Title
Mitigating Renewable Variability through Control and Optimization Techniques

Author
Subramanian, Anand

Publication Date
2013

Peer reviewed|Thesis/dissertation

Mitigating Renewable Variability through Control and Optimization Techniques

by

Anand Subramanian

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering - Mechanical Engineering
and the Designated Emphasis
in
Computational Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kameshwar Poolla, Chair
Professor Pravin Varaiya
Professor Andrew Packard

Fall 2013

Mitigating Renewable Variability through Control and Optimization Techniques

Abstract

Mitigating Renewable Variability through Control and Optimization Techniques

by

Anand Subramanian

Doctor of Philosophy in Engineering - Mechanical Engineering
and the Designated Emphasis
in
Computational Science and Engineering

University of California, Berkeley

Professor Kameshwar Poolla, Chair

Motivated by environmental and energy security concerns, many states and countries have enacted legislation calling for increased renewable adoption in the electricity sector. Unlike conventional fossil fuel-based electricity sources, renewable resources such as wind and solar power are characterized by intermittent, uncertain, and non-dispatchable output. Successfully addressing this variability in renewable generation is a prerequisite for deep renewable penetration. The prevailing paradigm in most power systems, exemplified by programs such as the Participant Intermittent Renewable Program (PIRP) in California, is one where the system operator accepts all renewable generation and absorbs the attendant variability through operating reserves. This approach works today at modest renewable generation levels but will result in untenable increases in reserve requirements tomorrow when renewables serve a sizeable fraction of total electric load. Hence, meeting ambitious renewable penetration targets will require the implementation of a number of variability mitigation strategies across the entire spectrum of power system operations. In this dissertation, we identify and explore areas in which control and optimization techniques can help provide some of these solutions.

One such mitigation strategy is the coordinated aggregation of deferrable loads and storage, in which load is tailored to match variable supply. We investigate methods of exploiting these demand-side capabilities by developing and evaluating algorithms for the real-time scheduling of these flexible resources. We find that the benefits of coordinated aggregation can be achieved at modest levels of both deferrable load participation and flexibility. We also provide an analytical framework for understanding how these resources influence the costs of meeting load requirements incurred in wholesale electricity markets. Specifically, we explore the interplay between deferrable load scheduling and cost-minimizing procurements of bulk power and reserve capacity made in the day-ahead forward market.

Another solution we consider, specific to wind power, involves curtailing generator output in certain situations. We explore how a wind power producer – subject to financial penalties for imbalances from contracted amounts – might leverage power curtailment capability to mitigate financial risk arising from price and production uncertainty. In particular, we analytically quantify the economic benefit derived from curtailment as an explicit function of expected prices, and compute empirical estimates of this curtailment benefit using price data from the various power system operators.

All my life, I came to see the ocean,
Today, I come with a few drops of my own in hand.

This thesis is dedicated to:
my mother, Prabamani Subramanian,
my father, Ayalur Krishnaiyer Subramanian,
my sister, Anagha Subramanian,
and my other half, Priyanka Rajagopalan.

Contents

List of Figures

List of Tables

Acknowledgments

I would be remiss if I did not mention the following people without whom this work would not be possible.

First, I'd like to thank my research advisor and friend, Kameshwar Poolla. I first met him as a entry-level graduate student hoping to work on control applications in semiconductor metrology. In the past five years, we've worked on that and much, much more. While I've learned a variety of technical skills from him in that time, perhaps the most instructive lesson was to always maintain perspective when conducting research and to never lose sight of the 'big picture'. I am sure that these life lessons coupled with his approach to technical problem-solving will leave a lasting impression upon me. I will always think fondly of my time working with him.

Next, I'd like to thank the other professors on my dissertation committee. Working with Pravin Varaiya has been in a word, amazing. His ability to provide deep insights that lead to novel research directions is truly remarkable. This work owes much to his stellar intellectual acumen. Andy Packard has been instrumental in giving me the tools necessary to create much of what follows in this document. However, one can argue that his defining contribution to my personal development has been through the many conversations we've had on late nights in 5105 Etcheverry Hall – on topics ranging from Jeff Tedford's contract buyout to Muggle Quidditch.

This dissertation would be lacking if not for the contributions of many of my colleagues. Eilyan Bitar is the most important person in this regard. A few years my senior at Berkeley, he has been a steadfast companion on this journey. His input has been invaluable. In fact, the analysis of wind curtailment (Chapter 5) is a direct result of ideas he helped formulate on a summer day in 2011. In addition, Joshua Taylor warrants special mention for his contributions to Chapter 4. This work should be viewed as an early endorsement of their technical prowess in what I expect to be long and prolific academic careers. More generally, it has been a pleasure to work with my lab-mates at the BCCI over the years. These people enriched my graduate career immensely, be it through on-the-spot MATLAB/LaTeX debugging or more substantive discussions over cups of coffee at Cafe Nefeli. I must also single out Manuel Garcia, who contributed to the simulations validating DER scheduling (Chapter 3).

Special mention goes to the following professors. Duncan Callaway helped teach me some of the many nuances of power system operation which I hope has been reflected adequately in this work. Pramod Khargonekar, during his time at Berkeley, helped teach me, as he has countless others, the crucial lesson of identifying and focusing on specific problems while conducting research. Tyrone Vincent, with whom I first worked on state estimation problems in semiconductor metrology, taught me the value of a methodical, organized approach to problem solving.

I must also mention Edwin Liu who was my manager while I interned at Quanta Technology. He gave me a bird's eye view of all things 'smart grid' and helped me understand the mindset of a distribution network engineer.

We finally come to people who had to put up with me regularly outside my time in lab: my parents, my sister Anagha, and my fiancée Priyanka. Everything I am is a direct result of the sacrifices and efforts my parents have made over the course of my life. My little sister's encouragement and appreciation for her elder brother has been truly incredible. Finally, Priyanka's unwavering support through this entire process, a small part of which is her relentless proof-reading of this dissertation, is something I am lucky to have.

Chapter 1

Introduction

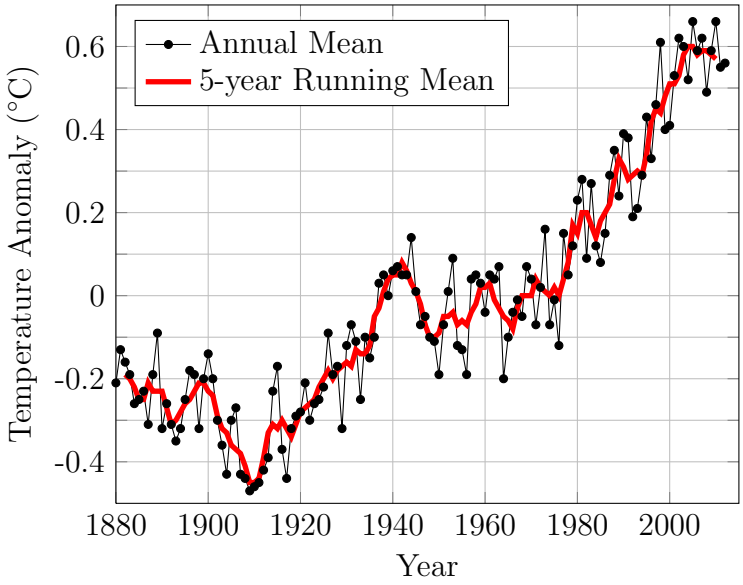
1.1 Climate Change

Man-made climate change, also known as global warming, is a real and serious threat to mankind's long-term survival. The explanation for this phenomenon is simple - greenhouse gas emissions resulting from human activity raise the average temperatures of the world's oceans, surface, and atmosphere. These changes in global temperature have the potential to disrupt all aspects of human life.

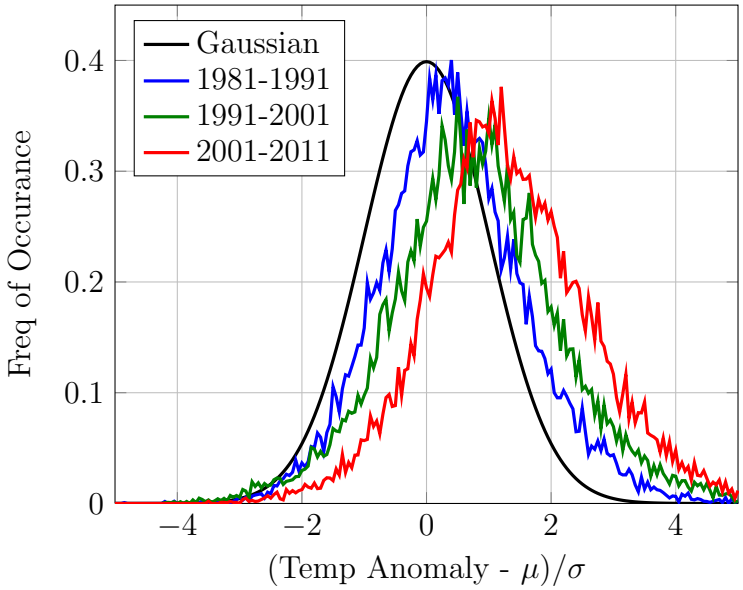
There is clear and conclusive evidence that the average temperature of the earth's surface, oceans, and atmosphere is rising. Figure 1.1a, which shows time-series data of the annual temperature average, indicates a 0.7°C increase in average global temperature over the past century. Moreover, these increases in *mean* global temperature have been accompanied by increases in temperature *variance* (Figure 1.1b). This increased temperature variation over the past two decades is concerning, as it suggests increased frequency of extreme weather events such as heat waves and droughts. In short, these are disconcerting temperature trends - trends that will likely cause a host of problems such as disruption in supply of food [101] and fresh water [76], coastal habitat inundation [74, 100], increased malnutrition, and other negative health consequences [89, 75].

There is broad scientific consensus that these changes in global temperature are caused by *consistent* increases in CO_2 emissions [10, 98]. As Figure 1.2 illustrates, the past *50 years* have witnessed an increase in atmospheric CO_2 concentration of 25% or 90 parts per million (ppm). This is a staggering increase compared to the past 800,000 years of atmospheric CO_2 concentration variation in two aspects [82]. The pace of this increase is unparalleled - comparable changes in atmospheric CO_2 concentration have occurred over time-scales of thousands of years. The magnitude of this increase (90 ppm) is also alarming - before this recent increase, atmospheric CO_2 concentration had stayed within a 110 ppm range (170-290 ppm) for 800,000 years.

This increase in greenhouse gas emissions is largely attributed to anthropogenic (man-made) activity. Specifically, this increase was caused by the burning of fossil fuels, such as



(a) Global mean land-ocean temperature from 1880 to present. These temperatures are expressed as anomalies relative to the average temperature over a 1951-1980 base range. (Version of Fig.1A in [65]) (Data from: `http://data.giss.nasa.gov/gistemp/graphs_v3/`)



(b) Distributions of Northern Hemisphere summer temperature anomalies by decade. Anomalies normalized with respect to 1951-1980 mean (μ) and standard deviation (σ). (Version of Fig. P1A in [64])

Figure 1.1: Figures outlining increases in both mean and variance of global temperature

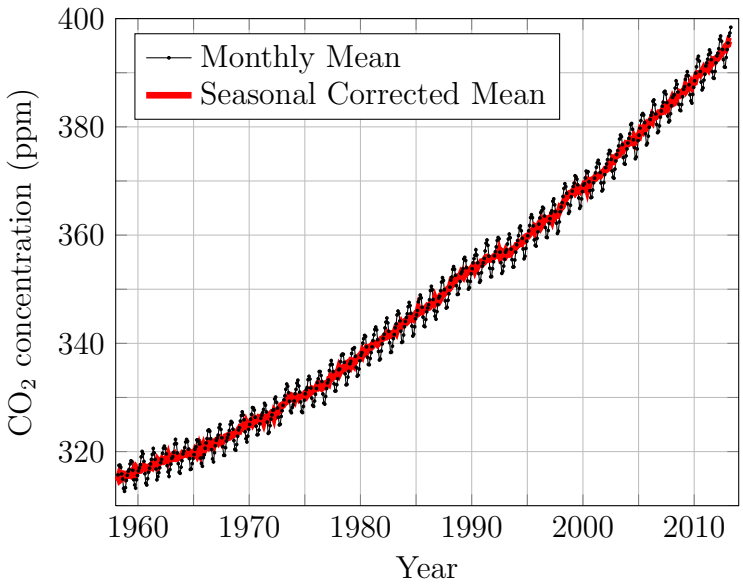


Figure 1.2: Atmospheric CO_2 concentration measured at Mauna Loa observatory over time. The black line shows monthly averages while the red line shows the monthly average correction for seasonal CO_2 variation. (Data from NOAA Earth System Research Laboratory: `http://www.esrl.noaa.gov/gmd/ccgg/trends/`)

coal, oil, and natural gas, since the industrial revolution. Today, the majority of energy needs are still met using fossil fuels. For instance, fossil fuel energy sources supplied 80% of U.S energy consumption in 2012 [11]. As the electricity sector makes up 40 % of total U.S. energy consumption (Figure 1.4), any attempt at tackling anthropogenic climate change will require replacing fossil fuel-based sources of electricity with carbon-neutral alternatives.

1.2 Renewable Integration

Increased adoption of renewable resources in the electricity sector is a key component of an overall strategy to curb greenhouse gas emissions. In order to incentivize the installation of renewable generation capacity, many countries have set renewable energy targets. These targets are legislative requirements on electricity grid operators to serve a specified proportion of load solely with renewable energy. Examples of these targets for electric sectors include Australia and Germany, which aim to have renewable energy penetrations of 20% by 2020 [45] and 50% by 2030 [56] respectively. These ambitious renewable energy targets are typically coupled with other forms of support such as tax credits, feed-in tariffs, guaranteed grid access, and other extra-market mechanisms. Such government subsidies to encourage renewable penetration are necessary to help cover the significant capital outlays associated with installing new generation capacity.

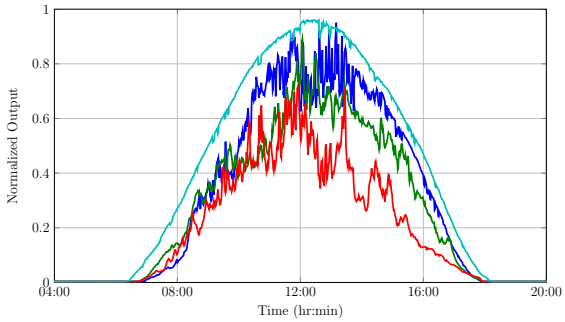
The United States has no such national target but individual U.S. states have created their own targets as part of legislative mandates called renewable portfolio standards (RPS). For instance, California's RPS calls for 33% renewable energy penetration by 2020 [32]. As of 2012, the three largest investor-owned utilities in California collectively served only 19.8% of electricity needs with renewables [32]. Clearly, the RPS calls for a significant increase from current levels.

Figure 1.4 illustrates that the total proportion of electricity needs supplied by renewable energy in the U.S. is around 12%. Additionally, the majority of this energy is derived from hydroelectric power sources. Since 1880, these energy sources have long been an integral component of the overall U.S. energy portfolio [49]. Hydroelectric power offers competitive rates of return, particularly in developing countries such as China [59]. However, the construction of new generating facilities is severely constrained by geographic considerations - hydroelectric power stations must be built on sites where a large river flows over a sudden elevation decrease. Moreover, recent awareness of the environmental impacts on water supply and river ecology has actually resulted in the decommissioning of numerous dams within the U.S. [47, 125]. Therefore, while hydroelectric power may displace fossil fuels in developing countries such as China and Brazil, it is unlikely to help meet renewable energy targets in the U.S and Western Europe.

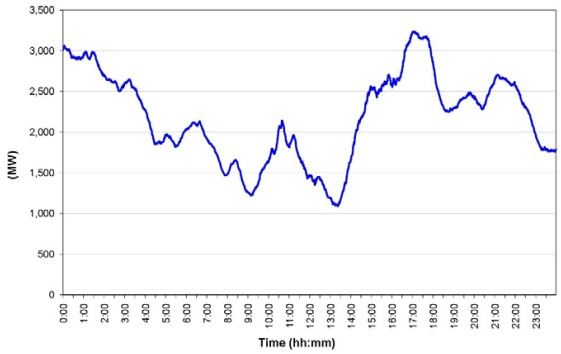
If not hydroelectric power, which type of renewable resource will enable governments to meet their RPS targets? Geothermal power involves extracting heat from deep within the earth's crust to power conventional steam turbines. As these types of power plants must be located at sites where energy can be procured with minimal drilling, the installation of new geothermal energy capacity, like hydroelectric power, is constrained by geography. Consequently, wind and solar energy are emerging as the main modalities with which the RPS targets of the developed world will be met. These resources are plentiful and, crucially, do not suffer from the type of geographical constraints that limit hydroelectric and geothermal generation. Indeed, there has been significant growth over the past few years in wind and solar energy production [34]. However, wind and solar power still only make up a tiny fraction ($< 2\%$) of the overall U.S. energy portfolio (Figure 1.4). Deep penetration of wind and solar power can only be achieved by overcoming a number of technological and market hurdles that limit their adoption.

Variability: There are three crucial ways in which wind and solar power are different from energy produced by conventional sources such as coal or natural gas.

1. **Intermittency**: Sudden fluctuations in power output occurring at short time-scales.
2. **Uncertainty**: Inability to accurately predict or anticipate fluctuations in power output in advance of the delivery interval.
3. **Non-dispatchability**: Inability to increase or decrease generation dispatch as required.



(a) Typical daily power profiles of aggregate solar production from 30 solar photovoltaic (PV) installations across California. Output normalized with respect to maximum generating capacity.



(b) Power profile of aggregate wind generation across ERCOT (Energy Reliability Council of Texas) footprint for a day in 2011. (Figure Source: [124])

Figure 1.3: Figures displaying solar and wind intermittency issues.

We will use the term *variability* as an umbrella term for these fundamental characteristics of renewable generation.

Intermittency in both solar and wind power stems from variations in weather conditions. Overcast conditions prevent solar installations from achieving their rated capacities. Moreover, as Figure 1.3a illustrates, solar power profiles exhibit significant fluctuations due to passing cloud cover. In the case of wind power, atmospheric conditions that dictate power output levels are subject to rapid, random changes. Of particular concern to system operators are wind *ramps* - large, sudden changes in output occurring over short (sub-hour) time-scales. This intermittent behavior is evident even in aggregate wind profiles over large regions (Figure 1.3b). These wind ramps can create unanticipated power imbalances within the grid, thereby compromising overall grid stability [124]. Accounting for these variations requires forecasting these ramps over multi-hour or day-long prediction windows. This is a difficult problem [31].

Grid operators, who are accustomed to dispatchable sources of generation, must modify their operating practices to accommodate generation with intrinsically uncontrollable output. We view variability as the greatest obstacle to the large-scale integration of wind and solar generation required to meet renewable energy targets.

Levelized Cost: Currently, wind and solar generation are more expensive than their fossil-fuel counterparts. The U.S Energy Information Agency estimates the levelized system cost for onshore wind and solar photovoltaic (PV) power plants entering service to be $87 and $144 per MWhr, respectively, in 2008 [4]. The levelized system cost reflects all costs associated with operating a generating facility, including capital and operating costs, additional transmission infrastructure investment, and estimates of capacity factors. While these costs have decreased significantly over the past few years (corresponding estimates two years ago were

$97 and $211 per MWhr, respectively [3]), they still exceed similar figures for conventional combined cycle natural gas ($67/MWhr) and conventional coal ($100/MWhr). Increased adoption of wind and solar power is contingent on addressing these economic realities, especially in view of the recent glut of natural gas production within the U.S., which will continue to depress natural gas operating costs [116]. Hence, large-scale renewable integration will require various forms of government subsidies in the short-term, and technological innovation that reduces the cost of renewables in the long-term.

