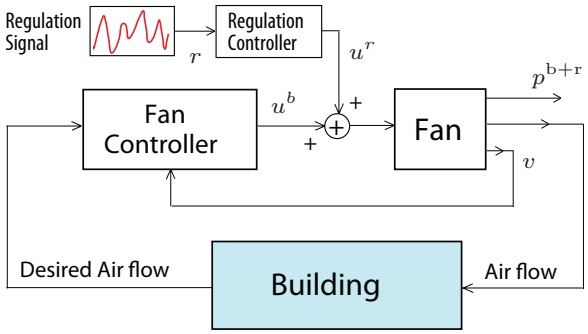


Figure 2.7: Modifications to the BAS control architecture to enable regulation provision from building HVAC loads.

by the fan is $\Delta p(t) := p^{b+r}(t) - p^b(t)$. Clearly, changing the fan speed from the nominal value determined by the building's existing control system will change the air flow through the building thus impacting the environment. The goal is to design the regulation controller so that $\Delta p(t)$ tracks $r(t)$ while causing little change in the building's indoor environment (measured by the deviation of the zonal temperatures from their set points).

Regulation Signal for Commercial Buildings

A major concern in engaging commercial buildings in regulation provision is the possibility of causing discomfort to the occupants or damaging the HVAC equipment. It is argued that discomfort as well as equipment stress can be avoided if the bandwidth of the regulation signal is suitably constrained. The considerations in determining this bandwidth are discussed here along with the control strategy implemented to extract regulation service.

The bandwidth of the regulation signal sent to buildings should be chosen based on the following factors. First, high frequency content in fan speed command perturbation $u^r(t)$ is desirable up to a certain upper limit. This is because the thermal dynamics of a commercial building have low-pass characteristics due to its large thermal capacitance. Consequently, high frequency changes in the air flow cause little change in its indoor temperature. Additionally, the VFD and fan motor have large bandwidth so that high frequency changes in the signal $u^r(t)$ lead to noticeable change in the fan speed and, consequently, fan power. Both effects are desirable, since the goal is to affect the

fan power consumption without affecting the building's temperature. However, an extremely high frequency content in $u^r(t)$ is not desirable as it might cause wear and tear of the fan motor. Likewise, $u^r(t)$ should not have very low frequency content. Otherwise, even if the magnitude of $u^r(t)$ is small, it may cause significant change in the mass flow rate, which in turn can produce a noticeable change in the temperature of the building. Furthermore, a large enough change in the temperature will cause the zonal controllers to try to change air flow rate to reverse the temperature change. In effect, the building's existing control system will try to reject the disturbance caused by u^r. Being a feedback loop, this disturbance rejection property is already present in the building control system.

In short, the frequency content of the disturbance $u^r(t)$ should lie in a particular band $[f_{\mathsf{low}},\ f_{\mathsf{high}}]$, where the gain of the closed loop transfer function from u^r to fan speed v is sufficiently large while that of the transfer function from u^r to temperature T is sufficiently small. This bandwidth may depend on several factors like thermal capacity of the building and so on; some insights on determining this bandwidth are provided in [101, 102]. The parameters $f_{\mathsf{low}}, f_{\mathsf{high}}$ are the key design variables to construct a suitable regulation signal for the buildings.

2.3.2 Regulation Provision by F-building

This section outlines simulation experiments which test the performance of the developed regulation controller for tracking a regulation signal. The BAS operation for a fictitious (F) building comprising of 4 stories and 44 zones is simulated for these tests. Each story has 11 zones constructed by cutting away a section of Pugh Hall. The HVAC system for F building consists of a single AHU and zonal controllers for each of its zones. The F building is meant to mimic the section of Pugh Hall serviced by one of the three AHUs that services 41 zones. Thermal parameters are identified for this building and used in the experiments.

For the purposes of simulations described in Section 2.3.2, the regulation signal $r(t)$ is constructed by passing raw ACE data from PJM [103] through a fifth-order Butterworth filter with passband $[1/600, 1/8]$ Hz. The choice of the

passband is by trial-and-error. The filtered ACE data is then scaled so that the magnitude of $r(t)$ is less than or equal to 5 kW – a conservative estimate of the regulation capacity of F-building.

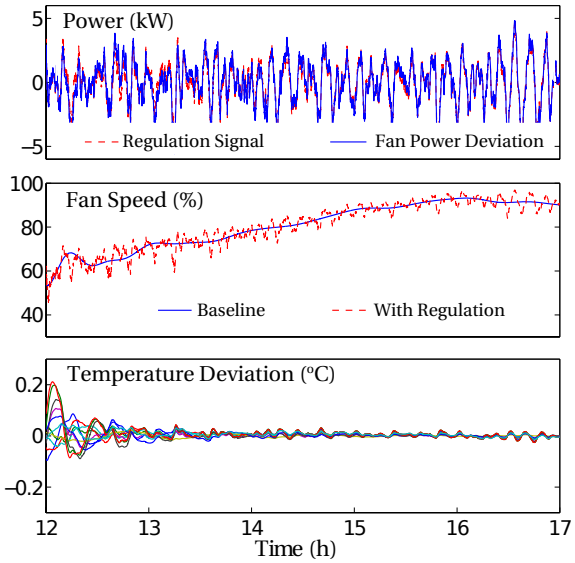


Figure 2.8: The impacts of tracking a regulation signal on fan power, fan speed and zone temperatures.

To unambiguously determine performance of the control scheme, two simulations are performed: a benchmark simulation with $u^r(t) \equiv 0$ to compute the nominal fan power consumption $p^b(t)$ and another with $u^r(t)$ driven by regulation $r(t)$ to determine power $p^{b+r}(t)$ consumed and, consequently, the deviation in fan power $\Delta p(t)$. Simulations were conducted for different sample paths of $r(t)$ and different initial conditions for the zone temperatures.

The results of one such experiment are depicted in Figure 2.8. The top plot depicts how well the fan power deviation $\Delta p(t)$ tracks the regulation signal $r(t)$. The deviation in the fan speed caused by tracking the regulation signal is depicted in the middle plot. Although the baseline fan speed is time-varying, the regulation controller designed with a constant baseline speed assumption performs well. Finally, the bottom plot depicts the deviation of the temperatures of the individual zones from their set points. Observe that the maximum deviations are at the beginning of the simulation; this is because of i) initial conditions, and ii) to provide the same amount of regulation, the

fan speed deviation from nominal speed at lower speed is larger than that at higher speed. Nevertheless, the temperature deviations after transient are very small, which will most likely be unnoticed by the occupants.

The passband of the bandpass filter corresponds to the bandwidth of regulation that the building can provide and is designed based on additional simulations not reported here. These simulations emphasize how the regulation reference signal that can be successfully tracked by the proposed fan speed control mechanism is restricted in a certain bandwidth that depends on the closed loop dynamics of the building. In case of the F-building, it was observed that if the regulation signal contained frequencies lower than 1/600 Hz (corresponding to period of 10 minutes), the zonal controllers would compensate for the indoor temperature deviations in the zones by modifying air supply requirements, thus nullifying the speed deviation command of the regulation controller. This resulted in a poor regulation tracking performance. The upper band limit was estimated to be 1/8 Hz to avoid stress on the mechanical parts of the supply fan.

Remark. The bandwidth of regulation that the building can provide is a key component in defining the "service" that the building provides. This bandwidth along with the total capacity of regulation that can be provided should be communicated to the grid operator. The grid operator can then construct an appropriate regulation signal whose frequency content and amplitude respect the limitations imposed by the close-loop dynamics of the building. Theoretical underpinnings for determining these parameters are contained in [101].

2.4 Concluding Remarks

The simulation studies described in this chapter demonstrate the value of ancillary services from DR-based mechanisms and propose a control method to harness ancillary services from commercial buildings. The first set of studies characterize the ancillary service needs of wind-integrated systems and demonstrate the effectiveness of DR-based reserve capacity to counter the volatility and uncertainty of wind resources. Results also indicate that load leveling can prove beneficial for systems with wind generation. A feedforward controller to

extract frequency regulation service from commercial building HVAC loads is also described. The second set of studies demonstrate that the impact of this controller on building environment is minimal while manipulating the power consumption to track the regulation signal.

An important takeaway for this chapter is the need to redesign operational practices to accommodate increasing deployment of renewable and demand response resources. The following chapters provide some control techniques for the design of such tools.

Chapter 3

ADP and Learning-based Control

This chapter begins with an introduction to tools from Markov decision theory, approximate dynamic programming (ADP) and reinforcement learning (RL). A modeling framework for the power grid and its resources is also presented. In this dissertation, these models and tools are applied to power system control problems to obtain approximately optimal control policies. This chapter demonstrates their application to devise control strategies for extracting ancillary services from flexible resources; the case of energy storage and thermal loads is considered.

A discrete time domain is chosen for the models to reflect the nature of the power system control architecture wherein control decisions are taken at discrete time intervals. The focus is on time scales of the order of a few seconds, minutes or hours; consequently governor dynamics are ignored. A unifying control-oriented model for the different power system resources is presented. When viewed through this modeling lens, energy storage resources and loads with virtual storage capabilities appear to have similar structure and underlying dynamics.

The numerical examples presented in this chapter correspond to control on medium to fast time scales. These examples illustrate how ADP and RL can be successfully applied to synthesize control policies for newer generation ancillary service providers such as storage and DR resources. Examples of control synthesis driven by real observations of the system are presented to showcase the capabilities of RL techniques to tune the control parameters to the statistics of the system.

3.1 A Markovian Framework

Markov decision processes (MDPs) provide a theoretical setting to study many of the problems considered in this work. A brief introduction to MDPs and the associated solution techniques is provided here.

3.1.1 Markov Decision Processes

An MDP can be thought of as modeling decision-making under uncertain and dynamic conditions. The decisions concern a system that is evolving in time as a result of its inherent dynamics, manifestation of different sources of uncertainty and actions taken over time. The system condition at any given time can be described by its "state," which evolves with time over a state space. Associated with each state and action is a cost function. The goal is finding the best *policy* to achieve some objective with respect to the cost.

Mathematically, an MDP may be described using state space models. Adopting the notation from [104, 105], a discrete-time, controlled MDP model may be expressed in the recursive form

$$X(t+1) = f\big(X(t), U(t), W(t)\big) \tag{3.1}$$

where $\boldsymbol{X}$ is the state process, $\boldsymbol{U}$ the control process and $\boldsymbol{W}$ is the disturbance process which is assumed to be i.i.d. (independent and identically distributed). The states and actions evolve on the spaces denoted by X and U spaces respectively. Furthermore, actions may be subject to state-dependent constraints: $\mathsf{U}(x)$ is used to denote the set of control actions that satisfy the state-dependent constraints when the state is $x \in \mathsf{X}$. The controlled transition law for the MDP is given by

$$\begin{aligned} P_u(x, \mathbb{A}) &= \mathsf{P}\left\{X(t+1) \in \mathbb{A} \mid X(t) = x, U(t) = u\right\} \\ &= \mathsf{P}\left\{f\left(x, u, W(1)\right) \in \mathbb{A}\right\} \ , \end{aligned}$$

for arbitrary $x \in \mathsf{X}$, $u \in \mathsf{U}(x)$, $\mathbb{A} \subset \mathsf{X}$ (Borel measurable).

A policy $\boldsymbol{\phi}$ is a sequence of functions $\{\phi^t\}$ that define the control actions as

$$U(t) = \phi^t\big(X(0), \ldots, X(t-1), X(t)\big)\,,$$

such that $U(t) \in \mathsf{U}(X(t))$ for each t. The policy $\boldsymbol{\phi}$ is Markov if ϕ^t depends only on $X(t)$ for each $t \geq 0$. A *stationary policy* is a Markov policy $\boldsymbol{\phi}$ such that $\phi^t = \phi$ is independent of t. Finally, a *randomized stationary policy* is a probabilistic mapping from state to action space: For each $x \in \mathsf{X}$, $\phi(x) := \{\phi_u(x) : u \in \mathsf{U}\}$ such that when $X(t) = x$,

$$\phi_u(x) = \mathsf{P}\left\{U(t) = u | X(0), \ldots, X(t), U(0), \ldots, U(t-1)\right\} \qquad t \geq 0$$

with $\phi_u(x) = 0$ if $u \notin \mathsf{U}(x)$. Without loss of generality, control inputs are restricted to those defined by a stationary policy, possibly randomized.

Optimality for the model is based on a one-step cost function $c : \mathsf{X} \times \mathsf{U} \to \mathbb{R}_+$: The goal is to minimize either *discounted* or *average* cost. The work described in this chapter focuses on the discounted-cost optimal control problem. The average-cost optimal control problem is presented in Section 4.1.1.

For the MDP model (3.1), the optimal discounted cost is defined as

$$g^*(x) = \inf_{\boldsymbol{U}} \sum_{t=0}^{\infty} \delta^t \mathsf{E}\left[c\left(X(t), U(t)\right)\right]\,, \qquad X(0) = x\,, \tag{3.2}$$

where $\delta \in (0, 1)$ is the discount parameter. Under general conditions, the minimum value function g^* exists and satisfies the dynamic programming (DP) equation

$$g^*(x) = \min_{u \in \mathsf{U}(x)} \left\{c(x, u) + \delta \mathcal{P} g^*\left(x, u\right)\right\}\,, \tag{3.3}$$

where the DP operator $\mathcal{P}$ denotes the expectation,

$$\mathcal{P}h\left(x, u\right) = \mathsf{E}\left[h\left(X(t+1)\right) \mid X(t) = x, U(t) = u\right]\,, \tag{3.4}$$

for any $x \in \mathsf{X}, u \in \mathsf{U}$ and function $h \colon \mathsf{X} \to \mathbb{R}$. The minimizer u^* in (3.3) defines an optimal state feedback policy $\phi^*(x)$.

To simplify the math, the following notation is adopted: under a fixed policy

ϕ, the cost and DP operator are defined as

$$c_\phi(x) := c\big(x, \phi(x)\big) \qquad \text{and} \qquad \mathcal{P}_\phi h\,(x) := \mathcal{P}h\,\big(x, \phi(x)\big)\,. \tag{3.5}$$

This notation is extensively used throughout this dissertation. Under this policy ϕ, the DP equation reduces to the following form:

$$g(x) = c_\phi(x) + \delta \mathcal{P}_\phi g\,(x)\ , \tag{3.6}$$

where g is the discounted cost under policy ϕ. This degenerate version of the DP equation is referred to as *Poisson's equation*, which finds many uses in the theory and solution techniques for discounted cost control.

3.1.2 Algorithms

Well-known techniques for solving MDPs include algorithms such as value iteration and policy iteration. For large-scale problems, approximation techniques or learning algorithms are often employed to reduce computational complexity. A brief review of the ADP and RL techniques used in this work is provided. All algorithms are described in the context of the discounted cost problem (3.2).

Value Iteration Algorithm (VIA)

Value iteration is a successive approximation technique to solve the DP equation. The algorithm is initialized with $V_0 : \mathsf{X} \to \mathbb{R}_+$ and the value function is successively approximated as

$$V_{n+1}(x) = \min_{u \in \mathsf{U}(x)} \{c(x,u) + \delta \mathcal{P} V_n\,(x,u)\} \qquad x \in \mathsf{X},\ n \geq 0\,. \tag{3.7}$$

The feedback law $\phi_n^* : \mathsf{X} \to \mathsf{U}$ is defined to be the minimizer in (3.7).

$$\phi_n^*(x) \in \arg\min_{u \in \mathsf{U}(x)} \{c(x,u) + \delta \mathsf{P} V_n(x,u)\}\,, \quad x \in \mathsf{X}\,.$$

Policy Iteration Algorithm (PIA)

In PIA, a sequence of deterministic stationary policies are obtained, with increasingly improved performance, in the sense that the corresponding discounted costs are non-increasing. The algorithm is initialized with a policy ϕ_0 and then the following operations are performed in the kth stage of the algorithm:

(i) Given the policy ϕ_k, find the solution g_k to the Poisson's equation

$$c_{\phi_k} + \delta \mathcal{P}_{\phi_k} g_k \ = g_k \, . \tag{3.8}$$

(ii) Update the next policy using

$$\phi_{k+1}(x) \in \arg\min_{u \in \mathsf{U}(x)} \{c(x,u) + \delta \mathcal{P} g_k \, (x,u)\} \ , \quad x \in \mathsf{X} \, , \tag{3.9}$$

and return to step (i) with $k = k + 1$.

PIA can be conveniently integrated with the learning techniques described next.

TD-Learning

TD-learning is a technique for approximating value functions of MDPs within a parameterized class. Suppose $\{g^\theta\}$ is a linearly parameterized family of approximations, where $g^\theta = \sum \theta_i \varphi_i$ for basis functions $\{\varphi_i : 1 \leq i \leq d\}$. Then, the goal is to find θ^* that such that $g^\theta \approx g$, where g is the discounted cost for some policy ϕ. An ergodic norm defined under a fixed policy is chosen for TD-learning with the error criterion chosen as $\|g - g^\theta\|^2$. If the norm is defined by an "inner product," a least-squares problem can be formulated. The resulting least-squares TD-learning algorithm is described in [104, 105] and can be used to successively approximate the basis weights θ^*. The TD-learning algorithm can be used to compute the solution g_k to the Poisson's equation (3.8) in PIA, thereby avoiding the need to explicitly solve the linear system of equations.

SARSA

The application of PIA is computationally difficult because of the policy update formula (3.9); the computation of $\mathcal{P}\, g$ itself may be difficult if the state space is large. The following approach can be considered to avoid integration: Let H denote the function of two variables,

$$H(x,u) = c(x,u) + \delta \mathcal{P} g\,(x,u)\,.$$

If H can be directly estimated, then the policy update can be obtained by minimizing $H(x,u)$ over u, for each state $x \in \mathsf{X}$. Techniques analogous to TD-learning can be applied to approximate H based on the following proposition.

Proposition 1. *Under a fixed stationary policy ϕ,*

(i) the state-control process $\Phi(t) = (X(t), U(t))$ is also a Markov chain, and,

(ii) the function H solves the Poisson's equation for the Markov chain $\Phi(t)$ and cost function $c(x,u)$.

Proof. Part (i) is obvious: The process $\{\Phi(t)\}$ evolves according to a controlled stochastic model of the recursive form (3.1).

To see (ii), suppose g is the solution to Poisson's equation for operator P_ϕ with cost c_ϕ. Then,

$$H_\phi(x) := H\left(x, \phi(x)\right) = c_\phi(x) + \delta \mathsf{P}_\phi g(x) = g(x)\,.$$

Substituting this back into the definition of H gives

$$H(x,u) = c(x,u) + \delta \mathsf{P} H_\phi(x,u)\,.$$

Thus, H solves Poisson's equation for the state-control process. □

The above proposition places the analysis for SARSA within the setting of TD-learning. A natural parameterization for the approximation is of the form

$$H^\theta = c + \theta^{\scriptscriptstyle T} \psi\,,$$

where $\psi\colon \mathsf{X}\times\mathsf{U}\to\mathbb{R}^d$. Given a basis $\{\varphi_i : 1\le i\le d\}$ intended for application in TD-learning, the SARSA basis $\{\psi_i : 1\le i\le d\}$ may be chosen as

$$\psi_i(x,u) = \mathcal{P}\varphi_i\,(x,u) \qquad x\in\mathsf{X}\ ,\ u\in\mathsf{U}\,. \tag{3.10}$$

If the integration $\mathcal{P}\varphi_i$ is difficult to compute, then an approximation may be used.

Q-learning

The goal of Q-learning is to learn a function on $\mathsf{X}\times\mathsf{U}$ analogous to SARSA. However, unlike SARSA where the function H is learned for a fixed policy, Q-learning learns the so-called *Q-function* for the optimal control policy. The Q-function is defined as follows:

$$H^*(x,u) = c(x,u) + \delta\mathcal{P}g^*\,(x,u)\,.$$

The main difference between SARSA and Q-learning is that SARSA restricts exploration on state-action space since learning is restricted to fixed policy whereas Q-learning naturally calls for exploration. More details on Q-learning are provided in Chapter 5.

3.2 Power Node Modeling Framework

Models used in the development of control policies for a power grid and its resources should capture the associated reliability needs, the underlying dynamics and the system uncertainties. The "power node" model of [75] – wherein each resource connected to the grid is viewed as an abstract single lumped unit with characteristic parameters – is adapted to develop a modeling framework that captures the uncertainty and dynamics in the system. Our modeling framework is based on control-oriented modifications to the power node model and is used for analysis in this dissertation. A brief description of the models is provided here.

The models described here are concerned with active power control and provision of ancillary services under category **(AS1)** described in Section 1.1.

Recall that this ancillary service category concerns the management of supply-demand balance on time scales ranging from seconds to hours. Each system resource – be it a generator, a load or a storage unit – provides some degrees of freedom in managing the supply-demand balance. Based on the nature of control offered by the resource, it is classified into the following three categories:

- controllable, where the power generation or consumption is completed controlled, possibly in response to needs of the grid;
- curtailable, where the power generation or consumption is not directly controlled but can be curtailed, possibly for reliability reasons; and,
- uncontrollable, where the resource has no flexibility.

For example, traditional energy sources from coal or natural gas are *controllable* whereas renewable generation from wind and solar resources is *curtailable* – the turbines can be turned off. Likewise, electric demand is either controllable or curtailable. Resources with storage capabilities are assumed to be inherently controllable with respect to the charging and discharging.

In the power node model, each resource – be it a generator, load, storage unit or a combination unit such as a solar generator with thermal storage – constitutes an abstraction of its specific unit characteristics. The modeling abstraction is presented here followed by a detailed description of the power node model.

The power grid is represented as an undirected graph. Directed graphs may be useful in modeling a power grid with DC interconnects, but this extension is not considered in present work. Each node of the graph corresponds to a grid resource and each link represents a transmission line. There are N nodes, indexed as $\{1, \ldots, N\}$, and L transmission lines, indexed as $\{1, \ldots, L\}$. The network is assumed to be connected.

Each node interacts with the grid via its injection of power into the grid or withdrawal of power from the grid. This injection or withdrawal depends on control actions being employed at that node. At each time interval t, there is at most one possible control action at a node n:

- ramping of power outputs $\Delta P_{\mathsf{G}n}(t)$ or consumption $\Delta P_{\mathsf{D}n}(t)$ for controllable resources,

- curtailment of generation $C_{\mathsf{G}n}(t)$ or load $C_{\mathsf{D}n}(t)$ for curtailable resources,
- power $P_{\mathsf{S}n}(t)$ withdrawn/injected for charging/discharging by storage resources.

The associated states for these resources are outputs of controllable/curtailable generators $P_{\mathsf{G}n}(t)$, the demand of controllable/curtailable loads $P_{\mathsf{D}n}(t)$ and the energy stored by the storage resource $E_{\mathsf{S}n}(t)$. The resulting dynamics are presented in Table 3.1.

Table 3.1: Simplified dynamics of typical resource-types

Resource	controllability	Dynamics
Generator	controllable	$P_{\mathsf{G}n}(t+1) = P_{\mathsf{G}n}(t) + \Delta P_{\mathsf{G}n}(t)$
	curtailable	$P_{\mathsf{G}n}(t) = G_n(t) - C_{\mathsf{G}n}(t)$
Load	controllable	$P_{\mathsf{D}n}(t+1) = P_{\mathsf{D}n}(t) + \Delta P_{\mathsf{D}n}(t)$
	curtailable	$P_{\mathsf{D}n}(t) = D_n(t) - C_{\mathsf{D}n}(t)$
storage	controllable	$E_{\mathsf{S}n}(t+1) = E_{\mathsf{S}n}(t) - \alpha_{\mathsf{S}n} P_{\mathsf{S}n}(t)$

Note that $G_n(t)/D_n(t)$ represents the external generation/demand at node n; for uncontrollable resources, the curtailments $C_{\mathsf{G}n}(t)$ and $C_{\mathsf{D}n}(t)$ take value *zero* at all times. Also, $\alpha_{\mathsf{S}n}$ represents the conversion efficiency for the storage resource. The dynamics of a generic power node with storage, demand and generation capabilities are described as follows:

$$E_{\mathsf{S}n}(t+1) = E_{\mathsf{S}n}(t) - \alpha_{\mathsf{S}n} P_{\mathsf{S}n}(t) + [G_n(t) - C_{\mathsf{G}n}(t)] - [D_n(t) - C_{\mathsf{D}n}(t)] \ . \tag{3.11}$$

The associated N-dimensional vectors are denoted by suppressing the nodal notation while boldface notation is used to denote sample paths. For instance, $P_{\mathsf{G}}(t)$ is the N-dimensional generation vector while $\boldsymbol{P}_{\mathsf{G}} := \{P_{\mathsf{G}}(t) : t \geq 0\}$ represents sample path of generation.

The dynamics at power node n are subject to many constraints. These constraints may be static or dynamic in nature, depending on the specific resources at that node. The constraints include

- capacity limits, ramping constraints and minimum up/down-time limitations enforced by the generation resources;

- capacity limits, ramping constraints and limitations imposed by the end-uses being served by the demand process; and
- charging/discharging constraints and limits on energy storage.