Distributed Generation: A significant proportion of new renewable capacity may be in the form of small, distribution-level assets such as rooftop solar PVs and small-scale wind power plants. These are examples of distributed energy resources (DERs), which are discussed in detail in Section 1.3. These resources are referred to as *distributed generation* (DG) and certain states have called for specific DG penetration targets. For instance, Arizona has mandated that distributed renewable generation must serve 4.5% of total electricity load by 2025 [12].

The installation of sizeable DG capacity poses operational challenges for medium- and low-voltage distribution networks. Distribution networks are traditionally operated as radial networks linking the substation, the point of connection with the high-voltage transmission network, to individual residential, commercial, and industrial loads. Utilities, which operate and maintain distribution networks, work under the assumption of *unidirectional* power transfer within this network from the source node to the leaves. However, with the adoption of DG at different points within this radial network, this assumption is no longer valid. Therefore, increased adoption of distributed renewables requires changes in the operating protocols for distribution networks [84]. Additionally, grid operators, who are typically accustomed to dealing with large-scale generation facilities, must be cognizant of distributed generation capabilities when operating the bulk transmission network.

The three challenges outlined here are, in our estimation, the main obstacles to renewable integration. In this work, we will not address the cost concerns associated with wind and solar power. Instead, we choose to focus on ways to *mitigate the variability in renewable output* while simultaneously respecting the *distributed nature* of such generation resources.

1.3 DER Proliferation

The emergence of distributed generation is part of a broader deployment of distributed energy resources (DERs). DERs refer to small energy devices that are placed within the distribution, rather than bulk transmission, network. Apart from renewables, this also includes flexible loads such as plug-in electric vehicles and thermostatically controlled loads, microgeneration in the form of small-scale natural gas and diesel generators, small-scale storage devices, and advanced power electronics. As with renewable generation, the interest in DERs is spurred

by consistent government policies identifying DER adoption and integration as a key policy goal [8].

The widespread deployment of DERs offers opportunities for greenhouse gas mitigation. This is evident with the increased adoption of plug-in electric vehicles (EVs). By displacing petroleum, these vehicles can greatly reduce greenhouse gas emissions. This impact is heightened if EV energy needs are met with carbon-free renewables.

DERs also offer substantial benefits to power system operations. While these resources are small, they offer, in aggregate, a range of capabilities for power system operation including but not limited to relieving transmission and distribution network congestion, improving system reliability, and reducing the need for additional generation to meet increasing load requirements [48, 1]. Because of their small size, DERs are often quicker and easier to deploy than traditional utility-scale resources, making them the method of choice to address various grid needs.

In terms of aiding renewable penetration, flexible loads and storage can absorb some of the variability associated with renewable generation by adjusting their power delivery profiles. Consider the case of increased EV penetration. As typical power demand for an individual EV ranges from 2-17 kW [81], charging high numbers of EVs would require capacity upgrades in distribution networks. However, mechanisms and schemes to schedule or incentivize charging at particular times can obviate the need for costly retrofits. In fact, coordinating EV charging with renewable generation availability can help address the key issue of variability in renewable generation. Clearly, exploiting the operational flexibility offered by DERs requires their efficient integration into power system operations.

The electric grid is currently configured to transmit power produced by utility-scale generation resources to static, inflexible loads. Current operating protocols – from coordinated generation dispatch to maintaining system reliability – have been designed with this paradigm in mind. Specifically, the distribution network is managed purely as a conduit for power from the bulk transmission network to individual loads. Effective DER integration demands more nuanced distribution network management than this traditional model can offer. Managing DERs to achieve desirable power system objectives will require different communication, monitoring, and control infrastructures than existing ones. However, truly realizing the potential of DERs requires fundamental changes in distribution network operating principles.

1.4 The Scourge of Variability

As mentioned earlier, wind and solar generation exhibit significant variability in their output. Currently, renewable power producers (RPPs) currently benefit from many forms of government support aimed at increasing the share of electricity generation. California's Participating Intermittent Resource Program (PIRP) is an example of legislation offering such subsidies [67]. A central component of PIRP, and many similar policies, is *priority dispatch*. Priority dispatch creates an operating paradigm in which the grid operator is obliged to

accept *all* renewable power production subject to certain contractual constraints. Consequently, RPPs do not have to participate in traditional electricity markets in which they would have to directly compete with other generation sources for grid access. In most cases, RPPs are also exempt from typical penalties for generation imbalances that grid operators typically levy on conventional power producers. Under PIRP for instance, RPPs only pay penalties for the net monthly deviations from hourly generation forecasts [87].

Such extra-market support is problematic as grid operators must *continuously balance load and generation*. Typically, power imbalances resulting from variability in loads or generation are counterbalanced with reserve generation purchased in dedicated ancillary service (AS) markets [55]. The costs associated with reserve procurement are allocated among the load-serving entities (i.e. utilities) based on their relative share of energy demands. The load-serving entities, then pass these costs on to consumers as a component of retail electricity rates. In the past, fluctuations and uncertainty in loads represented the primary source of variability in power system operations. Generation was dispatchable and, with the exception of rare contingency events such as outages, exhibited minimal uncertainty. Hence, the practice of socializing reserve costs among consumers made for fair policy.

However, this method of reserve cost allocation is no longer justified when a significant proportion of the load requirements is met with intermittent renewables. Recall that the grid operator must accept all renewable generation, effectively treating it as a negative load. Maintaining power balance in the face of the persistent variability in generation requires substantial increases in reserve requirements. For instance, recent studies in California project that, in order to meet the state's target of 33% renewable penetration, the maximum load-following down reserve requirement, used when generation exceeds load, will increase from 3,247 MW to 5,579 MW for the fall season [68]. These studies also forecast similar increases for various other reserve metrics [30, 68]. As the majority of reserve generation is typically provided by fast-ramping natural gas turbines, greenhouse gas emissions resulting from providing these additional reserves defeat the environmental benefit of renewables. Moreover, the practice of socializing these additional reserve costs among consumers – effectively an implicit subsidy for renewable generation – is untenable. Indeed, there have been significant policy discussions about the appropriate allocation of these costs at high renewable penetrations [121, 120].

As the current modus operandi is unsustainable, what changes to grid operations are necessary to successfully mitigate renewable variability? In the near-term, one solution is to discontinue the preferential, extra-market treatment of RPPs, instead forcing such generators to participate in typical wholesale electricity markets. This is already occurring in various countries. In the United Kingdom, large wind power producers (WPPs) must now participate in conventional two-settlement electricity markets. In particular, they must provide bids in the form of supply functions in forward markets, and they are subject to ex-post financial penalties for deviations from contracted positions agreed upon ex-ante [2]. In the United States, PJM, the operator of the majority of the Eastern grid, and the New York ISO (NYISO) have instituted similar deviation penalties [55]. Additionally, the Midwest ISO (MISO) modified its market rules in 2011, allowing WPPs to offer power in real-time markets

and respond *voluntarily* to ISO dispatch signals [109].

Faced with imbalance penalties, WPPs will have to devise strategies for market participation to manage the quantity risk emanating from their power output. In this setting, the ability to curtail generation, a property of many recently built wind turbines, offers a promising method of avoiding certain financial penalties. Cost-minimizing strategies for WPPs bidding into electricity markets must take into account the benefits offered by curtailment capability.

In the long-term, renewable participation in electricity markets has implications for the markets themselves. The current market structure, like many power system operations, was designed with large dispatchable generation in mind. This typically consists of one forward market, typically cleared a day ahead of the delivery window, and an additional market for imbalance settlement cleared closer to the delivery window. As forecast uncertainty in renewable generation grows with the prediction horizon, the introduction of additional intra-day markets could help leverage better forecasts over shorter horizons.

These strategies will help address variability concerns for large WPPs that directly participate in electricity markets. However, different mechanisms are required to combat the variability introduced by smaller, distributed renewables. Integrating these small-scale resources can be achieved by leveraging the capabilities present in DERs [33]. We envision a system in which the variability in such generation is handled through intelligent control of local assets such as flexible loads and storage. This vision requires the development of algorithms for distribution network operation that tailor load to match variable generation. Moreover, implementing such schemes at the distribution network level will localize the effects of variability, thereby demanding less of the bulk transmission network. Eliciting load participation to help mitigate renewable variability will require new models for adequately compensating DERs for their services. Quantifying the impact of coordinated load aggregation on reserve requirements is a key step in developing any such pricing or other incentive schemes.

1.5 Dissertation Organization

Effective mitigation of renewable variability calls for a number of solutions across the entire spectrum of power system operation. In this dissertation, we identify three particular areas where techniques from control and operation can play a role in enabling large-scale renewable integration. First, we develop and benchmark real-time scheduling algorithms (inspired by receding horizon control and processor scheduling) for exploiting the flexibility offered by load and storage to manage variability in distributed renewables. Second, we use optimization techniques to quantify the impact of load deferrability on ex-ante power and capacity procurements. Third, we formulate and solve a profit maximization problem to understand how a wind power producer with curtailment capability might optimally participate in electricity markets.

In Chapter 2, we offer a brief description of aspects of power system operation pertinent to this dissertation. Specifically, we introduce the reader to the structure of wholesale electricity markets employed by many grid operators. We also detail the various types of reserves (procured on different time-scales) that ensure reliable grid operation.

In Chapter 3, we develop and analyze algorithms for the real-time scheduling of a resource cluster - a diverse collection of distributed energy resources. We focus on the impact *coordinated aggregation* of such DERs can have on reducing operating reserve costs. We first show that an optimal causal scheduling policy does not exist. As a result, we focus on various causal but necessarily sub-optimal scheduling heuristics. We develop four such causal algorithms drawing on ideas from the processor time allocation literature. We then develop a receding horizon control-based scheduling strategy explicitly configured to minimize reserve costs. We evaluate algorithm performance both in the metrics of reserve energy and capacity via simulation. We find that the majority of the operational benefits of DER scheduling can be realized at low levels of deferrable load participation and flexibility.

In Chapter 4, we study the impact of load deferrability on forward market operations. We find cost-minimizing ex-ante procurement policies in the cases of loads with and without deferrability. In order to obtain analytical characterizations of these policies, we assume the net load is normally distributed. First, we focus on the scenario where loads are inflexible. We find that, assuming no reserve capacity procurement, the optimal bulk power purchase is a quantile on procurement prices. Assuming a single, symmetric capacity for both up and down reserves, we can express this optimal ex-ante procurement policy on a partition of forward market and operating reserve prices. We then investigate the case with deferrable loads. Using an aggregate model for load deferrability, we derive a threshold policy for real-time reserve scheduling through dynamic programming. We follow with heuristics for load scheduling expressly aimed at minimizing reserve capacity requirements. Crucially, we offer a mathematical framework to link the effect of load scheduling decisions made within the delivery window, to forward market decisions made ex-ante. Finally, we quantify the reductions in ex-ante procurements offered by deferrability via simulation studies.

In Chapter 5, we quantify the operational benefits afforded to a wind power producer (WPP) by curtailment capability. We work within a stylized two-settlement electricity market framework in which the WPP is subject to imbalance penalties for deviations from a contracted position agreed upon in a day-ahead market. In this context, curtailment is a voluntary action of a profit-maximizing WPP. Working within this framework, we find optimal curtailment and contract offer policies. Specifically, we analytically express the optimal contract offer on a partition of expected imbalance prices. We then explicitly quantify the financial benefit of curtailment capability, and show that curtailment always *increases* the expected profit. We conclude with empirical calculations, using wind and time series price data, of this curtailment benefit at various locations within the NYISO and MISO footprints.

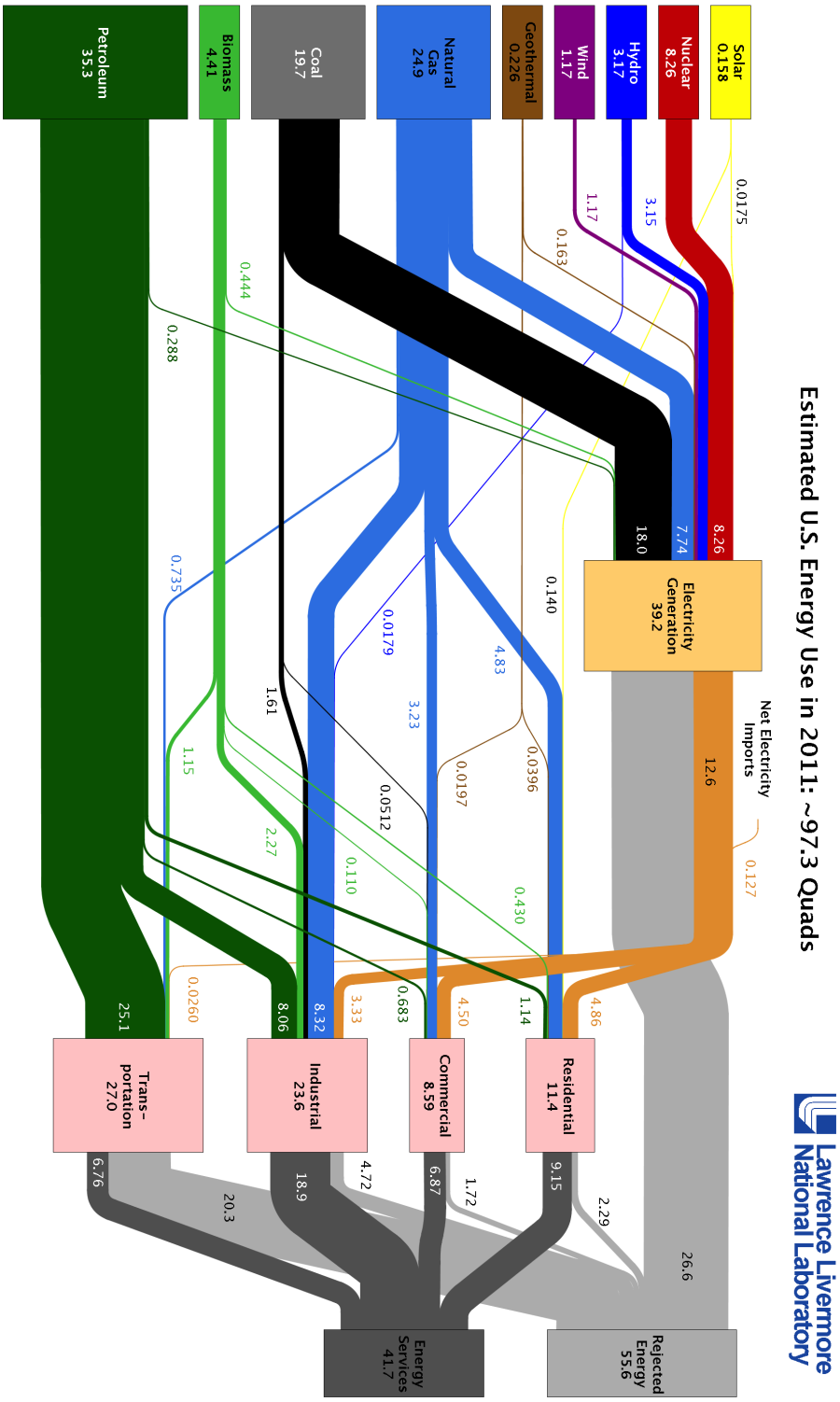


Figure 1.4: Sankey diagram showing breakdown of U.S. energy consumption by both energy source and demand sector. 1 quad is the equivalent of 2.93×10^{10} kWh or 1.055×10^{18} J. (Image Source: Lawrence Livermore National Laboratory, 2012)

Chapter 2

Background: Power Systems Operations

2.1 Electricity Markets

2.1.1 Introduction

The creation of wholesale electricity markets was a direct consequence of the deregulation of the power industry in the 1990s. Prior to this, the electricity industry was made up of highly regulated, vertically-integrated utilities that generated, transmitted, and distributed electricity to ratepayers. This regulated monopoly structure was replaced by one in which independent entities supply and consume electricity, the rationale being that the competition fostered in a deregulated environment reduces ratepayer prices [22]. Hence, wholesale electricity markets were created to facilitate the sale of electricity between interested suppliers and consumers.

Every market must be designed with reference to defining characteristics of the exchanged commodity. This is particularly true for electricity. Electrons are instantaneously delivered from power supplier to consumer with flow over a network governed by certain physical laws (Kirchoff's Voltage Law, AC Power Balance). There are also no cost-effective solutions for utility-scale energy storage. As consumers expect uninterrupted power on demand, electricity markets must be designed to help maintain near instantaneous power balance between electricity supply and demand.

The majority of electricity markets are mediated by a grid operator, also known as an independent system operator (ISO), charged with ensuring reliable power system operation over a geographic territory called a control area. These responsibilities include secure operation of the transmission network and instantaneous load balance at all times. To ensure that electricity transactions are made with these network and delivery constraints in mind, the ISO operates energy markets either in the form of power exchanges or bilateral markets. To ensure grid reliability requirements are met, the ISO also operates a set of ancillary service (AS) markets to procure various forms of reserve capacity.

While specific energy market rules and regulations vary across control areas, most consist of a series of forward markets enacted at different times leading upto the delivery window, and a market immediately preceding the delivery window for clearing energy imbalances [27, 28, 37]. Specifically, almost all ISOs operate a day-ahead (DA) market for scheduling electricity production and consumption over hour-long intervals, and a real-time (RT) market for reconciling imbalances between the procured DA schedules and actual loads while satisfying all applicable reliability criteria.

2.1.2 Day-Ahead (DA) Market

If an ISO operates the DA market as a centralized power exchange, it typically takes the form of a *uniform price auction*. We first outline the basic operating principles used in this type of market. We follow with some of the other main aspects of DA market operation.

Market Clearing: The market begins with the ISO soliciting individual offers to produce electricity from suppliers (generators), and bids to buy from consumers (load-serving entities) for each hour of the delivery date. These offers and bids are in the form of price-quantity functions relating the market clearing price to the quantity a supplier is willing to sell or consumer is willing to buy. The ISO receives offers and bids until a specified deadline in the late morning or early afternoon on the day before the intended power delivery. After this deadline, the ISO constructs aggregate supply and demand curves and clears the market at the intersection of these curves. Suppliers that offer to sell at prices below the clearing price are selected (unit commitment) and are given financially binding production schedules (economic dispatch). Consumers who have indicated, through their bids, a willingness to pay more than the clearing price, also receive binding schedules for electricity consumption. Under assumptions of limited market participant power, clearing the market in this manner results in prices and quantities with a number of desirable economic properties [19].

However, there is no guarantee that power flows resulting from the most economically optimal dispatch of generation and loads will satisfy line constraints. Hence, transmission network and reliability considerations demand departures from the previously outlined operating procedure. To enforce these constraints, the ISO uses a full network model to compute power flows and consequently evaluate bids and offers. A number of considerations go into determining the most economical bids and offers. First, the ISO removes from consideration offers that wield unfair market power as a result of their relative position in the network. The ISO then solves a series of constrained non-linear optimization problems to find the most economical allotment of bids and offers that satisfy network constraints and reliability constraints associated with pre-specified contingencies such as generator of line outages. This process is known as security constrained economic dispatch (SCED). The computed power schedules are then relayed to suppliers and consumers, who are now aware of their hourly generation and consumption commitments on the delivery date.

Bilateral Contracts: This market structure also accommodates suppliers and consumers who would rather engage directly with one another. In particular, market participants can engage in *bilateral contracts* where they independently negotiate a transaction quantity and price for power delivery between two bus locations within the network. As transmission capacity is constrained, the transacting parties must inform the ISO of the physical aspects of these contracts (transaction quantity, bus locations, etc.). The ISO can then factor these transactions into the SCED process and approve contracts that satisfy network, reliability, and security constraints. This mechanism can be used to create long-term power contracts agreed to months or even years ahead of the delivery date [66]. Bilateral contracting offers both suppliers and consumers flexibility in setting prices and volumes and the ability to manage exposure to uncertain prices. A bilateral market is one that attempts to match all load and generation solely with these types of contracts. Further details on bilateral contracts and markets can be found in [66].