These constraints are compactly represented as

$$(\boldsymbol{E}_{\mathsf{S}n}, \boldsymbol{P}_{\mathsf{S}n}, \boldsymbol{P}_{\mathsf{G}n}, \boldsymbol{P}_{\mathsf{D}n}, \boldsymbol{W}_{\mathsf{G}n}, \boldsymbol{W}_{\mathsf{D}n}) \in \mathbf{X}_n \,. \tag{3.12}$$

The nodal control actions are subject to system-wide constraints which are collectively summarized below:

$$(\boldsymbol{E}_{\mathsf{G}}, \boldsymbol{E}_{\mathsf{D}}, \boldsymbol{W}_{\mathsf{G}}, \boldsymbol{W}_{\mathsf{D}}) \in \mathbf{X}_{\mathsf{sys}} \,. \tag{3.13}$$

Again, equation (3.13) includes many static and dynamic constraints, which are listed below:

- power flow limitations imposed by transmission constraints
- supply-demand balance constraint (when dealing with slower time scales)
- frequency limitations (when dealing with fast time scales)

In the special case of thermal loads such as HVACs and other heating/cooling loads, the dynamics are similar to those of a pure storage resource. For an air-conditioning (cooling) load, the dynamics are expressed as

$$\Theta_n(t+1) - \Theta_n(t) \;=\; Q_{\mathsf{H}n}(t) - Q_{\mathsf{C}n}(t) - \tau_n \left[\Theta_n(t) - \Theta_{n,out}(t)\right] \,, \tag{3.14}$$

where $\Theta_n(t)$ is the temperature, $\Theta_{n,out}(t)$ is the outside temperature, τ_n is the heat transfer co-efficient, $Q_{\mathsf{H}n}(t)$ is the heat gain from solar radiation and internal gains due to human activities while $Q_{\mathsf{C}n}(t)$ is the cooling provided by the air conditioning unit at time t. Note that $Q_{\mathsf{C}n}(t) = \nu_{\mathsf{C}n} E_{\mathsf{D}n}(t)$, where $\nu_{\mathsf{C}n}$ and $E_{\mathsf{D}n}(t)$ are the efficiency and energy consumption of the air conditioner.

The power node models can be cast in a control framework to find appropriate control policies for operations. Representative examples are presented in the following sections and in future chapters.

3.3 Coordinating Combined Wind-Storage Resource Operations

This section provides an illustration of the control synthesis on slow to medium time scales. A simple example is considered to show how energy storage can be used in conjunction with volatile wind generation to meet specific demand requirements. It is assumed that the storage resource acts as a buffer between the wind farm and the grid injecting power into the grid in a controlled manner. The storage discharge is controlled so that the power injected into the grid meets specific requirements. In the first study, the storage discharge is controlled so the grid injection has low volatility, i.e., energy discharge is nearly steady. In the second study, the storage discharge is controlled so that the grid injection can be used to meet a time-varying exogenous demand.

The storage control problem in both examples is formulated as an MDP; its solution provides a state feedback control policy which indicates the optimal amount of storage discharge as a function of the state of the system. Under assumptions on the underlying disturbance process, the optimal solution of the MDP is computed and used as a benchmark for testing the quality of the approximate solutions obtained from TD-learning and SARSA. Insights from these restrictive models are then used to construct an architecture for tuning the RL techniques with real world data. The following subsections provide details on the problem formulation and solution schemes.

3.3.1 Smoothening Variability of Wind Generation

Since the dynamics of a single power node are under consideration here, the node n notation is suppressed. The storage unit is charged using wind generation only. Thus, the power node model of (3.11) can be simplified as

$$E_{\mathsf{S}}(t+1) = E_{\mathsf{S}}(t) - \alpha_{\mathsf{S}} U(t) + [G(t) - C_{\mathsf{G}}(t)]$$

where the control action $U(t)$ denotes the *controlled* storage discharge $P_{\mathsf{S}}(t)$. The following assumptions are adopted:

- conversion efficiency α_{S} is 1.

- the constraint set $\mathbf{X}$ only consists of the storage energy capacity constraint.
- wind generation is curtailed only to avoid violation of the capacity constraint.

In the light of these assumptions, the storage dynamics can be further reduced to the following equations:

$$\begin{aligned} E_{\mathsf{S}}(t+1) &= \min\left\{E_{\mathsf{S}}(t) - U(t) + G(t), E_{\mathsf{S}}^{\mathsf{max}}\right\} \\ C_{\mathsf{G}}(t+1) &= \max\left\{E_{\mathsf{S}}(t) - U(t) + G(t) - E_{\mathsf{S}}^{\mathsf{max}}, 0\right\} \end{aligned}$$

with $\boldsymbol{G}$ representing the stochastic wind generation process and $E_{\mathsf{S}}^{\mathsf{max}}$ denoting the storage capacity.

MDP formulation

The dynamics of the node can be cast as a one-dimensional MDP by defining the state as

$$X(t+1) := E_{\mathsf{S}}(t) - U(t) + G(t)\,.$$

The state process evolves according to the recursion

$$X(t+1) = \min\left\{X(t), E_{\mathsf{S}}^{\mathsf{max}}\right\} - U(t) + G(t) \quad \text{for } t \geq 0 \tag{3.15}$$

on state space X. The control actions are defined on action space U; they are subject to state-dependent constraints since storage discharge cannot exceed the energy stored. That is,

$$U(t) \leq \min\left\{X(t), E_{\mathsf{S}}^{\mathsf{max}}\right\}\,.$$

The energy stored and the curtailed wind generation at time t are recovered from the state as $\min\left\{X(t), E_{\mathsf{S}}^{\mathsf{max}}\right\}$ and $\max\left\{X(t) - E_{\mathsf{S}}^{\mathsf{max}}, 0\right\}$, respectively. The control actions are restricted to stationary policies. In this way, an MDP is defined with controlled transition law

$$\mathsf{P}_u(x_0, x_1) := \mathsf{P}\left\{X(t+1) = x_1 \mid X(t) = x_0, U(t) = u\right\}\,,$$

which can be computed based on the knowledge of the distribution of $G(t)$.

In this problem, the discounted cost optimality criterion (3.2). The cost function is assumed to have the following structure:

$$c(x,u) \;=\; c_{\mathsf{S}}(x) + \gamma\,(u-\bar{u})^2 \;\;. \tag{3.16}$$

The state-dependent cost $c_{\mathsf{S}}(x)$ can be used to capture the cost of storing energy (which may be relevant for batteries) as well as the cost of wind generation curtailment (when $x > E_{\mathsf{S}}^{\mathsf{max}}$). Additionally, $c(x,u)$ penalizes the deviations of control action u from mean storage discharge $\bar{u}$ (which, in turn, is equal to the mean wind energy). The optimal policy $\phi^*(x)$ is chosen as a minimizer to the associated DP equation (3.3).

Numerical Experiments

All variables are normalized with respect to the storage capacity so that $E_{\mathsf{S}}^{\mathsf{max}} = 1$ *per unit* (p.u.). The penalty factor $\gamma = 100$ and the storage cost is taken as $c_{\mathsf{S}}(x) \;=\; 100\,(x-0.95)_+^2$, where $(\cdot)_+$ denotes the non-negative projection. The wind generation process $\boldsymbol{G}$ is assumed to be i.i.d. with

$$G(t) \;=\; \bar{g} \;+\; N_{\mathsf{G}}(t)\,,$$

where $N_{\mathsf{G}}(t)$ uniformly distributed on the interval $[-E_{\mathsf{G}}^{\mathsf{max}}/2, E_{\mathsf{G}}^{\mathsf{max}}/2]$. The length of the interval $E_{\mathsf{G}}^{\mathsf{max}}$ and mean wind energy $\bar{g}$ are used as parameters to simulate different wind generation profiles.

A discount factor of $\delta = 0.95$ is assumed for simulation. The MDP is solved using VIA for different values of $\bar{g}$ and $E_{\mathsf{G}}^{\mathsf{max}}$. Figure 3.1 illustrates the optimal control policy as a function of energy stored for varying degrees of volatility in the wind generation (characterized by the associated coefficient of variation $c_v = \frac{\sqrt{\mathsf{Var}\,[G(t)]}}{\mathsf{E}[G(t)]}$) for mean wind energy $\bar{g} = 0.08$ p.u. Observe that the control policy is nonlinear and minimally impacted by the variability of $\boldsymbol{G}$.

Polynomial functions in x of the fourth order are used to construct basis φ for TD-learning. The basis ψ for SARSA is obtained from the TD-learning basis φ using (3.10). The TD-learning and SARSA algorithms are used with PIA to derive approximate MDP solutions. A simple projected, linear feedback

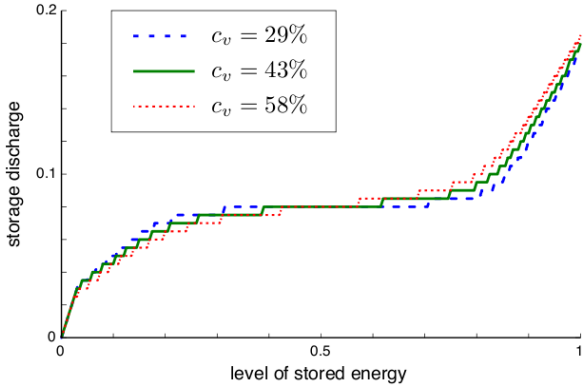


Figure 3.1: Optimal discharging policy as a function of level of stored energy for different $\boldsymbol{G}$ profiles.

policy is used to initialize the PIA. The policy approximations obtained from TD-learning and SARSA are illustrated in Figure 3.2.

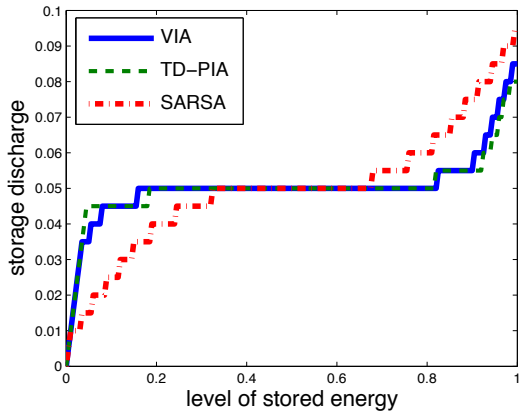


Figure 3.2: Approximations of the optimal discharging policy as a function of level of stored energy.

The policy approximated by TD-learning is a very good approximation for the optimal policy obtained by VIA. The SARSA approximation was constructed using only the knowledge of the mean $\bar{g}$ of the wind generation process $\boldsymbol{G}$. A better approximation of control policy can be obtained if more information is used in the SARSA construction. Furthermore, the SARSA approximation is a good fit in the state space region corresponding to maximum invariant probability mass resulting in nearly identical performance of approximate policy from SARSA as compared to that of the optimal policy

obtained from VIA.

Experiments driven by real data

As mentioned earlier, RL algorithms can accommodate data from the real world. In what follows, SARSA approximations combined with PIA are used to synthesize control policy for system dynamics driven with actual wind generation data.

Scaled 5-minute wind generation data from NREL [97] is used in the SARSA approximation to learn the value function. The performance of control policy for storage capacity $E_{\mathsf{S}}^{\mathsf{max}} = 2$, which is $2000\bar{g}$, is demonstrated in Figure 3.3. From these experiments, it can be concluded that while volatile wind generation output can be transformed into a base-load type generation unit, a very large amount of storage is needed for such a steady output.

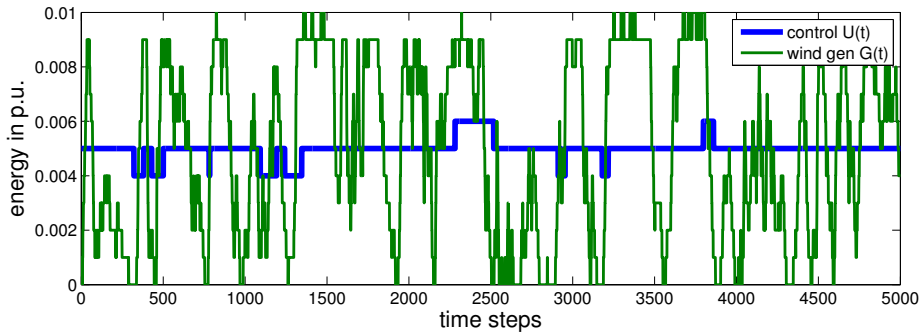


Figure 3.3: Appropriate control on a large storage unit can transform the volatile wind generation into a base-load type of unit with nearly steady output.

Under the assumption that the deviations of the storage discharge are managed by dispatching ancillary service, ancillary service costs can be computed. These are plotted in Figure 3.4 for different storage sizes. A careful consideration of such trade-offs between storage capacity and ancillary service provision can help in making investment decisions for sizing of storage units.

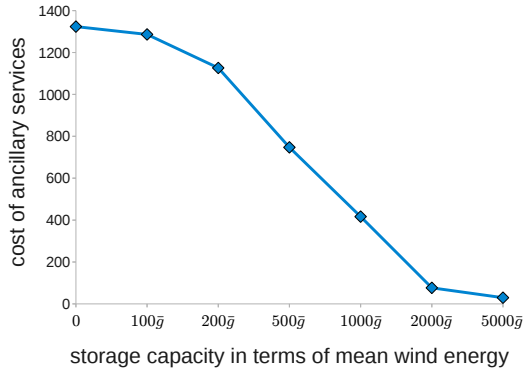


Figure 3.4: Impacts of storage sizing on ancillary service requirements.

3.3.2 Meeting Exogenous Demand

Here, the problem of using wind generation and energy storage to meet demand is considered, extending the setting considered in Section 3.3.1. That is, the energy discharged from the storage unit is supplied to an exogenous demand instead of providing a nearly steady grid injection as considered earlier.

MDP formulation

The control action is as before: the amount of energy discharged from the storage resource, which has to track the exogenous demand. The dynamical system is extended to incorporate two states: state $X(t)$ as defined previously with dynamics described by (3.15) and the externally driven demand $D(t)$ whose dynamics are assumed to modeled by an uncontrolled Markov chain with transition probabilities

$$\mathsf{P}(d_0, d_1) := \mathsf{P}\left\{D(t+1) = d_1 \mid D(t) = d_0\right\} \ .$$

The objective is to find the optimal control policy that minimizes discounted costs for the following cost structure

$$c\left([x, d], u\right) \ = \ c_{\mathsf{S}}(x) \ + \ \gamma \left(u - d\right)^2 \ . \tag{3.17}$$

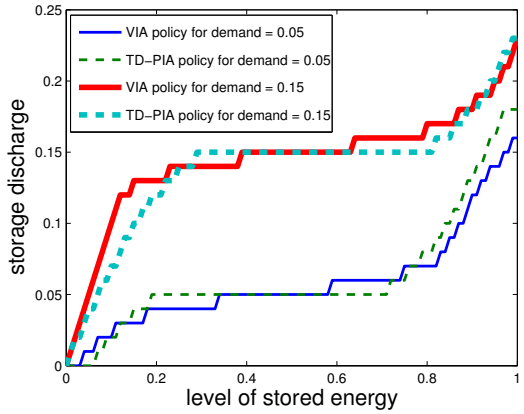


Figure 3.5: Optimal policy and its TD-approximation for matching time-varying demand.

Numerical Experiments

The case of a Markov demand process $\boldsymbol{D}$ is considered for illustrative purpose. The demand is assumed to switch between two values $\{0.05, 0.15\}$ so that the transition probability matrix given as

$$\mathsf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The wind generation is modeled as in Section 3.3.1. Suppose $\bar{g} = 0.1$ and $E_{\mathsf{G}}^{\max} = 0.2$ while other simulation parameters remain the same. The control policy and its TD-approximation are shown in Figure 3.5. Observe that the approximation is a good fit for the optimal policy.

The approximation techniques can be combined with real-data as shown in Section 3.3.1 using SARSA technique. The technique may be employed to size storage resources for meeting time-varying demand on a renewable generator. Further investigations are needed to explore these aspects.

3.4 Frequency Regulation from Thermal Loads

This section provides an illustration of control synthesis on fast times scales. A simplified example of thermal loads providing frequency regulation service

is considered. Specifically, it is assumed that an incentive is in place to encourage flexible thermal loads to provide corrective responses to grid frequency deviations. The goal is to synthesize a control policy to provide the corrective grid responses without sacrificing comfort: that is, power consumption of the loads is manipulated to minimize frequency as well as temperature deviations.

For simplicity, only cooling loads are considered. Furthermore, all cooling loads in the grid are assumed to be identical with respect to their thermal and electrical characteristics. Finally, the constraint set $\mathbf{X}_{\mathsf{sys}}$ is assumed to only consist of constraints imposed by frequency deviations; that is, the transmission network constraints are ignored.

The frequency dynamics of the system are expressed as

$$F(t+1) - F(t) = \beta^{-1} \left\{ 1^T E_{\mathsf{G}}(t) - 1^T E_{\mathsf{D}}(t) \right\} ,$$

where β represents the aggregate governor response of the entire system [75]. Each variable in the above expression can be decomposed into two parts: the predicted component and the deviation component. For instance, the actual frequency at time t is given by $F(t) = \hat{f}(t) + \Delta F(t)$, where $\hat{f}(t)$ is the predicted frequency while $\Delta F(t)$ is the frequency deviation.

The predicted values are used in scheduling and economic dispatch algorithms, which ensure supply-demand balance for the predicted conditions, thereby maintaining $\hat{f}(t) = 60$ Hz (that is, the nominal frequency) for $t \geq 0$. Therefore, the frequency dynamics can be simplified in terms of deviation components:

$$\Delta F(t+1) - \Delta F(t) \;=\; \beta^{-1} \left\{ 1^T \Delta E_{\mathsf{G}}(t) - 1^T \Delta E_{\mathsf{D}}(t) \right\} .$$

At each node n, the deviations in grid injections can be brought about by variations in renewable generation and/or equipment outages. On the other hand, the deviations in the energy drawn from the grid at node n can arise due to deviations in heating/cooling (thermal) demand at that node, denoted by $E^{\mathsf{th}}_{\mathsf{D}n}$, as well as deviations in the non-thermal demand, denoted by $E^{\mathsf{nth}}_{\mathsf{D}n}$. The frequency dynamics can then be reformulated to explicitly consider the impacts of deviations in thermal demand such that the following recursion

holds:

$$\Delta F(t+1) \;=\; \Delta F(t) + \beta^{-1}\left\{\Delta N_{\mathsf{F}}(t) - J\Delta E^{\mathsf{th}}_{\mathsf{D}n}(t)\right\}\;, \tag{3.18}$$

where J is the number of responsive cooling loads on the system and $\Delta N_{\mathsf{F}}(t) = 1^{T}\Delta E_{\mathsf{G}}(t) - 1^{T}\Delta E^{\mathsf{nth}}_{\mathsf{D}}(t)$ is used to collectively represent the excursions from predicted supply and non-thermal demand.

Using similar mathematical development, the thermal dynamics of heating and cooling loads can also be reduced to deviations in heat gains, ambient temperature and so on. For a cooling load modeled by (3.14), these dynamics take the form

$$\Delta\Theta_n(t+1) - \Delta\Theta_n(t) \;=\; \Delta Q_{\mathsf{H}n}(t) - \Delta Q_{\mathsf{C}n}(t) - \tau_n\left[\Delta\Theta_n(t) - \Delta\Theta_{n,out}(t)\right]\;.$$

The deviations in heat gains can be collectively represented as $\Delta N_{\mathsf{H}}(t) = \Delta Q_{\mathsf{H}n}(t) + \;\tau_n \cdot \Delta\Theta_{n,out}(t)$. Substituting this and $\Delta Q_{\mathsf{C}n}(t) = \nu_{\mathsf{C}n}\Delta E^{\mathsf{th}}_{\mathsf{D}n}(t)$, the temperature deviations evolve according to the recursion

$$\Delta\Theta_n(t+1) = (1-\tau_n)\Delta\Theta_n(t) - \nu_{\mathsf{C}n}\Delta E^{\mathsf{th}}_{\mathsf{D}n}(t) + \Delta N_{\mathsf{H}}(t)\;. \tag{3.19}$$

MDP formulation

The dynamics in (3.18) and (3.19) can be cast as a two-dimensional MDP by defining the state as $X(t+1) = [\Delta\Theta_n(t), \Delta F(t)]^{T}$. The state process evolves according to the recursion

$$X(t+1) \;=\; AX(t) + BU(t) + V(t)\;, \tag{3.20}$$

where control $U(t)$ in the cooling demand $\Delta E^{\mathsf{th}}_{\mathsf{D}n}(t)$ and

$$A = \begin{bmatrix} 1-\tau_n & 0 \\ 0 & 1 \end{bmatrix} \quad , \quad B = \begin{bmatrix} -J\beta^{-1} \\ -\eta_{\mathsf{C}n} \end{bmatrix} \quad \text{and} \quad V(t) = \begin{bmatrix} \Delta N_{\mathsf{H}}(t) \\ \beta^{-1}\Delta N_{\mathsf{F}}(t) \end{bmatrix}\;.$$

The state and action spaces are denoted by X and U, respectively. Based on the knowledge of the underlying distribution of $V(t)$, the controlled transition law $\mathsf{P}_u(x_0, x_1)$ can be computed.

The control objective is to minimize the discounted costs, with the one-step

cost $c(x, u)$ governed by temperature and frequency deviations and a cost for control. In the simulations described below, the cost is assumed to be

$$c(x, u) \;=\; x^T x + \tfrac{1}{2}u^2 \,. \tag{3.21}$$

Numerical experiments

The optimal control policy for the MDP is computed using the VIA for

$$A = \begin{bmatrix} 0.995 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} -0.3 \\ -0.1 \end{bmatrix}$$

discount factor $\delta = 0.95$ and noise process assumed to be i.i.d. with mean-*zero*, uniform distribution. The states – temperature and frequency deviations – are measured in °C and 0.1 Hz, respectively.

Since δ is so close to 1, a closed form solution can be obtained for a relaxation of the above problem. Specifically, the constraints are relaxed and the noise process are replaced by their means, which in this case, are zero. With $\delta \approx 1$, the control problem for the relaxed model is the Linear Quadratic Regulator (LQR) problem for which closed-form solution exists [106].

In Figure 3.6, the control policies computed from the solution of both MDP and LQR problems are shown side-by-side. Superimposed on these state-feedback plots are the nominal trajectories of the controlled state process for the same initial condition and noise perturbations. In spite of some differences in the control policies, the performances of MDP and LQR solutions – as evidenced from the sample path behavior – are nearly identical.

From the identical performance evidenced from the plots, it can be inferred that the control policy for the fluid model may indeed be used to control the MDP with marginal loss in performance. Naturally, fluid approximations can serve as good candidates for constructing TD and SARSA approximations [105, 107].

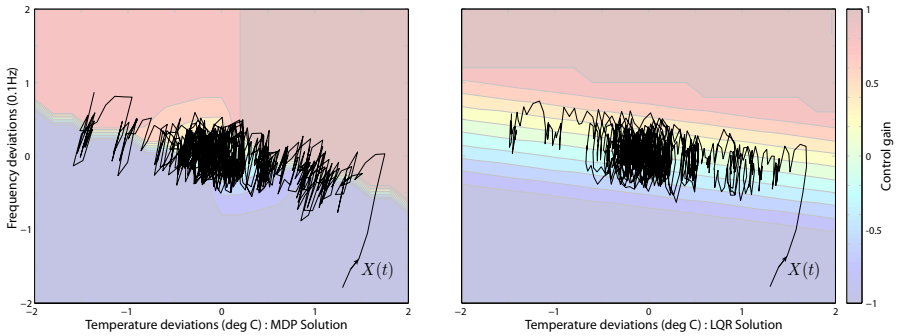


Figure 3.6: Plots of state-feedback control policies obtained from MDP and LQR solutions and the corresponding sample path trajectories for the same initial conditions and noise perturbations.

3.5 Concluding Remarks

This chapter presents two examples to illustrate application of ADP and RL techniques to power system control problems. Numerical studies provide a proof of concept on how RL techniques driven by real-world data are successfully applied to tune to the underlying statistics. Thus, the need to adopt artificial assumptions regarding the uncertainties inherent to such control problems can be avoided.

ADP and RL can be successful provided the approximation architecture employed is a reasonably good fit for the problem. In other words, a good basis is needed for successful approximation. In the examples considered in this chapter, relaxations of the MDP model where used to construct a basis for approximation.

In more complex settings such as power networks, the choice of the basis functions can be more challenging. In the following chapters, a solution is proposed, which combines the best features of ADP/RL and MPC. The future chapters discuss this combination of control approaches.

Chapter 4

Stability and Approximate Optimality

When dealing with approximate solutions to MDPs, the following concerns need to be addressed: (a) How good is the quality of the approximation? (b) Does the approximation provide a stabilizing control policy? (c) How does the system perform under this control policy? This chapter closely examines the above issues under the *average cost* optimality criterion for MDPs. The average cost optimization framework is considered since Lyapunov techniques can be used to address the stability issues in this context. Also, the criterion naturally extends to total cost optimization.

The *Bellman error* provides a quantification of the mismatch in an approximation to the DP equations. This error is used to characterize stability of the control policies associated with the approximation. Furthermore, the Bellman error is used to establish performance bounds for the control policy.

The chapter begins with a definition of the Bellman error for average cost optimization for MDPs. Fluid model-based approximations for MDPs are introduced and bounds on the Bellman error associated with such approximations are established in Theorem 6. Also, closed form solutions to value function for a special class of linear MDPs are obtained in Proposition 4. A stability criterion based on the Bellman error is proposed in Theorem 9. Finally, performance bounds for the control policy from the approximations are established in Theorem 11.

4.1 Bellman Error

This section concerns with approximate solutions to an MDP that may be obtained from either ADP or RL or any other technique. Our concern lies in the Bellman error associated with these approximate MDP solutions. It is shown how the approximate MDP solutions solve the DP equation *exactly* for the cost function perturbed by the Bellman error.

4.1.1 Average Cost Optimality for MDPs

As in Chapter 3, the system dynamics are modeled as evolving in discrete time. Recall that $\mathsf{X} \subseteq \mathbb{R}^{\ell_x}$ and $\mathsf{U} \subseteq \mathbb{R}^{\ell_u}$ denote the state and control input spaces, respectively; lower case notation is used for deterministic variables; their stochastic counterparts are denoted by upper case and boldface notation is used to denote the respective sample paths.

To simplify analysis, an additive-noise is assumed for the MDP.

Assumption 4.1. The system dynamics are as follows:

$$X(t+1) = f\left(X(t), U(t)\right) + W(t)\,, \tag{4.1}$$

where $\boldsymbol{X}$ is the state process, $\boldsymbol{U}$ the control process and $\boldsymbol{W}$ is an i.i.d. sequence that takes values on $\mathsf{W} \subseteq \mathbb{R}^{\ell_x}$, with zero mean and finite covariance Σ_W.

The set $\mathsf{U}(x)$ denotes the set of feasible inputs when state is $x \in \mathsf{X}$. The controlled transition law for the MDP is given by

$$P_u(x, \mathbb{A}) = \mathsf{P}\left\{f(x,u) + W(1) \in \mathbb{A}\right\}\,,$$

for arbitrary $x \in \mathsf{X}$, $u \in \mathsf{U}(x)$, $\mathbb{A} \subset \mathsf{X}$ (Borel measurable).

A *fluid* model associated with the stochastic model (4.1) can be obtained by setting

$$\bar{f}(x,u) = \mathsf{E}\left[X(t+1) \mid X(t) = x,\ U(t) = u\right]\,.$$

Under assumption 4.1, it follows that $\bar{f}(x,u) = f(x,u)$ and the fluid model dynamics can be described by the following nonlinear state-space model de-

scribed as below:

$$x(t+1) = f\left(x(t), u(t)\right) \ , \tag{4.2}$$

with $\boldsymbol{x}$ and $\boldsymbol{u}$ representing state and input trajectories. The fluid model may be used to approximate MDP solutions [104, 105].

The optimality for either the stochastic or fluid model is based on the cost function c defined on the state-action space. Stability of the optimal policies is guaranteed under the following assumption:

Assumption 4.2. The function $c(x, u)$ is non-negative. Further, the function $\inf_u c(\,\cdot\,, u)$ is *coercive*. That is, each sub-level set $S_n \subset \mathsf{X}$ is bounded, where

$$S_n = \{x : c(x, u) \le n \quad \text{for some } u\}, \qquad n \ge 1.$$

For the MDP model (4.1), the optimal long-run *average cost* is defined as

$$\eta^* := \inf_{\boldsymbol{U}} \left\{ \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathsf{E}\left[c\left(X(t), U(t)\right)\right] \right\} \ , \tag{4.3}$$

which, under general conditions, is unique and independent of the initial condition [104]. The associated *average cost optimality equation* (ACOE) is expressed as follows:

$$h^*(x) + \eta^* = \min_{u\in\mathsf{U}(x)} \left\{ c(x, u) + \mathcal{P}h^*\left(x, u\right) \right\} \ . \tag{4.4}$$

Recall the DP operator $\mathcal{P}$ is as defined in (3.4):

$$\mathcal{P}g\left(x, u\right) = \mathsf{E}\left[g\left(X(t+1)\right) \mid X(t) = x, U(t) = u\right] \ , \tag{4.5}$$

for some $x \in \mathsf{X}$, $u \in \mathsf{U}$ and $g\colon \mathsf{X} \to \mathbb{R}$. The function $h^* : \mathsf{X} \to \mathbb{R}$ is called the *relative value function* and is typically unique up to a constant [104]. The minimizer u^* in (4.4) defines an optimal state feedback policy $\phi^*(x)$.

The *total cost* optimality criterion is often employed for the fluid model (4.2). The associated infinite-horizon value function is defined as follows:

$$J^*(x) = \min_{\boldsymbol{u}} \sum_{t=0}^{\infty} c\left(x(t), u(t)\right) \ , \ x(0) = x \in \mathsf{X} \, . \tag{4.6}$$

If the value function is finite valued on X, it satisfies the following DP equation

$$J^*(x) = \min_{u \in \mathsf{U}(x)} \left\{ c(x,u) + \mathcal{K}J^*\,(x,u) \right\}\,, \tag{4.7}$$

where the DP operator $\mathcal{K}$ is defined as in MDP theory: For any function $g\colon \mathsf{X} \to \mathbb{R}$, $\mathcal{K}g$ denotes the function on $\mathsf{X} \times \mathsf{U}$ given by

$$\mathcal{K}g\,(x,u) = g(\bar{f}(x,u))\,, \quad x \in \mathsf{X}\,, u \in \mathsf{U}\,. \tag{4.8}$$

The minimizing control input u^* in (4.7) defines an optimal state feedback policy $\bar{\phi}^*(x)$. Under certain conditions, total cost optimal solutions for the fluid model may be used to approximate average cost optimal solutions for the MDP model.

4.1.2 Approximate MDP Solutions

In this section, a pair (h, η) which approximately solves the ACOE (4.4) is considered. The state feedback policy associated with this approximation is denoted by ϕ and obtained using

$$\phi(x) \in \arg\min_{u \in \mathsf{U}(x)} \left\{ c(x,u) + \mathcal{P}h\,(x,u) \right\}\,, \quad x \in \mathsf{X}\,. \tag{4.9}$$

The Bellman error is a measure of the mismatch in the DP equation with approximate solutions plugged in place of the optimal solution. It is defined as follows.

Definition 2. For any (h, η) where $h\colon \mathsf{X} \to \mathbb{R}$ and $\eta \in \mathbb{R}$, the Bellman error $\mathcal{E}^{\mathsf{BE}}$ is defined as error in the ACOE:

$$\mathcal{E}^{\mathsf{BE}}(x) := h(x) + \eta - \min_{u} \left\{ c(x,u) + \mathcal{P}h\,(x,u) \right\}\,. \tag{4.10}$$

Simple algebraic manipulations allow us to use the Bellman error to construct another cost function such that the approximations (h, η) are optimal solutions to the ACOE corresponding to this cost function. In other words, the following proposition holds:

Proposition 3. *Given any pair (h, η) such that $h\colon \mathsf{X} \to \mathbb{R}$ and $\eta \in \mathbb{R}$, the ACOE is solved for the perturbed cost function $\hat{c} = c + \mathcal{E}^{\mathsf{BE}}$.*

In the following sections, the role of Bellman error in understanding the stability and performance of the approximate MDP solutions is more concretely defined.

4.2 ACOE for Fluid Models

Many of the problems considered in this dissertation can be modeled as an MDP with linear dynamics:

$$X(t+1) = AX(t) + BU(t) + \bar{w} + W(t)\,, \tag{4.11}$$

and quadratic costs. In the examples considered in this dissertation, $\bar{w}$ is non-zero and the constraint set $\mathsf{U}(x)$ is convex for each state $x \in \mathsf{X}$. The corresponding fluid model is a linear system subject to a constant disturbance,

$$x(t+1) = Ax(t) + Bu(t) + \bar{w}\,. \tag{4.12}$$

Under certain relaxations, a closed form solution for the fluid model optimal control problem can be found. This solution is a special case of the LQR problem and provides a starting point in constructing value function approximations for MDPs of the form (4.11).

4.2.1 Exact Solution for ACOE

In this section, a closed form solution for the fluid model problem is obtained by relaxing the input constraints. In fact, optimal solution satisfies the following DP equation:

$$h_0^*(x) + \eta_0^* = \min_u \left\{ c(x,u) + \mathcal{K} h_0^*\,(x,u) \right\}\,, \tag{4.13}$$

which is precisely the ACOE (4.4) for the MDP without noise; that is, $\Sigma_W = 0$. The subscript 0-notation for the relative value function and average cost of the fluid model is chosen to emphasize the absence of noise.

The average-cost optimal control problem corresponding to the ACOE (4.13) provides an analytical platform to study fluid models for which the infinite-horizon value function $J^*(x)$ is *infinite*. Equivalently, the deterministic system *does not* have an equilibrium in the conventional sense:

(C1) There exist no $x^e \in \mathsf{X}$ and $u^e \in \mathsf{U}(x^e)$ for which

$$x^e = Ax^e + Bu^e + \bar{w} \quad \textit{and} \quad c(x^e, u^e) = 0\,.$$

With the input constraints are relaxed, $\mathsf{U}(x) = \mathsf{U} = \mathbb{R}^{\ell_u}$. The ACOE (4.13) admits a solution under the following mild conditions:

(C2) The cost is quadratic: $c(x,u) = x^{\scriptscriptstyle T}Qx + u^{\scriptscriptstyle T}Ru$, with matrices $Q \geq 0$, $R > 0$ of appropriate dimensions.

(C3) The matrix Q can be expressed as $Q = C^{\scriptscriptstyle T}C$ for a matrix C which satisfies:

$$(A, B) \text{ is stabilizable, and } (A, C) \text{ is detectable}\,.$$

Proposition 4. *Under conditions **(C1)-(C3)**, the optimal control problem for the fluid model of* (4.12) *admits a solution to the ACOE* (4.13) *in which*

(i) *The optimal average cost is a solution to the quadratic program,*

$$\begin{aligned} \eta_0^* = \min_{\{x,u\}} \quad & c(x,u) \\ s.t. \quad & x = Ax + Bu + \bar{w}\,. \end{aligned} \tag{4.14}$$

(ii) *The relative value function is quadratic,*

$$h_0^*(x) = x^{\scriptscriptstyle T}M_*x + m_*^{\scriptscriptstyle T}x\,, \tag{4.15}$$

with $M_* \geq 0$ being the solution to the algebraic Riccati equation (ARE) obtained with $\bar{w} = 0$:

$$M_* = Q + A^{\scriptscriptstyle T}M_*A - A^{\scriptscriptstyle T}M_*B\left(R + B^{\scriptscriptstyle T}M_*B\right)^{-1}B^{\scriptscriptstyle T}M_*A\,. \tag{4.16}$$

On denoting K_* as the usual optimal Kalman gain, the vector m_* is the unique solution to

$$m_* = (A - BK_*)^{\intercal} m_* + 2(A - BK_*)^{\intercal} M_* \bar{w} \,. \tag{4.17}$$

The proof of the proposition is based on consideration of the finite-horizon control problem,

$$J_T^*(x) = \min_{\boldsymbol{u}_0^{T-1}} \sum_{t=0}^{T-1} c(x(t), u(t)) + J_0(x(T)) \,, \tag{4.18}$$

where $\boldsymbol{u}_0^{T-1} = \{u(0), u(1), \ldots, u(T-1)\}$, $x(0) = x$ and J_0 represents the terminal cost satisfying:

(C4) J_0 is quadratic: $J_0(x) = x^{\intercal} M_0 x + m_0^{\intercal} x + a_0$.

The sequence of value functions can be recursively defined through value-iteration:

$$J_{T+1}^*(x) = \min_u \left\{ c(x,u) + \mathcal{K} J_T^*(x,u) \right\} . \tag{4.19}$$

Proposition 5. *For each $T \geq 0$, the finite-horizon value function J_T^* is quadratic,*

$$J_T^*(x) = x^{\intercal} M_T x + m_T^{\intercal} x + a_T \,,$$

with the parameters satisfying the recursion

$$\begin{aligned} M_{T+1} &= Q + A^{\intercal} M_T A - A^{\intercal} M_T B Z_T B^{\intercal} M_T A \,, \\ m_{T+1} &= A_T^{\intercal} m_T + 2 A_T^{\intercal} M_T \bar{w} \,, \\ a_{T+1} &= a_T + \eta_T \,, \end{aligned} \tag{4.20}$$

where

$$\begin{aligned} Z_T &= (R + B^{\intercal} M_T B)^{-1} \,, \\ K_T &= Z_T B^{\intercal} M_T A \,, \\ A_T &= A - B K_T \,, \\ \textit{and,} \qquad \eta_T &= \bar{w}^{\intercal} M_T A_T A^{-1} \bar{w} + m_T^{\intercal} A_T A^{-1} \bar{w} - \tfrac{1}{4} m_T^{\intercal} B Z_T B^{\intercal} m_T \,. \end{aligned} \tag{4.21}$$

Proof. The proof is by induction. It is true by assumption for $T = 0$ since $J_0^* = J_0$. Assume J_T^* is of the same form as J_0 for any $T \geq 1$; that is,

$$J_T^*(x) = x^{\scriptscriptstyle T} M_T x + m_T^{\scriptscriptstyle T} x + a_T \,.$$

On substitution in the DP equation (4.19), $J_{T+1}^*(x)$ can be computed as

$$\begin{aligned} J_{T+1}^*(x) = \min_u \Big\{ & x^{\scriptscriptstyle T} Q x + u^{\scriptscriptstyle T} R u + (Ax + Bu + \bar{w})^{\scriptscriptstyle T} M_T (Ax + Bu + \bar{w}) \\ & + m_T^{\scriptscriptstyle T} (Ax + Bu + \bar{w}) + a_T \Big\} \,. \end{aligned}$$

The minimizing input u^* is found as affine state feedback,

$$u^* = -K_T x - \left[K_T A^{-1} \bar{w} + \tfrac{1}{2} Z_T B^{\scriptscriptstyle T} m_T \right] \,. \tag{4.22}$$

Substituting u^* in (4.19) gives the desired result. □

The update equation for M_T in (4.20) is precisely the Riccati equation update [108]. Therefore, under condition **(C3)**,

(i) The sequence of matrices $\{M_T\}$ converges to the unique positive-semi-definite solution, M_∞, to the ARE obtained with $\bar{w} = 0$.

(ii) The sequence of gains $\{K_T\}$ converges to a limiting gain K_∞ and $\{Z_T\}$ converges to a limiting matrix Z_∞.

(iii) The closed loop system matrix $(A - BK_\infty)$ is stable with all eigenvalues strictly within the unit circle [108].

These conclusions are applied to prove Proposition 4.

Proof of Proposition 4: Take $M_* = M_\infty$, so that $K_* = K_\infty$. Stability of $(A - BK_*)$ implies that under the control obtained as the minimizer in (4.13), the resulting input-state trajectory $(u^*(t), x^*(t))$ converges to a constant. Finite-horizon optimality can be used to show that the constant must be a solution to (4.14).

To establish the ACOE (4.13), the finite-horizon relative value function is considered: For $x \in \mathsf{X}, T \geq 0$ by, define

$$h_T(x) = J_T^*(x) - J_T^*(0) = x^{\scriptscriptstyle T} M_T x + m_T^{\scriptscriptstyle T} x \,.$$

The sequence of functions $\{h_T\}$ is convergent and (4.19) implies that its limit solves the ACOE (4.13). □

A salient characteristic of the DP equation (4.13) is that it holds even for the case where the infinite-horizon value function $J^*(x)$ is finite, that is, if the fluid model has an equilibrium $x^e \in \mathsf{X}$ and $u^e \in \mathsf{U}(x)$ such that $c(x^e, u^e) = 0$. In this case, ACOE admits a solution which satisfies

$$\eta_0^* = 0 \qquad \text{and} \qquad h_0^*(x) = J^*(x) \quad \text{for } x \in \mathsf{X}\,.$$

In other words, the solution to the ACOE satisfies the total cost optimality equation (4.7). Therefore, the ACOE (4.13) can be used as the DP equation under both average and total cost optimality for fluid models.

Concrete examples and applications of Proposition 4 for approximating MDP solution are presented in Section 4.5. The application of the closed form solution in Proposition 4 for constructing basis for RL is discussed in chapters 5 and 6.

4.2.2 Fluid Model-Based Approximations for MDPs

Since the fluid value function may be used as an approximation for the MDP solution, the associated Bellman error is studied in this section. In particular, the objective here is to find bounds on the Bellman error associated with this approximation.

Theorem 6. *Consider an MDP satisfying assumptions 4.1 and 4.2. Suppose* (h_0^*, η_0^*) *solves the ACOE* (4.13) *for the corresponding fluid model. If the second derivative of* h_0^* *is uniformly bounded, then the Bellman error for the pair* (h_0^*, η_0^*) *is bounded as follows:*

$$\inf_{(x,u)} \tfrac{1}{2} \operatorname{tr}\left[\nabla^2 h_0^*\left(f(x,u)\right) \cdot \Sigma_W\right] \le \mathcal{E}^{\mathsf{BE}}(x) \le \sup_{(x,u)} \tfrac{1}{2} \operatorname{tr}\left[\nabla^2 h_0^*\left(f(x,u)\right) \cdot \Sigma_W\right]. \tag{4.23}$$

Proof. To prove the desired, a Taylor series approximation is used to bound the difference between $\mathcal{P}g$ and $\mathcal{K}g$ for a C^2 function $g\colon \mathsf{X} \to \mathbb{R}$. Using definition

(4.5) and assumption 4.1, for any $(x, u) \in \mathsf{X} \times \mathsf{U}$,

$$\mathcal{P}g\,(x,u) = \mathsf{E}\left[g\left(X(t+1) \mid X(t) = x, U(t) = u\right)\right] = \mathsf{E}\left[g\left(f(x,u) + W(1)\right)\right]$$
$$= g\left(f(x,u)\right) + \nabla g\left(f(x,u)\right)^{\scriptscriptstyle T} \mathsf{E}\left[W(1)\right] + \tfrac{1}{2}\operatorname{tr}\left[\nabla^2 g\left(f(x,u)\right) \cdot \Sigma_W\right] + \ldots$$

Applying mean value theorem gives

$$\left|\mathcal{P}g\,(x,u) - \mathcal{K}g\,(x,u)\right| \le \sup_{(x,u)} \tfrac{1}{2}\operatorname{tr}\left[\nabla^2 g\left(f(x,u)\right) \cdot \Sigma_W\right]. \tag{4.24}$$

The Bellman error for the pair (h_0^*, η_0^*) is

$$\mathcal{E}^{\mathsf{BE}}(x) = h_0^*(x) + \eta_0^* - \min_u \left\{c(x,u) + \mathcal{P}h_0^*\,(x,u)\right\}.$$

Plugging in (4.13) and using inequality (4.24) for the function h_0^* establishes the desired upper bound. Analogous treatment is used to obtain a lower bound. □

The bounds established in Theorem 6 provide a quantification of the "goodness of fit" for using fluid model-based approximations to MDP solutions.

4.3 Stability

Two formulations of stability are used in the dissertation. The first is based on average cost and the second is based on total cost.

Definition 7. A policy ϕ is *ac-stabilizing* if for each initial condition $X(0) = x$, the average cost,

$$\eta(x) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-1} \mathsf{E}[c(X(t), U(t))]\,, \tag{4.25}$$

is finite.

Definition 8. The policy is *tc-stabilizing* if $\eta \equiv 0$ and, moreover, the total

cost J is finite for each initial condition $X(0) = x$, where

$$J(x) = \sum_{t=1}^{\infty} \mathsf{E}[c(X(t), U(t))] \,. \tag{4.26}$$

We do not expect J to be finite valued unless $\Sigma_W = 0$, in which case tc-stability is a property of the fluid model. The focus of the dissertation is on ac-stability, so the prefix "ac" will be dropped if there is no risk of confusion.

It follows from the Comparison Theorem of [109] that a policy ϕ is ac-stabilizing for the stochastic model (4.1) if there exists a function $V \colon \mathsf{X} \to \mathbb{R}_+$ and finite constant $\bar{\eta}$ such that the *Poisson's inequality* holds

$$\mathcal{P}_\phi V \,(x) \le V(x) - c_\phi(x) + \bar{\eta} \,. \tag{4.27}$$

In this case, $\eta(x) \le \bar{\eta}$ for each x.

In the following theorem, bounds on the Bellman error are used to establish sufficient conditions for stability of policy obtained from ADP, RL or other approximation techniques.

Theorem 9. *Consider the pair (h, η) satisfying $h \colon \mathsf{X} \to \mathbb{R}_+$ and $\eta < \infty$ as approximate solution to the ACOE* (4.4) *for the MDP model* (4.1)*. Suppose ϕ is the resulting control policy obtained using* (4.9)*. If there exists a finite $\epsilon > 0$ and $n < \infty$ such that the Bellman error for the approximate solution satisfies the following condition:*

$$\frac{\mathcal{E}^{\mathsf{BE}}(x)}{c_\phi(x)} \ge -1 + \epsilon \,, \qquad \textit{for } \|x\| \ge n \,. \tag{4.28}$$

Then the Poisson's inequality holds under policy ϕ.

Proof. Observe that for each finite n

$$\frac{\mathcal{E}^{\mathsf{BE}}(x)}{c_\phi(x)} \ge -1 + \epsilon \qquad \Longrightarrow \qquad c_\phi(x) + \mathcal{E}^{\mathsf{BE}}(x) \ge \epsilon c_\phi(x) \quad \text{for } \epsilon > 0 \,.$$

The Bellman error defined in (4.10) can be expressed in terms of the policy ϕ as follows:

$$\mathcal{E}^{\mathsf{BE}}(x) = h(x) + \eta - c_\phi(x) - \mathcal{P}_\phi h \,(x) \,.$$

Then, rearranging and invoking (4.28) gives

$$\mathcal{P}_\phi h\,(x) \le h(x) + \eta - \epsilon c_\phi(x)\,.$$

Thus, Poisson's inequality (4.27) holds for $V = \epsilon^{-1}h$ and $\bar{\eta} = \epsilon^{-1}\eta$. □

Theorem 9 defines a stability criterion for control policies associated with approximate MDP solutions. The conditions in this theorem extend to the specific case of a fluid model by setting $\Sigma_W = 0$. In particular, a Bellman error for an approximate solution to the ACOE (4.13) is defined in a manner analogous to (4.10)

$$\mathcal{E}_0^{\mathsf{BE}}(x) := h(x) + \eta - \min_u \{c(x,u) + \mathcal{K}h\,(x,u)\}\ . \tag{4.29}$$

Bounds on this error provide sufficient conditions for the stability of the fluid model, as stated below.

Corollary 10. *Consider the pair (h,η) satisfying $h\colon \mathsf{X} \to \mathbb{R}_+$ and $\eta < \infty$ as approximate solution to the ACOE* (4.13) *for a fluid model* (4.2)*. Let ϕ denote the resulting control policy and $\mathcal{E}_0^{\mathsf{BE}}$ denote the associated Bellman error. If there exists finite $\epsilon > 0$ for each $n < \infty$ such that the Bellman error for the approximate solution satisfies the condition*

$$\frac{\mathcal{E}_0^{\mathsf{BE}}(x)}{c_\phi(x)} \ge -1 + \epsilon \qquad \textit{for } \|x\| \ge n\,, \tag{4.30}$$

then policy ϕ is ac-stabilizing for the given dynamics.

Recall that in the context of total cost optimality for the fluid model, the ACOE is solved for the pair $(J^*, 0)$. Then, the Poisson's inequality takes the following form: A policy ϕ is tc-stabilizing for the fluid model (4.2) if there exists a function $V\colon \mathsf{X} \to \mathbb{R}_+$ such that

$$\mathcal{K}_\phi V\,(x) \le V(x) - c_\phi(x)\,. \tag{4.31}$$

The Poisson's inequality in the above form is used to provide a tc-stability criterion for the fluid model.

Stability of the control policies resulting from the approximation is of particular importance if the approximate value function is used as a terminal cost in a receding horizon, predictive control framework. The precise connections between the stabilizing properties of the approximation techniques and those of predictive controllers are established in Chapter 6.

4.4 Performance Bounds

The performance of a control policy ϕ obtained from ADP or RL and applied to the stochastic model (4.1) may be characterized in terms of the resulting invariant distribution and the associated costs. In this section, the steady-state mean of c under the policy ϕ is used as a measure for the control performance. It is expressed as

$$\eta_\phi = \mathsf{E}^\phi \left[c(X, U) \right] \ ,$$

where (X, U) denotes the stationary realization under the invariant distribution for the given policy ϕ.

Our interests lie in comparing the expected cost η_ϕ of the policy ϕ with the optimal average costs η^*, which can be expressed as the steady-state mean of c under the optimal policy ϕ^*:

$$\eta^* = \mathsf{E}^{\phi^*} \left[c(X, U) \right] \ .$$

In the following theorem, bounds on the Bellman error are used to establish bounds on the comparative performance of η_ϕ with respect to η^*.