Nodal Pricing: ISOs use nodal (or in some cases zonal) pricing to generate efficient prices that accurately reflect transmission constraints. Under nodal pricing, the ISO computes a distinct price, called the locational marginal price (LMP), at each bus within the network. Specifically, the LMP at a bus is the incremental system cost of meeting an additional MW of demand at that bus. The system cost is comprised of both the marginal energy costs of generation as well as transmission congestion charges. These LMPs can be interpreted as the Lagrange multipliers associated with nodal power constraints in the economic dispatch optimization. In practice, the ISO finds LMPs as a by-product of solving optimization problems as part of SCED. Market participants are subject to LMPs at their respective nodes in both DA and RT markets.

As power constraints directly affect LMPs, variations in network power flow patterns can cause significant price volatility. For instance, a small change in power flows that result in a previously uncongested line reaching its rated limit can trigger large perturbations in prices. Moreover, prices at disparate nodes in the network are linked with one another making price forecasting a difficult proposition [20]. To hedge against the financial risks created by LMPs, market participants typically procure instruments called financial transmission rights (FTR). By purchasing a FTR, the market participant effectively insures against rapid rises in congestion costs. There are auctions dedicated to the sale and purchase of FTRs [7].

Capacity Products: In addition to quantifying demand through consumer bids, the ISO also employs forecasts of total demand in DA operations. Based on these load forecasts, ISOs procure additional generation capacity to meet any remaining demand unaccounted for in consumer bid-based market clearing. This process has different names that vary across control areas – for instance, it is called residual unit commitment in the CAISO markets [29]. Under CAISO rules, generators that commit to providing capacity as part of residual unit commitment must participate in the subsequent real-time (RT) market.

The demand forecast is also used to determine the capacity requirements for reserves over varying time-scales. This is done in ancillary service (AS) markets where ISOs procure capacities of different reserve products such as load-following, regulation, and contingency reserves. Section 2.2 contains detailed descriptions of these reserve products. In AS markets reserve capacities are sold as call options – suppliers receive an ex-ante payment for making generation available for dispatch. These generators receive an additional payment if reserves are dispatched during the delivery window. The rules outlining reserve capacity requirements, which are set by federal and regional regulatory authorities, are often expressed as percentages of load forecasts [95].

2.1.3 Real-Time (RT) Market

As the delivery window approaches, the ISO may operate additional forward markets (e.g. an hour-ahead market) to take advantage of more reliable load and generation forecasts. Regardless of the number of forward markets, the ISO will operate a real-time (RT), or spot market to reconcile differences between electricity generation and consumption schedules agreed to ex-ante and actual system conditions. At this stage, the ISO also determines the reserve dispatch necessary to maintain instantaneous load balance. Typically, the RT market is cleared separately for each 5 minute delivery window.

Market participants use the RT market to modify their ex-ante contracted positions to account for output deviations caused by demand uncertainty, and unplanned generation or transmission outages. Similar to the DA market, the ISO generates nodal prices that reflect both generation and congestion costs. The market participants settle output deviations with respect to DA schedules at these RT market prices. Prices in RT markets are typically more volatile than those seen in DA markets [37]. This is a direct consequence of the smaller volumes transacted in RT markets as compared to DA markets. With these limited volumes, there is an increased chance of an imbalance between supply and demand causing dramatic LMP price spikes and troughs. As mentioned earlier, market participants can insulate themselves from sudden price variations using FTRs.

2.2 Reserves

Many key components of power system operation, such as load requirements, generator output, and transmission network, are subject to rapid and unpredictable changes. In the face of such variability, wholesale energy markets alone are not sufficient to maintain reliable system operation. System operators, who are ultimately responsible for maintaining reliability, must procure additional capacity that can assist in balancing load and generation when requested. Reserves, or ancillary services, refer to these additional generation and demand-side resources procured to counteract power imbalances on various time-scales. In this section, we only discuss reserves required to counteract real power imbalances and do not discuss other ancillary services, such as reactive power supply for voltage support.

Reserve requirements are set by federal and regional regulatory bodies such the North American Electric Reliability Corporation (NERC) and the Western Electricity Coordination Council (WECC). These reliability criteria outline distinct procurement targets for various types of reserves. In determining these targets, reserves are classified in many ways. Indeed, the definitions used to distinguish between different reserve products vary across control areas. To differentiate between the various forms of reserve, we will focus on three properties of the power imbalance underlying the need for reserves.

1. **Imbalance direction**: Do power imbalances correspond to system over- or under-generation conditions?
2. **Time-scale**: What response speed and duration can reserves exhibit to adequately counteract imbalances?
3. **Imbalance cause**: What is the primary cause of the power imbalance? Specifically, are they caused by events that occur periodically in the course of normal operation, or are they the result of a rare contingency?

2.2.1 Imbalance Direction

Based on the imbalance direction, all reserves can be classified into one of two categories: *up* and *down reserves*. **Up reserves** refer to additional generation used to meet load requirements when scheduled generation alone is insufficient. In addition to generators that can increase output on demand, loads that can momentarily reduce consumption can also provide up reserves. In contrast, **down reserves** refer to those procured in situations where scheduled generation exceeds load. This service can be provided either by generators able to curtail output, or loads that can increase consumption. Historically, there has always been a greater need for up reserves as outages or disconnections of generators represented the primary source of large-scale power imbalances [51]. Sudden disconnection of large loads occur far less often. However, increased adoption of renewables promises to alter the relative importance of up and down reserves.

2.2.2 Time-Scale

Reliability requirements typically characterize different types of reserves based on their response speed and duration. Figure 2.1 shows the various types of reserves deployed to meet real power imbalances at different time-scales. Clearly, effective load balancing requires reserve scheduling on a variety of time frames. This ranges from addressing random load variations on the order of seconds, to counteracting more protracted under-generation situations caused by transmission or generator outages. Moreover, temporal characteristics of a given power imbalance limit the choice of resource that can provide relief via reserves. For instance, a natural gas-fired peaking plant, with fast ramping capabilities, is better suited to providing quick responses to load fluctuations than a coal-fired power plant.

Frequency Responsive Reserves: This refers to fast reserve capability which responds to imbalances on time-scales of a few seconds. In typical power systems, synchronous generators are the primary source of this type of reserve. As the name 'frequency response' suggests, these generators use local measurements of system frequency to address real power imbalances. This type of reserve is also known as governor control or primary reserves.

The relationship between frequency and real power is best illustrated with an example. Consider a synchronous generator in an under-generation (i.e. load exceeds generation) scenario. Serving the additional load reduces the kinetic energy present in the generator's rotating machinery. This causes rotor speed or frequency to fall below nominal levels. Conversely, an over-generation scenario causes frequency to exceed the nominal system frequency. Generators providing frequency response reserves sense these deviations and correct them by adjusting their actuator setpoints (i.e. opening steam valves to increase generator output). To relate deviation measurements to actuator inputs, generators typically use a proportional controller with a small (0.010-0.040Hz) deadband around the nominal system frequency [123]. This deadband prevents unnecessary control actions in response to very small frequency deviations – control actions that would otherwise reduce the operating life of the generator. Due to these deadbands, frequency response is often used only to counteract sudden, large contingency events rather than normally occurring variations. However, this is not true in certain isolated power networks like those in large islands (e.g.: UK and Ireland), where this type of reserve is used to balance small, chronic load fluctuations [51].

Automatic Generation Control (AGC): This refers to a fast form of reserve capability which counteracts imbalances on the order of seconds to 5 minutes. In large interconnected systems, these reserves are provided by generators that automatically receive output setpoints every few seconds from an operator balancing load requirements within that unit's control area. These dispatch instructions are based on a metric known as the Area Control Error (ACE) which reflects both the overall control area power imbalance as well as deviations in system frequency from nominal levels. This centralized response results in a coordinated response that effectively maintains grid reliability. The automated nature of reserve dispatch distinguishes these reserves from frequency response reserves, in which generators respond to imbalances autonomously. Clearly, resources providing regulation capability must be outfitted with appropriate communication channels to receive dispatch signals.

Spinning Reserves: This refers to reserves that are synchronized to grid frequency (i.e. on-line) and capable of increasing its output within 10 minutes of a dispatch instruction [37]. Generators providing spinning reserves operate below their rated limits and are thus capable of increasing their output upon request. While regulation and frequency response reserves are also provided by synchronous generators and therefore are technically 'spinning reserves', the term spinning reserves usually refers to other synchronous generators that can

respond on time-scales of a few minutes. Due to the absence of start-up constraints, various energy resource modalities can provide spinning reserves.

Spinning reserves are necessary to continuously meet loads in the event of sudden outages of load fluctuations on the order of minutes. Moreover, these reserves assume additional importance in the context of renewable variability where sudden ramps in power output threaten to create lasting power imbalances. In this context, having generators ready to modify their output on command without any start-up delays is a foolproof way to ensure reliable grid operation. Indeed, reliability criteria in many regions make specific mention of spinning reserves. Examples include the WECC, which calls for spinning reserve to account for at least 50% of the total operating reserve requirement [95], and the mid-Atlantic region of PJM, where spinning reserves capacity must equal 75% to 100% of the largest on-line generator's output [85]. Spinning reserve can also be provided by responsive loads willing to curtail or increase load and providing metering information to audit their actions.

Non-Spinning Reserves: This refers to reserves that are initially off-line but able and available to come on-line and generate its rated capacity within 10 minutes of a dispatch instruction [37]. Moreover, these units must be able to maintain that desired output for at least two hours. Unlike spinning reserves, which can be provided by a variety of energy resources, non-spinning reserves can only be offered by generators which can ramp up from off-line to full production mode within a few minutes. As a result, hydro-electric, natural gas, and diesel power plants are the main types of generators capable of providing non-spinning reserve capability [105]. Demand response from responsive loads can also provide these reserves.

As they operate on similar time-scales to spinning reserves, both non-spinning and spinning reserves can respond to similar situations. Indeed, they are sometimes jointly referred to as secondary reserves. However, due to the start-up costs associated with bringing non-spinning reserves online, system operators prefer to exhaust all avenues of correcting imbalances using spinning reserves first. As a result, non-spinning reserves are intended for use primarily when there is a sustained power imbalance [51].

Supplemental Reserves: This refers to reserves that need to provide generation within 30 minutes of receiving a dispatch command. In that time, generators providing supplemental reserves must be synchronized with system frequency and be able to provide an agreed upon amount of generation. Supplemental reserves relieve non-spinning and spinning reserves of their duties thereby enabling those reserve generators to return to their original positions. This ensures reserve response capability for a subsequent reliability event. These reserves are also referred to as tertiary reserves.

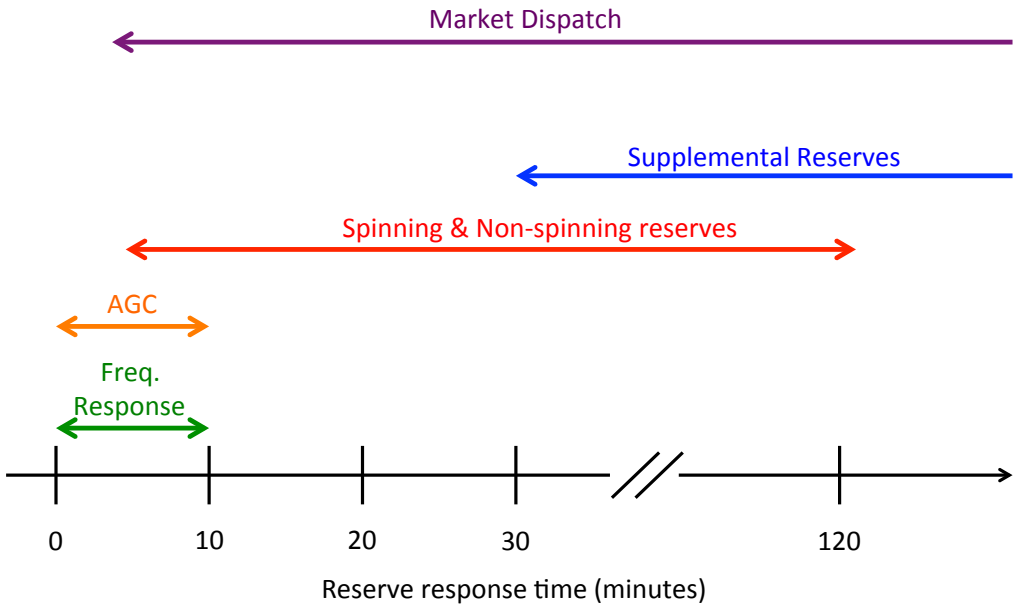


Figure 2.1: Figure showing relative time-frames on which different types of reserve are deployed. (Version of Figure 3 in [51])

2.2.3 Imbalance Cause

We classify power imbalances into two broad categories, with each reflecting its underlying cause: customary and abnormal imbalances. Customary imbalances are those that occur frequently and continuously in the course of normal power system operation. This includes load deviations from forecasted values used in forward markets, which require response on the order of minutes to hours, to small fluctuations on the order of seconds caused by load variability. Such power imbalances are common, and system operators typically address them on most days. Abnormal imbalances correspond to severe events that occur rarely under normal operating conditions. Examples of severe events include single or multiple-phase transmission line faults and generator outages. Unlike customary imbalances, system operators seldom observe these phenomena. These delineations, detailed extensively in [51], help elucidate key reserve concepts in the context of renewable integration.

Customary Imbalances

Reserve characteristics differ depending on the time-scale on which customary imbalances occur. There are two classes of reserves that address customary imbalances: regulation and load-following reserves. Regulation reserves tackle near-instantaneous load balancing needs while load-following reserves address the needs of more slowly developing customary imbalances.

Regulation Reserves: These reserves primarily address small variations in load and generation occurring on time-scales under 5 minutes – phenomena which occur too quickly for market-based recourse. Due to the need for rapid response, regulation reserves are provided by generators providing AGC and frequency responsive reserves [37]. For most large systems consisting of interconnected control areas, system operators prefer AGC-based regulation reserves over autonomous governor control because AGC offers better control of power flows between control areas. Moreover, frequency deviations resulting from minor imbalances in such large systems are too small for effective employment of frequency responsive reserves. As these are not major considerations for smaller, isolated power systems, regulation reserves in those systems are often provided by frequency responsive generation. Notice that reserves on other time-scales, like non-spinning and supplemental reserves, are not quick enough to deal with these rapid imbalances.

Regulation reserves are used to address the variable nature of renewables. This is particularly true for distributed solar PVs, which can exhibit intermittent output when affected by random weather fluctuations occurring on rapid time-scales such as passing cloud cover. Errors in short-term (under 5 minutes) forecasts of renewable generation also create a need for regulation reserves. These forecasts are used when dispatching generation in the RT market. Hence, forecast errors can cause inaccurate generation setpoints resulting in power imbalances.

Load-Following Reserves: These reserves help balance load and generation on longer time-scales than regulation reserve. Typically, these imbalances represent deviations from conditions used to dispatch resources in the forward market, and are larger than those addressed via regulation reserves. If the system operator expects such imbalances, it dispatches load-following reserves directly or as part of the RT market clearing process [51]. Load-following reserves are typically provided by a combination of spinning and non-spinning reserves. For protracted imbalances, supplemental reserves may also be called to perform load-following.

Renewable variability significantly increases the need for load-following reserves. These reserves are the primary mechanism to deal with intra-hour fluctuations in renewable output that can only be forecast after forward market clearing. As these fluctuations are sizeable, this greatly increases the amount of load-following reserve necessary to maintain reliability. Moreover, uncertainty in the hourly wind and solar forecasts used in forward markets also result in imbalances that must ultimately be addressed with this type of reserve. In addition to increased reserve capacity requirements, resources providing reserve capacity will have to modify their operating setpoints more frequently to battle the consistent intermittency in renewable output. This places additional strain on the resources providing load-following reserves.

Abnormal Imbalances

As with customary imbalances, reserves required to counteract abnormal imbalances can be classified into two groups based on their response speed. Contingency reserves address sudden, instantaneous abnormal imbalances while ramping reserves deal with abnormal imbalances occurring on slower time-frames.

Contingency Reserves: These reserves respond to instantaneous, critical system events that can compromise overall grid reliability. Canonical examples of these abnormal events are sudden losses of large amounts of generation due to line or generator outages. Such events require a concerted response from many reserves on various time-scales. Initially, the resulting power imbalance triggers frequency responsive reserves. This is typically followed by the deployment of spinning and non-spinning reserves to make up for the generation shortfall. Finally, the system operator may call on supplemental reserves to replace the other forms of reserve. Particulars of the response time required for each of these steps vary between control areas [51].

Renewable variability only has a limited impact on contingency reserves [55]. Certainly, momentary disconnection of large-scale solar thermal or wind facilities would classify as a contingency in the same way it does for any other form of generation. Recently, there has been also some discussion of using contingency reserves in response to wind ramps [51].

Ramping Reserves: These reserves respond to rare, large-scale changes in power balance that are not instantaneous. For instance, ramps (sudden, sizeable changes in load or renewable output) occur on the order of a few minutes to hours and can, in certain cases, be forecast in advance. Resources responding to such events must be able to respond quickly to events.

The concept of ramping reserves is relatively new. Indeed, most system operators, with the exception of those in California (CAISO) and the Midwest (MISO), have not broached the subject of specific ramping reserves [55]. They instead use load-following reserves to attend to such behavior. Creating a dedicated ramping reserve product will explicitly reward resources for their ability to change output quickly. These reserves assume added importance in the context of renewable variability. Indeed, MISO envisions ramping reserves as an ancillary service targeted at addressing the reliability concerns posed by wind ramps [55].

Chapter 3

Real-time Scheduling of DERs

3.1 Motivation

Increasingly, renewable energy sources deployed at the distribution level such as rooftop solar photovoltaics (PVs) or small-scale wind farms make up a sizeable fraction of newly installed renewable generation capacity [53, 78]. This represents a paradigm shift for grid operators generally accustomed to dealing with large, conventional sources of generation enjoying direct power connections to the bulk transmission network.

Grid operations must address the inherent variability associated with such distributed generation. As discussed in Section 1.3, distribution level assets offer the potential of absorbing some of this variability. Fully realizing the operational benefits afforded by these resources requires *coordinated aggregation* of their capabilities. Coordinated aggregation refers to the intelligent control of resources such as deferrable loads and available storage to match variable generation. By directly accounting for and managing generation variability, it promises to effectively mitigate the increased reserve costs of deep renewable penetration.

Coordinated aggregation schemes broadly fall into two classes:

1. **Indirect load control (ILC)**: DERs respond, in real-time, to price or appropriate proxy signals that induce desired resource behavior. In this scenario, the price signals would indicate generation conditions.

2. **Direct load control (DLC)**: DERs cede physical control of devices to operators who determine and execute appropriate actions based on knowledge of operational system conditions.

In both cases, the central problem is management of a *resource cluster*. This is a diverse collection of networked resources at the distribution level including renewable and micro-generation, deferrable and non-deferrable loads, and electricity storage. In DLC, each resource cluster is managed by a dedicated *Cluster Manager* (CM) which coordinates operation of the various distribution level assets. Under this paradigm, the CM:

1. participates in ex-ante markets to procure generation to meet the cluster's load requirements,

2. performs real-time resource scheduling and reserve procurement, and

3. aggregates the cluster's capabilities and presents them to the system operator (SO) as a dispatchable resource.

This hierarchical architecture of CM-based coordinated aggregation is necessary as centralized, system-wide aggregation of DERs involves prohibitive computational costs [60], and falls outside the scope of existing SO business models. We focus on the CM's real-time scheduling function and analyze its impact on *operating reserves*. In this context, operating reserves refers to load-following reserves, procured in real-time energy markets, and regulation reserves on time scales of a few minutes. In particular, this does not include any contingency reserves dispatched in response to rare and severe events.