Theorem 11. *Suppose ϕ and ϕ^* denote the policies obtained from the approximate and optimal solutions to the ACOE (4.4) respectively. Furthermore, suppose the Bellman error $\mathcal{E}^{\mathsf{BE}}$ for the approximate MDP solution satisfies the following conditions:*

- *$\mathcal{E}^{\mathsf{BE}}(x)/c_{\hat{\phi}}(x)$ is uniformly bounded over X for $\hat{\phi} = \phi$ and ϕ^*; and,*
- *for some $\epsilon_1, \epsilon_2 > 0$*

$$-1 + \epsilon_1 \leq \frac{\mathcal{E}^{\mathsf{BE}}(x)}{c_{\hat{\phi}}(x)} \leq -1 + \epsilon_2 \qquad \text{for } c_{\hat{\phi}}(X) \leq n \ , \ \hat{\phi} = \phi \text{ and } \phi^* \ . \tag{4.32}$$

Finally, if the following condition holds for some large n:

$$\mathsf{E}\left[c(X,U)\mathbb{I}_{\{c(X,U)>n\}}\right] \le o(1) \qquad \textit{under policies } \phi, \phi^* \,, \tag{4.33}$$

then the following bound holds true:

$$\eta_\phi \le \frac{\epsilon_2}{\epsilon_1}\eta^* \,. \tag{4.34}$$

Proof. To get the desired result, (4.33) is used to bound expected cost under policy ϕ as follows:

$$\begin{aligned}
\eta_\phi &= \mathsf{E}^\phi\left[c(X,U)\right] \\
&= \mathsf{E}^\phi\left[c(X,U)\,\mathbb{I}_{\{c(X,U)\le n\}}\right] + \mathsf{E}^\phi\left[c(X,U)\,\mathbb{I}_{\{c(X,U)>n\}}\right] \\
&\le \mathsf{E}^\phi\left[c(X,U)\,\mathbb{I}_{\{c(X,U)\le n\}}\right] + o(1) \,.
\end{aligned}$$

Then, using (4.32),

$$\eta_\phi \le \frac{1}{\epsilon_1}\mathsf{E}^\phi\left[(c+\mathcal{E}^{\mathsf{BE}})\mathbb{I}_{\{c(X,U)\le n\}}\right] + o(1) \,.$$

Observe that $(c+\mathcal{E}^{\mathsf{BE}})\mathbb{I}_{\{c(X,U)\le n\}} \uparrow (c+\mathcal{E}^{\mathsf{BE}})$ as $n\to\infty$. Furthermore, $\epsilon_1 > 0$ implies $c+\mathcal{E}^{\mathsf{BE}} \ge 0$. Thus, the monotone convergence theorem applies, and therefore, $\mathsf{E}^\phi\left[(c+\mathcal{E}^{\mathsf{BE}})\mathbb{I}_{\{c(X,U)\le n\}}\right] \to \mathsf{E}^\phi\left[(c+\mathcal{E}^{\mathsf{BE}})\right]$ as $n\to\infty$. Therefore, taking the limit as $n\to\infty$ on both sides of the second inequality

$$\eta_\phi \le \frac{1}{\epsilon_1}\mathsf{E}^\phi\left[(c+\mathcal{E}^{\mathsf{BE}})\right] .$$

Optimality of the policy ϕ for the ACOE with the perturbed cost $c+\mathcal{E}^{\mathsf{BE}}$ follows from Proposition 3. Therefore,

$$\begin{aligned}
\eta_\phi &\le \frac{1}{\epsilon_1}\mathsf{E}^{\phi^*}\left[(c+\mathcal{E}^{\mathsf{BE}})\right] \\
&= \frac{1}{\epsilon_1}\mathsf{E}^{\phi^*}\left[(c+\mathcal{E}^{\mathsf{BE}})\mathbb{I}_{\{c(X,U)\le n\}}\right] + \mathsf{E}^{\phi^*}\left[(c+\mathcal{E}^{\mathsf{BE}})\mathbb{I}_{\{c(X,U)>n\}}\right] \,.
\end{aligned}$$

Again, using the bounds in (4.32) and (4.33) gives

$$\eta_\phi \;\leq\; \frac{\epsilon_2}{\epsilon_1}\mathsf{E}^{\phi^*}\left[c(X,U)\,\mathbb{I}_{\{c(X,U)\leq n\}}\right] + o(1) + \frac{1}{\epsilon_1}\mathsf{E}^{\phi^*}\left[\mathcal{E}^{\mathsf{BE}}(X)\mathbb{I}_{\{c(X,U)>n\}}\right] .$$

Since $\mathcal{E}^{\mathsf{BE}}(x)/c_{\hat{\phi}}(x)$ is uniformly bounded over the state space, applying the dominated convergence theorem gives the desired result. □

Theorem 11 characterizes the performance of the approximate policy in terms of the expected costs under the optimal policy. If the error thresholds ϵ_1 and ϵ_2 are sufficiently tight, then the system performance under the approximate control policy is sufficiently close to that under the optimal policy.

4.5 Approximations for the PNNL Model

This chapter introduces Bellman error as a metric to quantify the goodness of the approximation, as well as a tool to analyze the stability and the performance for a control policy obtained from approximate solutions to MDPs. In this section, numerical results on a test system illustrate how the metric may be applied to practical power system control problems. As an example, the dispatch problem for a representative power system – the PNNL model – is considered.

4.5.1 The PNNL Model

PNNL has developed a microgrid test system model in the DIgSILENT software. The schematic in Figure 4.1 depicts the system: it is derived from the IEEE 34 bus test feeder. The IEEE distribution system is modified to accommodate detailed dynamic models of households, diesel generators, wind turbine generators and a battery energy storage system(BESS). The modifications are listed in [110].

The numerical studies described in this chapter, as well as the next chapters 5 and 6, use a simplified model of the PNNL test system. The PNNL model used in this dissertation is assumed to consist of a diesel generator, a BESS, a wind power plant and a mix of residential loads which constitute the total

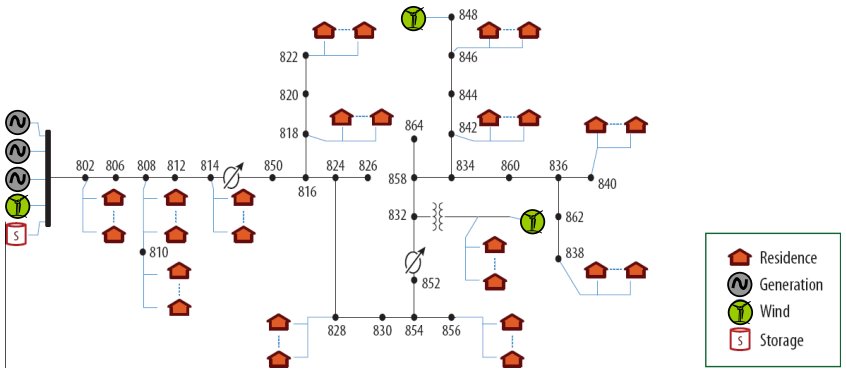


Figure 4.1: The PNNL microgrid testbed.

demand. The wind plant data is obtained from [97] and an aggregate load of 1500 houses is generated using [111].

The charging/discharging of BESS compensates for variability in the net load (total load minus wind generation) and is governed by a threshold policy: The BESS is charged if the net load is less than the threshold and discharged if it is greater than the threshold. Indeed, the threshold value can be viewed as power demanded by the net load and BESS. It is supplied by the diesel generator. In the event of a shortfall or surplus, an expensive balancing service is deployed.

The diesel generator's fuel costs are assumed to be quadratic. The BESS operational costs are cast as proxy costs which penalize deviations of its state of charge (SOC) from a specified reference value. The balancing service is assumed to be procured from an expensive ancillary service resource, whose operation is independent of the other resources in the system. This resource can be thought of either as the microgrid's interaction with a larger interconnected power network or as fast-responding generation source/load sink which is run only to manage the shortfalls/surpluses in system generation. In either case, economics dictate minimum reliance on this service.

The capacity and ramping limits of the resources are simulation parameters that are varied to study control synthesis in constrained environments. The specific values used are discussed in sections describing the corresponding numerical results.

4.5.2 Problem Formulation

The control objective in the numerical studies is to find the least-cost control strategies for determining the outputs of the diesel generator and BESS so as to meet the net load demand on the system. Factors such as system losses and frequency/voltage dynamics are disregarded and simplified cost structures are adopted for generation and storage resources. The goal is control synthesis, for which a simplified model is frequently justifiable.

The problem formulation is adapted from [112]. The state of the system $X(t)$ at time t is described by the output of the diesel generator $P_{\mathsf{G}}(t)$, the threshold of BESS $P_{\mathsf{thr}}(t)$, its SOC $\xi_{\mathsf{S}}(t)$ and the amount of balancing service required $P_{\mathsf{bal}}(t)$ at that time. That is,

$$X(t) = \left[P_{\mathsf{G}}(t), P_{\mathsf{thr}}(t), \xi_{\mathsf{S}}(t), P_{\mathsf{bal}}(t)\right]^T .$$

The control actions $U(t)$ available at this time are the ramping in the generation output $\Delta P_{\mathsf{G}}(t)$ and change in the BESS threshold $\Delta P_{\mathsf{thr}}(t)$:

$$U(t) = \left[\Delta P_{\mathsf{G}}(t), \Delta P_{\mathsf{thr}}(t)\right]^T .$$

The main sources of uncertainty are the output of the wind plant $G(t)$ and the residential load $D(t)$. These impact the power supplied by the BESS:

$$P_{\mathsf{S}}(t) = D(t) - G(t) - P_{\mathsf{thr}}(t) ,$$

where $P_{\mathsf{S}}(t) > 0$ indicates discharging of the BESS and $P_{\mathsf{S}}(t) < 0$ indicates its charging. For the balancing service, $P_{\mathsf{bal}}(t) > 0$ indicates excess generation while $P_{\mathsf{bal}}(t) < 0$ indicates a generation deficit.

System dynamics: The changing set-points result in the following dynamics:

$$\begin{aligned}
P_{\mathsf{G}}(t+1) &= P_{\mathsf{G}}(t) + \Delta P_{\mathsf{G}}(t) , \\
P_{\mathsf{thr}}(t+1) &= P_{\mathsf{thr}}(t) + \Delta P_{\mathsf{thr}}(t) , \\
\xi_{\mathsf{S}}(t+1) &= \xi_{\mathsf{S}}(t) - \alpha_{\mathsf{S}}\big(D(t) - G(t) - P_{\mathsf{thr}}(t)\big) , \\
P_{\mathsf{bal}}(t+1) &= P_{\mathsf{G}}(t+1) - P_{\mathsf{thr}}(t+1) , \\
&= P_{\mathsf{G}}(t) - P_{\mathsf{thr}}(t) + \Delta P_{\mathsf{G}}(t) - \Delta P_{\mathsf{thr}}(t) .
\end{aligned} \tag{4.35}$$

Here, the parameter α_{S} represents the conversion factor for the BESS with

$$\alpha_{\mathsf{S}} = \frac{\nu_{\mathsf{S}}}{E_{\mathsf{S}}^{\mathsf{max}}} \Delta t \,,$$

where ν_{S} and $E_{\mathsf{S}}^{\mathsf{max}}$ denote the efficiency and energy capacity of the storage device and Δt represents the time step duration in hours. The dynamics in (4.35) can be cast in a linear form as

$$X(t+1) = AX(t) + BU(t) + DV(t) \,, \tag{4.36}$$

where $V(t) = \big[G(t), D(t)\big]^{\,T}$ is the disturbance process. This model resembles the MDP model described in (4.11).

The states and inputs are constrained so that

$$X^{\mathsf{min}} \leq X(t) \;\; \leq X^{\mathsf{max}} \quad \text{and} \quad \mathsf{U}^{\mathsf{min}} \leq U(t) \;\; \leq U^{\mathsf{max}} \tag{4.37}$$

for each t where the limits X^{min}, X^{max}, U^{min} and U^{max} are determined by capacity and ramping limits as follows:

$$\begin{aligned}
X^{\mathsf{min}} &:= \big[P_{\mathsf{G}}^{\mathsf{min}}, P_{\mathsf{thr}}^{\mathsf{min}}, \xi_{\mathsf{S}}^{\mathsf{min}}, P_{\mathsf{bal}}^{\mathsf{min}}\big]^{\,T} \,, \\
X^{\mathsf{max}} &:= [P_{\mathsf{G}}^{\mathsf{max}}, P_{\mathsf{thr}}^{\mathsf{max}}, \xi_{\mathsf{S}}^{\mathsf{max}}, P_{\mathsf{bal}}^{\mathsf{max}}]^{\,T} \,, \\
U^{\mathsf{min}} &:= \big[\Delta P_{\mathsf{G}}^{\mathsf{min}}, \Delta P_{\mathsf{thr}}^{\mathsf{min}}\big]^{\,T} \,, \\
U^{\mathsf{max}} &:= [\Delta P_{\mathsf{G}}^{\mathsf{max}}, \Delta P_{\mathsf{thr}}^{\mathsf{max}}]^{\,T} \,.
\end{aligned}$$

The parameters defining the constraints on states and actions for the numerical results presented here are described in Table 4.1.

Cost function: The dispatch problem is set up to minimize the fuel costs of generators, deviations of SOC from a reference value $\xi_{\mathsf{S}}^{\mathsf{ref}}$, balancing service needed, and the mechanical wear and tear on generators caused by ramping. A cost function is formulated to take into account these diverse costs and takes the form of a weighted sum,

$$\begin{aligned}
c\big(X(t), U(t)\big) = \gamma_1 \big(\, a P_{\mathsf{G}}^2(t) + b P_{\mathsf{G}}(t) + c \, \big) + \gamma_2 \big(\, \xi_{\mathsf{S}}(t) - \xi_{\mathsf{S}}^{\mathsf{ref}} \, \big)^2 \\
+ \gamma_3 P_{\mathsf{bal}}^2(t) + \gamma_4 \Delta P_{\mathsf{G}}^2(t) + \gamma_5 \Delta P_{\mathsf{thr}}^2(t) \,,
\end{aligned} \tag{4.38}$$

Table 4.1: Simulation parameters for PNNL model

Resource	Constraint specifications
Diesel generator	$P_{\mathsf{G}}^{\mathsf{min}} = 0$, $P_{\mathsf{G}}^{\mathsf{max}} = 5$ GW, $\Delta P_{\mathsf{G}}^{\mathsf{min}} = -\infty$, $\Delta P_{\mathsf{G}}^{\mathsf{max}} = \infty$ GW
BESS	$P_{\mathsf{thr}}^{\mathsf{min}} = -\infty$, $P_{\mathsf{thr}}^{\mathsf{max}} = \infty$ $\xi_{\mathsf{S}} \in [0,1]$, $E^{\mathsf{max}} = 3.6$ GWh
Disturbances	$\mathsf{E}\left[D(t)\right] = 2.27$ GW , $\mathsf{E}\left[G(t)\right] = 0.27$ GW

where the weight γ_i determines the relative importance of the i^{th}objective and $\sum_i \gamma_i = 1$. The cost can reformulated in a quadratic form,

$$c\big(x,u\big) = (x - x^{\mathsf{ref}})^T Q (x - x^{\mathsf{ref}}) + u^T R u + \text{some constant}\,, \tag{4.39}$$

where x^{ref} is a reference state.

The usual objective for the dispatch problem is to minimize the cost defined in (4.39) over a specified time horizon, subject to system dynamics (4.36) and state/input constraints (4.37). For the purposes of control synthesis, the time horizon in the dispatch problem is taken to be infinite and the performance of the resulting control policies is studied.

4.5.3 LQR-based Approximate Solution

A fluid model corresponding to the MDP model (4.36) is obtained as follows:

$$x(t+1) = Ax(t) + Bu(t) + D\bar{v}\,. \tag{4.40}$$

The linear dynamics in (4.40) are in the exact same form as (4.12) with $\bar{w} = D\bar{v}$. Since the costs incurred are quadratic in nature, the fluid model obtained by relaxing the constraints on the states and control inputs corresponds to the LQR model described in Section 4.2.1.

In the numerical results reported here, a closed form solution for the LQR relative value function $h_0^*(x)$ and the average cost η_0^*, obtained from Proposition 4, is used to approximate the optimal value function $h^*(x)$ and optimal

average cost η^* for the MDP model. The control policy corresponding to the approximation is an affine state feedback of the form (4.22), with the gain K_T and matrix Z_T replaced by their limiting values K_* and Z_* respectively.

The Bellman error $\mathcal{E}^{\mathsf{BE}}(x)$ for this approximate MDP solution is computed and studied to investigate the stability and performance when the LQR control policy is applied to the MDP model (4.40). Figure 4.2 shows a plot of the ratio $\dfrac{\mathcal{E}^{\mathsf{BE}}(x)}{c_\phi(x)}$ as a function of the norm $\|x\|$ for the state values observed under the LQR policy. Observe that the LQR control policy satisfies the conditions of 9 and is, hence, stabilizing for the MDP model. However, its performance is not good, and leaves scope for improvement.

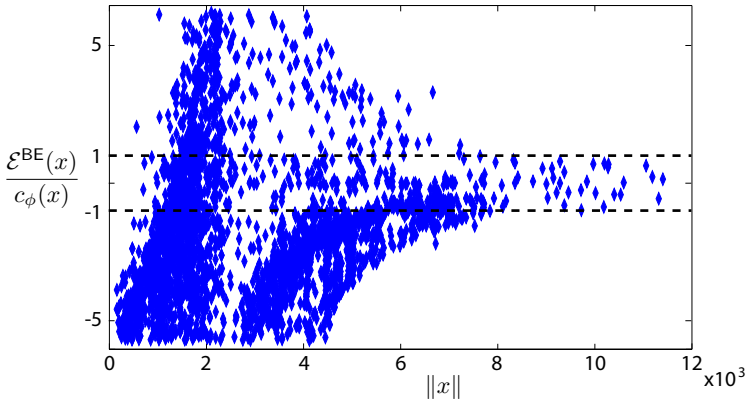


Figure 4.2: Bellman error ratio for the LQR-based approximate MDP solution.

In the next chapter, it is shown how the LQR-based approximation can be improved by using RL techniques.

4.6 Concluding Remarks

In this chapter, an analytical framework is introduced to study the performance of approximate solutions to the MDP and fluid models. The Bellman error, defined as a mismatch in the DP equation as a result of the approximation, is a key component in the analysis. Theorems 9 and 11 show how bounds on the Bellman error can be used to establish sufficient conditions for the stability as well as performance bounds for the control policy.

4.6 Concluding Remarks

Another important aspect discussed in this chapter is the use of fluid model-based approximations for solving MDPs. Theorem 6 provides bounds on the resulting Bellman error. A practical example considered in Section 4.5 illustrates how the fluid model may be constructed from the MDP model, but it also emphasizes the need to improve fluid model-based approximations using techniques from RL. In the next chapter, Q-learning algorithms are devised and deployed for this purpose.

Chapter 5

Parameterized Q-learning Algorithms

In this chapter, RL techniques are devised to approximately solve the DP equations. One concern in RL applications is the issue of *exploration*. The ergodic norm used to define the approximation error criterion in TD learning or SARSA algorithms is for a fixed policy. This precludes the possibility of exploring the state-action space. The Q-learning algorithm provides a work-around regarding this: a randomized stationary policy is employed to sample the state-action space and, thus, allow exploration.

In this chapter, two parameterized Q-learning algorithms are devised for the control of nonlinear state space models under the average cost optimality criterion. These algorithms overcome the curse of dimensionality associated with Watkin's original Q-learning technique. The first Q-learning algorithm is a discrete-time counterpart of the Q-learning algorithm devised in [82]: an approximation criterion based on Bellman error is used for learning. The second algorithm is devised based on linear programming approach to solve an MDP; it is an RL variant of the ADP algorithm of [113].

The choice of basis plays an important role in obtaining a good approximation. Simplified models (e.g., fluid or diffusion model) are employed for basis construction in this dissertation. The Q-learning algorithms are developed based on the fluid model approximation of MDP. This relaxation allows us to employ a stochastic approximation of the steepest descent algorithm to recursively estimate the basis weights for the parameterized Q-function.

Practical applications of the proposed Q-learning algorithms for power system control purposes are provided via numerical studies on test systems.

Specifically, the dispatch problem for the PNNL model introduced in Section 4.5 is revisited. Issues such as the choice of basis for the parameterization and tuning the algorithm parameters are illustrated via this example. Finally, the performance of the control policies obtained from the Q-learning algorithms is compared against LQR feedback policy studied in Section 4.5. Simulation results demonstrate how the control performance can be improved when the Q-function is approximated as a combination of the fluid value function and penalty functions that take into account state/action constraints.

5.1 Q-learning for Deterministic Systems

The *Q-function* used in Q-learning is a real-valued function defined on $\mathsf{X} \times \mathsf{U}$. For average cost optimization, it is defined as the function appearing in the braces of the ACOE.

Recall the MDP modeled under assumptions 4.1 and 4.2:

$$X(t+1) = f\left(X(t), U(t)\right) + W(t)\,,$$

and the corresponding ACOE:

$$h^*(x) + \eta^* = \min_{u \in \mathsf{U}(x)} \left\{ c(x,u) + \mathcal{P}h^*\left(x,u\right) \right\}\,.$$

Then, the Q-function for the MDP model is defined as

$$H^*(x,u) := c(x,u) + \mathcal{P}h^*\left(x,u\right). \tag{5.1}$$

The goal of Q-learning is to approximate this function H^*.

A fluid model approximation of the MDP is used in algorithm development. Recall that the fluid model dynamics take the form

$$x(t+1) = f\left(x(t), x(t)\right)\,,$$

and the associated ACOE is given as

$$h_0^*(x) + \eta_0^* = \min_{u} \left\{ c(x,u) + \mathcal{K}h_0^*\left(x,u\right) \right\}\,.$$

Then, the Q-function for the fluid model is defined by

$$H_0^*(x,u) := c(x,u) + \mathcal{K}h_0^*(x,u)\,. \tag{5.2}$$

This chapter provides Q-learning algorithms to approximate H_0^*. The approximation is with respect to a specific norm. The details for the approximation are described below.

Approximation architecture: Similar to [82], a parameterized family of real-valued functions on $\mathsf{X} \times \mathsf{U}$, denoted by $\{H^\theta(x,u) : \theta \in \mathbb{R}^d, x \in \mathsf{X}, u \in \mathsf{U}\}$, is considered. The goal of Q-learning is to find parameters θ so that $H^\theta \approx H_0^*$.

A natural parameterization for H^θ is of the form

$$H^\theta(x,u) = c(x,u) + \theta^T\psi(x,u)\,, \tag{5.3}$$

where $\psi : \mathsf{X} \times \mathsf{U} \to \mathbb{R}^d$ is the basis and $\theta \in \mathbb{R}^d$ is the parameter to be learned. Given a basis $\{\varphi_i : 1 \le i \le d\}$ intended for TD learning, a basis for Q-learning may be chosen as the functions on $\mathsf{X} \times \mathsf{U}$,

$$\psi_i(x,u) = \mathcal{K}\varphi_i\,(x,u) = \varphi_i\,(f(x,u)) \;\text{ for }\; 1 \le i \le d\,. \tag{5.4}$$

For average cost optimization, an additional parameter $\hat{\eta}$ that approximates the optimal average cost η_0^* is also used.

Ergodic environment for learning: As mentioned in the introduction, the usual Q-learning algorithms for MDPs apply a randomized stationary policy to allow sufficient sampling of the state-action space [114]. Similar assumptions are adopted here so that the fluid model dynamics provide a stationary and ergodic realization.

In the context of this dissertation, ergodicity implies as $T \to \infty$

$$\frac{1}{T}\int F\,(x(t),u(t))\,dt \to \int F(x,u)\varpi(dx,du)$$

Ergodicity for the fluid model is achieved by design: control actions are "randomly" chosen by perturbing a stabilizing state-feedback policy ϕ with an excitation signal $\boldsymbol{\zeta}$. The following is assumed:

Assumption 5.1. The input is of the form

$$u(t) = \bar{\phi}\big(x(t)\big) + \zeta(t) \,. \tag{5.5}$$

The controlled system admits a stationary and ergodic realization $(\boldsymbol{X}, \boldsymbol{U})$ with marginal distribution ϖ.

A Hilbert-space setting is adopted for approximation, based on the corresponding ergodic norm. For measurable functions $F, G : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$, the inner product and norm are defined as follows:

$$\begin{aligned} \langle F, G \rangle &:= \int F(x,u) G(x,u) \varpi(dx, du) \,, \\ \|F\|^2 &:= \int F^2(x,u) \varpi(dx, du) \,. \end{aligned}$$

In terms of the stationary realization $(\boldsymbol{X}, \boldsymbol{U})$,

$$\langle F, G \rangle = \mathsf{E}_{\varpi}[F(X(t), U(t)) G(X(t), U(t))] \,,$$

where the expectation is independent of time. Under the ergodicity assumption, these expectations can be approximated from a sample path trajectory on the $\mathsf{X} \times \mathsf{U}$ space.

The discussion in this section and the algorithm development in the following sections can be extended to total cost optimality criterion. We stick to average cost optimization.

5.2 Bellman Error-Based Q-learning

A natural criterion for choosing the parameter θ is to minimize the actual error $\|H_0^* - H^\theta\|$, which is the viewpoint taken in TD-learning. However, this criterion is intractable. Hence, the Bellman error minimization criterion is adopted for the first Q-learning algorithm.

On denoting

$$\underline{H}_0^*(x) = \min_{u \in \mathsf{U}(x)} H_0^*(x,u) \,, \tag{5.6}$$

the DP equation implies that $\underline{H}_0^* = h_0^* + \eta_0^*$. The DP equation is thereby transformed into a fixed point equation in H_0^*:

$$H_0^*(x,u) = c(x,u) + \mathcal{K}\underline{H}_0^*(x,u) - \eta_0^* \,. \tag{5.7}$$

The Bellman error is defined to be the error in the fixed point equation (5.7).

The Q-learning algorithm devised here minimizes the *mean-square Bellman error*, which is defined as

$$\begin{aligned} \mathcal{E}^{\mathsf{mse}}(\theta,\hat{\eta}) &:= \tfrac{1}{2}\,\|H^\theta - \big(c + \mathcal{K}\,\underline{H}^\theta - \hat{\eta}\big)\|^2 \\ &= \tfrac{1}{2}\,\mathsf{E}\Big[\Big(H^\theta\big(X(t),U(t)\big) - \big[c\big(X(t),U(t)\big) + \underline{H}^\theta\big(X(t+1)\big) - \hat{\eta}\big]\Big)^2\Big], \end{aligned} \tag{5.8}$$

where the function $\underline{H}^\theta \colon \mathsf{X} \to \mathbb{R}$ is defined analogous to (5.6):

$$\underline{H}^\theta(x) = \min_{u\in\mathsf{U}(x)} H^\theta(x,u) \,. \tag{5.9}$$

If $\mathcal{E}^{\mathsf{mse}}(\theta^*,\hat{\eta}^*) = 0$, then the fixed point equation (5.7) holds in a mean-square sense. Consequently, the DP equation (4.13) is solved a.e. $[\varpi]$.