Our contributions are as follows. We develop four resource scheduling algorithms inspired by well-known scheduling heuristics in the context of processor time allocation – earliest deadline first (EDF), least laxity first (LLF), deadline prioritized adjusted scheduling (DPAS), and laxity prioritized adjusted scheduling (LPAS). We also formulate a receding horizon control (RHC) scheduling algorithm that explicitly accounts for operating reserve costs. We then present simulation studies to assess the performance of these resource scheduling policies in the metrics of reserve energy and capacity costs. Through these simulations, we also quantify the marginal benefits of deferrable load penetration and of load scheduling flexibility. We conclude that resource scheduling under *any* algorithm offers compelling reserve energy cost reductions while only RHC-based scheduling consistently offers significant reserve capacity cost reductions. We conclude that, while reserve cost reductions offered by coordinated aggregation depend on the statistical nature of the renewable generation process, *the majority of these benefits can be realized at modest levels of deferrable load penetration and load deferrability.*

The development and analysis of coordinated aggregation, and more generally demand response (DR) [5] schemes is an active research area. Numerous recent studies have delved into aspects of DR, such as its impact on electricity markets [23, 117] and DR program implementation analysis [42, 43]. A detailed treatment of all DR research directions is beyond the scope of this study. Instead, we focus on the sub-area of coordinated aggregation, where studies can be broadly classified based on the mode of load control. Recent research in ILC has focused on modeling [79] and developing algorithms [38, 44, 40] for consumer response to electricity price signals. There have also been studies focused on the economic efficiency [21, 115] and stability of price signals [111], as well as the practical issues associated with implementing ILC programs [14]. Current research in DLC focuses on developing and analyzing tractable algorithms for coordinated aggregation [77, 70, 90]. A wide variety of algorithms have been explored in the literature. Gan et al. [61] develop a distributed scheduling protocol for electric vehicle charging. Papasiviliou and Oren [99] use approximate dynamic programming to couple wind generation with deferrable loads. Closer in spirit to

our work are the following studies. In [35], the authors develop laxity-based heuristics for accepting and completing electric vehicle charging requests in a parking garage. In [36], the authors determine electricity consumption schedules for a residential consumer's deferrable loads in the face of price uncertainty and constraints on deferrable load energy delivery. We also note that a number of recent studies [83, 71, 60, 86] have suggested the use of RHC approaches for resource scheduling. Our contribution is not in suggesting the usage of RHC control strategies, but rather in *the development of specific cost functions which explicitly minimize operating reserve costs.*

The remainder of this chapter is organized as follows. In Section 3.2, we outline the problem formulation in which we describe generation and load models, introduce necessary terminology, and outline a scheduling policy optimization problem. In Section 3.3, we develop and describe various algorithms for resource scheduling and reserve procurement algorithms. Section 3.4 contains results from simulations studies comparing the different algorithms and quantifying the marginal benefit of deferrability. We conclude and outline future research directions in Section 3.5.

3.2 Problem Formulation

We focus on the problem of meeting a resource cluster's load requirements over an operating window. Without loss of generality, let this operating window be $[0, T]$. We assume load balancing occurs N times within this window at times indexed $k \in \{1, 2, \ldots, N\}$. These balancing times k occur periodically within the operating window every $\Delta t = \frac{T}{N}$ hours.

At each time k, the CM performs two functions. First, the CM allocates generation to the various deferrable and non-deferrable loads. This is called *resource scheduling.* Second, the CM determines the amount of reserves (r_k) required to meet load requirements. As we ignore the impact of rare and severe events, all energy imbalances are managed solely with operating reserves. This amount of operating reserves is determined at each time k and is constant for the following Δt time interval, until the next opportunity to balance load and generation. This is called *reserve scheduling.*

3.2.1 Generation Modeling

A CM, tasked with meeting load requirements of a resource cluster, will have access to a wide variety of generation resources. In our analysis, there are three categories of generation resources chosen based on generation variability, and CM operational considerations. Specifically, the CM can procure generation from:

1. Renewable generation
2. Bulk power
3. Operating reserves

Renewable generation

Renewable generation refers to all power generated from within the CM's resource cluster. The CM must accept all such generation to meet load requirements. This includes rooftop solar PVs, small-scale wind farms, and residential natural gas or diesel generators. Let

$$\mathbf{w} = \{w_k\}_{k=1}^{N} \tag{3.1}$$

denote the sequence of renewable generation realizations during the operating window. We assume this generation is free (zero marginal cost) from the CM's perspective but exhibits significant variability. The CM must absorb this variability through either resource scheduling or scheduling operating reserves.

Bulk power

Bulk power refers to generation procured from the bulk transmission system through previously agreed-upon contracts in forward markets. We assume there is no uncertainty associated with bulk power delivery. In reality, bulk power delivery is fairly certain and deviations from bulk power schedules are only caused by generator failures or transmission line outages. Let

$$\mathbf{b} = \{b_k\}_{k=1}^{N} \tag{3.2}$$

denote the sequence of bulk power deliveries at different points within the operating window. In line with typical power system operations [27, 41] [28, pg. 7-6], we model $\mathbf{b}$ as constant over hour-long blocks. The CM purchases separate amounts of bulk power for each hour within the operating window.

Operating Reserves

Operating reserves refer to generation dispatched to ensure load requirements are met at each balancing time k. This refers to generation procured in real-time energy and ancillary service markets cleared every δt hours. This generation ensures that load requirements are met exactly at each balancing time k. Let

$$\mathbf{r} = \{r_k\}_{k=1}^{N} \tag{3.3}$$

denote the sequence of operating reserves dispatched during the operating interval. We assume the amount of operating reserves dispatched is constant over intervals of length Δt. We do not model additional operating reserves for load-balancing at finer time-scales. In reality, these minor energy imbalances are dealt with via additional ancillary services such as frequency regulation [27]. Notice that both up ($r_k > 0$) or down ($r_k < 0$) reserves can be modeled within this framework.

These are two costs associated with operating reserves: (1) energy and (2) capacity.

1. The CM pays a price p_r (\$/MWhr) for each unit of reserves dispatched for load-balancing. We assume energy costs are symmetric - the same price applies to both up and down reserves.

2. The CM pays the same price p_C(\$/MW) for the maximum instantaneous reserve dispatch (up or down) required for each unit of reserve capacity needed for adequate load-balancing. We assume capacity costs are symmetric.

While we not consider the case of asymmetric reserve energy and capacity costs, the results and algorithms developed here can be readily extended to handle asymmetric prices.

3.2.2 Resource Modeling

A resource cluster consists of a wide variety of heteregenous loads. These may include: (1) loads that offer no scheduling flexibility, (2) loads whose energy consumption can be deferred such as thermostatically controlled loads (TCLs) and plug-in electric vehicles (EVs), and (3) residential-scale energy storage. Detailed load models can be constructed to various degrees of fidelity and load modeling of deferrable loads is, in its own right, an active research area. Various types of TCL models have been developed for demand response such as hybrid linear systems [88], non-linear systems [39], and circuit-equivalent models [97]. There has also been work focusing on EV battery modeling [107]. For our purposes, it suffices to categorize loads as either:

1. Static loads

2. Deferrable loads

3. Storage

Static loads

Static loads require a specified power at each time and thus have no flexibility associated with their power demands. Let the power profile

$$\mathbf{l}^S = \left\{ {l^S}_k \right\}_{k=1}^{N} \tag{3.4}$$

be the aggregate power requirement of all such loads within the operating window. The CM must ensure that adequate generation is present to satisfy this requirement at each balancing time k.

Deferrable loads

Deferrable loads only require delivery of an certain amount of energy over a specified time interval. The energy needs of these loads are modeled as *tasks*.

Definition 3.1. *A **task** T_i is fully characterized by an energy requirement (E_i) that must be delivered over a service interval $\{a_i, \ldots, d_i\}(a_i \subseteq \{1, \ldots, N\}, d_i \subseteq \{2, \ldots, N+1\})$ while respecting a maximum power transfer rate limit (m_i).*

Let p_{ik} be the power delivered to Task T_i at time k. We can express this tasks' requirements as

$$\sum_{k=a_i}^{d_i-1} p_{ik}\,\Delta t = E_i, \quad 0 \le p_{ik} \le m_i \;\; \forall k \in \{a_i, \ldots, d_i - 1\}. \tag{3.5}$$

With this model (3.5), we have made the following assumptions regarding tasks:

A1 Tasks are *pre-emptive*, that is task servicing can be interrupted at any time within the service interval.

A2 The power delivered to a task admits values on the continuous interval $[0, m_i]$.

The ability to continuously modulate power delivery (Assumption A2) is uncommon in traditional distribution networks where discrete power delivery levels are more prevalent. However, this is a reasonable approximation in determining power allocations. Crucially, it enables the computation of power allocation schedules using convex optimization methods. For practical implementations of these algorithms, we suggest rounding p_{ik} to the nearest discretized power level.

Thus, the degree of deferrability of each task T_i is succinctly parametrized by (E_i, m_i, a_i, d_i). For an EV, E_i corresponds to a user-specified state-of-charge increase over a charging interval $[a_i, d_i]$. The rate limit m_i corresponds to constraints imposed by the charging equipment or power delivery infrastructure. For a TCL, we can construct a simple duty cycle model where a certain amount of energy E_i is required every W hours. This energy requirement E_i depends on exogenous variables such as ambient temperature, and user comfort. In this case, the service interval can be expressed as $[nW, (n+1)W]$ where n is an integer. The rate limit m_i is, once again, determined by the power delivery infrastructure.

We now define some key quantities which will be used in this chapter.

Definition 3.2. *The **energy state** of a task T_i, parametrized by (E_i, m_i, a_i, d_i), at time k is:*

$$e_{ik} = E_i - \sum_{n=a_i}^{k-1} p_{in}\Delta t \;\; \forall k \in \{a_i + 1, \ldots, d_i\}, \tag{3.6}$$

where p_{ik} is the power delivered task T_i at time k. Let $e_{ia_i} = E_i$.

The energy state is the remaining energy requirement at the *start* of the k^{th} time-step.

Definition 3.3. *A task T_i is **active** at time k if $a_i \le k < d_i$, and $e_{ik} > 0$. Let $\mathbb{T} = \{T_i\}_{i=1}^{M_T}$ denote the collection of all tasks within this operating window. We define $\mathbb{A}_k (\subseteq \mathbb{T})$ as the set of active tasks at time k.*

An active task is one that can and needs to be serviced at a given time.

Let the power profile

$$\mathbf{l}^D = \{l_k^D\} \tag{3.7}$$

be the total power delivered to deferrable loads over the operating window. Clearly, the power delivered to deferrable load at time k is characterized solely by power allocations to active tasks.

$$l_k^D = \sum_{i \in \mathbb{A}_k} p_{ik} \tag{3.8}$$

where p_{ik} is the power delivered to Task T_i at time k.

A special case of this profile is the *nominal deferrable load profile*.

Definition 3.4. *The **nominal rate** of a task T_i, parametrized by (E_i, m_i, a_i, d_i), is:*

$$q_i = \frac{E_i}{d_i - a_i} \tag{3.9}$$

Servicing a task at this nominal rate over the entire service interval guarantees task completion.

Definition 3.5. *The **nominal deferrable load profile** $\mathbf{l}^D = \{l_k^D\}$ is the deferrable load profile resulting from servicing each task at its nominal rate. For a set of tasks $\mathbb{T}$, this quantity can be readily computed:*

$$l_k^D = \sum_{i \in \mathbb{A}_k} \frac{E_i}{d_i - a_i} \quad \forall k \in \{1, \ldots, N\} \tag{3.10}$$

The nominal deferrable load profile is used extensively in our analysis. Apart from its role in the development of two real-time scheduling algorithms, it serves as a baseline for computing the reserve reductions possible through resource scheduling.

Storage

We model electricity storage as *devices*.

Definition 3.6. *A **device** is characterized by maximum (E_j^+) and minimum (E_j^-) energy capacities, and maximum charging (m_j^+) and discharging (m_j^-) rate limits. Let p_{jk} be the power delivered to Device D_j at time k. Device requirements can then be expressed as:*

$$E_j^- \leq \sum_{n=1}^{k} p_{jn}\, \Delta t \leq E_j^+, \quad -m_j^- \leq p_{jk} \leq m_j^+ \;\; \forall k \in \{1, \ldots, N\}. \tag{3.11}$$

Each device D_j can be characterized by $(E_j^+, E_j^-, m_j^-, m_j^-)$. For a storage device with energy capacity E and an initial state of charge E_0, the maximum and minimum energy capacities can be easily computed ($E_j^+ = E - E_0$, $E_j^- = -E_0$). Let $\mathbb{D} = \{D_j\}_{j=1}^{M_D}$ denote the collection of all devices within this operating window. As the service interval for a device is the entire operating window, devices are always active.

Clearly, tasks and devices differ in their energy requirements. For a task, the energy requirement is a single equality constraint in the space of power allocations. For a device, the corresponding requirement is a set of $2N$ linear inequality constraints on the space of power allocations.

3.2.3 Scheduling Policies

The CM must satisfy all load requirements over the entire operating window. Before any resource or reserve scheduling, the CM attempts to meet static load requirements with renewable generation and bulk power. The remaining generation, used to satisfy deferrable load requirements, is called *available generation*.

Definition 3.7. ***Available generation*** *at time k is the renewable and bulk power remaining after satisfying static load requirements. Let*

$$\mathbf{g} = \{g_k\}_{k=1}^{N}, \quad g_k = w_k + b_k - {l^S}_k \tag{3.12}$$

denote the available generation profile over the operating window.

Scheduling policies dictate the allocations of available generation to tasks and devices. They also determine the reserve generation necessary to fulfill all load requirements.

Definition 3.8. *A* ***scheduling policy*** *σ is an algorithm that computes reserves, and allocations of available generation and reserves to complete all task requirements over the operating window. Specifically, the scheduling policy is a function of an available generation profile $\mathbf{g}$, and collections of tasks $\mathbb{T}$ and devices $\mathbb{D}$.*

$$\sigma(\mathbf{g}, \mathbb{T}, \mathbb{D}) = (\mathbf{r}, \{\mathbf{p}_k\}_{k=1}^{N}), \tag{3.13}$$

where $\mathbf{p}_k$ is the set of power allocations at time k to all active tasks $\{p_{ik}\}_{i \in \mathbb{A}_k}$, and devices $\{p_{jk}\}_{j \in \mathbb{D}}$. As this policy must ensure that generation and load are balanced within the operating window, it must satisfy:

$$g_k + r_k = \sum_{i \in \mathbb{A}_k} p_{ik} + \sum_{j \in \mathbb{D}} p_{jk} \;\; \forall k \in \{1, \ldots, N\}. \tag{3.14}$$

Certain available generation profiles can be scheduled to meet all task requirements without the need for any reserves. This property is referred to as *feasibility*.

Definition 3.9. *An available generation profile* $\mathbf{g}$ *is* ***feasible*** *with respect to a set of tasks* $\mathbb{T}$ *if there exists some scheduling policy* σ *that completes all tasks without dispatching any reserve generation. Specifically, for this policy*

$$\sigma(\mathbf{g}, \mathbb{T}, \mathbb{D}) = (\mathbf{0}, \{\mathbf{p}_k\}_{k=1}^{N}), \tag{3.15}$$

where $\mathbf{p}_k$ *ensures that all task requirements are met*

$$e_{id_i} = 0 \ \ \forall T_i \in \mathbb{T}. \tag{3.16}$$

A scheduling policy σ is *causal* if its allocations at time k depends only on the *information state* $\mathcal{I}_k$ at time k.

Definition 3.10. *The* ***information state*** $\mathcal{I}_k$ *at time* k *consists of:*

1. *Task parameters of tasks active at time* k*:* (E_i, m_i, a_i, d_i) $\forall T_i \in \mathbb{A}_k$*,*
2. *Device parameters of all devices:* $\left(E_j^-, E_j^+, m_j^-, m_j^+\right)$ $\forall D_j \in \mathbb{D}$*,*
3. *Energy states of tasks active at time* k*:* e_{ik} $\forall T_i \in \mathbb{A}_k$*,*
4. *Realized values of available generation:* $\{g_n\}_{n=1}^{k}$*.*

3.2.4 Cost Metric

The performance of different scheduling policies is assessed by comparing the cost of procuring reserve generation. As mentioned previously, there are two components to this cost:

$$J(\mathbf{r}) = p_r \sum_{k=1}^{N} |r_k| \Delta t \ + \ p_C \max_k |r_k| \tag{3.17}$$

The first term is an energy cost; it penalizes the total amount of reserve generation procured to meet load requirements. The second term is a capacity cost; it captures the non-energy costs associated with making generators that provide reserve generation available. In practice, such reserve *capacity* is procured either through bilateral contracts, where the CM enters into long-term agreements with select generating facilities, or through the clearing of ex-ante capacity markets [26]. This capacity cost also discourages undesirable peaks in power delivery that may place stress on distribution networks. As our primary focus is to quantify the impact of real-time scheduling on reserve costs, we do not incorporate the procurement costs associated with bulk power.

3.2.5 Policy Optimization

Our objective is to develop scheduling policies that reduce reserve costs as captured by the metric (3.17). This is a *functional* optimization problem (3.18) over the function space of scheduling policies.

$$
\begin{aligned}
\min_{\sigma} \quad & J(\mathbf{r}) && (3.18)\\
\text{subject to:} \quad & (\{\mathbf{p}_k\}_{k=1}^N, \mathbf{r}) = \sigma(\mathbf{g}, \mathbb{T}, \mathbb{D}),\ \forall k \in \{1, \dots, N\} && \text{(a)}\\
& \forall k, g_k + r_k = \sum_{i \in \mathbb{A}_k} p_{ik} + \sum_{j \in \mathbb{D}} p_{jk}, && \text{(b)}\\
& \forall T_i \in \mathbb{T},\ \sum_{k=a_i}^{d_i - 1} p_{ik} \Delta t \ = \ E_i, && \text{(c)}\\
& \qquad 0 \le p_{ik} \le m_i\ \forall k \in \{a_i, \dots, d_i - 1\}, && \\
& \forall D_j \in \mathbb{D},\ E_j^- \le \sum_{n=1}^{k} p_{jn} \Delta t \ \le \ E_j^+, && \text{(d)}\\
& \qquad -m_j^- \le p_{jk} \le m_j^+\ \forall k \in \{1, \dots, N\}. &&
\end{aligned}
$$

Constraints

1. **Scheduling Policy**: (3.18-a) defines the power allocations to tasks and devices and reserve generation for a candidate scheduling policy.
2. **Balancing**: (3.18-b) ensures that generation and load are balanced at all times within the operating interval.
3. **Tasks and Devices**: (3.18-c) and (3.18-d) ensures the scheduling policy meets all task and device requirements respectively.

Causality

As the CM can make decisions solely based on current information, we seek *causal* scheduling policies. Unfortunately, an optimal causal scheduling policy does not exist. We show this using an adversarial argument.

Theorem 3.1. *The optimal scheduling policy with respect to (3.18) is not causal.*

Proof. By counterexample. Consider an operating interval with two tasks and no devices:

Task T_1: $E_1 = 2$ kWh, $m_1 = 2$ kW, $a_1 = 0$, $d_1 = 2$,

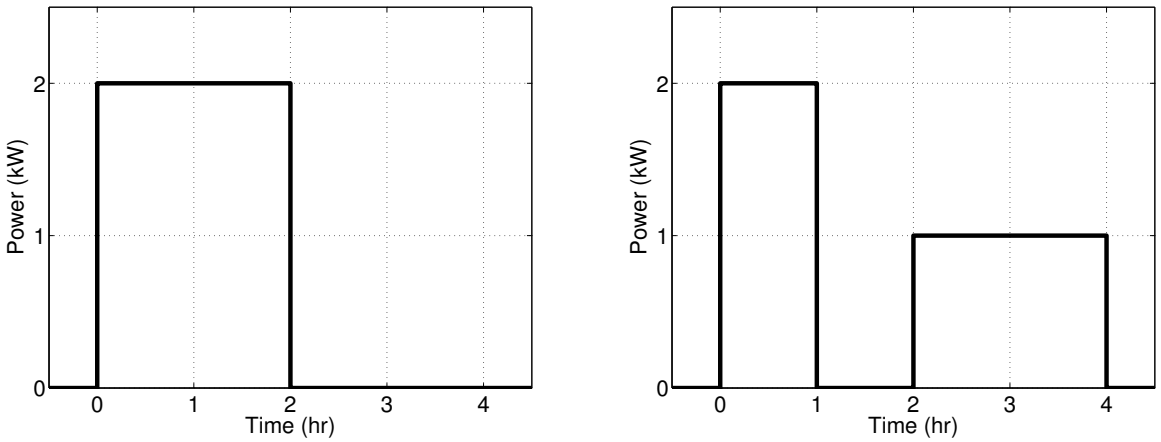
Task T_2: $E_2 = 2$ kWh, $m_2 = 1$ kW, $a_2 = 0$, $d_2 = 4$.

Consider the two available generation profiles $\mathbf{g}^A$ and $\mathbf{g}^B$ shown in Figure 3.1. Both $\mathbf{g}^A$ and $\mathbf{g}^B$ are *feasible*. Figure 3.2 shows the power allocations of these profiles that completes tasks T_1 and T_2 without any need for reserves. We first argue that these power allocations are *unique*.

For profile $\mathbf{g}^A$, completion of task T_2 requires 2 kWh with a rate limit of 1 kW. Since $g_k^A > 0$ only when $k = \{0, 1\}$, we must service task T_2 at its rate limit. The remaining power must be allocated entirely to task T_1.

For profile $\mathbf{g}^B$, completion of task T_1 requires 2 kWh with a rate limit of 2 kW. This task must be serviced on the interval $[0, 2)$. Since $g_1^B = 0$, task T_1 must be serviced at its rate limit at time $k = 0$. The remaining power, containing 2 kWh of energy, must be fully allocated to task T_2.

Notice that the two available generation profiles $\mathbf{g}^A, \mathbf{g}^B$ are *identical* on $t \in [0, 1)$. Therefore, any *causal* policy σ must offer identical allocations under either power profile at time $k = 0$. If σ is optimal, it must complete both tasks without any need for reserves under $\mathbf{g}^A$ *and* $\mathbf{g}^B$. But completion of both tasks solely with the feasible available generation profiles requires *different* allocations at time $k = 0$. □



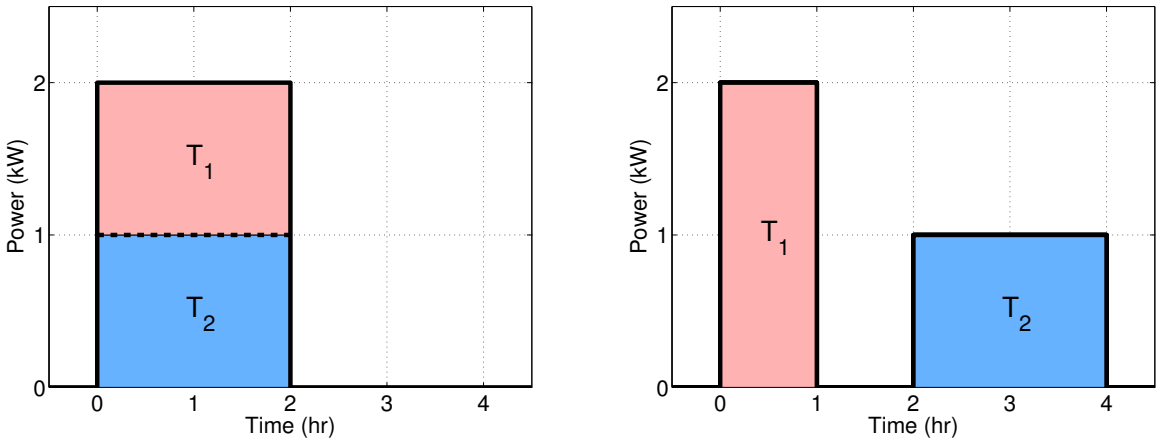
(a) Profile $\mathbf{g}^A$ (b) Profile $\mathbf{g}^B$

Figure 3.1: Available generation profiles $\mathbf{g}^A$ and $\mathbf{g}^B$ described in proof of Theorem 3.1

This result forces us to be content with causal heuristics for resource scheduling that are sub-optimal with respect to the policy optimization (3.18).

3.3 Scheduling Algorithms

In this section, we describe four causal scheduling heuristics for resource scheduling. We also develop a policy (zero-laxity) for procuring and allocating reserve generation based on one of



(a) Profile $\mathbf{g}^A$ (b) Profile $\mathbf{g}^B$

Figure 3.2: Power allocations to tasks that show feasibility of profiles $\mathbf{g}^A$ and $\mathbf{g}^B$

these heuristics. We conclude by formulating a receding horizon control (RHC) for resource and reserve scheduling inspired by the policy optimization problem (3.18). We specifically concentrate on the following scheduling problem: available generation at time k (g_k) must be allocated to active tasks $\mathbb{A}_k$ and devices $\mathbb{D}$.

3.3.1 Earliest Deadline First (EDF)

Policy

The **EDF scheduling policy** allocates available generation g_k to the task T_I with the most imminent deadline (with ties broken arbitrarily).

$$I = \underset{i \in \mathbb{A}_k}{\operatorname{argmin}}\, d_i \tag{3.19}$$

Available generation in excess of the rate limit for task T_I (m_I), is allocated to the active task with the next most imminent deadline. This process continues until either all generation g_k is expended or all active tasks are serviced at their rate limits.

Device scheduling under EDF is as follows. If all generation is expended, active tasks not serviced at their rate limits receive power from devices. This process continues either until all active tasks are serviced at their rate limits or until device constraints (discharge rate (m_j^-) or minimum energy level (E_j^-)) limit any further power service. If all active tasks are being serviced at their rate limits, excess available generation is allocated to devices $\mathbb{D}$. This process continues either until all available generation is expended or until device constraints (charging rate (m_j^+) or maximum energy capacity (E_j^+)) limit any further service.

Discussion

EDF is a well-known scheduling algorithm for processor time allocation (PTA). The objective in PTA is to schedule a collection of computational tasks, on single or multiple processors, in a way that completes all task requirements. Tasks, in this context, are characterized by a processing time requirement, an arrival, and a deadline.

There is extensive work analyzing the performance of EDF for PTA applications. In the case of single processors, it is known that EDF is an optimal scheduling policy for a single processor in the following sense: if some scheduling policy can meet all deadlines of a set of pre-emptive tasks, scheduling those tasks under EDF will also meet all deadlines [16]. We call scheduling policies that satisfy this property **optimal at preserving feasibility**. Liu and Layland, in a seminal paper [80], showed that the EDF scheduling policy is optimal at preserving feasibility for periodic tasks on a single processor. Dertouzos and Mok showed that this result does not apply to the multiprocessor setting [46]. Liu and Layland also derived necessary and sufficient conditions for ascertaining scheduling feasibility - whether or not a set of periodic tasks scheduled under EDF are completed - in the single processor case [80]. More recent work has centered around developing similar conditions that guarantee task completion under EDF in the multiprocessor setting [63, 13].

The resource scheduling problem described in this chapter is analogous to PTA with available generation playing the part of available processing time. However, there are three primary differences between resource scheduling and PTA:

1. Total available processing time is constant for a given processor while it's analog in resource scheduling, available generation is not constant and indeed highly variable, over an operating window.

2. Tasks in resource scheduling have explicit rate limits that must be respected during scheduling. As task requirements in PTA are defined solely in terms of processing time, there are no comparable scheduling constraints on PTA.

3. At any given time, available processing capacity is fully devoted to a single task. In contrast, multiple tasks can be concurrently scheduled and serviced in resource scheduling.

Owing to these differences, many of the performance guarantees derived for EDF in the PTA literature, such as optimality, do not apply to resource scheduling. At best, we can show that given a feasible available generation profile, the allocation under EDF will meet all task requirements without any need for reserves only in the *absence of rate constraints*.

Theorem 3.2. *Ignoring task rate limits, EDF is optimal at preserving feasibility.*

Proof. Let $\mathbf{g}$ be a *feasible* generation profile with respect to a set of tasks $\mathbb{T}$. Extensions to cases with devices Let σ correspond to the associated scheduling policy that completes all task requirements and σ_{EDF} correspond to the EDF scheduling policy. The proof technique

closely mirrors that used in [16] except we focus on scheduling available generation rather than processing time. We perturb the power allocations defined by σ to show σ_{EDF} also completes all task requirements.

If σ and σ_{EDF} are identical, the claim follows trivially. If σ and σ_{EDF} are distinct, let k be the first instance when the allocations under the two policies differ. Over the interval $[k_1\Delta t, (k_1+1)\Delta t]$, assume σ assigns energy ΔE to task T_1 while σ_{EDF} assigns ΔE to task T_2.

Since σ completes all tasks, it must complete task T_2. Hence, there exists some time interval $[k_2\Delta t, (k_2+1)\Delta t]$ before the deadline for T_2 over which σ assigns ΔE to task T_2. Define a new scheduling policy $\hat{\sigma}$ that schedules T_2 over $[k_1\Delta t, (k_1+1)\Delta t]$ and T_1 over $[k_2\Delta t, (k_2+1)\Delta t]$. Clearly, $\hat{\sigma}$ completes all task requirements as $k_2 \leq d_2 \leq d_1$. The last inequality reflects the choice of σ_{EDF} over the interval $[k_1\Delta t, (k_1+1)\Delta t]$. This procedure is repeated iteratively, with $\hat{\sigma}$ replacing σ on each iteration. The resulting policy at the end of this process is σ_{EDF} thus proving the claim. □

3.3.2 Least Laxity First (LLF)

The EDF policy allocates available generation solely on the basis of task deadlines. A more nuanced algorithm would also consider remaining energy requirements allocating power. LLF is one such scheduling algorithm that explicitly takes task energy states into account. In lieu of deadlines, this algorithm makes decisions based on *laxity*.

Definition 3.11. *The **laxity** of task T_i at time k is:*

$$\phi_{ik} = (d_i - k) - \frac{e_{ik}}{m_i} \quad \forall k \in \{a_i, \ldots, d_i\}, \tag{3.20}$$

Laxity is the difference between the time remaining to service a task $(d_i - k)$ and the minimum time required to complete task requirements $\left(\frac{e_{ik}}{m_i}\right)$. This measure succinctly captures the degree of deferrability present in scheduling a task. Tasks with larger laxities offer greater scheduling flexibility.

Policy

The **LLF scheduling policy** allocates available generation g_k to the task T_I with the smallest laxity (with ties broken arbitrarily).

$$I = \underset{i \in \mathbb{A}_k}{\operatorname{argmin}}\, \phi_{ik} \tag{3.21}$$

Available generation in excess of the rate limit for task T_I (m_I), is allocated to the active task with the next smallest laxity. This process continues until either all generation g_k is expended or all active tasks are serviced at their rate limits. The device scheduling component under LLF is identical to that under EDF.

Discussion

LLF is also a well analyzed dynamic scheduling algorithm in the context of PTA [46, 69]. Mok [92] showed that the LLF policy, like EDF, is optimal at preserving feasibility when scheduling pre-emptive tasks on a single processor. Certain studies have explored the use of an LLF-based scheduling policy for resource scheduling. In the case of a parking garage, Chen et al. analyzed the performance of LLF-based scheduling of EV charging tasks [35]. They specifically investigate the case where the scheduler has the option to admit or reject task requests. In the case of residential load management, Barker et al. consider the ability of 'least slack first', another term for LLF, scheduling at managing peak demand [15].

3.3.3 Deadline Prioritized Adjusted Scheduling (DPAS)

The DPAS policy, like EDF, also makes scheduling decisions based on task deadlines. However, instead of allocating all available generation to tasks, as EDF does, DPAS only allocates deviations of available generation ($\mathbf{g}$) from the nominal deferrable load profile ($\mathbf{l}^D$).

Policy

Let

$$x_k = g_k - l_k^D \tag{3.22}$$

be the *residual available generation* – generation available if all active tasks were served at constant rates.

If additional generation is available ($x_k > 0$), the DPAS policy first commits to meeting the nominal deferrable load profile. Specifically, all active tasks T_i are serviced at their nominal rates. The policy then begins allocating the residual available generation x_k to the task with most imminent deadline T_I where I is defined by (3.19). If task T_I is serviced at its rate limit, excess generation is allocated to the task with the next most imminent deadline. This process continues until all residual available generation (x_k) is expended or all active tasks are serviced at their rate limits, whichever comes first.

If available generation fails to meet the nominal load profile ($x_k < 0$), the DPAS policy cannot meet the nominal deferrable load profile. The policy then prioritizes the active tasks according to their deadlines, servicing as many tasks at their nominal rates as possible. Specifically, the policy allocates available generation to the task with the most imminent deadline T_I where I is defined by (3.19). Any available generation in excess of the nominal rate q_I of task T_I is allocated to the task with the next most imminent deadline. This process continues until all available generation is expended.

Device scheduling under DPAS is similar to that under EDF with the only difference being the following. If available generation fails to meet the nominal load profile, active tasks not serviced at their nominal rates receive power from devices. Power allocation from devices is also prioritized by deadline and continues until all either all active tasks are serviced at their nominal limits or until device constraints (discharge rate (m_j^-) or minimum energy

level (E_j^-)) limit any further power service. After enforcing this rule, the device scheduling component for DPAS is identical to that under EDF.

3.3.4 Laxity Prioritized Adjusted Scheduling (LPAS)

The LPAS scheduling policy is identical to DPAS with laxity, rather than deadlines, forming the basis for all task allocation decisions. We state this algorithm fully for completeness.

Policy

Define the residual available generation x_k according to (3.22). If additional generation is available ($x_k > 0$), the LPAS policy first commits to meeting the nominal deferrable load profile. Specifically, all active tasks T_i are serviced at their nominal rates. The policy then begins allocating the residual available generation x_k to the task with least T_I where I is defined by (3.21). If task T_I is serviced at its rate limit, excess generation is allocated to the task with the next smallest laxity. This process continues until all residual available generation (x_k) is expended or all active tasks are serviced at their rate limits, whichever comes first.

If available generation fails to meet the nominal load profile ($x_k < 0$), the LPAS policy cannot meet the nominal deferrable load profile. The policy then prioritizes the active tasks according to their laxities, servicing as many tasks at their nominal rates as possible. Specifically, the policy allocates available generation to the task with the smallest laxity T_I where I is defined by (3.21). Any available generation in excess of the nominal rate q_I of task T_I is allocated to the task with the next smallest laxity. This process continues until all available generation is expended.

Device scheduling under LPAS is similar to that under LLF with the only difference being the following. If available generation fails to meet the nominal load profile, active tasks not serviced at their nominal rates receive power from devices. Power allocation from devices is also prioritized by laxity and continues until all either all active tasks are serviced at their nominal limits or until device constraints (discharge rate (m_j^-) or minimum energy level (E_j^-)) limit any further power service. After enforcing this rule, the device scheduling component for LPAS is identical to that under LLF.

3.3.5 Zero-Laxity (ZL)

The scheduling algorithms described so far (EDF, LLF, DPAS, and LPAS) determine allocations of available generation to tasks and devices. A resource scheduling policy must also determine reserve generation procurement decisions and allocate this additional generation to various tasks during the operating window. We use task laxities to develop a causal reserve scheduling heuristic. This policy paired with any of the described algorithms for allocating available generation constitutes an complete resource scheduling policy.

Policy

The first component of the policy ensures that static load requirements are satisfied. If static load exceeds the sum of bulk power and renewable generation ($g_k < 0$), the CM procures sufficient reserves to account for this deficit ($r_k = -g_k$).

The second component of the policy focuses on meeting task requirements. A task cannot be completed if the minimum time required to satisfy its energy requirement exceeds the time until deadline. The zero-laxity policy ensures that no task laxities become negative. Specifically, a task T_i is infeasible at time k in and only if $\phi_{ik} < 0$. The CM procures reserve generation to service active tasks with laxities within a threshold $\epsilon > 0$. For each such task T_i, the CM procures adequate reserve generation to ensure the total power allocated to this task is m_i.

There is no device scheduling component to the zero-laxity heuristic.

3.3.6 Receding Horizon Control (RHC)

Receding horizon control (RHC), also known as model predictive control (MPC), is a widely-used and effective strategy for state and input constrained control problems [62]. RHC involves solving finite horizon optimization problems successively at each time-step to determine appropriate control actions.

Policy

In the context of resource scheduling, an optimization problem is solved at each time-step k to yield cost-minimizing allocations of available and reserve generation over some time horizon $\{k, \ldots, k+H\}$. These decisions are based on renewable generation and static load forecasts and information from all active tasks and devices. Tasks are then scheduled based only on computed allocation decisions for the first time-step k. This process is repeated at the next time-step $k+1$ with an updated set of active tasks, devices, and generation forecasts on a different horizon $\{k+1, k+1+H\}$. Determining power allocations in this iterative fashion enables the incorporation of updated task information and more precise generation forecasts.

Consider the optimization problem solved for RHC scheduling at each time-step k. Without loss of generality, we state the RHC optimization problem when $k = 1$. Let M_T and M_D be the number of active tasks and devices respectively. We define the following quantities to help with our exposition:

Definition 3.12. *The **horizon length** H is the the number of Δt time-steps between k and the largest deadline in the set of active tasks.*

$$H = \max_{i \in \mathbb{A}_k} d_i. \tag{3.23}$$

Definition 3.13. *The **generation forecast** $\hat{\mathbf{g}}$ refers to forecasted values of available generation through the horizon.*

$$\hat{\mathbf{g}} = \{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_H\}. \tag{3.24}$$

Each optimization problem attempts to meet all tasks requirements by allocating only forecasted generation. We now describe the decision variables and objective function of the RHC optimization problem.

Decision Variables

1. $G \in \mathbb{R}_+^{M_T \times H}$: G_{it} is the amount of available generation assigned to task T_i at time t,
2. $R \in \mathbb{R}_+^{M_T \times H}$: R_{it} is the amount of reserve generation assigned to task T_i at time t,
3. $S \in \mathbb{R}^{M_D \times H}$: S_{jt} is the amount of power assigned to device D_j at time t.

Objective Function

$$\begin{aligned} C(G,R,S) = \quad & \alpha_E \sum_{t=1}^{H} \Big(\sum_{i=1}^{M_T} R_{it} \Big) + \Big(\hat{g}_t - \sum_{i=1}^{M_T} G_{it} - \sum_{j=1}^{M_D} S_{jt} \Big) \\ + \; & \alpha_C \Big(\max_t \sum_{i=1}^{M_T} R_{it} + \max_t \Big(\hat{g}_t - \sum_{i=1}^{M_T} G_{it} - \sum_{j=1}^{M_D} S_{jt} \Big) \Big) \\ + \; & \sum_{t=1}^{H} \sum_{i \in \mathbb{A}_t} (N - \phi_{it})^2 \end{aligned} \tag{3.25}$$

The first and second terms capture up and down reserve *energy* costs respectively while the third and fourth terms capture up and down reserve *capacity* costs respectively. The fifth term maximizes task laxities at subsequent time-steps within the horizon H. Effectively, this incentivizes earlier allocations of available generation to help maintain adequate task flexibility for all tasks throughout the time horizon. The parameters α_E, and α_C negotiate the relative importance of the objective function components.

Optimization Problem At each time k, we solve the following optimization problem (3.26).

$$\min_{G\geq 0, R\geq 0, S} \quad C(G,R,S) \tag{3.26}$$

$$\text{subject to:} \quad \forall\, t,\ \sum_{i=1}^{M_T} G_{it} + \sum_{j=1}^{M_D} S_{jt} \ \leq\ \hat{g}_t, \tag{a}$$

$$\forall\, T_i,\ \sum_{t=1}^{d_i} G_{it} + R_{it} \ =\ E_i, \tag{b}$$

$$\forall\, T_i,\ \begin{cases} 0 \leq G_{it} + R_{it} \leq m_i \Delta t, & \forall\, t : t < d_i \\ G_{it} = 0, R_{it} = 0, & \forall\, t : t \geq d_i \end{cases} \tag{c}$$

$$\forall\, D_j, t,\ E_j^- \ \leq \sum_{n=1}^{t} S_{jn} \ \leq\ E_j^+, \tag{d}$$

$$\forall\, D_j, t,\ -m_j^- \ \leq\ S_{jt} \ \leq\ m_j^+, \tag{e}$$

$$\forall\, T_i, t,\ \phi_{it} = d_i - t - \frac{e_{it}}{m_i}, e_{it} = E_i - \sum_{n=1}^{t} G_{in} + R_{in}. \tag{f}$$

Constraints

1. **Generation**: (3.26-a) ensures that the sum of all power allocated to tasks and devices cannot exceed available generation forecast ($\hat{g}_t$) at any time t. Additionally, when satisfying tasks using energy stored in devices, (3.26-a) guarantees total power delivered to tasks does not exceed total energy discharged from devices.