The Q-learning algorithm devised here to minimize $\mathcal{E}^{\mathsf{mse}}$ is a stochastic approximation algorithm intended to approximate the steepest descent. First, an expression for the gradient is obtained. The function $H^\theta(x,u)$ defined in (5.3) is affine in θ with

$$\nabla H^\theta(x,u) = \psi(x,u) \,.$$

Letting $u_{x,\theta}^*$ denote the minimizer in (5.9),

$$\nabla\underline{H}^\theta(x) := \underline{\psi}^\theta(x) := \psi(x,u_{x,\theta}^*) \,.$$

The gradient of $\mathcal{E}^{\mathsf{mse}}$ with respect to θ is thus

$$\begin{aligned} \nabla_\theta\,\mathcal{E}^{\mathsf{mse}}(\theta,\hat{\eta}) &= \langle H^\theta - \big(c + \mathcal{K}\,\underline{H}^\theta - \hat{\eta}\big), \psi - \mathcal{K}\,\underline{\psi}^\theta\rangle \\ &= \mathsf{E}\left[\Delta(X,U;\theta,\hat{\eta})\right] \,, \end{aligned} \tag{5.10}$$

where

$$\Delta(X,U;\theta,\hat{\eta}) = \Big[H^{\theta}\left(X(t),U(t)\right) - c\left(X(t),U(t)\right) - \underline{H}^{\theta}\big(X(t+1)\big) + \hat{\eta}\Big] \times \Big[\psi\big(X(t),U(t)\big) - \underline{\psi}^{\theta}\big(X(t+1)\big)\Big] .$$

Justification of the interchange of derivative and expectation operations is possible under general conditions (e.g., if ϖ has compact support and $\underline{H}^{\theta}$ is continuously differentiable).

The steepest descent algorithm is thus

$$\theta(t+1) = \theta(t) - \gamma(t)\nabla_{\theta}\,\mathcal{E}^{\mathsf{mse}}(\theta,\hat{\eta}) ,$$

where the gradient is defined in (5.10). The stochastic approximation algorithm is obtained to recursively estimate θ^* by removing the expectation in (5.10):

$$\theta(t+1) = \theta(t) - \gamma(t)\Delta\left(x(t),u(t);\theta(t),\hat{\eta}(t)\right) . \tag{5.11}$$

The gain sequence $\{\gamma(t)\}$ is chosen such that standard conditions for stochastic approximation are satisfied [115, 116]. A typical form of $\{\gamma(t)\}$ is

$$\gamma(t) = \frac{(1+t)^{-1}}{1+\|\theta\|^{p}} ,$$

where p is chosen such that $\dfrac{\Delta(x,u;\theta,\eta)}{1+\|\theta\|^{p}}$ is Lipschitz.

The update equation from $\hat{\eta}(t)$ is derived based on the first order optimality conditions for $(\theta^*,\hat{\eta}^*)$: For a fixed θ, the optimizing $\hat{\eta}$ satisfy

$$\hat{\eta}^*(\theta) = \mathsf{E}\big[c\left(X(t),U(t)\right) + \underline{H}^{\theta}\left(X(t+1)\right) - H^{\theta}\left(X(t),U(t)\right)\big] .$$

Thus, it is convenient to adopt a two-time-scale approach to minimize $\mathcal{E}^{\mathsf{mse}}$. The Q-learning algorithm updates $\theta(t)$ using (5.11) while the update equation

for $\hat{\eta}(t)$ is then taken to be the sample mean:

$$\hat{\eta}(t) = \frac{1}{t}\sum_{\tau=1}^{t}\left[c(x(\tau),u(\tau)) + \underline{H}^{\theta(\tau)}(x(\tau+1)) - H^{\theta(\tau)}(x(\tau),u(\tau))\right]. \quad (5.12)$$

This approach is used in the numerical results reported in this dissertation.

5.3 Linear Programming-Based Q-learning

A major concern in the Q-learning algorithm described in Section 5.2 is that the objective function appearing in (5.8) is not convex. The linear programming approach to ADP [81,113] provides insights on how to construct a convex Q-learning algorithm.

The average cost optimality problem can be cast as a linear program (LP) as follows:

$$\max\ \eta_0 \quad \text{s.t.}\ \ c(x,u) + \mathcal{K}h_0\,(x,u) \geq h_0(x) + \eta_0\,, \qquad \text{for all } x,u\,. \quad (5.13)$$

This is an infinite dimensional LP for a general state space, where the variables are $\eta_0 \in \mathbb{R}$ and $h_0 \colon \mathsf{X} \to \mathbb{R}$. In [113], a parameterized family $\{h^\theta\}$ is used for ADP based on the following LP formulation:

$$\max\ \eta_0 \quad \text{s.t. } c(x,u) + \mathcal{K}h_0^\theta\,(x,u) \geq h_0^\theta(x) + \eta_0\,, \qquad \text{for all } x,u\,.$$

If the parameterization is linear $h_0^\theta = \theta^{\scriptscriptstyle T}\varphi$, then this is an LP in the variables (η_0,θ). In the present work, a Q-function representation is sought.

Proposition 12. *The LP formulation of the ACOE in* (5.13) *is equivalent to the following convex program in* (η_0, H_0)*:*

$$\max\ \eta_0 \quad \textit{s.t.}\ \ c(x,u) + \mathcal{K}\underline{H}_0\,(x,u) \geq H_0(x,u) + \eta_0\,, \qquad \textit{for all } x,u\,, \quad (5.14)$$

where $H_0 \colon \mathsf{X}\times\mathsf{U} \to \mathbb{R}$ *and* $\underline{H}_0$ *is defined as in* (5.6)*:*

$$\underline{H}_0(x) = \min_{u\in\mathsf{U}(x)} H_0(x,u)$$

Proof. Observe that the LP in (5.13) can be recast as an LP in (η_0, h_0, H_0):

$$\begin{aligned}\max\ \eta_0 \quad \text{s.t.}\ \ & c(x,u) + \mathcal{K}h_0\,(x,u) \geq H_0(x,u) + \eta_0 \\ & \textit{and}\ \ H_0(x,u) \geq h_0(x)\,, \qquad\qquad \text{for all } x, u\,.\end{aligned}$$

Defining $\underline{H}_0$ and substituting it in place of h_0 in the first constraint allows us to relax the second constraint. This way, the LP in (η_0, h_0, H_0) can be reformulated as convex program (5.14) in variables (η_0, H_0). □

The convex optimization problem (5.14) provides the foundation to construct a convex Q-learning algorithm. A parameterization of the form H^θ defined in (5.3) is considered as candidate approximation to H_0^*, the solution to the LP (5.14). The corresponding approximate LP is

$$\max\ \hat{\eta} \quad \text{s.t.}\ \ c(x,u) + \mathcal{K}\underline{H}^\theta\,(x,u) \geq H^\theta(x,u) + \hat{\eta}\,, \qquad \text{for all } x, u\,,$$

for the variables $\hat{\eta} \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$. A relaxation of the approximate LP is used to construct an error criterion for the convex Q-learning algorithm:

$$\min_{\hat{\eta},\theta} \Big\{ -\hat{\eta} + \tfrac{1}{2}\kappa \sum_{(x,u)} \varpi(x,u)\big[H^\theta(x,u) + \hat{\eta} - \big(c(x,u) + \mathcal{K}\underline{H}^\theta\,(x,u)\big)\big]_+^2 \Big\}\,,$$

where ϖ is the invariant distribution on $\mathsf{X} \times \mathsf{U}$, $[\cdot]_+$ denotes a non-negative projection of the argument and κ is a penalty parameter. For any $\kappa > 0$, this is a convex function of θ since $\underline{H}^\theta$ is concave in θ. Then, the following error criterion is adopted for the convex Q-learning algorithm:

$$\begin{aligned}\mathcal{E}^{\mathsf{alp}}(\theta,\hat{\eta}) \ &:=\ -\hat{\eta} + \tfrac{1}{2}\kappa\,\mathsf{E}\Big\{\big[H^\theta + \hat{\eta} - (c + \mathcal{K}\underline{H}_0\,)\big]_+^2\Big\}\,, \qquad (5.15)\\ &= -\hat{\eta} + \tfrac{1}{2}\kappa\,\mathsf{E}\Big\{\big[H^\theta\,(X(t),U(t)) + \hat{\eta} - \big(c\,(X(t),U(t)) + \underline{H}^\theta\,(X(t+1))\big)\big]_+^2\Big\}\,.\end{aligned}$$

A steepest descent or Newton-Raphson algorithm can be devised to compute $(\theta^*, \hat{\eta}^*)$ that minimizes $\mathcal{E}^{\mathsf{alp}}$.

Here, a stochastic approximation of the steepest descent algorithm is used to recursively estimate $(\theta^*, \hat{\eta}^*)$ that minimize $\mathcal{E}^{\mathsf{alp}}$ for a given value of κ. The

gradients of interest are

$$\begin{aligned}\nabla_\theta \mathcal{E}^{\mathsf{alp}} &= \kappa\, \mathsf{E}\left\{ \left[H^\theta + \hat{\eta} - (c + \mathcal{K}\underline{H}^\theta)\right]_+ \cdot (\psi - \mathcal{K}\underline{\psi}^\theta) \right\} \\ &= \kappa \mathsf{E}\left[\Delta_\theta\left(X, U; \theta, \hat{\eta}\right)\right] , \qquad (5.16) \\ \nabla_{\hat{\eta}} \mathcal{E}^{\mathsf{alp}} &= -1 + \kappa\, \mathsf{E}\left\{ \left[H^\theta + \hat{\eta} - (c + \mathcal{K}\underline{H}^\theta)\right]_+ \right\} \\ &= -1 + \kappa \mathsf{E}\left[\Delta_{\hat{\eta}}\left(X, U; \theta, \hat{\eta}\right)\right] , \qquad (5.17)\end{aligned}$$

where

$$\begin{aligned}\Delta_\theta(X, U; \theta, \hat{\eta}) = &\Big[H^\theta\left(X(t), U(t)\right) + \hat{\eta} - c\left(X(t), U(t)\right) - \underline{H}^\theta\big(X(t+1)\big)\Big]_+ \\ &\times \Big[\psi\big(X(t), U(t)\big) - \underline{\psi}^\theta\big(X(t+1)\big)\Big] , \\ \Delta_{\hat{\eta}}(X, U; \theta, \hat{\eta}) = &\Big[H^\theta\left(X(t), U(t)\right) + \hat{\eta} - c\left(X(t), U(t)\right) - \underline{H}^\theta\big(X(t+1)\big)\Big]_+ .\end{aligned}$$

Then, analogous to the stochastic approximation algorithm in (5.11), $(\theta^*, \hat{\eta}^*)$ can be estimated using the following update equations:

$$\theta(t+1) = \theta(t) - \gamma_1(t)\kappa \mathsf{E}\left[\Delta_\theta\left(x(t), u(t); \theta(t), \hat{\eta}(t)\right)\right] \qquad (5.18)$$

$$\hat{\eta}(t+1) = \hat{\eta}(t) - \gamma_2(t)\big(-1 + \kappa \mathsf{E}\left[\Delta_{\hat{\eta}}\left(x(t), u(t); \theta(t), \hat{\eta}(t)\right)\right]\big) . \qquad (5.19)$$

As before, the gain sequences $\{\gamma_1(t)\}$ and $\{\gamma_2(t)\}$ are chosen such that the resulting difference equations for θ and $\hat{\eta}$ satisfy standard conditions [115,116].

5.4 Q-learning for the PNNL Model

The practical applications of the Q-learning algorithms devised in this chapter are demonstrated for the purpose of coordinating the dispatch of resources in a representative power system. The numerical studies reported in this section are conducted on the PNNL test system.

5.4.1 Overview of the PNNL Model

The PNNL model used in our studies consists of a diesel generator, a BESS, a wind power plant and a group of residential loads. Real world wind power

data from [97] and simulated load demand from [111] are used in our numerical studies. The control objective is to find the least-cost dispatch strategies for the diesel generator and BESS so as to meet the net load demand on the system.

The state of the system $X(t)$ at time t is described by the output of the diesel generator, the threshold of BESS, the SOC and the balancing service required at time t while the control input $U(t)$ is described by the ramping in the generation output and change in the BESS threshold at time t. The system dynamics follow the recursion

$$X(t+1) = AX(t) + BU(t) + DV(t)\,, \tag{5.20}$$

and are subject to state-action constraints:

$$X^{\mathsf{min}} \leq \ X(t) \ \leq X^{\mathsf{max}} \quad \text{and} \quad \mathsf{U}^{\mathsf{min}} \leq \ U(t) \ \leq U^{\mathsf{max}}$$

for each t. The constraint parameters are specified in Table 4.1 (on page 81).

Optimality is based on a quadratic cost,

$$c\big(x,u\big) = (x - x^{\mathsf{ref}})^{T} Q (x - x^{\mathsf{ref}}) + u^{T} R u + \text{some constant}\,, \tag{5.21}$$

where x^{ref} is a reference state. The objective for the dispatch problem is to minimize this cost over a specified time horizon, subject to system dynamics (5.20) and state/input constraints (4.37). In the context of control synthesis, the time horizon is taken to be infinite and Q-learning techniques are used to find the control policies for this system.

5.4.2 Implementation of Q-learning

In the numerical experiments reported in the following sections, Q-learning algorithms devised in sections 5.2 and 5.3 were applied to two separate system models. First, Q-learning was conducted on a fluid model constructed from a mean-field approximation of the MDP model (5.20) to gain insights on the structure of the problem. The dynamics of the fluid model are assumed to

follow the recursion:

$$\bar{x}(t+1) = A\bar{x}(t) + B\bar{u}(t) + D\bar{v}\,, \tag{5.22}$$

where $\bar{v}$ is the mean of the disturbance process, and, $\bar{x}(t)$ and $\bar{u}(t)$ are the states and control actions of the mean-field model which are subject to state-action constraints:

$$X^{\mathsf{min}} \leq \ \bar{x}(t) \ \leq X^{\mathsf{max}} \quad \text{and} \quad \mathsf{U}^{\mathsf{min}} \leq \ \bar{u}(t) \ \leq U^{\mathsf{max}} \tag{5.23}$$

Q-learning was also performed on the MDP model, where the model dynamics were simulated using sample path trajectories of the disturbance process constructed from actual measured data.

Basis selection: Three functions are used to construct the basis ψ for Q-learning on both models. The basis functions are obtained using the approach (5.4). That is,

$$\psi_i(x,u) = \varphi_i\left(Ax + Bu + D\bar{v}\right) \qquad \text{for } i = 1,2,3\,;$$

where $\{\varphi_i : 1 \leq i \leq 3\}$ is the basis used for TD-learning. The functions φ_i are obtained by using fluid value functions and penalty functions: the fluid value functions are computed by solving the optimality equations for a simpler idealized system model (much like the approach taken in Section 3.4 and [82, 105]) while the penalty functions take into account the state/action constraints ignored in the fluid model approximations (similar to basis construction adopted in Section 3.3).

Specifically, in this example, one fluid value function and two penalty functions are used to construct the basis $\{\varphi_i\}$. An LQR-relaxation of the MDP model is considered: the dynamics take the form (5.22), costs are quadratic in nature as seen in (5.21) and state/action constraints in (5.23) are relaxed. The value function for this fluid model is computed using Proposition 4, as shown in Section 4.5. The LQR-like fluid value function $h_0^*(x)$ thus computed is used as the first basis function $\varphi_1(x)$. The other two basis functions take into account the capacity constraints on diesel generation and battery SOC.

That is, the second basis φ_2 is designed to penalize movement of the state trajectory towards the generation boundary. And the third basis φ_3 models the higher overall costs anticipated when the SOC approaches its capacity constraints. The Q-learning algorithm optimizes the weights associated with these functions to approximate the Q-function.

Excitation input: For Q-learning on both the fluid and MDP model, the randomized policy is constructed as described in (5.5) with the stabilizing policy $\bar{\phi}$ obtained by projecting the LQR state-feedback policy, solution of (4.13), onto the constrained action space. The excitation signal $\boldsymbol{\zeta}$ is obtained through a quasi-Monte-Carlo approach [116] with

$$\zeta(t) = \sum_{j=1}^{n} \gamma_j \sin(\omega_j t) \,,$$

where $\{\gamma_j\}$ are constants, $\{\omega_j\}$ are various frequencies and n is an integer. In the numerical results reported here, $n = 5$.

The choices of the basis and excitation signal are fine-tuned to ensure appropriate sampling of the state-action space and reduction in the Bellman error.

5.4.3 Numerical Experiments on the Fluid Model

In this section, the numerical results from applying the two Q-learning algorithms to the dispatch problem for the fluid model in (5.22) are discussed. The performance of the resulting control policies is also compared.

Bellman error-based Q-learning: Many numerical experiments were conducted for different choices of excitation $\boldsymbol{\zeta}$ and boundary penalty functions φ_2 and φ_3. The quality of each of the resulting approximations was judged based on mismatch in the fixed point equation (5.7) for the different sampled state-action pairs as well as the Bellman error $\mathcal{E}_0^{\mathsf{BE}}$, defined in (4.29), associated with the approximate value function.

The choice of parameters for the excitation signal impacts the quality of the approximation. The parameters $\{A_j\}$ and $\{\omega_j\}$ should be chosen in such a way that the state space is appropriately sampled. The plots in Figure 5.1

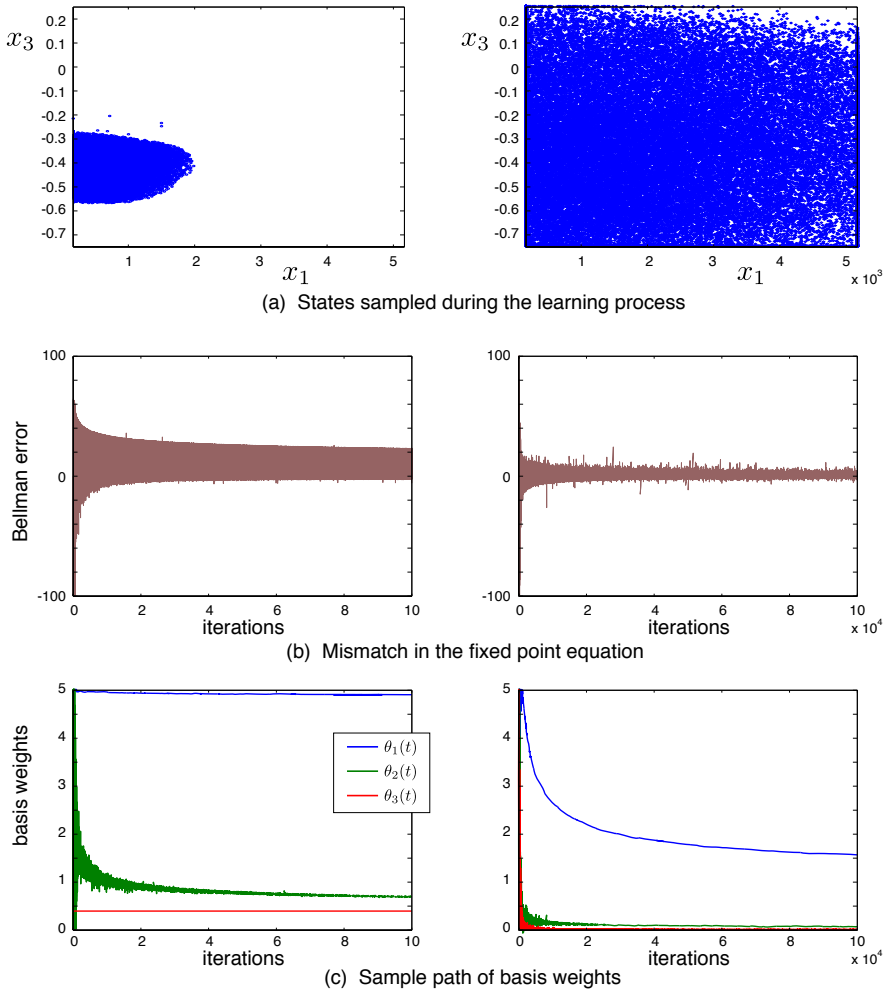


Figure 5.1: Comparing the results from Q-learning for a choice of low versus high amplitude excitation signal.

motivate this observation. The plots show the sampled state space, error in the fixed point equation and the sample paths of the basis weights $\{\theta_i\}$ for two different choices of $\boldsymbol{\zeta}$ – the plots corresponding to a low amplitude $\boldsymbol{\zeta}$ are shown on the left while those for high amplitude $\boldsymbol{\zeta}$ are on the right. Inadequate sampling of the state space – as evidenced from Figure 5.1(a) – results in a higher mismatch in the fixed point equation as shown in Figure 5.1(b). In fact, the insufficient sample of state space due to low exploration can hamper the learning of basis weights: observe the absence of tuning for parameter θ_3 in Figure 5.1(c).

The final choices for the exploration parameters $\{A_j\}$ and $\{\omega_j\}$ and basis functions $\{\varphi_i\}$ were found by trial-and-error such that the resulting approximations have negligible Bellman error.

Linear programming-based Q-learning: Similar to the first Q-learning algorithm, the LP-based Q-learning algorithm was applied to the fluid model under different choices of excitation $\boldsymbol{\zeta}$ and boundary penalty functions φ_2 and φ_3. As seen with the Bellman error-based Q-learning, the excitation parameters $\{A_j\}$ and $\{\omega_j\}$ have significant impact on the quality of the resulting approximations, which was quantified by the Bellman error and the mean-square mismatch in the fixed point equation.

In addition to choice of excitation parameters and basis functions, the quality of the approximation in LP-based Q-learning also depends on the penalty parameter κ. Figure 5.2 emphasizes the convergence of the basis weights $\{\theta_i\}$ for higher κ, as would be expected from a typical barrier penalty method in the non-linear programming. Convergence was observed for $\kappa \geq 100$. The quality of the approximation, as judged by the Bellman error, also improved with higher κ. Figure 5.3 emphasizes this point by demonstrating the decrease

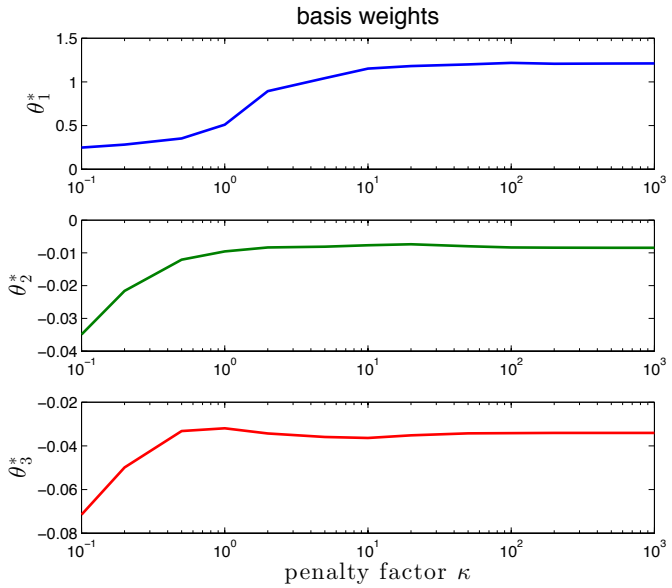


Figure 5.2: Convergence of basis weights θ_i's for large values of penalty factor κ.

in the average cost and the mean square Bellman error for approximations corresponding to higher values of κ.

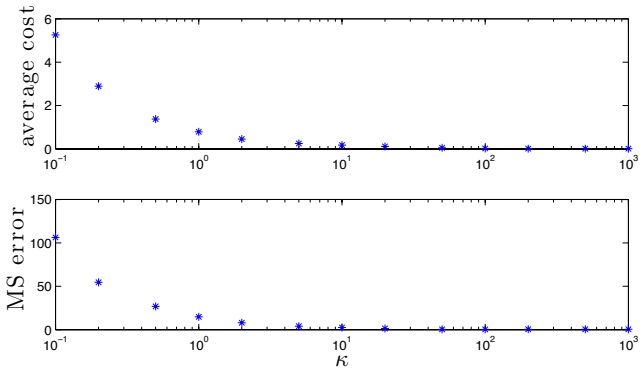


Figure 5.3: Impact of choice of κ on mean-square Bellman error and average cost.

One concern in using stochastic approximation algorithm defined in (5.18) and (5.19) is the high gain introduced in the algorithm for higher values of κ. This can lead to high variance and slow down the algorithm. In our experiments, introducing a scaling by $1/\kappa$ in the update equations improved convergence. Additionally, the averaging scheme proposed by Polyak and Juditsky in [117] was also used to accelerate convergence. A comparison of the resulting sample paths for $\{\theta_i\}$ and their average $\{\bar{\theta}_i\}$ is shown in Figure 5.4.

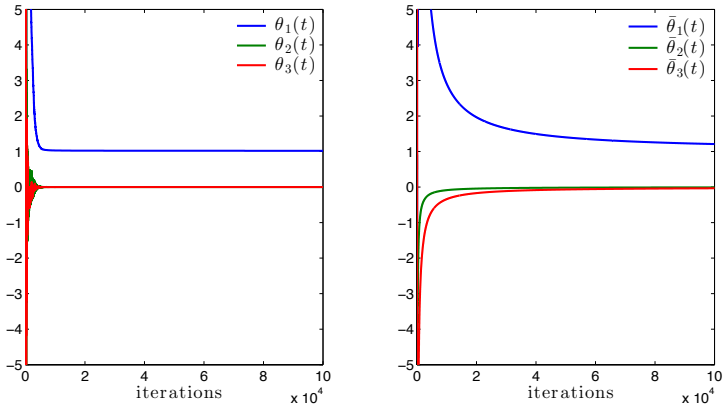


Figure 5.4: Sample path of θ_i's and the associated running averages $\bar{\theta}_i$'s for the Polyak averaging scheme.

Table 5.1: Performance Comparison for the Fluid Model

Control policy	η
LQR state feedback	0.041
Bellman error-based Q-learning	0.060
LP-based Q-learning ($\kappa = 1000$)	0.040

Comparing performance of controllers: The performances of the control policies obtained from the two Q-learning algorithms are compared against that of the control policy of the unconstrained fluid model, that is, the LQR state feedback policy. The running costs for different state-input trajectories corresponding to many different initial conditions are computed and averaged to estimate the average cost for each policy; Table 5.1 lists these average costs for the three policies. The control policy obtained from the LP-based Q-learning algorithm ($\kappa = 1000$) provides the least-cost performance.

The Bellman error defined in (4.29) is used to study the quality of approximation and stability of the controller. In Figure 5.5, the ratio $\dfrac{\mathcal{E}_0^{\mathsf{BE}}(x)}{c_\phi(x)}$ is plotted against the norm of the state for state trajectories corresponding to the same initial condition but controlled under the three different approximate control policies. The ratio is plotted for the observed state values based on the policy being applied to the system. As prescribed in Corollary 10, stability is guaranteed if $\dfrac{\mathcal{E}_0^{\mathsf{BE}}(x)}{c_\phi(x)} > -1$ for large values of $\|x\|$. From the plot, it can be seen that the all control policies satisfy this condition and are, hence, stabilizing. However, the LQR-policy does not result in a low Bellman error nor does the ratio of $\mathcal{E}_0^{\mathsf{BE}}(x)/c_\phi(x)$ under this policy provide the tight bounds necessary for near-optimal performance, as prescribed in Theorem 11.

5.4.4 Numerical Experiments on the MDP Model

The Q-learning algorithms devised in this chapter are also applied to the stochastic system model. That is, for the purposes of learning, the dynamics are assumed to follow the recursion:

$$x(t+1) = Ax(t) + Bu(t) + Dv(t)\,,$$

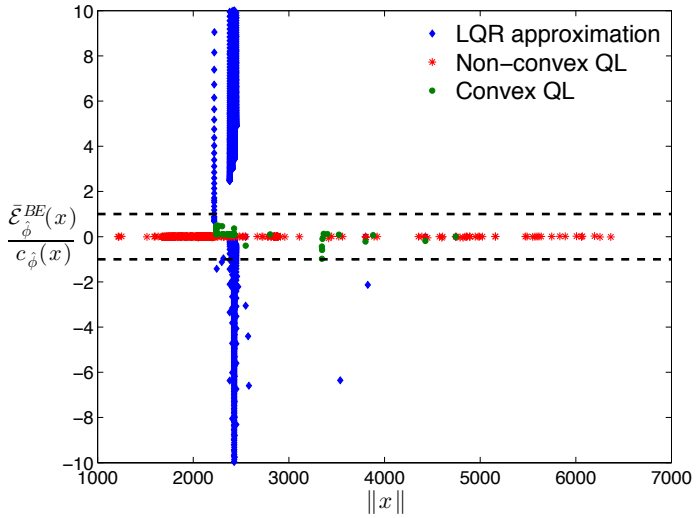


Figure 5.5: Bellman error ratio for three approximations applied to the mean-field model.

where $v(t)$ correspond to actual measurements of the disturbances obtained from real data. The convergence of the algorithms for this model is not guaranteed but many of our experiments did exhibit convergence. In Figure 5.6, the sample paths of the basis weights for one such experiment with Q-learning based on Bellman error reduction are shown. The Polyak averaging scheme [117] was found to be quite useful in these experiments.

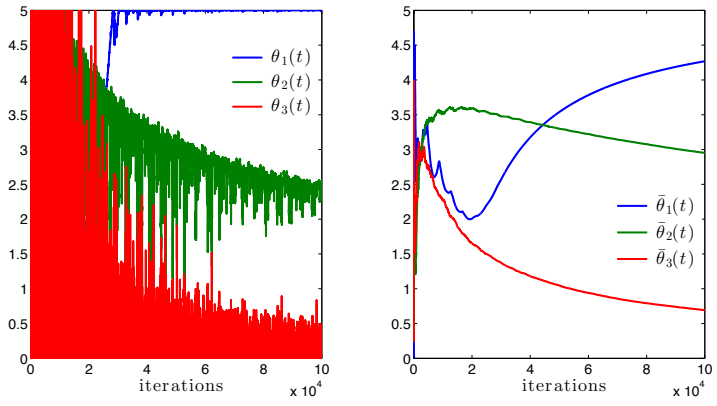


Figure 5.6: Sample path of θ_i's and the associated running averages $\bar{\theta}_i$'s for the Polyak averaging scheme.

The Bellman error defined in (4.10) provides a means to study the quality

of approximation, stability of the controller and its performance. As seen in the case of fluid model, the quality of approximation can be improved through appropriate choice of exploration signal parameters and basis functions. In Figure 5.7, the ratio $\frac{\mathcal{E}^{\mathsf{BE}}(x)}{c_\phi(x)}$ is plotted against $\|x\|$ for state trajectories corresponding to the same initial condition but controlled under the four different control policies:

- LQR state feedback policy from the unconstrained fluid model relaxation
- policy obtained from Bellman error-based Q-learning for the fluid model
- policy obtained from LP-based Q-learning for the fluid model
- policy obtained from Bellman error-based Q-learning for the stochastic model

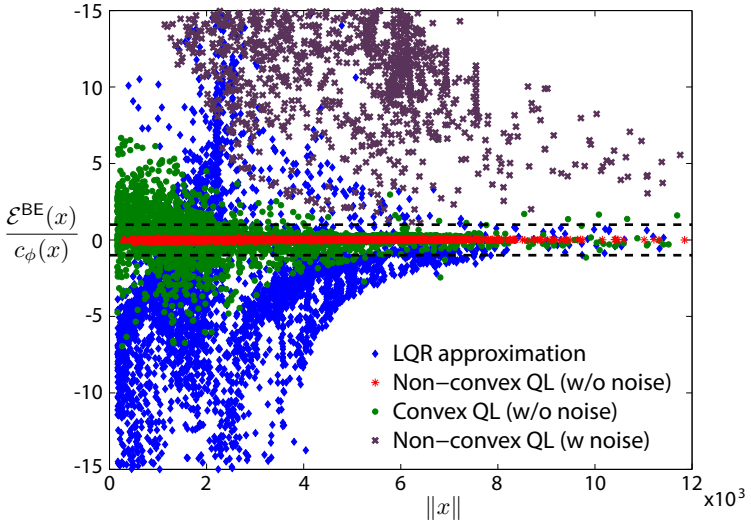


Figure 5.7: Bellman error ratio for the three approximations applied to the stochastic system.

From the plot in Figure 5.7, it can be seen that the ratio $\mathcal{E}_0^{\mathsf{BE}}(x)/c_\phi(x)$ for the different control policies satisfies the sufficient conditions of stability established in Theorem 9. That is, $\frac{\mathcal{E}_0^{\mathsf{BE}}(x)}{c_\phi(x)} > -1$ for large x and hence stability is guaranteed for all four policies considered here. However, the Bellman error is low only for policy corresponding to the Bellman error-based Q-learning on the fluid model. The average costs for different control policies are tabulated

Table 5.2: Performance Comparison for Stochastic System

Control policy	η
LQR state feedback (unconstrained fluid model)	0.0751
Bellman error-based Q-learning (fluid model)	0.1123
LP-based Q-learning ($\kappa = 1000$) (fluid model)	0.0720
Bellman error-based Q-learning (MDP model)	0.1801

in Table 5.2. The control policy obtained from the LP-based Q-learning ($\kappa = 1000$) on the fluid model provides the least-cost performance.

5.5 Concluding Remarks

In this chapter, two new Q-learning algorithms are introduced for the control of non-linear state space models under the average cost optimality criterion. A practical application of the algorithm to the dispatch problem on the PNNL model is illustrated. Implementation issues such as choice of basis and impacts of exploration on the quality of the approximation are discussed.

The performance of the control policies obtained from the two Q-learning algorithms is compared against the policy obtained from the LQR solution, as described in Section 4.5. Based on our experiments, it can be inferred that the control performance can be improved if the Q-function is approximated as a combination of the fluid value function and penalty functions that take into account constraints on the states and actions.

The Q-learning algorithms devised in this chapter can be integrated into a predictive control framework for improving performance of the predictive controller. The precise connections are established in the next chapter.

Chapter 6

Q-MPC for Control in Power Networks

The previous chapters apply RL algorithms for controlling resources in a power grid. However, the basis selection for RL applications in a network setting can be particularly challenging. A possible solution is to use RL techniques in conjunction with other controllers. In this chapter, the Q-learning algorithms devised in the previous chapter are used in a model predictive control (MPC) framework to control resources in a power network.

MPC is a popular approach due to its ability to handle complex constraints on states and inputs [118]. It has recently been recommended as a control mechanism to dispatch resources in power grids due to its flexibility in incorporating complex inter-temporal constraints on resources, along with short-term forecasts of renewable supply and electricity demand [76, 112]. However, MPC may not be effective without careful design. In particular, an inappropriate choice for the terminal cost function can result in poor performance or even an unstable closed-loop system. Although a large prediction horizon can offset these effects, it comes at the expense of correspondingly higher computational cost. It is well known that all these drawbacks are resolved if the infinite-horizon value function that solves the DP equations is chosen as the terminal cost function [118, 119]. An approximation of this value function can also improve performance for short prediction horizons [120, 121].

The Q-learning algorithms devised in Chapter 5 are used to approximate the infinite-horizon value function and, thus, define the terminal cost for MPC. The marriage of these two control approaches results in a new *Q-MPC* approach to control for both stochastic and deterministic systems. It admits a stabilizing

policy under mild conditions, regardless of the time horizon. This chapter presents the theoretical underpinnings of the Q-MPC approach.

The computational efficiency of Q-MPC is investigated via its application to the economic dispatch problem in power networks. The numerical results reported here conclusively demonstrate the computational efficacy of Q-MPC as compared to other typical MPC implementations in the sense that good performance is obtained even for small time-horizons.

6.1 Model Predictive Control

This section surveys aspects of MPC that will impact the design of the Q-MPC approach introduced later. Design aspects that guarantee stability, improve performance and reduce computational complexity by reducing the required time horizon T are discussed.

6.1.1 MPC for Deterministic Systems

Recall the fluid model introduced in Section 4.1.1:

$$x(t+1) = f\left(x(t), u(t)\right) \ , \tag{6.1}$$

with $\boldsymbol{x}$ and $\boldsymbol{u}$ representing state and input trajectories. The MPC algorithm applied to this model minimizes, at each time step τ, the finite-horizon cost

$$J_T(x) = \sum_{t=0}^{T-1} c\big(x(t), u(t)\big) + J_0\big(x(T)\big) \ , \tag{6.2}$$

where T is the prediction horizon, J_0 is the terminal cost and $x(0) = x = x(\tau)$; that is, x is the state measured at the current time step. The optimization is subject to system dynamics (6.1) and state/action constraints which are compactly represented as

$$x(t) \in \mathsf{X} \qquad \text{and} \qquad u(t) \in \mathsf{U}\left(x(t)\right) \qquad \text{for each } t = 0, 1, \ldots, T-1.$$

Suppose the resulting minimizing control sequence is $\{u^*(0), u^*(1), \ldots, u^*(T-1)\}$. Then, the first element $u^*(0)$ of this sequence is implemented at the current time step τ and the algorithm proceeds to next time step. This procedure defines a state-feedback control law $\bar{\phi}_T^*(x)$, which maps the state x to $u^*(0)$. It is stabilizing under general conditions on J_0, c and f; see [118] for a survey. Our concern lies with how the terminal cost J_0 impacts the stability of the controller.

Recall from the discussion on tc-stability in Section 4.3, that a policy $\bar{\phi}$ is stabilizing for the fluid dynamics (6.1) if there exists a function $V\colon \mathsf{X} \to \mathbb{R}_+$ for which the Poisson's inequality is satisfied:

$$\mathcal{K}_{\bar{\phi}} V\,(x) \le V(x) - c_{\bar{\phi}}(x)\,. \tag{6.3}$$

This fact is used to establish stability for a MPC feedback law. The following theorem provides sufficient conditions on the terminal penalty function J_0 to lead to stable MPD feedback law.

Theorem 13. *Suppose $J_0\colon \mathsf{X} \to \mathbb{R}_+$ satisfies the Poisson's inequality for some policy $\bar{\phi}_0$. Then, the MPC policy $\bar{\phi}_T^*$ is tc-stable for any $T \ge 1$.*

Proof. The central idea is to show that the MPC value function $\bar{\phi}_T^*$ satisfies (6.3) for the feedback policy $\bar{\phi}_T^*$. Since J_0 satisfies the Poisson's inequality for the policy $\bar{\phi}_0$,

$$\mathcal{K}_{\bar{\phi}_0} J_0\,(x) \le J_0(x) - c_{\bar{\phi}_0}(x)\,. \tag{6.4}$$

Recall that the DP equation for a control horizon of $T \ge 1$ is

$$J_T^*(x) = \min_u \left\{ c(x,u) + \mathcal{K} J_{T-1}^*\,(x,u) \right\}\,,$$

with $J_0^* = J_0$. Substituting the minimizing control action, which is also the MPC feedback policy $\bar{\phi}_T^*(x)$, gives

$$J_T^*(x) = c_{\bar{\phi}_T^*}\,(x) + \mathcal{K}_{\bar{\phi}_T^*} J_{T-1}^*\,(x)\,. \tag{6.5}$$

Then, rearranging the terms gives

$$\mathcal{K}_{\bar{\phi}_T^*} J_T^*\,(x) - J_T^*(x) + c_{\bar{\phi}_T^*}\,(x) = \mathcal{K}_{\bar{\phi}_T^*} J_T^*\,(x) - \mathcal{K}_{\bar{\phi}_T^*} J_{T-1}^*\,(x)\,.$$

To show that J_T^* is satisfies (6.3), it suffices to show that $J_T^* \leq J_{T-1}^*$. This can be proved using induction if $J_1^* \leq J_0^*$. Substituting $T = 1$ in (6.5),

$$J_1^*(x) = c_{\bar{\phi}_1^*}(x) + \mathcal{K}_{\bar{\phi}_1^*} J_0^*(x) \leq c_{\bar{\phi}_0}(x) + \mathcal{K}_{\bar{\phi}_0} J_0^*(x) \leq J_0^*(x)\,,$$

where the last inequality follows from (6.4) and the definition $J_0^* = J_0$. Thus, $J_T^* \leq J_{T-1}^*$ for any $T \geq 1$ so that

$$\mathcal{K}_{\bar{\phi}_T^*} J_T^*(x) - J_T^*(x) + c_{\bar{\phi}_T^*}(x) \leq 0\,,$$

which implies that J_T^* is satisfies condition (6.3) for the MPC policy $\bar{\phi}_T^*$. Hence, policy $\bar{\phi}_T^*$ is tc-stable. □

Thus, from Theorem 13 it can be inferred that if the terminal cost satisfies Poisson's inequality, then the MPC policy for each prediction horizon T is stabilizing. The next theorem concerns the choice of J_0.

Theorem 14. *Suppose J^* is the infinite-horizon value function given by*

$$J^*(x) = \min_{\boldsymbol{u}} \sum_{t=0}^{\infty} c\left(x(t), u(t)\right) \;,\; x(0) = x \in \mathsf{X}\,,$$

and assumed to be finite valued on X. If J^* is chosen as the terminal cost function J_0, then the resulting feedback law $\bar{\phi}_T^*$ is tc-stabilizing for the given dynamics. Furthermore, $\bar{\phi}_T^*$ is independent of T.

Proof. Recall that J^* satisfies the DP equation

$$J^*(x) = \min_{u \in \mathsf{U}(x)} \left\{ c(x,u) + \mathcal{K} J^*(x,u) \right\}\,,$$

and the minimizing control input defines the optimal policy $\bar{\phi}^*$. Clearly, J^* satisfies (6.3) for the minimizing policy $\bar{\phi}^*$. Then, stability follows from Theorem 13.

It follows from the principle of optimality that for any T,

$$J^*(x) = \min_{\boldsymbol{u}_0^{T-1}} \Big(\sum_{t=0}^{T-1} c(x(t), u(t)) + J^*(x(T)) \Big)\,,$$

and in this minimization, the optimizing $u^*(0)$ is independent of T. Consequently, the MPC policy $\bar{\phi}_T^*$ is independent of T if $J_0 = J^*$. □

Thus, the choice of J^* as terminal cost J_0 results in a stable control policy and leads to reduced computational complexity. Such desirable features also exist for the relative function h_0^* if average cost optimality for the deterministic model is being considered. In the MPC applications considered in this chapter, approximations of J^* or h_0^* are used to define J_0 in an attempt to ensure stability and simultaneously reduce the time horizon in MPC.

6.1.2 MPC for Stochastic Systems

This section concerns with MPC applied to a stochastic system. For simplicity, the MDP model introduced in Section 4.1.1 is adopted for analysis. That is, the system dynamics are assumed to follow the recursion

$$X(t+1) = f\left(X(t), U(t)\right) + W(t)\,, \tag{6.6}$$

where $\boldsymbol{X}$ is the state process, $\boldsymbol{U}$ the control process and $\boldsymbol{W}$ is an i.i.d. sequence with zero mean and finite covariance Σ_W taking values on $\mathsf{W} \subseteq \mathbb{R}^{\ell_x}$. The MPC algorithm applied to this system minimizes, at each time step τ, the expected cost over a finite horizon:

$$V_T(x) = \mathsf{E}\Big[\ \sum_{t=0}^{T-1} c\big(X(t), U(t)\big) + V_0\big(X(T)\big)\ \Big]\,. \tag{6.7}$$

Here, V_0 denotes the terminal cost and $X(0) = x$, state at time τ. As before, the minimization is subject to the system dynamics and the constraints on the states and actions. The solution admits a state feedback policy $\phi_T^*(x)$ similar to the deterministic case.

Similar to the deterministic case, the stability of the MPC for stochastic systems hinges on the choice of the terminal cost V_0. Again, Poisson's inequality is used to establish the stability for a stochastic MPC feedback policy. Recall that a policy ϕ is stabilizing if there exists a function $V : \mathsf{X} \to \mathbb{R}_+$ that satisfies

the Poisson's inequality

$$\mathcal{P}_\phi V\,(x) \le V(x) - c_\phi(x) + \varepsilon\,, \tag{6.8}$$

for some constant ε.

Theorem 15, the stochastic counterpart of Theorem 13, provides conditions on the terminal cost V_0 to ensure stability of the MPC feedback policy for the stochastic model of (6.6).

Theorem 15. *If there exists a policy ϕ_0 and a finite constant $\bar{\varepsilon}$ such that Poisson's inequality holds for the terminal cost V_0:*

$$\mathcal{P}_{\phi_0} V_0\,(x) \le V_0(x) - c_{\phi_0}(x) + \bar{\varepsilon}\,, \tag{6.9}$$

then the MPC policy ϕ_T^* is ac-stable for any $T \ge 1$.

The result is proved by establishing that the MPC value function V_{T-1}^* satisfies Poisson's inequality for the MPC policy ϕ_T^* for each $T \ge 1$. The proof is based on manipulating the recursion

$$\begin{aligned} V_{T+1}^*(x) &= \min_u \left\{ c(x,u) + \mathcal{P} V_T^*\,(x,u) \right\} \\ &= c_{\phi_{T+1}^*}(x) + \mathcal{P}_{\phi_{T+1}^*} V_T^*\,(x)\,, \end{aligned} \tag{6.10}$$

which holds for $T \ge 1$ with $V_0^* = V_0$. The proof uses arguments similar to those employed by Chen and Meyn in [119]: progressively tighter bounds are obtained for the differences in successive value functions,

$$\varepsilon_T(x) := V_{T+1}^*(x) - V_T^*(x)\,,$$

and these bounds are used to establish the desired result. The following lemma establishes the bounds on ε_T.

Lemma 16. *For each $T \ge 0$,*

(i) $\varepsilon_{T+1}(x) \le \mathcal{P}_{\phi_{T+1}^*} \varepsilon_T\,(x)$ *for all* $x \in \mathsf{X}$.

(ii) Let $\bar{\varepsilon}_T = \sup_{x \in \mathsf{X}} = \varepsilon_T(x)$. *Then,* $\bar{\varepsilon}_{T+1} \le \bar{\varepsilon}_T \le \bar{\varepsilon}$.

Proof. To prove (i), the definition of $\varepsilon_{T+1}(x)$ is invoked:

$$\begin{aligned}\varepsilon_{T+1}(x) &= V^*_{T+2}(x) - V^*_{T+1}(x)\\ &= c_{\phi^*_{T+2}}(x) + \mathcal{P}_{\phi^*_{T+2}} V^*_{T+1}(x) - V^*_{T+1}(x)\\ &\le c_{\phi^*_{T+1}}(x) + \mathcal{P}_{\phi^*_{T+1}} V^*_{T+1}(x) - V^*_{T+1}(x)\\ &\le c_{\phi^*_{T+1}}(x) + \mathcal{P}_{\phi^*_{T+1}} \left(V^*_T + \varepsilon_T\right)(x) - V^*_{T+1}(x)\,.\end{aligned}$$

Using (6.10) gives $\varepsilon_{T+1}(x) \le \mathcal{P}_{\phi^*_{T+1}} \varepsilon_T(x)$. Result (ii) follows from (i):

$$\bar{\varepsilon}_{T+1} = \sup_{x\in\mathsf{X}} \varepsilon_{T+1}(x) \le \sup_{x\in\mathsf{X}} \mathcal{P}_{\phi^*_{T+1}} \varepsilon_T(x)\,.$$

Clearly, $\bar{\varepsilon}_T$ is an upper bound for the RHS, which implies $\bar{\varepsilon}_{T+1} \le \bar{\varepsilon}_T$. Finally, using $T = 0$ in (6.10) gives

$$\begin{aligned}V^*_1(x) &= c_{\phi^*_1}(x) + \mathcal{P}_{\phi^*_1} V^*_0(x)\\ &\le c_{\phi_0}(x) + \mathcal{P}_{\phi_0} V^*_0(x)\\ &\le V_0(x) + \bar{\varepsilon}\,,\end{aligned}$$

where the last inequality follows from (6.9). By invoking definition, $\varepsilon_0(x) \le \bar{\varepsilon}$ for all x and result (ii) is proved. ☐

The bounds in Lemma 16 are used in proving Theorem 15.

Proof of Theorem 15: Subtracting $V^*_T(x)$ from both sides of (6.10) and invoking the definition of $\varepsilon_T(x)$ and $\bar{\varepsilon}_T$ gives

$$\begin{aligned}\mathcal{P}_{\phi^*_{T+1}} V^*_T(x) &= V^*_T(x) - c_{\phi^*_{T+1}}(x) + \varepsilon_T(x)\\ &\le V^*_T(x) - c_{\phi^*_{T+1}}(x) + \bar{\varepsilon}_T\,.\end{aligned}$$

Since Lemma 16 establishes that $\bar{\varepsilon}_T$ is finite for each $T \ge 0$, MPC policy ϕ^*_{T+1} satisfies Poisson's inequality for $V = V^*_T$ for each $T \ge 0$. The desired result follows. ☐

Under certain conditions, a function that satisfies Poisson's inequality (6.3) for the fluid model may satisfy its counterpart (6.8) for the stochastic model

and, hence, provide a suitable terminal cost for the stochastic MPC problem. The conditions are outlined in the theorem below.

Theorem 17. *Consider an MDP model satisfying assumption 4.1. Suppose J satisfies* (6.3) *for some policy $\bar{\phi}$. Furthermore, suppose J has bounded second derivative. Then, J satisfies Poisson's inequality* (6.8). □

The theorem is proved by considering a Taylor series approximation for the stochastic DP operator $\mathcal{P}$ and using mean value theorem to bound this approximation by the deterministic DP operator $\mathcal{K}$.

Proof. To establish that J satisfies (6.8), a Taylor series approximation of stochastic DP operator $\mathcal{P}$ is considered. Recall (4.24) under assumption 4.1 with J plugged in as g so that for some $(x, u) \in \mathsf{X} \times \mathsf{U}$,

$$\mathcal{P}J\,(x,u) = J\,(f(x,u))\,\tfrac{1}{2}\operatorname{tr}\left[\nabla^2 J\,(f(x,u))\cdot\Sigma_W\right] + \dots\ .$$

Using the mean value theorem to bound the approximation gives

$$\mathcal{P}J\,(x,u) \le J\,(f(x,u)) + \tfrac{1}{2}\operatorname{tr}\left[\nabla^2 J\,(f(\bar{x},\bar{u}))\cdot\Sigma_W\right], \tag{6.11}$$

where $(\bar{x},\bar{u}) = \arg\max \operatorname{tr}\left[\nabla^2 J\,(f(x,u))\cdot\Sigma_W\right]$. Since J satisfies (6.3) for a policy $\bar{\phi}$,

$$\mathcal{K}_{\bar{\phi}}J\,(x) \le J\,(x) - c_{\bar{\phi}}\,(x)\ .$$

Then, modifying (6.11) for the specific policy $\bar{\phi}$ gives

$$\mathcal{P}_{\bar{\phi}}J\,(x) \le J\,(x) - c_{\bar{\phi}}\,(x) + \tfrac{1}{2}\operatorname{tr}\left[\nabla^2 J\,(f(\bar{x},\bar{u}))\cdot\Sigma_W\right],$$

which is precisely the condition (6.8) due to boundedness of $\nabla^2 J$ and Σ_W. □

Theorems 15 and 17 indicate that a function satisfying the Poisson inequality for the fluid model is indeed a good candidate for the terminal cost V_0 under some general conditions. Indeed, J^* defined in (4.6) is one such candidate function, provided the derivative bound is met.

In summary, the infinite-horizon fluid value function J^* may be a good candidate for both J_0 and V_0 from a stability and computational efficiency

perspective. Although computation of J^* is intractable in most cases, an approximation of J^* may result in a stabilizing and computationally efficient controller (see [119] for theory and [120, 121] for examples). Learning techniques described in Chapter 5 may be used for such value function approximations.

6.2 Q-MPC Algorithm

Q-learning gives an approximation $H_0^{\theta^*}(x, u)$ to the Q-function $H_0^*(x, u)$ and, consequently, $\underline{H}^{\theta^*}(x)$ which approximates the infinite-horizon value function – the optimal terminal cost function for MPC. This links the two solution techniques, giving birth to the Q-MPC approach. In the deterministic case, Q-MPC uses the following modified objective:

$$J_T^*(x) = \min_{\boldsymbol{u}_0^{T-1}} \sum_{t=0}^{T-1} c\big(x(t), u(t)\big) + \underline{H}^{\theta^*}\big(x(t)\big) \,. \tag{6.12}$$

An analogous expression for the stochastic case can be derived.