2. **Task: Total Energy Requirement**: (3.26-b) ensures each task's energy requirement (E_i) is met through allocation of available and reserve generation.

3. **Task: Rate Limits**: (3.26-c) ensures power is only allocated to active tasks. Moreover, the allocation for a task is non-negative and bounded by the task rate limit (m_i).

4. **Device: Capacity**: (3.26-d) enforces maximum (E_j^+) and minimum (E_j^-) energy levels on the device energy state at all times within the horizon.

5. **Device: Rate Limits**: (3.26-e) enforces discharge ($-m_i^-$) and charging (m_i^+) rate limits on allocations to and from devices.

6. **Laxity**: (3.26-f) is used to compute task laxities ϕ_{ik} which are present in the objective function.

Discussion

We remark that this optimization problem is a quadratic program (QP) - quadratic cost and linear equality and inequality constraints - which must be solved at each time-step k. As such, the computational requirements associated with RHC scheduling are substantial when compared to those under any of the other policies described in Section 3.3. However, our simulation studies reveal that RHC scheduling offers significant reductions in reserve capacity costs.

RHC, unlike the other scheduling algorithms described in Section 3.3, can explicitly incorporate forecasts of available generation. As a result, this is a non-causal scheduling policy with respect to the information state $\mathcal{I}_k$. We showed earlier in this chapter (Theorem 3.1) that the optimal scheduling policy with respect to our functional optimization problem (3.18) is not causal. A similar argument can be used to show that RHC, based only on active task information, is not that optimal non-causal policy.

Theorem 3.3. *A RHC scheduling policy is not optimal with respect to (3.18).*

Proof. We use an adversarial argument. Specifically, we consider an available generation profile that is feasible with respect to two sets of tasks. We will show that the RHC scheduling policy, even armed with perfect forecasts of this profile, fails to complete both sets of tasks without any need for reserves.

Consider the following tasks over the operating interval $[0, 4]$:

Task T_1: $E_1 = 2$ kWh, $m_1 = 2$ kW, $a_1 = 0$, $d_1 = 2$,

Task T_2: $E_2 = 2$ kWh, $m_2 = 1$ kW, $a_2 = 0$, $d_2 = 4$,

Task T_3: $E_3 = 2$ kWh, $m_3 = 10$ kW, $a_3 = 1$, $d_3 = 2$,

Task T_4: $E_4 = 2$ kWh, $m_4 = 10$ kW, $a_4 = 2$, $d_4 = 4$.

Consider the available generation profile $\mathbf{g}^C$ shown in Figure 3.3. Figure 3.4 demonstrates that this is a feasible profile with respect to the sets of tasks (T_1, T_2, T_3) and (T_1, T_2, T_4).

These power allocations are *unique*. We first focus on the set of tasks (T_1, T_2, T_3). Completion of task T_3 requires 2 kWh over the interval $[1, 2)$. The remaining power available over the interval $[1, 2)$, totaling 2 kWh, must be allocated entirely to task T_1. As a result, task T_2 must be serviced on the interval $[2, 4)$ and completion of this task requires all power available over this interval.

For the set of tasks (T_1, T_2, T_4), completion of task T_4 requires all available generation on the interval $[2, 4)$. As remaining power is available only on the interval $[0, 2)$, task T_2 must be serviced at its rate limit ($m_2 = 1$ kW) over this interval. The remaining power available, totaling 2 kWh, must be allocated to task T_1.

For both sets of tasks, RHC-based scheduling over $[0, 1)$ is done knowing the tasks parameters of tasks T_1 and T_2, and forecasts of the profile $\mathbf{g}^C$. As a result, the RHC scheduling policy must offer *identical* allocations for either set of tasks over $[0, 1)$. However, completing

each set of tasks requires *different* allocations over $[0, 1)$ depending on the nature of task arriving after this interval. An optimal scheduling policy must complete both sets of tasks without any need for reserves. Clearly, an RHC scheduling policy, even armed with perfect generation forecasts, cannot do so. □

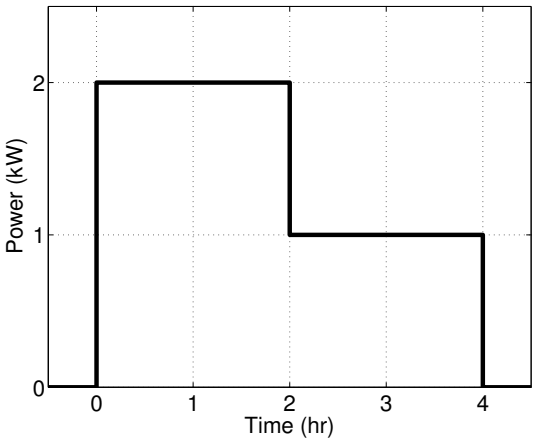
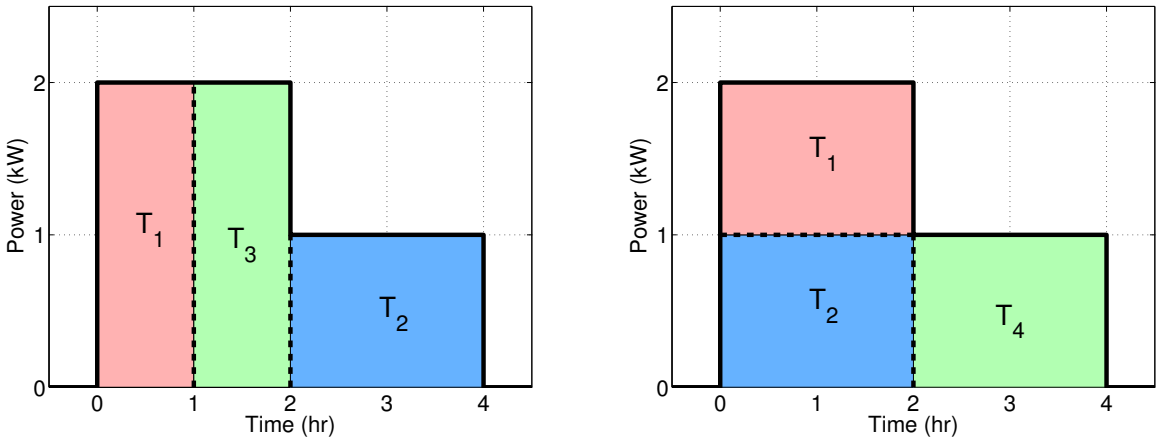


Figure 3.3: Available generation profile $\mathbf{g}^C$ described in proof of Theorem 3.3



(a) Allocation to (T_1, T_2, T_3) (b) Allocation to (T_1, T_2, T_4)

Figure 3.4: Power allocations to each set of tasks that show feasibility of profile $\mathbf{g}^C$

The use of RHC approaches to resource scheduling is an active research area. Hug-Glanzmann uses an RHC approach to coordinate the operation of storage devices, intermittent generation, and load [71], while Galus et al. [60] focus on balancing wind generation by

modifying EV charging schedules. Recent work has also focused on creating decentralized RHC strategies, which are more computationally tractable, for charging large numbers of deferrable loads [86, 83]. Our contribution to this literature is not in the use of RHC control strategies, but rather in the development of specific cost functions inspired by PTA methods which explicitly minimize reserve costs.

3.4 Simulation Results

In this section, we demonstrate the value of resource scheduling at reducing reserve generation costs. We create synthetic test cases where a CM manages a resource cluster with significant solar PV and EV penetration. In each test case, the CM attempts to mitigate the added variability associated with renewables by judiciously determining EV charging schedules. We apply the scheduling algorithms described in Section 3.3 to meet load requirements at (Δt) 10 minute intervals over a 24 hour operating window. We quantify performance of these algorithms in two cost metrics: (1) reserve energy and (2) reserve capacity. All reported results are averaged over 50 test cases.

3.4.1 Test Case Description

Load

The resource cluster consists of two types of loads: static and deferrable. In order to create static load profiles that are representative of typical power usage statistics, we use time-series data of the total power demand of the California ISO balancing area. This data, sampled at 10 minute intervals, is normalized to a peak load of 2 MW to create realistic profiles mimicking the power requirements of the resource cluster. We generate load profiles by randomly selecting one day from 8 days of June 2012 data. Let the power profile $\mathbf{l} = \{l_k\}_{k=1}^{N}$ correspond to this load profile normalized to 2 MW.

We assume that a specified fraction of the total load energy requirement, α, is deferrable. For a given value of α, the static load profile $\mathbf{l}^S$ can be readily computed by scaling the power profile $\mathbf{l}$ appropriately:

$$\mathbf{l}^S = (1 - \alpha)\mathbf{l}. \tag{3.27}$$

Deferrable loads are modeled as tasks. We ensure that the total energy requirement for such loads reflects the choice of α by enforcing constraint (3.28) when generating task energy needs.

$$\sum_{i=1}^{M_T} E_i \;=\; \alpha \sum_{k=1}^{N} l_k \Delta t. \tag{3.28}$$

In this test case, we focus exclusively on EVs. Accordingly, task parameters (arrivals, deadlines, and energy needs) are randomly generated based on typical EV charging characteristics [94]. These parameters are chosen to ensure initial task feasibility ($E_i \leq (d_i - a_i)m_i$).

We use a constant rate limit m_i for all tasks consistent with the SAE J1772 AC Level 1 EV charging standard [113]. This is a reasonable assumption as EV battery characteristics and distribution network constraints are similar across tasks.

For this collection of tasks, we compute the nominal deferrable load profile $\mathbf{l}^D$ as defined by (3.10). This is used in bulk power procurement decisions and serves as a baseline for computing reserve cost reductions.

Generation

We use time-series data of solar PV generation output to create renewable generation profiles. This data, obtained from the PV integrator SolarCity, represents the aggregate real-power output, sampled at 1 minute intervals, from 30 different but proximate PV installations in California. We normalize this time-series data to a peak output of 750 kW. We generate renewable generation profiles by randomly selecting one day from 40 days of February and March 2012 data. We choose this time period as PV generation exhibits more variability in spring rather than summer.

As the CM determines load allocations every 10 minutes, scheduling decisions are made with a sub-sampled version of the renewable generation process. Effectively, we assume, when determining allocations, that the amount of renewable generation at each balancing time is constant over the following 10 minute interval. However, reserve requirements for load balancing are computed using the actual renewable generation profile sampled every minute.

DA Market Clearing

In this analysis, we assume the CM has one opportunity to purchase power *ex-ante*. We simulate this bulk power purchase based on hourly forecasts of load and renewable generation data. We create these forecasts according to the following procedures.

1. **Static Loads**: We use day-ahead load forecasts, corresponding to load time-series data, from the California ISO. These are hourly forecasts of the aggregate demand used by the California ISO in their day-ahead market operations. We normalized these forecasts according to the same procedure (3.27) used to generate $\mathbf{l}^S$. Let $\widetilde{l^S}_k$ refer to the forecasted static load at time k.

2. **Deferrable Loads**: We create forecasts of the nominal deferrable load profile $\mathbf{l}^D$ by adding *zero-mean Gaussian noise* to hourly averages of $\mathbf{l}^D$. We assume this noise process has a standard deviation of 3% of the hourly average of the nominal deferrable load. Let $\widetilde{l^D}_k$ refer to the forecasted static load at time k. Concretely, this computation

can be expressed as (3.29).

$$\begin{aligned} \widetilde{l^D}_k &= \frac{1}{|\mathbb{H}|}\sum_{t\in\mathbb{H}} {l^D}_t + \epsilon_k, \quad \forall k \in \{1,2,\ldots,N\} \\ \epsilon_k &\sim \mathcal{N}\left(0, \left(\frac{0.03}{|\mathbb{H}|}\sum_{t\in\mathbb{H}} {l^D}_t\right)^2\right) \end{aligned} \tag{3.29}$$

where $\mathbb{H}$ is the set of balancing time indices corresponding to the hour containing time k.

3. **Renewable Generation**: We create hourly forecasts of renewable generation by averaging data corresponding to past days. For each hour of within the operating window, we calculate the average over the past 5 days of the mean generation for that hour-long interval. Let $\widetilde{w}_k$ correspond to the DA renewable generation forecast at time k.

Based on these forecasts, the bulk power procured for each time-step k is:

$$B_k = \widetilde{l^S}_k + \widetilde{l^D}_k - \widetilde{w}_k \tag{3.30}$$

We remark that, by construction, the bulk power purchase B_k is constant over hour-long intervals.

RHC Forecasts

We create synthetic renewable generation forecasts for RHC scheduling by adding Gaussian noise to the renewable generation profiles. Specifically, the renewable generation forecast at time t ($\widehat{w}_t$) made at time k is described by (3.31).

$$\begin{aligned} \widehat{w}_t &= w_k + \sum_{n=k+1}^{t} \epsilon_n, \quad \forall t \in \{k+1,\ldots,k+H\} \\ \epsilon_n &\sim \mathcal{N}\left(0, \sigma_n^2\right) \end{aligned} \tag{3.31}$$

where H is the horizon length at time k over which the RHC scheduling policy determines allocations. We ensure the noise variance σ_n^2 increases *linearly* with the prediction window length.

Finally, we assume that this rule only applies to all t corresponding to times between 06:00 and 20:00 in the day. Outside this interval, we assume there is no renewable generation ($\hat{w}_t = 0$). This is an entirely reasonable assumption as solar PV is the only source of renewable generation in these test cases.

3.4.2 Comparison of Scheduling Algorithms

Figure 3.5 illustrates the performance of the five scheduling algorithms in reducing operating reserve requirements for a test case. In this test case, the load requirement slightly exceeds the amount of generation available. Each sub-figure compares the load profiles achieved by one of these resource scheduling algorithms to the baseline load and generation profiles. The generation profile shown in each of these sub-figures is the sum of bulk and renewable power. Clearly, these plots demonstrate that the CM, through resource scheduling, can modify loads profiles to more closely match generation and thereby, reduce the need for reserves. This is true for *any of the resource scheduling algorithms* described in Section 3.3.

Qualitatively, the load profiles under the various scheduling policies vary significantly towards the end of the operating interval (16:00-24:00 hours). Specifically, scheduling under any of the non-RHC algorithms results in greater up reserve requirements during this period than those under RHC scheduling. In fact, RHC is characterized by balanced reserve procurement over the entire operating window (Figure 3.5e). This is a direct consequence of explicit reserve capacity minimization in the RHC scheduling algorithm - the RHC policy schedules tasks with reserve generation earlier in the operating window to offset possible increases in reserve capacity later in the window. The other algorithms, which schedule reserves only as a last resort under the zero-laxity heuristic, do not exhibit this behavior.

Figures 3.5b and 3.5d, which correspond to the LLF or LPAS scheduling algorithms respectively, also indicate an initial spike in reserve procurement. This is an artifact of using a zero-laxity reserve policy in conjunction with a laxity-based resource scheduling policy. In these cases, the CM procures reserves only when task laxities near 0. As the laxities of all tasks are *identical* under LLF, serving a particular task with reserve generation implies *all active tasks* must be served with reserves.

This behavior is evident only in cases of overall generation deficit (total load requirement exceeds the amount of generation) and depends on the size of the deficit. Figure 3.6, which shows load profiles in a test case with severe generation deficit, illustrates the impact of generation deficit size on the magnitude of the initial reserve procurement spike. Clearly, larger generation deficits accentuate this characteristic of LLF and LPAS scheduling. We remark that RHC-based scheduling again exhibits balanced reserve procurement resulting in a load profile that tracks changes in the generation profile. (Figure 3.6c). Indeed, the advantages of RHC over the other scheduling algorithms are more apparent in such test cases.

Table 3.1 shows average percentage decreases in four cost metrics (up reserve energy, down reserve energy, up reserve capacity, and down reserve capacity) for each of the five algorithms described in Section 3.3. Clearly, resource scheduling under *any* of the considered algorithms reduces reserve *energy* costs. Moreover, the choice of algorithm only has a limited impact on energy cost reductions. In contrast, the scheduling algorithms offer significantly different performance in the metric of reserve *capacity*. RHC is a more potent algorithm at mitigating capacity costs, particularly those associated with down reserve. This agrees with intuition, as reserve capacity is explicitly penalized in the RHC objective function. It is also apparent

		Scheduling Algorithms				
		EDF	**LLF**	**DPAS**	**LPAS**	**RHC**
Energy	**Up Reserves**	38.0	39.6	41.0	41.2	40.7
	Down Reserves	29.2	28.9	30.0	29.0	32.9
Capacity	**Up Reserves**	22.6	17.3	22.6	17.8	27.8
	Down Reserves	3.1	-2.3	2.8	-2.0	11.8

Table 3.1: Comparison of scheduling algorithms showing *percentage reductions* in 4 reserve cost metrics (compared to the base case of no coordination)

that the deadline-based algorithms (EDF and DPAS) offer better performance than their laxity-based counterparts (LLF and LPAS). We attribute the limited up reserve capacity reductions achieved by LLF and LPAS (17.3% and 17.8%) to the large instantaneous reserve procurements resulting from using the zero-laxity heuristic on tasks with identical laxities.

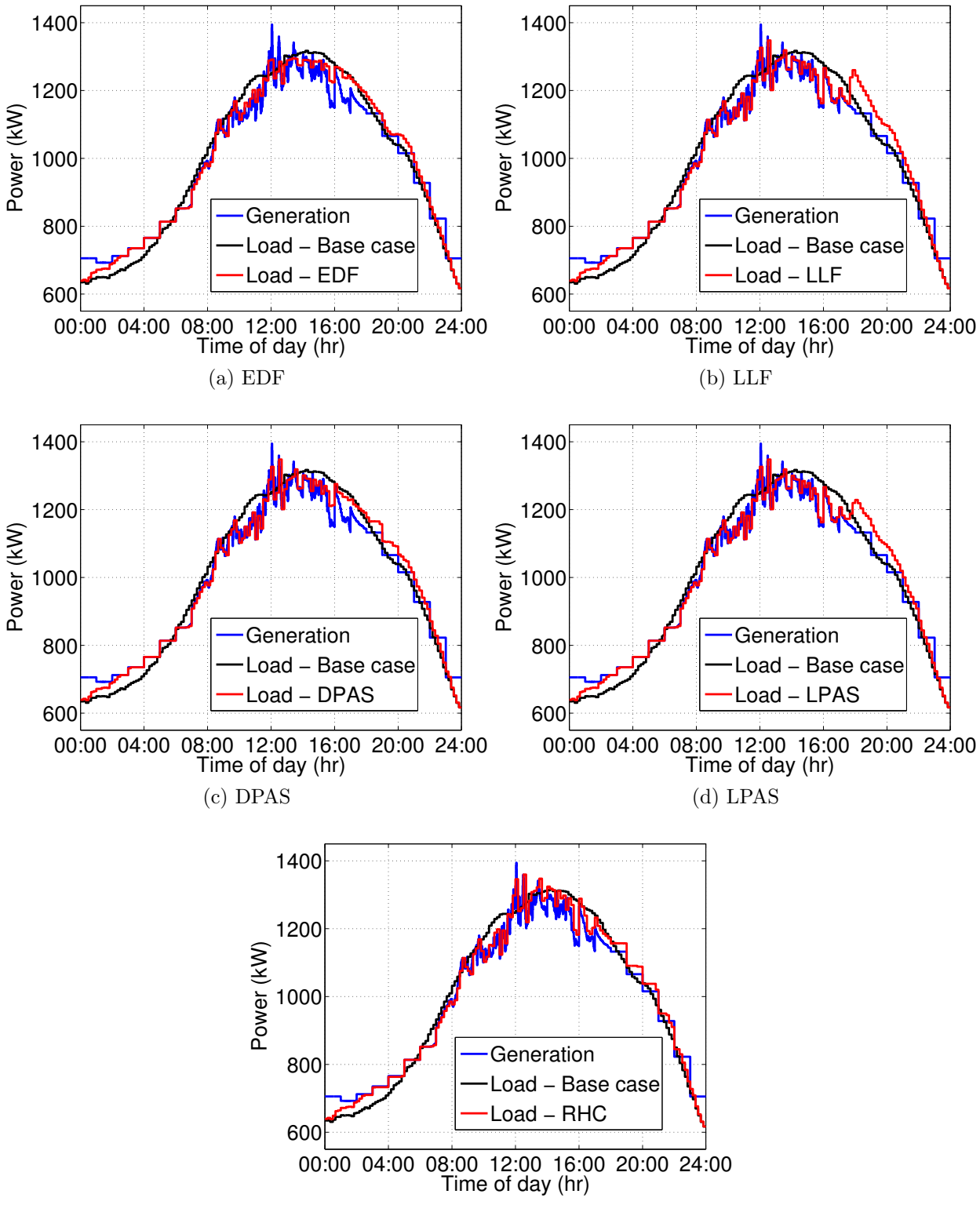
3.4.3 Impact of Deferrable Load Penetration

We now investigate the marginal benefit of scheduling additional deferrable loads. Let α be the proportion of total load (in terms of energy) that is deferrable. We vary this proportion and compute the reserve cost reductions enabled by RHC-based scheduling.