In this way, Q-learning is integrated into the MPC framework to enhance the performance of the control algorithm and guarantee stability even for small prediction horizons.

Stability is guaranteed by design. For example, by construction, it can be assumed that $\underline{H}^{\theta}$ satisfies (6.3) for each $\theta \in \mathbb{R}_+^d$ satisfying $\min_i \theta_i \geq 1$. The Q-learning algorithm can then be modified to include a projection of θ onto this domain.

Furthermore, the Q-learning algorithm is amenable to online tuning. The approximate, optimal terminal cost can be periodically updated to adapt to any changes in the system environment, since such changes are typically on time scales slower than that of the actual dynamics of the system.

6.3 MPC for the PNNL Model

The Q-MPC approach was applied to the economic dispatch problem for the PNNL model. Simulation results are described here, providing a comparison of the Q-MPC approach against three other MPC implementations.

6.3.1 Overview of the PNNL Model

Recall that the PNNL system consists of a diesel generator and a BESS with an expensive ancillary service support to manage supply-demand mismatch. These resources are deployed to meet the net residential load demand. The objective for the economic dispatch problem is to find the least-cost control strategies for the diesel generator and BESS to meet the net load demand on the system.

Recall that the states for the PNNL model are the diesel generator output, the threshold of BESS, the SOC and the balancing service deployed at time t. The control actions at that time are the ramping in the generation output and change in the BESS threshold. The dynamics for the PNNL system are in a linear form

$$X(t+1) = AX(t) + BU(t) + DV(t)\,, \tag{6.13}$$

and are subject to state-action constraints:

$$X^{\mathsf{min}} \leq X(t) \;\; \leq X^{\mathsf{max}} \quad \text{and} \quad \mathsf{U}^{\mathsf{min}} \leq U(t) \;\; \leq U^{\mathsf{max}}$$

for each t. The constraint parameters are specified in Table 4.1 (on page 81).

Optimality is based on a quadratic cost,

$$c\big(x,u\big) = (x - x^{\mathsf{ref}})^{T} Q (x - x^{\mathsf{ref}}) + u^{T} R u + \text{some constant}\,, \tag{6.14}$$

where x^{ref} is a reference state. The objective for the dispatch problem is to minimize this cost over a specified time horizon, subject to system dynamics (5.20) and state/input constraints (4.37). For Q-learning, the time horizon is taken to be infinite, but for MPC, time horizon is treated as a simulation parameter to investigate computational complexity of the Q-MPC approach.

6.3.2 MPC Implementation

A predictive model for the system dynamics is defined as follows:

$$\hat{x}(t+1) = A\hat{x}(t) + B\hat{u}(t) + D\hat{v}(t)\,. \tag{6.15}$$

At each step, the actual values of the state x and the disturbances v measured from the system are used to initialize these dynamics at $\hat{x}(0)$ and $\hat{v}(0)$. Predictions for the disturbances for the horizon, $\hat{\boldsymbol{v}}_1^{T-1} := \{\hat{v}(1), \ldots, \hat{v}(T-1)\}$, are obtained using autoregressive integrated moving average (ARIMA) models for wind generation and aggregate load.

Given predictions $\hat{\boldsymbol{v}}_1^{T-1}$, the deterministic MPC algorithm described in Section 6.1.1 is employed to compute the feedback law $\bar{\phi}_T^*$ subject to the dynamics in (6.15) and constraints in (4.37). Then, the control action $\bar{\phi}_T^*(x)$ is applied to the system at current time step. In the numerical experiments, the system evolves according to dynamics in (4.36), so that $X(t+1)$ is defined and the procedure is repeated to obtain $U(t+1)$. More details on the MPC set-up and the forecasting techniques used are available in [112].

6.3.3 Numerical Results

In the numerical results reported here, three implementations of MPC are considered:

- benchmark MPC with $J_0(x) = 0$,
- LQR-based MPC with $J_0(x) = h_0^*(x)$, and,
- Q-MPC with $J_0(x) = \underline{H}_0^{\theta^*}(x)$.

The first choice of terminal cost is used to motivate the importance of J_0 by inducing its absence: this is chosen as the benchmark for comparison. The second choice of J_0 is motivated by the fluid model approximation described in Section 4.5 and is referred to as the LQR-MPC approach. Finally, the last choice of J_0 used the value function approximation obtained from the Bellman error-based Q-learning and is the promised Q-MPC approach.

The numerical studies use control steps of 10 minutes over a simulation period of 24 hours so that $T^{\mathsf{sim}} = 144$. The performance of each algorithm is compared in terms of total cost

$$J_{\mathsf{tot}}^* = \sum_{t=1}^{T^{\mathsf{sim}}} c\big(x^*(t), u^*(t)\big) \,,$$

where $x^*(t)$ and $u^*(t)$ are based on the MPC state-feedback policy.

The three MPC algorithms are applied to dispatch resources in two modifications of the PNNL model, with the associated capacity and ramping constraints defined in Table 6.1. Observe that system B is more constrained than system A, with twice the wind generation capacity.

Table 6.1: Test system description

System	$P_{\mathsf{G}}^{\max}$	$\Delta P_{\mathsf{G}}^{\max}$	$E_{\mathsf{S}}^{\max}$
A	5 MW	n/a	3.6 MWh
B	3 MW	1 MW	3.6 MWh

Figure 6.1 illustrates some of results obtained from numerical studies, all based on data obtained in a typical summer day. In all cases, the Q-MPC algorithm outperforms the other three MPC algorithms, irrespective of the choice of T. In fact, for very low values of $T < 4$, the cost of the Q-MPC algorithm is nearly $\frac{1}{6}$ that of the benchmark MPC implementation for system A. In general, the benchmark MPC results in very high costs for $T < 5$, demonstrating the need to choose J_0 appropriately.

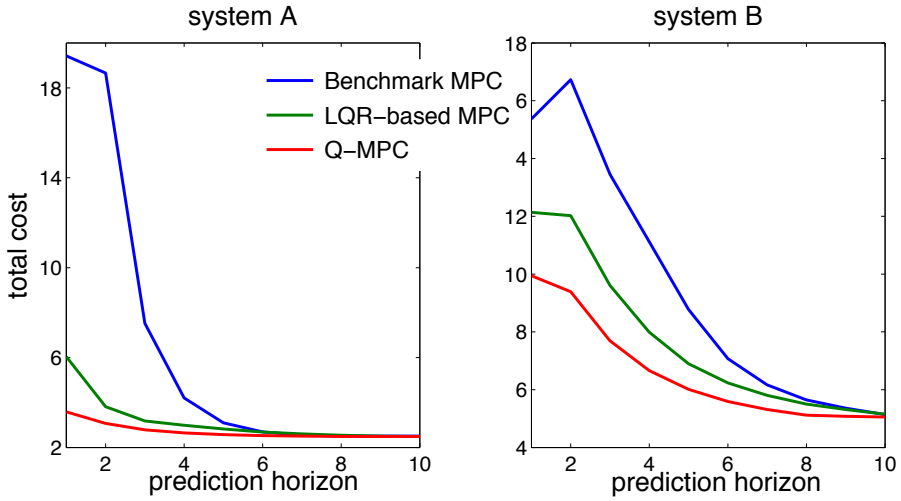


Figure 6.1: Total costs as a function of the prediction horizon for each of the test cases.

A comparison of the performances of the three MPC-based dispatch algorithms across different days allows us to test the effectiveness of Q-MPC for

different load and wind generation patterns, which can vary significantly depending on weather and other external factors. Qualitatively similar results are obtained after simulating dispatch for different days: the Q-MPC is a clear winner for small prediction horizons, particularly for $T \leq 4$.

6.4 Control in Network Settings

The Q-MPC algorithm has practical impact on how the economic dispatch is usually implemented in a power network. This connection is established here and Q-MPC is applied to dispatch resources in constrained power networks. Numerical experiments based on two test systems are reported here. In both systems, the network impacts are modeled by DC power flow model [1].

6.4.1 Economic Dispatch Problem Formulation

This section casts the usual economic dispatch problem solved by systems operators in the MPC framework. For the purposes of illustration, a case of controllable generation and uncertain demand is considered. That is, the generators are the only controllable resources available to the operator to meet the exogenous demand.

Recall the power node model of Section 3.2: $P_{\mathsf{G}n}(t)$ denotes generation at node n while $P_{\mathsf{D}n}(t)$ is the demand at that node. The usual economic dispatch problem is formulated as a cost minimization problem over a finite horizon:

$$\min_{\{P_{\mathsf{G}n}(t)\}} \quad \sum_{t=\tau}^{\tau+T} \sum_{n\in\mathcal{N}} C_{\mathsf{G}n}\left(P_{\mathsf{G}n}(t)\right)$$

$$\text{s.t.} \quad \left.\begin{array}{c} \sum_n P_{\mathsf{G}n}(t) = \sum_n \hat{P}_{\mathsf{D}n}(t) \\ P_{\mathsf{G}n}^{\mathsf{min}} \leq P_{\mathsf{G}n}(t) \leq P_{\mathsf{G}n}^{\mathsf{max}} \\ \Delta P_{\mathsf{G}n}^{\mathsf{min}} \leq P_{\mathsf{G}n}(t+1) - P_{\mathsf{G}n}(t) \leq \Delta P_{\mathsf{G}n}^{\mathsf{max}} \\ F^{\mathsf{min}} \leq H\left(P_{\mathsf{G}}(t) - P_{\mathsf{D}}(t)\right) \leq F^{\mathsf{max}} \end{array}\right\} \begin{array}{c} \text{for } t=\tau, \\ \tau+1,\ldots, \\ \tau+T \end{array} \tag{6.16}$$

where $\hat{P}_{\mathsf{D}n}(t)$ is the predicted demand at time t. The first constraint corresponds to the supply-demand balance constraint while the second and third constraints represent the capacity and ramping limits on the generators. The last constraint models the power flow limitations imposed by thermal limits on transmission lines: DC power flow models are used for this purpose, with H representing the injection shift factor matrix which correlates nodal injections to line flows [1] and $\left[F^{\mathsf{min}}, F^{\mathsf{max}}\right]$ represent the line flow limits.

To view the economic dispatch problem in (6.16) as a control problem, a choice of states and actions needs to be specified. Suppose the net nodal injections are chosen to define the state $X(t)$ of the system and the ramping of the generators constitutes the action $U(t)$ for the same. Then, in the MPC framework, the usual economic dispatch problem of (6.16) is equivalent to solving an MPC cost minimization problem with prediction horizon $T+1$ and terminal cost $J_0 = 0$. That is, the usual economic dispatch problem is the benchmark case shown to be computationally expensive for the PNNL model in Section 6.3. Since Q-MPC was effective for the PNNL system, it is applied to solve the economic dispatch problem for two other test systems in the following sections.

6.4.2 Three-bus System/Texas Model

The three-bus model considered in present work is the simplest loop network that can be studied. As shown in Figure 6.2, it consists of two thermal generators at nodes 1 and 2 along with a group of residential loads and a wind power plant at node 3. The residential load data is generated in [111] while the wind plant data is obtained from [97].

Problem formulation: The control objective is to find the least-cost dispatch of the two generators such that generation and transmission constraints are met with supply equal to demand. Adopting notation of the power node model introduced in Section 3.2, $P_{\mathsf{G}i}(t)$ and $\Delta P_{\mathsf{G}i}(t)$ are used to denote the output and ramping of the generator at node i at time t for $i = 1, 2$. And $P_{\mathsf{D}3}(t)$ is used to denote the net load at node 3.

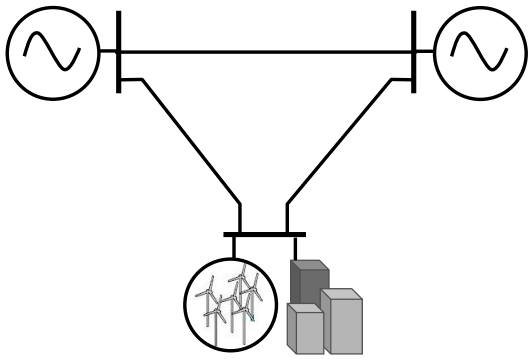


Figure 6.2: The three-bus/Texas model.

The dynamics of the system are described as follows:

$$\begin{aligned} P_{\mathsf{G1}}(t+1) &= P_{\mathsf{G1}}(t) + \Delta P_{\mathsf{G1}}(t)\,, \\ P_{\mathsf{G2}}(t+1) &= P_{\mathsf{G2}}(t) + \Delta P_{\mathsf{G2}}(t)\,, \\ P_{\mathsf{D3}}(t) &= D_3(t) - G_3(t)\,. \end{aligned} \tag{6.17}$$

where the $D_3(t)$ and $G_3(t)$ represent the externally driven disturbances processes dependent on the residential demand and wind power generation, respectively.

The net power injection at each node at time t is used to describe the state of the system at that time, so that

$$X(t) = [P_{\mathsf{G1}}(t), P_{\mathsf{G2}}(t), -P_{\mathsf{D3}}(t)]^T\,.$$

The control actions are the ramping of the generators' outputs,

$$U(t) = [\Delta P_{\mathsf{G1}}(t), \Delta P_{\mathsf{G2}}(t)]^T\,.$$

The dynamics of the system can be described by the linear MDP model:

$$X(t+1) = AX(t) + BU(t) + W(t)\,, \tag{6.18}$$

where the disturbance process takes the form

$$W(t) = [0, 0, -D_3(t) + G_3(t)]^T\,.$$

In addition to the usual constraints on states and actions imposed by the capacity and ramping limits on generation, the states and actions are also subject to network constraints. The constraints are represented as follows:

$$\begin{aligned} X^{\min} \le X(t) \le X^{\max} \\ U^{\min} \le U(t) \le U^{\max} \\ 1^T X(t) = 0 \\ HX(t) \le F^{\max}, \end{aligned} \tag{6.19}$$

where the equality in the third constraint is a consequence of the supply-demand balance while the last constraint represents the transmission constraints. The matrix H is the injection shift factor matrix whose element in the ℓ^{th} row and i^{th} represents the fraction of the power injection at node i that is diverted via line ℓ. The vector $F^{\max}$ represents line capacity limits.

A quadratic cost structure is adopted for generation and ramping costs are imposed. Then,

$$\begin{aligned} c(x,u) &= \left(a_1 p_{\mathsf{G}1}^2 + b_1 p_{\mathsf{G}1} + c_1\right) + \left(a_2 p_{\mathsf{G}2}^2 + b_2 p_{\mathsf{G}2} + c_2\right) + \gamma_1 \Delta p_{\mathsf{G}1}^2 + \gamma_2 \Delta p_{\mathsf{G}2}^2 \\ &= (x - x^{\mathsf{ref}})^T Q (x - x^{\mathsf{ref}}) + u^T R u + \text{ some constant}, \end{aligned} \tag{6.20}$$

where x^{ref} is a reference state.

In this way, the MDP model is defined using the dynamics in (6.18), the constraints in (6.19) and the costs in (6.20).

Q-learning for the* 3*-bus system: The Q-learning architecture described in Section 5.4.2 is applied to approximate the value function for the MDP defined above. Again, a fluid model is used for learning. The basis functions are also adopted in a manner analogous to that employed in Section 5.4.2, with one addition: more basis functions are introduced to consider the impacts of transmission constraints. Two basis architectures are considered. In the first basis, a separate penalty function for each transmission constraint is considered. In the second basis, an aggregate penalty function for all transmission constraints is considered; the parameters for the aggregate penalty are tuned to reduce the Bellman error in the approximation.

MPC for the* 3*-bus system: The MPC architecture described in Sec-

tion 6.3.2 is also applied to the economic dispatch problem for the 3-bus system. Four implementations of MPC are considered:

- Benchmark MPC with $J_0(x) = c(x, 0)$,
- LQR-MPC with $J_0(x) = h_0^*(x)$,
- Q-MPC with $J_0(x) = \underline{H}_0^{\theta^*}(x)$ derived using the extended basis, and,
- Q-MPC with $J_0(x) = \underline{H}_0^{\theta^*}(x)$ derived using the aggregate basis.

Control steps of 10 minutes are used in the simulations. As before, the total dispatch costs are computed for different values of the prediction horizon T for typical summer day. The plot in Figure 6.3 emphasizes how the Q-MPC approaches provide close-to-optimal solutions for low values of T. Furthermore, the performance of the two Q-MPC control policies does not change much if an extended basis is used instead of an aggregate basis.

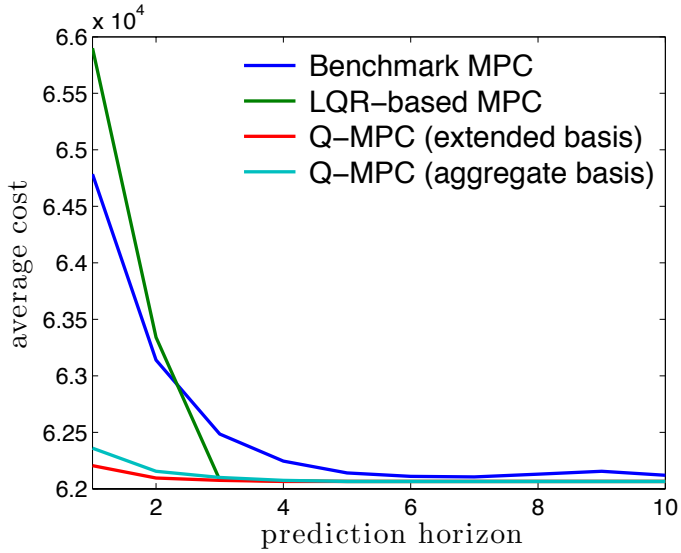


Figure 6.3: Total costs for different prediction horizons for dispatch of the 3-bus system.

6.4.3 Twelve-bus System/NYISO Model

The Q-MPC approach is also applied to the 12-bus system of [76] shown in Figure 6.4, with two minor modifications:

- the nodes with the renewable and solar power plants in the system are assumed to possess energy storage, and,

- generators incur ramping costs which need to be considered in the dispatch.

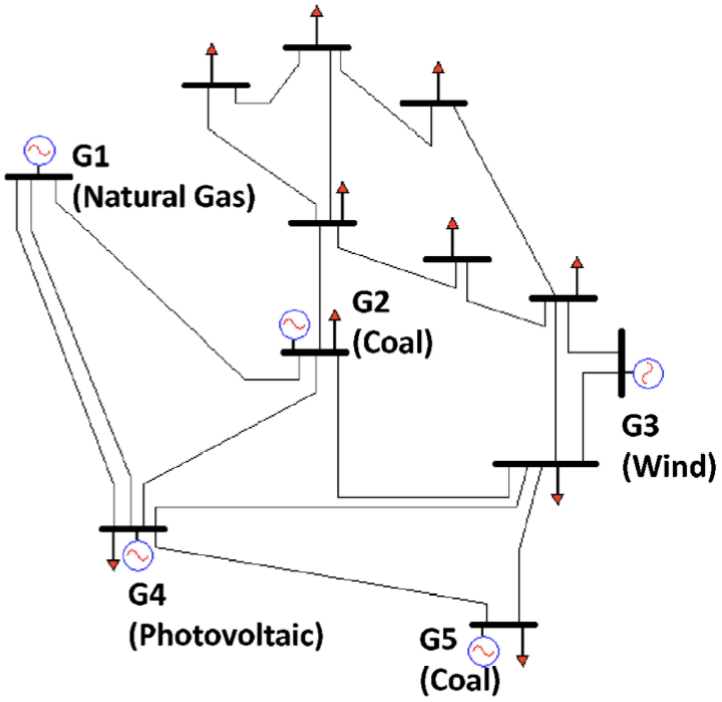


Figure 6.4: The twelve-bus system.

The MDP model for this system can be derived in much the same form as the PNNL and 3-bus system. The Q-learning implementation described in Section 5.4.2 is applied to a reduced order model of the system, wherein all the disturbances at a group of buses are clubbed together for computational tractability. The basis for Q-learning uses an aggregate penalty function to model the impacts of transmission constraints.

The MPC implementation follows the steps outlined in Section 6.3.2. Three MPC implementations are considered:

- Benchmark MPC with $J_0(x) = c(x, 0)$,
- LQR-MPC with $J_0(x) = h_0^*(x)$, and,
- Q-MPC with $J_0(x) = \underline{H}_0^{\theta^*}(x)$.

Control steps of 15-minutes are used in the simulations. The total dispatch costs are computed for different values of the prediction horizon T for typi-

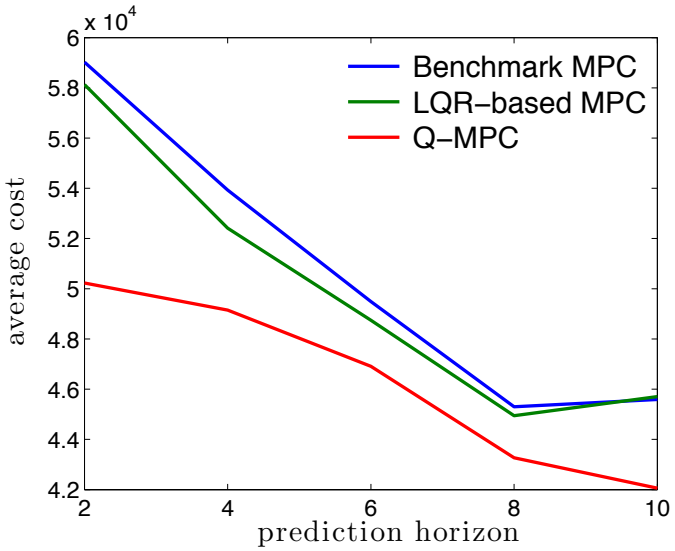


Figure 6.5: Total costs for different prediction horizons for dispatch of the 3-bus system.

cal summer day and plot in Figure 6.5. The plot serves to demonstrate the computational efficacy of the Q-MPC approach.

6.5 Concluding Remarks

MPC is a popular approach to control for constrained systems. It is shown here that performance and computation complexity can be improved significantly by applying new techniques from RL. These insights lead to the genesis of Q-MPC, a seamless integration of Q-learning algorithms devised in Chapter 5 with the standard MPC framework.

The computational efficacy of Q-MPC is examined in the context of the economic dispatch problem: our results indicate that the usual implementation of economic dispatch is computationally less efficient the proposed Q-MPC solution. Indeed, from simulations of the economic dispatch problem for three different test systems, it is observed that the Q-MPC solution provides least-cost dispatch strategies, especially for low values of prediction horizon. In conclusion, the Q-MPC may indeed be a better alternative than the usual MPC approach employed for economic dispatch.

Chapter 7

Conclusions

The main contributions of the dissertation are summarized here, along with proposed extensions and future research.

7.1 Summary

The main contribution of the research presented in this dissertation is the development of new techniques that can contribute towards an advanced control architecture for the future power grid. The need for advances in control for the power grid has increasingly become more critical. Many new resources such as renewable generation and flexible loads are being deployed in the power grid. Simultaneously, new communication and metering technology is also being put into practice. Effective use of these new resources and technologies will necessitate changes to the operational paradigms for the power system. The control techniques proposed in this thesis provide a platform for development of new operational decision-making tools.

An advantage of the two Q-learning algorithms presented in Chapter 5 is that they can be applied for control synthesis in settings for which the precise distribution of the uncertainty and its temporal statistics are not known. Our examinations of RL techniques like SARSA driven by real-world data in Chapter 3 have borne out remarkable success. The Q-learning algorithms proposed in this thesis have also been tuned to underlying statistics based on actual measurements of different test systems. Our techniques along with other popular RL techniques can be effective for control synthesis for power grids with wind and solar resources as well as flexible loads and DR, for which

the distribution of uncertainty and temporal statistics are often not known, but a great deal of energy data is available.

In addition to practical applications, the dissertation also makes important theoretical contributions towards the fields of Markov decision theory, approximate dynamic programming and reinforcement learning. The analytical architecture described in Chapter 4 provides a platform to examine stability and performance of approximate solutions to the average cost optimization in MDPs. Also, the two new parameterized Q-learning algorithms presented in Chapter 5 are a significant improvement over Watkin's Q-learning and are effectively applied to construct approximate solutions to MDPs modeled on power grid control problems. Finally, a new lens from Markov theory is used to examine the stability of MPC. In this setting, sufficient conditions for the stability of stochastic MPC implementation are provided; a result hitherto unknown to the MPC community. And the Q-learning algorithms are used to enhance the performance of the standard MPC algorithm: the marriage of these two techniques gives birth to the Q-MPC algorithm. The algorithm admits a stabilizing control policy under mild conditions. Its computational efficiency is proved via simulation studies: numerical results reported in this thesis demonstrate how Q-MPC provides close-to-optimal solutions for small prediction horizons.

7.2 Future Work

The research presented in this dissertation provides a different perspective on control synthesis in power grids: we argue for the use of control techniques that do not impose restrictive assumptions on the underlying system dynamics and statistics such as uncertain wind energy forecasts or demand response models. The techniques introduced here, along with other standard techniques from ADP and RL, serve as a starting point for the development of more advanced control architecture for future grids. However, several more questions need to be answered in order to design the best possible architecture to manage the grid and its component resources. These questions open up many avenues for future research; a few are listed here.

Control Synthesis In Network Settings: Applications of ADP and RL techniques to complex network settings lead to several concerns regarding the selection of basis functions. This is a consequence of the fact that analysis of detailed network models is computationally intractable and complicated by the following factors:

- realistic model of a power transmission network is highly complex, and,
- power flows across transmission lines are subject to strict constraints.

One possible solution for handling network constraints is to employ model reduction techniques. *Aggregation* techniques such as lumping together nodes on the basis of geographical proximity and/or similar resource characteristics are commonly used in power system studies [122]. *Workload relaxations* approximate complex networks by simpler workload models of reduced dimensionality; these techniques have been successfully applied for manufacturing systems and queuing networks [104] and, to a limited extent, to power grids as well [83]. These techniques may provide an architecture to construct a basis for RL techniques. The use of model reduction techniques for learning-based control merits further investigations.

Distributed Control for Power Grids: Increasing use of distributed resources such as small renewable generators may necessitate future power grids to move towards a more distributed control architecture. Such a move is supported by the Smart Grid vision [123], which places a greater emphasis on use of advanced communication, metering and information technologies in power grid operations and control. However, applications of RL to control synthesis in a distributed network setting require further investigation.

MPC has been applied to distributed control of generation for load frequency regulation [124]. In order to use Q-MPC for such applications, the issue of basis selection for distributed control must be dealt with. A possible solution is to use mean-field games framework to derive basis for distributed Q-learning, as applied in [82]. Likewise, ad-hoc policies like the max-weight policies [125] may be used to construct the basis for learning.

An important question in the context of distributed control is the information architecture. Reliance on too much global information introduces complexity, thus defeating the purpose of distributed control, and also increases

security risks, since the information can be manipulated more easily to create instability in the grid. On the other hand, too little global information and a heavy reliance on local information can have disastrous consequences as well: for instance, the "wave" phenomenon observed due to electromechanical oscillations [126].

Theoretical Advances in ADP and RL: Although the RL algorithms proposed in this dissertation have successfully been tuned using real world data, convergence for the learning process in the most general case remains open. Convergence properties of the proposed algorithms in settings where the i.i.d. assumption and stationarity of $\boldsymbol{W}$ may not hold, need to be examined. This may be of particular interest if Q-learning is to be driven by real-world data, as was done in the numerical experiments described in chapters 3 and 5. Likewise, implementation in a time-inhomogeneous environment can be examined and it will require modifications to the proposed approach.

References

[1] A. J. Wood and B. F. Wollenberg, *Power Generation, Operation and Control*, 2nd ed. New York, NY: John Wiley and Sons, Inc., 1996.

[2] S. Stoft, *Power System Economics: Designing Markets for Electricity*. New York, NY: Wiley-IEEE Press, 2002.

[3] "Greatest engineering achievements of the 20$^{\text{th}}$ century," http://www.greatachievements.org/, 2010.

[4] J. Endrenyi, *Reliability Modeling in Electric Power Systems*. John Wiley & Sons, 1979.

[5] B. Kirby, "Ancillary services: Technical and commercial insights," July 2007, prepared for Wartsila.

[6] REN 21, "Renewables 2012 global status report," http://www.ren21.net/REN21Activities/GlobalStatusReport.aspx, Paris: REN 21 Secretariat, Tech. Rep., 2012.

[7] J. Smith, M. Milligan, E. DeMeo, and B. Parsons, "Utility wind integration and operating impact state of the art," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 900 –908, aug. 2007.

[8] Y. Makarov, C. Loutan, J. Ma, and P. de Mello, "Operational impacts of wind generation on california power systems," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 1039 –1050, may 2009.

[9] S. Meyn, M. Negrete-Pincetic, G. Wang, A. Kowli, and E. Shafieepoorfard, "The value of volatile resources in electricity markets," in *Proceedings of the 49th Conference on Decision and Control*, 2010.

[10] H. Holttinen, A. Orths, P. Eriksen, J. Hidalgo, A. Estanqueiro, F. Groome, Y. Coughlan, H. Neumann, B. Lange, F. Hulle, and I. Dudurych, "Currents of change," *IEEE Power and Energy Magazine*, vol. 9, no. 6, pp. 47–59, 2011.

[11] M. G.-S. E.F. Camacho, T. Samad and I. Hiskens, "Control for renewable energy and smart grids," in *The Impact of Control Technology*, T. Samad and A. Annaswamy, Eds. IEEE Control Systems Society, 2011.

[12] L. Xie, P. M. S. Carvalho, L. A. F. M. Ferreira, J. Liu, B. Krogh, N. Popli, and M. Ilic, "Wind integration in power systems: Operational challenges and possible solutions," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 214–232, 2011.

[13] M. Negrete-Pincetic and S. Meyn, "Intelligence by design for the entropic grid," in *Proc. of the 2011 IEEE Power and Energy Society General Meeting*, July 2011, pp. 1–8, invited lecture for PES panel: Deploying Tomorrow's Electric Power Systems: Low Carbon, Efficiency and Security.

[14] J. DeCesaro, K. Porter, and M. Milligan, "Wind energy and power system operations: A review of wind integration studies to date," *The Electricity Journal*, vol. 22, no. 10, pp. 34–43, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1040619009002668

[15] "Flexible generation: Backing up renewables," EURELECTRIC, Tempe, AZ, Tech. Rep., October 2011.

[16] S.-J. Deng, S. Oren, and G. Gross, "Design and valuation of demand response mechanisms and instruments for integrating renewable generation resources in a smart grid environment," Power Systems Engineering Research Center, Tempe, AZ, Tech. Rep. 12-27, October 2012.

[17] "Gas turbine to help grow renewable energy," http://cleantechnica.com/2011/05/29/gas-turbine-to-help-grow-renewable-energy/, May 29, 2011.

[18] I. Gyuk, "Energy storage for a greener grid," *AIP Conference Proceedings*, vol. 1044, no. 1, pp. 376–392, 2008.

[19] B. Lee and D. Gushee, "Massive electricity storage," American Institute of Chemical Engineers, New York, NY, Tech. Rep., June 2008.

[20] P. Denholm, E. Ela, B. Kirby, and M. Milligan, "The role of energy storage with renewable electricity generation," National Renewable Energy Laboratory, Golden, CO, Tech. Rep. NRELTP-6A2-47187, January 2010.

[21] "Biggest power users provide gigawatts of smart grid flexibility," http://www.greentechmedia.com/articles/read/Biggest-Power-Users-Provide-Gigawatts-of-Smart-Grid-Flexibility, March 11, 2013.

[22] P. Steffes, "Grid-interactive renewable water heating: Analysis of the economic and environmental value," www.steffes.com/LiteratureRetrieve.aspx?ID=72241, September 2012.

[23] "Walmart Canada opens its first sustainable distribution centre," http://news.walmart.com/news-archive/2010/11/16/walmart-canada-opens-its-first-sustainable-distribution-centre, November 10, 2010.

[24] D. Todd, M. Caufield, B. Helms, A. Generating, I. Starke, B. Kirby, and J. Kueck, "Providing reliability services through demand response: A preliminary evaluation of the demand response capabilities of Alcoa Inc," *ORNL/TM*, vol. 233, 2008.

[25] "Agricultural demand response program in California helps farmers reduce peak electricity usage, operate more efficiently year-round." [Online]. Available: http://www.smartgrid.gov/

[26] "Energy department to launch new energy innovation hub focused on advanced batteries and energy storage," http://www.eurekalert.org/pub_releases/2012-02/ddoe-ed020712.php, February 7, 2012.

[27] "Boxcar energy: Brilliant new strategies for storing power," http://www.slate.com/articles/health_and_science/alternative_energy/2013/03/energy_storage_technology_batteries_flywheels_compressed_air_rail_storage.html, March 28, 2013.

[28] "Germany's rising share of renewable energy sources boosting energy storage segment," http://www.gtai.de/GTAI/Navigation/EN/_Meta/press,did=344222.html, September 28, 2011.

[29] U. S. Department of Energy, "Benefits of demand response in electricity markets and recommendations for achieving them," February 2006. [Online]. Available: http://www.oe.energy.gov/DocumentsandMedia/congress_1252d.pdf

[30] "New wave of direct load control update on DLC systems, technology," http://www.elp.com/index/display/article-display/1932499483/articles/utility-automation-engineering-td/volume-16/issue-7/features/new-wave-of-direct-load-control-update-on-dlc-systems-technology.html, July 2011.

[31] "Honeywell and Hawaiian electric to use demand response to integrate renewables and reduce fossil fuel dependence," http://honeywell.com/News/Pages/press-releases.aspx, February 2, 2012.

[32] "EnerNOC's DemandSMART chosen by BPA to showcase power of demand response to manage intermittent wind power," http://www.enernoc.com/press, February 1, 2011.

[33] P. Denholm and R. M. Margolis, "Evaluating the limits of solar photovoltaics (PV) in electric power systems utilizing energy storage and other enabling technologies," *Energy Policy*, vol. 35, no. 9, pp. 4424–4433, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030142150700095X

[34] "Exploring the impacts of Californias renewable portfolio standard," Pacific Northwest Utilities Conference Committee, Portland, OR, Tech. Rep., April 2010.

[35] M. Milligan, P. Donohoo, D. Lew, E. Ela, B. Kirby, H. Holttinen, E. Lannoye, D. Flynn, M. O'Malley, N. Miller, P. Eriksen, A. Gottig, B. Rawn, J. Frunt, W. Kling, M. Gibescu, E. Gomez Lazaro, A. Robitaille, and I. Kamwa, "Operating reserves and wind power integration : an international comparison," in *Proc. of the* 9^{th} *International Workshop on Large-Scale Integration of Wind Power Into Power Systems As Well As On Transmission Networks For Offshore Wind Power Plants*, October 2010, pp. 1–6.

[36] "Fast responding regulation service," http://www.ercot.com/mktrules/pilots/frrs/index, September 2012.

[37] R. Wiser and G. Barbose, "Renewables portfolio standards in the United States – a status report with data through 2007," eetd.lbl.gov/ea/ems/reports/lbnl-154e.pdf, Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-154E, April 2008.

[38] G. Barbose, R. Wiser, A. Phadke, and C. Goldman, "Reading the tea leaves: How utilities in the west are managing carbon regulatory risk in their resource plan," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-44E, March 2008.

[39] J. Apt, "The spectrum of power from wind turbines," *Journal of Power Sources*, vol. 169, no. 2, pp. 369–374, 2007.

[40] M. Lange, "On the uncertainty of wind power predictions - analysis of the forecast accuracy and statistical distribution of errors," *Transactions of the ASME-N-Journal of Solar Energy Engineering*, vol. 127, no. 2, pp. 177–184, 2005.

[41] U. Focken, M. Lange, K. Mönnich, H.-P. Waldl, H. G. Beyer, and A. Luig, "Short-term prediction of the aggregated power output of wind farms - a statistical analysis of the reduction of the prediction error by spatial smoothing effects," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 90, no. 3, pp. 231–246, 2002.

[42] A. Mills, "Implications of wide-area geographic diversity for short-term variability of solar power," Power Systems Engineering Research Center, Golden, CO, Tech. Rep. LBNL-3884E, October 2010.

[43] P. Ruiz, C. Philbrick, and P. Sauer, "Wind power day-ahead uncertainty management through stochastic unit commitment policies," in *IEEE/PES Power Systems Conference and Exposition, 2009. PSCE '09.*, 2009, pp. 1 –9.

[44] A. S. Kowli and S. P. Meyn, "Supporting wind generation deployment with demand response," in *Proc. of the 2011 IEEE Power and Energy Society General Meeting*, July 2011, pp. 1–8.

[45] Y. V. Makarov, L. S., J. Ma, and T. B. Nguyen, "Assessing the value of regulation resources based on their time response characteristics," Pacific Northwest National Laboratory, Richland, WA, Tech. Rep. PNNL-17632, June 2008.

[46] K. Vu, R. Masiello, and R. Fioravanti, "Benefits of fast-response storage devices for system regulation in ISO markets," in *Proc. of the 2009 IEEE Power and Energy Society General Meeting*, July 2009, pp. 1 –8.

[47] Federal Energy Regulatory Commission, "Demand response compensation in organized wholesale energy markets," Washington DC, March 15, 2011, Docket No. RM10-17-000; Order No. 745.

[48] ——, "Frequency regulation compensation in organized wholesale power markets ad10-11-000," Washington DC, October 20, 2011, Docket Nos. RM11-7-000 and AD10-11-000; Order No. 755.

[49] ERCOT, "Grid event," http://www.ercot.com/news/press_releases/show/253, February 2008.

[50] P. Behr, "Demand response helped some regions conserve electricity during heat wave," http://www.nytimes.com/cwire/2011/07/27/27climatewire-demand-response-helped-some-regions-conserve-89838.html, July 27, 2011.

[51] C. Gellings, "The concept of demand-side management for electric utilities," *Proceedings of the IEEE*, vol. 73, no. 10, pp. 1468–1470, October 1985.

[52] C. W. Gellings, *The Smart Grid: Enabling Energy Efficiency and Demand Response*. Lilburn, GA: Fairmont Press, 2009.

[53] F. Schweppe, R. Tabors, J. Kirtley, H. Outhred, F. Pickel, and A. Cox, "Homeostatic utility control," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-99, no. 3, pp. 1151 –1163, May 1980.

[54] S. Koch, M. Zima, and G. Andersson, "Potentials and applications of coordinated groups of thermal household appliances for power system control purposes," in *2009 IEEE PES/IAS Conference on Sustainable Alternative Energy (SAE)*, Sept. 2009, pp. 1 –8.

[55] D. S. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Conversion and Management*, vol. 50, no. 5, pp. 1389 – 1400, 2009.

[56] D. Callaway and I. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184–199, 2011.

[57] S. Kundu, N. Sinitsyn, S. Backhaus, and I. Hiskens, "Modeling and control of thermostatically controlled loads," in *2011 Power Systems Computation Conference*, 2011, pp. 1 –5.

[58] J. Mathieu, S. Koch, and D. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 430–440, Feb.

[59] C. Alvarez, A. Gabaldon, and A. Molina, "Assessment and simulation of the responsive demand potential in end-user facilities: application to a university customer," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 1223–1231, 2004.

[60] N. Motegi, M. Piette, D. Watson, S. Kiliccote, and P. Xu, "Introduction to commercial building control strategies and techniques for demand response," California Energy Commission, PIER, Berkeley, CA, Tech. Rep. LBNL-59975, 2006.

[61] J. Mathieu, P. Price, S. Kiliccote, and M. Piette, "Quantifying changes in building electricity use, with application to demand response," *IEEE Transactions on Smart Grid*, vol. 2, no. 3, pp. 507–518, 2011.

[62] A. Roscoe and G. Ault, "Supporting high penetrations of renewable generation via implementation of real-time electricity pricing and demand response," *IET Renewable Power Generation*, vol. 4, no. 4, pp. 369 –382, July 2010.

[63] A. Papavasiliou and S. Oren, "Supplying renewable energy to deferrable loads: Algorithms and economic analysis," in *Proc. of the 2010 IEEE Power and Energy Society General Meeting*, July 2010, pp. 1–8.

[64] M. Ilic, L. Xie, and J.-Y. Joo, "Efficient coordination of wind power and price-responsive demand–part i: Theoretical foundations," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 1875–1884, 2011.

[65] ——, "Efficient coordination of wind power and price-responsive demand–part ii: Case studies," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 1885–1893, 2011.

[66] H. T. Le, S. Santoso, and W. Grady, "Development and analysis of an ESS-based application for regulating wind farm power output variation," in *Proc. of the 2009 IEEE Power Energy Society General Meeting*, July 2009, pp. 1 –8.

[67] A. S. Kowli and S. P. Meyn, "Power node control for renewable integration," in *Proc. of the 2012 IEEE Power and Energy Society General Meeting*, July 2012.

[68] S. Dutta and T. Overbye, "Optimal storage scheduling for minimizing schedule deviations considering variability of generated wind power," in *2011 IEEE/PES Power Systems Conference and Exposition (PSCE)*, March 2011, pp. 1 –7.

[69] W. Kempton and J. Tomić, "Vehicle-to-grid power implementation: From stabilizing the grid to supporting large-scale renewable energy," *Journal of Power Sources*, vol. 144, no. 1, pp. 280–294, 2005.

[70] M. Caramanis and J. Foster, "Management of electric vehicle charging to mitigate renewable generation intermittency and distribution network congestion," in *Proc. of the 48th IEEE Conf. on Dec. and Control; held jointly with the 2009 28th Chinese Control Conference*, 2009, pp. 4717–4722.

[71] E. Zoulias and N. Lymberopoulos, "Techno-economic analysis of the integration of hydrogen energy technologies in renewable energy-based stand-alone power systems," *Renewable Energy*, vol. 32, no. 4, pp. 680–696, 2007.

[72] P. Denholm and R. Sioshansi, "The value of compressed air energy storage with wind in transmission-constrained electric power systems," *Energy Policy*, vol. 37, no. 8, pp. 3149–3158, 2009.

[73] F. Bouffard and M. Ortega-Vazquez, "The value of operational flexibility in power systems with significant wind power generation," in *Proc. of the 2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1–5.

[74] A. Ulbig and G. Andersson, "On operational flexibility in power systems," in *Proc. of the 2012 IEEE Power and Energy Society General Meeting*, 2012, pp. 1–8.

[75] K. Heussen, S. Koch, A. Ulbig, and G. Andersson, "Unified system-level modeling of intermittent renewable energy sources and energy storage for power system operation," *IEEE Systems Journal*, vol. PP, no. 99, p. 1, 2011.

[76] L. Xie and M. Ilic, "Model predictive economic/environmental dispatch of power systems with intermittent resources," in *Proc. of the 2009 IEEE Power and Energy Society General Meeting*, July 2009, pp. 1 –6.

[77] I. Kamwa, R. Grondin, and Y. Hebert, "Wide-area measurement based stabilizing control of large power systems-a decentralized/hierarchical approach," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 136–153, 2001.

[78] C. H. Hauser, D. E. Bakken, and A. Bose, "A failure to communicate: next generation communication requirements, technologies, and architecture for the electric power grid," *IEEE Power and Energy Magazine*, vol. 3, no. 2, pp. 47–55, 2005.

[79] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[80] H. Yu and D. P. Bertsekas, "Q-learning algorithms for optimal stopping based on least squares," in *Proc. European Control Conference (ECC)*, July 2007.

[81] C. Moallemi, S. Kumar, and B. Van Roy, "Approximate and data-driven dynamic programming for queueing networks," 2008, preprint available at http://moallemi.com/ciamac/research-interests.php.

[82] P. G. Mehta and S. P. Meyn, "Q-learning and Pontryagin's minimum principle," in *Proc. of the 48th IEEE Conf. on Dec. and Control; held*

jointly with the 2009 28th Chinese Control Conference, Dec. 2009, pp. 3598–3605.

[83] M. Chen, I.-K. Cho, and S. P. Meyn, "Reliability by design in distributed power transmission networks," *Automatica*, vol. 42, pp. 1267–1281, August 2006.

[84] I.-K. Cho and S. P. Meyn, "Efficiency and marginal cost pricing in dynamic competitive markets with friction," *Theoretical Economics*, vol. 5, no. 2, pp. 215–239, 2010.

[85] Y. Yan, S. Yang, F. Wen, and I. MacGill, "Generation scheduling with volatile wind power generation," in *International Conference on Sustainable Power Generation and Supply, 2009. SUPERGEN '09.*, 2009, pp. 1 –7.

[86] V. Hamidi, F. Li, L. Yao, and M. Bazargan, "Domestic demand side management for increasing the value of wind," in *China International Conference on Electricity Distribution*, Dec. 2008, pp. 1 –10.

[87] R. Sioshansi and W. Short, "Evaluating the impacts of real-time pricing on the usage of wind generation," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 516 –524, May 2009.

[88] M. Parvania and M. Fotuhi-Firuzabad, "Demand response scheduling by stochastic SCUC," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 89 –98, 2010.

[89] M. Klobasa, "Analysis of demand response and wind integration in Germany's electricity market," *Renewable Power Generation, IET*, vol. 4, no. 1, pp. 55 –63, 2010.

[90] D. Callaway and I. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184 –199, Jan. 2011.

[91] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, 8th ed., ser. Springer Series in Operations Research and Financial Engineering. New York, NY: Springer, 2006.

[92] G. Wang, M. Negrete-Pincetic, A. Kowli, E. Shafieepoorfard, S. Meyn, and U. Shanbhag, "Dynamic competitive equilibria in electricity markets," in *Control and Optimization Theory for Electric Smart Grids*, A. Chakrabortty and M. Illic, Eds. New York, NY: Springer-Verlag, January, 2012.

[93] G. Wang, M. Negrete-Pincetic, A. Kowli, E. Shafieepoorfard, S. Meyn, and U. V. Shanbhag, "Real-time prices in an entropic grid," in *3rd IEEE PES Conference on Innovative Smart Grid Technologies (ISGT 2012)*, Jan 2012, invited lecture for IEEE panel: Load modeling and control.

[94] L. Wu, M. Shahidehpour, and T. Li, "Stochastic security-constrained unit commitment," *Power Systems, IEEE Transactions on*, vol. 22, no. 2, pp. 800 –811, May 2007.

[95] S. Kazarlis, A. Bakirtzis, and V. Petridis, "A genetic algorithm solution to the unit commitment problem," *Power Systems, IEEE Transactions on*, vol. 11, no. 1, pp. 83 –92, Feb. 1996.

[96] "ISO New England," http://www.iso-ne.com.

[97] "Wind systems integration: Data resources," http://www.nrel.gov/wind/systemsintegration/data_resources.html.

[98] J. L. Mathieu, "Modeling, analysis, and control of demand response resources," Ph.D. dissertation, University of California at Berkeley, 2012.

[99] "Buildings energy data book." [Online]. Available: http://buildingsdatabook.eren.doe.gov/default.aspx

[100] "First 'small scale' demand-side projects in PJM providing frequency regulation," http://www.sacbee.com/2011/11/21/v-print/4070973/first-small-scale-demand-side.html, November 2011.

[101] H. Hao, A. Kowli, Y. Lin, P. Barooah, and S. Meyn, "Ancillary service for the grid via control of commercial building HVAC systems," 2013, to appear in Proc. of the American Control Conference.

[102] H. Hao, A. Kowli, P. Barooah, and S. Meyn, "Ancillary service through control of HVAC in commercial buildings," 2013, under review for the special issue on Control Theory and Technology in IEEE Transactions on Smart Grid.

[103] "PennsylvaniaNew JerseyMaryland ISO," http://www.pjm.com.

[104] S. P. Meyn, *Control Techniques for Complex Networks*. Cambridge: Cambridge University Press, 2007, pre-publication edition available online.

[105] D. Huang, W. Chen, P. Mehta, S. Meyn, and A. Surana, "Feature selection for neuro-dynamic programming," in *Reinforcement Learning and*

Approximate Dynamic Programming for Feedback Control, F. Lewis, Ed. Wiley, 2011.

[106] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems (2nd ed.)*. New York, NY: Springer-Verlag New York, Inc., 1998.

[107] W. Chen, D. Huang, A. A. Kulkarni, J. Unnikrishnan, Q. Zhu, P. Mehta, S. Meyn, and A. Wierman, "Approximate dynamic programming using fluid and diffusion approximations with applications to power management," in *Proc. of the 48th IEEE Conf. on Dec. and Control; held jointly with the 2009 28th Chinese Control Conference*, 2009, pp. 3575–3580.

[108] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1990.

[109] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: Cambridge University Press, 2009, published in the Cambridge Mathematical Library. 1993 edition online.

[110] S. Lu, M. A. Elizondo, N. Samaan, K. Kalsi, E. Mayhorn, R. Diao, C. Jin, and Y. Zhang, "Control strategies for distributed energy resources to maximize the use of wind power in rural microgrids," in *Proc. of the 2011 IEEE Power and Energy Society General Meeting*, July 2011, pp. 1–8.

[111] "GridLAB-D residential module user's guild," http://sourceforge.net/apps/mediawiki/gridlab-d/index.php?title=Residential_module_user%27s_guide.

[112] E. Mayhorn, K. Kalsi, M. A. Elizondo, W. Zhang, S. Lu, N. Samaan, and K. Butler-Purry, "Optimal control of distributed energy resources using model predictive control," in *Proc. of the 2012 IEEE Power and Energy Society General Meeting*, July 2012, pp. 1–8.

[113] D. P. de Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations Res.*, vol. 51, no. 6, pp. 850–865, 2003.

[114] D. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Cambridge, Mass: Atena Scientific, 1996.

[115] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Delhi, India and Cambridge, UK: Hindustan Book Agency and Cambridge University Press (jointly), 2008.

[116] D. Shirodkar and S. Meyn, "Quasi stochastic approximation," in *Proc. of the 2011 American Control Conference (ACC)*, July 2011, pp. 2429–2435.

[117] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.

[118] D. Mayne, J. Rawlings, C. Rao, and P. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0005109899002149

[119] R.-R. Chen and S. P. Meyn, "Value iteration and optimization of multiclass queueing networks," *Queueing Syst. Theory Appl.*, vol. 32, no. 1-3, pp. 65–97, 1999.

[120] R. Negenborn, B. D. Schutter, M. Wiering, and H. Hellendoorn, "Learning-based model predictive control for Markov decision processes," in *Proc. 16th World Congress of the International Federation of Automatic Control (IFAC)*, Prague, Czech Republic, July 2005.

[121] J. Lee, "Approximate dynamic programming approach for process control," in *2010 International Conference on Control Automation and Systems (ICCAS)*, Oct., pp. 459–464.

[122] A. Papaemmanouil and G. Andersson, "On the reduction of large power system models for power market simulations," in *2011 Power Systems Computation Conference*, 2011, pp. 1 –7.

[123] U.S. Department of Energy, "Smart grid," http://energy.gov/oe/technology-development/smart-grid, 2011.

[124] A. N. Venkat, I. A. Hiskens, J. B. Rawlings, and S. J. Wright, "Distributed mpc strategies with application to power system automatic generation control," *Control Systems Technology, IEEE Transactions on*, vol. 16, no. 6, pp. 1192–1206, 2008.

[125] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Automat. Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[126] A. Phadke and J. Thorp, "Electromechanical wave propagation," in *Synchronized Phasor Measurements and Their Applications*. Springer, 2008, pp. 223–243.