Figure 3.7 shows percentage reductions in up and down reserve metrics achieved by employing the RHC scheduling policy. It is apparent that additional amounts of deferrable load enable greater reductions in reserve *energy* metrics. Moreover, the marginal benefit of having additional deferrable loads clearly *decreases* with *increasing deferrable load penetration*. This phenomenon, while more visible with up reserves (Figure 3.7a), suggests that the primary impact of resource scheduling on reserve generation is evident even at low levels of deferrable load penetration. In fact, most of these deferrability benefits can be achieved even when only 10% of loads are deferrable ($\alpha = 0.1$).

At high deferrable load penetrations, energy reserve costs stem from surpluses or shortfalls in total energy procured to meet load over the entire operating window. These imbalances are caused by errors in mean values of the load and generation forecasts employed in ex-ante markets. Resource scheduling can only address the reserve costs stemming from *generation intermittency* and has limited impact at mitigating the costs associated with overall procurement imbalances.

Figure 3.7 exhibits no clear pattern with respect to capacity cost reductions. We remark that reserve capacity computations are very sensitive to high frequency fluctuations of renewable generation. Resource scheduling decisions, made every 10 minutes, do not account for fluctuations in renewable generation at shorter time-scales. However, reserve capacity is computed based on actual generation data sampled at these shorter time scales. Accordingly, achieving large reserve capacity cost reductions requires increases in the frequency at which



(a) EDF

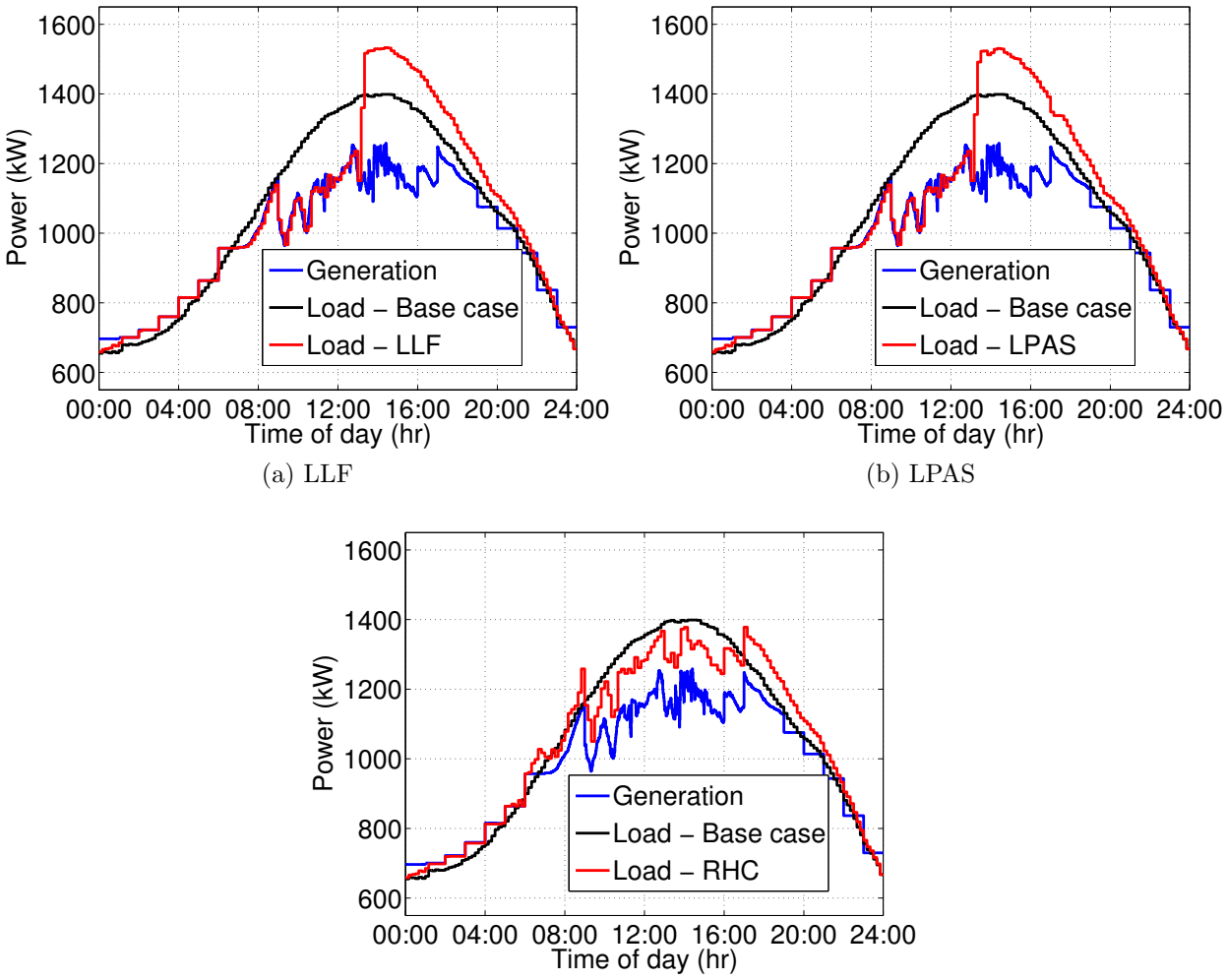
(b) LLF

(c) DPAS

(d) LPAS

(e) RHC

Figure 3.5: Load profiles comparing the impact of load scheduling under EDF, LLF, DPAS, LPAS, and RHC to no scheduling base case for a typical test case

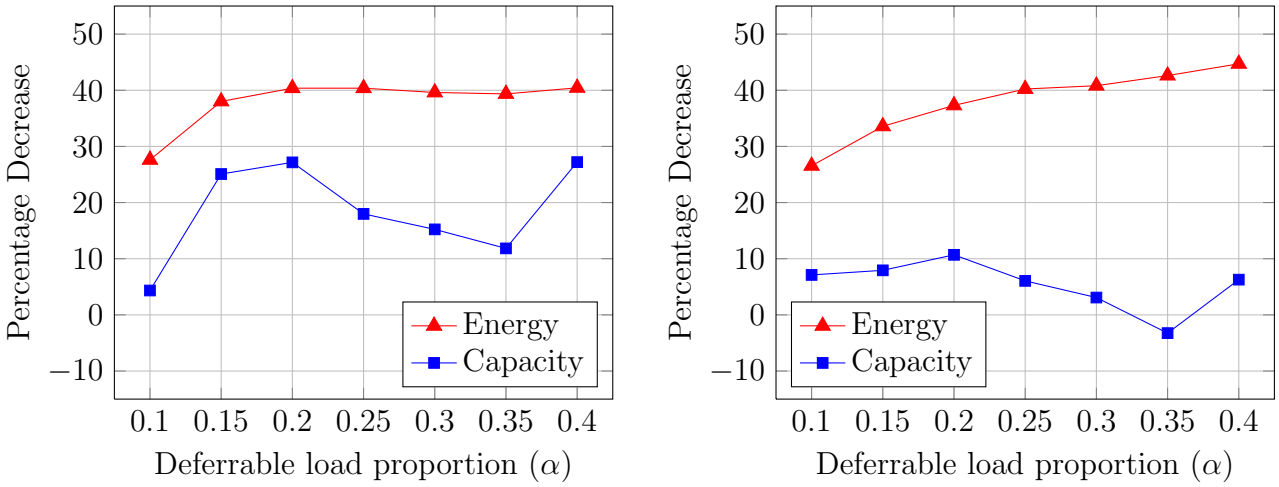


(a) LLF

(b) LPAS

(c) RHC

Figure 3.6: Load profiles comparing the impact of load scheduling under LLF, LPAS, and RHC to no scheduling base case for a test case involving severe generation deficit. These profiles highlight characteristics of reserve procurement under LLF and LPAS.



(a) Up Reserves

(b) Down Reserves

Figure 3.7: Percentage reductions in up (a) and down (b) reserve costs achieved by RHC-based scheduling at various levels of deferrable load penetration (α).

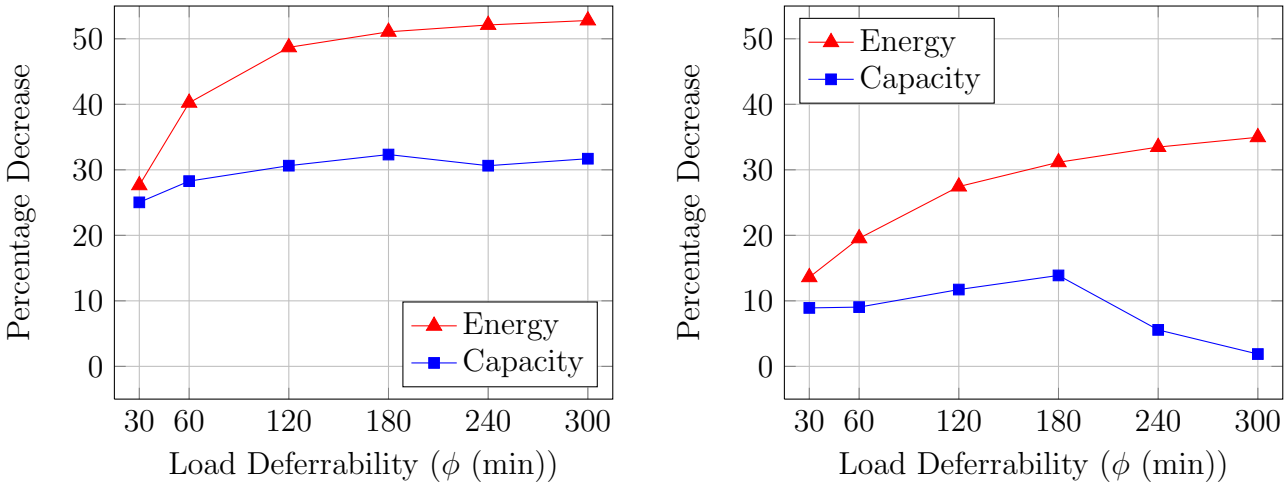
resource scheduling decisions are made.

3.4.4 Impact of Load Deferrability

Next, we investigate the impact of load deferrability, as captured by task laxity, on reserve requirements. In these simulations, all tasks have the same laxity (ϕ) upon arrival. For a fixed deferrable load proportion ($\alpha = 0.2$), we vary task laxity (ϕ) and compute the reserve energy and capacity cost reductions achieved through RHC-based scheduling.

Figure 3.8 shows percentage decreases in up, and down, reserve energy and capacity costs at various degrees of scheduling flexibility. Clearly, greater load deferrability has a compelling impact on reserve energy requirements. Specifically, energy costs for both up and down reserves can be reduced by a further 20% through additional scheduling flexibility. However, the concave nature of the energy curves of Figure 3.8 once again suggests that the marginal benefit of deferrability decreases with additional scheduling flexibility.

In contrast, increased flexibility only enables modest improvements in up reserve capacity and has an indeterminate impact on down reserve capacity (Figure 3.8a). We believe this is an artifact of the methodology used to compute capacity reductions in this study. Certainly, performing resource scheduling at finer time-scales can help address these issues. In short, these simulations demonstrate the value of additional scheduling flexibility in reducing reserve energy, and to a lesser extent, reserve capacity cost reductions.



(a) Up Reserves

(b) Down Reserves

Figure 3.8: Percentage reductions in up (a) and down (b) reserve costs achieved by RHC-based scheduling at various levels of task deferrability (ϕ) at $\alpha = 0.2$.

3.4.5 Practical Considerations

Finally, we examine two practical aspects of implementing such coordinated scheduling algorithms: computational requirements and the impact of switching on EVs.

Table 3.2 shows the average time taken by each scheduling algorithm to compute deferrable load schedules. These values are computed for different levels of deferrable load penetration (α). We performed all simulations using Matlab on a Dell XPS Studio 8100 machine with a processing speed of 2.93 GHz and 12.0 GB of RAM. In order to solve the quadratic programs associated with RHC scheduling, we use the Gurobi optimization solver.

As expected, the computational cost increases with the number of deferrable loads. Moreover, coordinated scheduling under RHC is far more computationally intensive than under any other scheduling algorithm. This is expected as the RHC algorithm determines schedules by solving a sequence of QPs – a far more time-intensive process than the sorting of task deadlines or laxities performed in the other four algorithms. We also observe that our implementations of the non-RHC scheduling algorithms exhibit *linear*, or better, time complexity.

We highlight the following with regard to RHC scheduling. Recall that there are 144 balancing times within the operating window. While the total computation time over the entire operating interval (24 hours) for RHC is high (692s for $\alpha = 0.4$), the amount of computation done at each balancing time is much shorter. Indeed, RHC-based scheduling is computationally tractable for problems of the size described in this chapter.

Coordinated scheduling algorithms may cause increased on-off switching of deferrable loads. Frequent switching of EV batteries is undesirable as it could adversely impact battery

α	EDF	LLF	DPAS	LPAS	RHC
0.1	0.06	0.05	0.05	0.05	41
0.2	0.11	0.10	0.09	0.08	126
0.3	0.16	0.15	0.12	0.13	388
0.4	0.20	0.20	0.16	0.16	692

Table 3.2: Comparison of scheduling algorithms showing average computation time (s) required for coordinated scheduling at different levels of deferrable load penetration (α)

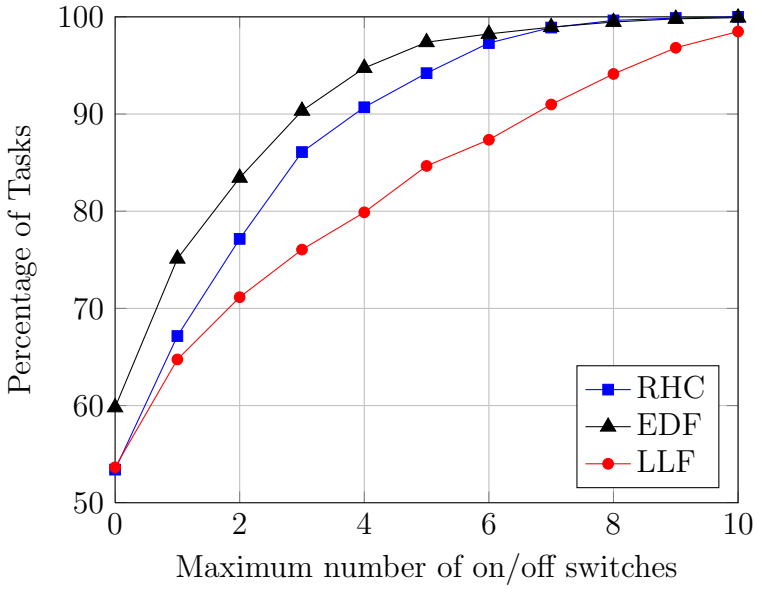


Figure 3.9: Cumulative distribution function on the maximum number of on-off switches under three scheduling algorithms: EDF, LLF, and RHC.

life and long-term operation [127]. Figure 3.9 shows a cumulative distribution function on the maximum number of on-off switches under RHC-based scheduling empirically computed over the 100 simulation test cases. One can see that most tasks (50%) are operated continuously once scheduled, and a small number (10%) are switched more than four times. Results for EDF- and LLF-based scheduling are qualitatively similar to those for RHC-based scheduling. We remark that compared to RHC, EDF exhibits less switching while LLF exhibits more.

Sporadic battery charging may have negative effects on battery health, though less so than frequent charging and discharging. If the state of health impacts can be quantified, it would be straightforward to extend the RHC framework to penalize frequent cycling in the objective function. Future work will involve addressing these issues.

3.5 Conclusions & Possible Extensions

In this chapter, we have introduced the notion of coordination aggregation of two classes of DERs: deferrable loads and storage. In particular, we have quantified the ability of scheduling algorithms for coordinated aggregation to mitigate operating reserve costs stemming from renewable variability. First, we modeled deferrable loads as tasks, and storage as devices. We then developed four algorithms (EDF, LLF, DPAS, and LPAS) for resource scheduling, and a laxity-based heuristic for reserve scheduling. We also developed an RHC algorithm for coordination aggregation with a *novel cost metric that explicitly handles operating reserve costs*. Through simulation studies, we showed that scheduling under *any* of these algorithms reduces reserve energy costs, while RHC-based scheduling also consistently reduces reserve capacity costs. Most importantly, we conclude that *the benefits of coordination aggregation can be realized at modest levels of deferrable load penetration and flexibility*.

Realizing the benefits of coordination aggregation requires technology infrastructure to exploit the capabilities of DERs. This requires research into the underlying communication and control framework for resource aggregation. How much bandwidth, latency, and reliability is required in these communication channels? What type of control architectures enable maximum leveraging of DER capabilities? These technological aspects demand exploration. Moreover, we are also developing aggregate deferrability metrics that succinctly capture the flexibility associated with a collection of tasks. For the SO, such metrics can convert the capabilities of a collection of deferrable loads into a consolidated dispatchable resource.

Realizing the benefits of coordination aggregation also requires mobilization of DERs capable of providing services to the power system. This motivates research on mechanisms that elicit and reward resource participation. Who pays for storage? How should loads be compensated for their flexibility? What are fair pricing mechanisms for flexibility? Are incentive mechanisms like lotteries, which exploit consumer biases, applicable in this setting? These economic aspects deserve detailed study.

Chapter 4

Impact of Deferrability on Ex-Ante Operations

4.1 Motivation

As demonstrated in Chapter 3, coordinated aggregation can facilitate the large-scale adoption of distributed renewable generation. By directly scheduling deferrable load and storage power profiles in direct response to variations in renewable generation, coordinated aggregation reduces the need to dispatch load-following reserves. Moreover, coordinated aggregation improves the forecastability of renewable generation in *forward* markets, enabling reductions in *ex-ante* reserve capacity and bulk power commitments. In these grid operating cost reductions lies the true value of coordinated aggregation.

Grid-level coordinated aggregation, which is performed by the system operator, is impractical for two reasons. First, there are steep communication infrastructure development and computational costs associated with centralized control of distribution side resources [60]. Second, direct control of resources falls outside the purview and business models of system operators. In fact, recent Federal Energy Regulatory Commission (FERC) rulings have highlighted some possible conflicts of interests arising from system operator-based control of resources [57]. Accordingly, we consider the case where coordinated aggregation is performed on *resource clusters*. A resource cluster is a diverse collection of networked distribution-side resources including renewable and micro-generation, deferrable loads, and electricity storage. Each cluster is managed by a Cluster Manager (CM). The CM ensures the cluster's load demands are met, aggregates the cluster's capabilities, and presents them to the system operator as a dispatchable resource.

In this chapter, we focus on policies for ex-ante operation that ensure a resource cluster's load requirements are met. Specifically, we attempt to quantify the maximum benefit in grid operating cost reductions afforded by CM-based coordinated aggregation of deferrable loads. Assuming a stylized two-stage market model, we compute cost-minimizing ex-ante procurement policies for bulk power and reserve capacity in the cases of loads having, and

not having, deferrable components. In this analysis, we use an aggregate deferrable load model consisting of a single energy need. We also characterize the optimal scheduling policy for this aggregate deferrable load component in the face of uncertainty in generation as a threshold policy via dynamic programming. In particular, we relate the optimal forward market decisions to real-time load scheduling decisions. We note that, owing to our aggregate deferrable load model, the computed reduction in operational costs is an *upper bound* on the value of coordinated aggregation in scheduling loads with arbitrary levels of flexibility.

The study of the practical aspects of implementing coordinated aggregation is an active research area. There has been considerable work in developing, and analyzing the performance of, scheduling algorithms for controlling deferrable loads in response to renewable generation [35, 118, 71]. Scheduling policies have also been formulated, via dynamic programming, for deferrable loads in the face of uncertain prices [112], and for energy storage [119]. In [99], the authors use approximate dynamic programming to couple load coordination with wind power. Other studies explore the appropriate communication and control architecture required to conduct coordinated aggregation [6, 73].

Additionally, there is well-developed literature detailing optimal ex-ante market policies for a variety of power system participants. Both [103] and [18] show that the optimal ex-ante contract for a wind power producer is a quantile on prices. In [122] and [108], the authors propose the introduction of numerous ex-ante markets to leverage better renewable generation forecasts, and find resulting optimal procurement policies. In [72], the authors find coordinated aggregation policies and day-ahead bulk power purchases that maximize social welfare. The contribution of this chapter to this field is the calculation of optimal ex-ante procurement decisions (bulk power and reserve capacity) with respect to the costs of grid operations, with and without coordinated aggregation.

The remainder of this chapter is organized as follows. Section 4.2 contains the problem formulation, where we describe generation procurement, model loads, and quantify grid operation costs. In Section 4.3, we present the optimal reserve scheduling policy under our aggregate deferrable load model. In Section 4.4, we present a series of analytical characterizations of the optimal forward market procurement policies when there are *no deferrable* loads. This is followed by Section 4.5, where we tackle the case with deferrability. Specifically, we offer intuition for forward market acquisition policies. In Section 4.6, we employ the theory developed in this chapter to numerically compute the benefit of deferrability to ex-ante operations in a set of test cases. We conclude and outline future work in Section 4.7.

4.2 Problem Formulation

We focus on the problem of procuring adequate generation to meet a resource cluster's load requirements over an operating window. Without loss of generality, let this operating window be $[0, T]$. As this chapter, like Chapter 3, focuses on meeting the load requirements of a resource cluster's, this problem formulation shares many common elements with that of Chapter 3.

Generation procurement occurs through a set of forward markets cleared in advance of this operating window, and the particular rules and regulations governing procurement vary across different balancing areas [27, 9, 28, 41]. We first describe the market structure considered in this chapter.

4.2.1 Market Structure

We assume the CM participates in a two-stage market system. This system consists of an *ex-ante* forward market and a real-time (RT) market for instantaneous balancing of supply and demand. Figure 4.1 illustrates the various stages within this market system at which a CM makes procurement decisions.

In the forward market, the CM acquires the following quantities through forward contracts: (1) bulk power (B), a constant amount of generation made available over the operating window, and (2) capacities for reserve generation (C^+, C^-). In the RT market, the ISO supplies the CM with *operating reserves* to balance demand and supply. We assume load balancing occurs at N equally spaced times within the operating window. Let $k \in \{1, \ldots, N\}$ denote these balancing times. The amount of operating reserves dispatched to the CM changes at each balancing time k and is constant over a time interval $\Delta t = \frac{T}{N}$. Crucially, this reserve dispatch is *constrained* by the amount of reserve capacity procured ex-ante.

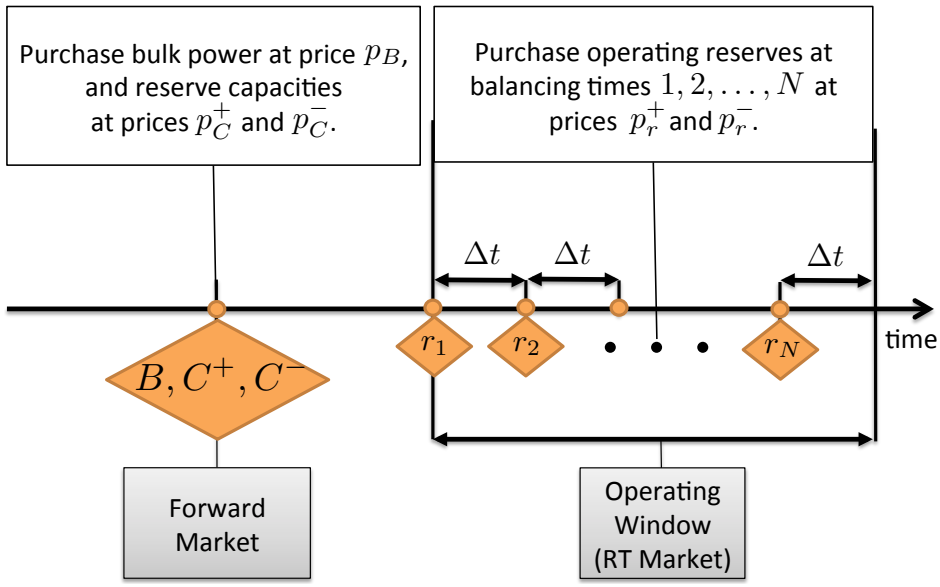


Figure 4.1: Timeline illustrating different stages at which the CM procures various generation components.

4.2.2 Generation Modeling

A CM can acquire generation from a number of sources to meet load requirements. In this chapter, we classify these sources into 3 categories, according to their variability and time of procurement.

1. Renewable generation
2. Bulk power
3. Operating reserves

Renewable generation

Renewable generation refers to all power generated within the CM's resource cluster. This refers to generation from many sources, such as rooftop solar PVs, small-scale wind farms, and residential natural gas or diesel generators. The CM must accept all such generation, and account for their inherent variability, in servicing load requirements.

To accommodate the variable nature of renewable generation, we model renewable generation as a discrete-time stochastic process. Let

$$\mathbf{w} = \{w_k\}_{k=1}^{N} \tag{4.1}$$

denote the sequence of renewable generation realizations during the operating window. Note that w_k admits values on the open interval $[0, \infty]$. For fixed $k \in \{1, \ldots, N\}$, w_k is a random variable with probability density function $f_{w_k}(w)$ and cumulative distribution function $F_{w_k}(w)$.

We assume the CM pays a price p_w (\$/MWhr) for all renewable generation from its resource cluster.

Bulk power

Bulk power refers to generation procured from the bulk transmission system through contracts agreed to ex-ante in the forward market. The CM agrees, and is contractually obliged, to accept this amount of power over the operating window. We assume this quantity, purchased in the forward market, is *constant* over the entire operating window. Working within this framework, let B refer to the constant amount of bulk power procured over the operating window $[0, T]$.

The cost of bulk power procurement explicitly depends on the forward market clearing price. We assume the CM, behaving as a *price-taker* in this market, pays a price p_B (\$/MW) for every unit of bulk power.

Operating reserves

Operating reserves refer to any additional generation dispatched to maintain instantaneous load balancing during normal, non-contingency conditions. This includes *non-event* reserves such as load-following and regulation. This does not include balancing and ramping reserves specifically procured in response to system contingencies. In this chapter, we restrict ourselves to reserves required to balance load requirements on time scales equal to, or less than, ΔT. Minor load imbalances on finer time-scales are dealt with via frequency-responsive generation and additional ancillary services [27].

Let

$$\mathbf{r} = \{r_k\}_{k=1}^{N} \tag{4.2}$$

refer to amounts of operating reserves dispatched throughout the operating window. Note that these dispatches can correspond to up ($r_k > 0$) or down ($r_k < 0$) reserves. This dispatch is *constrained* by reserve capacity procured ex-ante in the forward market. The CM procures separate capacities for up (C^+) and down (C^-) reserves.

There are two components to the cost of operating reserves:

1. **Energy**: Costs incurred within the operating interval on every unit of operating reserves dispatched. We assume the CM pays prices p_r^+, and p_r^- (\$/MWhr) on up, and down reserves respectively.

2. **Capacity**: Costs incurred in the forward market for reserve capacity procurement. We assume the CM pays prices p_C^+, and p_C^- (\$/MW) on up, and down reserve capacities respectively. We assume that these reserve capacity prices are positive.

$$p_C^+ > 0, p_C^- > 0 \tag{4.3}$$

At certain points within this chapter, we will simplify our analysis by assuming either *symmetric capacities* (4.4),

$$\begin{aligned} C^+ &= C^- = C, \\ p_C^+ &= p_C^- = p_C, \end{aligned} \tag{4.4}$$

or *symmetric reserve costs* (4.5),

$$\begin{aligned} C^+ &= C^- = C, \\ p_C^+ &= p_C^- = p_C, \\ p_r^+ &= p_r^- = p_r. \end{aligned} \tag{4.5}$$

We will also consider a modified formulation where there is *no reserve capacity*. Specifically, the CM only procures bulk power in the forward market, and has no constraints on operating reserve dispatch within the operating window.

4.2.3 Load Modeling

As discussed in Section 3.2.2, load modeling in resource clusters is, in its own right, an active research area. In this chapter, our primary focus is quantifying the benefit of deferrability. Accordingly, we decompose load requirements into two categories:

1. Static loads
2. Deferrable loads

Static loads

Static loads are characterized by their complete lack of flexibility in power consumption. We model their load requirements as a power profile:

$$\mathbf{s} = \{s_k\}_{k=1}^{N}, \tag{4.6}$$

that must be satisfied at each time k. Due to inherent randomness present in such loads owing to consumer behavior, we model the static load profile $\mathbf{s}$ as a scalar-valued discrete time stochastic process. In particular, we assume s_k is a random variable with time-varying probability density function $f_{s_k}(x)$ and cumulative distribution function $F_{s_k}(x)$.

Deferrable loads

Deferrable loads admit a range of admissible power profiles. We assume an **aggregate** model of all deferrable loads characterized by a total energy need L that needs to be satisfied over the operating window. We also impose a servicing rate limit m on the maximum instantaneous power that can be supplied.

Let

$$\mathbf{d} = \{d_k\}_{k=1}^{N}, \tag{4.7}$$

refer to the deferrable load profile. This profile must satisfy the following constraints:

$$\sum_{k=1}^{N} d_k \, \Delta t = L, \quad 0 \leq d_k \leq m \;\; \forall k \in \{1, 2, \ldots, N\}. \tag{4.8}$$

The remaining energy need at any point within the operating window is called the *energy state*.

Definition 4.1. *The **energy state** of deferrable loads at time k is the remaining energy requirement at the **start** of the k^{th} time-step.*

$$e_k = L - \sum_{\tau=1}^{k-1} d_\tau \Delta t \;\; \forall k \in \{1, 2, \ldots, N+1\} \tag{4.9}$$

Clearly, the terminal constraint $e_{N+1} = 0$ denotes satisfying the deferrable load requirement.

We admit that modeling deferrable loads in this fashion fails to capture individual power delivery constraints of the various constituent loads. Explicitly incorporating such constraints prevents an analytical characterization of the benefit of deferrability. Our calculations, using an aggregate deferrable load model, represent an *upper bound* on the benefit afforded by load flexibility.

Net load

Meeting load requirements at each time k creates the following constraint:

$$B + r_k + w_k \;=\; s_k + d_k, \quad \forall k \in \{1, 2, \ldots, N\} \tag{4.10}$$

Notice that the renewable generation and static load processes are the only random components in this analysis. In fact, the difference between static load and renewable generation fully captures the underlying randomness in the model.

Definition 4.2. *The **net load** process* $\mathbf{n}$ *is the difference between static load and renewable generation*

$$\mathbf{n} = \{s_k - w_k\}_{k=1}^{N} \tag{4.11}$$

This is a stochastic process. For each time k, the net load n_k has probability density function $f_k(x)$ and cumulative distribution function $F_k(x)$. We remark that these functions have support $\mathbb{R}$.

For each operating window, we define the time-averaged cumulative distribution function (CDF) $F(x)$ according to:

$$F(x) = \frac{\sum_{k=1}^{N} F_k(x)}{N} \tag{4.12}$$

In addition, we define the *quantile function* $F^{-1} : [0, 1] \to \mathbb{R}$ corresponding to the time-averaged CDF:

$$F^{-1}(y) = \inf \{x \in \mathbb{R} : F(x) \geq y\} \tag{4.13}$$

As one might expect, the net load process plays a crucial role in our results.

4.2.4 Cost Metrics

Working under the described model, the cost of meeting load requirements over the operating window is given by:

$$\begin{aligned}\Pi\left(B, C^+, C^-, \mathbf{r}, \mathbf{w}\right) &= p_B B + p_C^+ C^+ + p_C^- C^- \\ &\quad + p_r^+ \Delta t \sum_{k=1}^{N} [r_k]^+ + p_r^- \Delta t \sum_{k=1}^{N} [-r_k]^+ \\ &\quad + p_w \Delta t \sum_{k=1}^{N} w_k.\end{aligned} \tag{4.14}$$

where $[x]^+ = \max(x, 0)$. The first three terms in (4.14) correspond to forward market purchases. The fourth and fifth terms represent the cost of operating reserves in generation deficit (up reserve) and surplus (down reserve) scenarios respectively. The last term represents payments to renewable resources. We exclude this term in the following analysis as it is unaffected by coordinated aggregation decisions.

Due to the random nature of renewables and static loads, we are concerned with the *expected cost* where expectation is taken with respect to the net load process $\mathbf{n}$:

$$\begin{aligned}J\left(B, C^+, C^-, \mathbf{r}\right) &= p_B B + p_C^+ C^+ + p_C^- C^- \\ &\quad + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [r_k]^+ + p_r^- \Delta t \sum_{k=1}^{N} [-r_k]^+\right].\end{aligned} \tag{4.15}$$

Throughout the course of this chapter, we will analyze simplified versions of this cost function. First, consider the symmetric capacity case described by (4.4). The corresponding expected cost is:

$$J_{BC}\left(B, C, \mathbf{r}\right) = p_B B + p_C C + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [r_k]^+ + p_r^- \Delta t \sum_{k=1}^{N} [-r_k]^+\right]. \tag{4.16}$$

Second, consider the no reserve capacity case described in Section 4.2.2. The expected cost under this scenario is:

$$J_{B}\left(B, \mathbf{r}\right) = p_B B + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [r_k]^+ + p_r^- \Delta t \sum_{k=1}^{N} [-r_k]^+\right]. \tag{4.17}$$

4.2.5 Optimization Problem

We seek policies for bulk power (B), operating reserve capacity (C^+, C^-), and reserve dispatch ($\mathbf{r}$) that minimize the expected cost (4.15). As these quantities are decided upon at different stages within the procurement timeline, we formulate a two-stage optimization problem where the decision variables follow a *causal* information structure. Specifically, bulk power and reserve capacity decisions are made in the forward market (Stage I), in advance of the operating window throughout which reserve dispatch decisions are made (Stage II). Let

$\mathcal{I}_I$ and $\mathcal{I}_{II}$ be the information states representing all knowledge about the net load process at Stage I and during Stage II respectively. The expected costs at each stage can then be expressed as:

$$\begin{aligned} J_I\left(B, C^+, C^-\right) &= p_B B + p_C^+ C^+ + p_C^- C^- \\ &\quad + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [r_k{}^*]^+ + p_r^- \Delta t \sum_{k=1}^{N} [-r_k{}^*]^+ \,\middle|\, \mathcal{I}_I \right], \qquad (4.18) \\ J_{II}\left(\mathbf{r}\right) &= \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [r_k]^+ + p_r^- \Delta t \sum_{k=1}^{N} [-r_k]^+ \,\middle|\, \mathcal{I}_{II} \right], \qquad (4.19) \end{aligned}$$

where $\mathbf{r}^*$ in (4.18) is the optimal reserve dispatch policy with respect to (4.19). Having defined these cost functions, we can formulate the two-stage optimization problem by introducing appropriate load constraints. Load requirements are met if and only if $C^- \leq r_k \leq C^+$ at each balancing time k.

Stage I: Forward Market

We express load constraints as loss of load probabilities. Assuming the load must be satisfied with probability η, we can formulate the chance-constrained optimization problem (4.20) to find the cost minimizing ex-ante procurement policy:

$$\begin{aligned} \left(B^*, C^{+*}, C^{-*}\right) = \underset{B, C^+, C^-}{\operatorname{argmin}} \quad & J_I(B, C^+, C^-) \\ \text{subject to:} \quad & \mathbb{P}\left\{ r_k{}^* \in \left[C^-, C^+\right] \right\} \geq \eta, \forall k \\ & C^+ \geq 0, C^- \geq 0 \end{aligned} \qquad (4.20)$$

where $\mathbf{r}$ corresponds to the optimal reserve dispatch policy.

Stage II: Operating Window

We compute the optimal reserve scheduling policy by solving:

$$\begin{aligned} \mathbf{r}^*\left(B, C^+, C^-\right) = \underset{\mathbf{r}}{\operatorname{argmin}} \quad & J_{II}\left(\mathbf{r}\right) \\ \text{subject to:} \quad & C^- \leq r_k \leq C^+, \ \forall k \in \{1, 2, \ldots, N\} \\ & 0 \leq B - n_k + r_k \leq m, \ \forall k \in \{1, 2, \ldots, N\} \\ & e_{k+1} = e_k - (B - n_k + r_k)\Delta t, \ \forall k \in \{1, 2, \ldots, N+1\} \\ & e_1 = L, \ e_{N+1} = 0 \end{aligned} \qquad (4.21)$$

We remark that the optimal reserve dispatch policy is a function of the bulk power and reserve capacity contracted ex-ante (B, C^+, C^-).

4.3 Optimal Reserve Dispatch

4.3.1 Without Deferrability

Recall the constraint (4.10) that ensures load requirements are met through the operating interval. According to this constraint, the optimal reserve dispatch, in the absence of deferrable loads, *must* be:

$$r_k{}^* = n_k - B, \quad \forall k \in \{1, 2, \ldots, N\} \tag{4.22}$$

4.3.2 With Deferrability

Assuming the deferrable load model presented in Section 4.2.3, we find the optimal reserve dispatch policy over the operating window (Stage II) by solving the optimization problem (4.21). We do this via dynamic programming (DP).

Let $\mathbf{n}_k$ represent all values of the net load process ($\mathbf{n}$) from time k to the end of the operating window N. Similarly, let $\mathbf{r}_k$ refer to the corresponding sub-sequence for operating reserves ($\mathbf{r}$).

$$\mathbf{n}_k = \{n_i\}_{i=k}^N$$
$$\mathbf{r}_k = \{r_i\}_{i=k}^N$$

To simplify exposition, we assume, without loss of generality, that $\Delta t = 1$. As we are concerned with developing a DP solution, we need to define a *value function* for this problem. Let $V_k(e_k)$ denote this value function.

$$\begin{aligned} V_k(e_k) = \quad \min_{\mathbf{n}_k} \quad & \mathbb{E}_{\mathbf{n}_{k+1}}\left[p_r^+ \sum_{i=k}^{N} [r_i]^+ + p_r^- \sum_{i=k}^{N} [-r_i]^+\right] \\ \text{subject to:} \quad & C^- \le r_i \le C^+, \ \forall i \in \{k, k+1, \ldots, N\} \\ & 0 \le B - n_i + r_i \le m, \\ & e_{i+1} = e_i - (B - n_i + r_i), \\ & e_{N+1} = 0 \end{aligned} \tag{4.23}$$

Theorem 4.1 (Optimal reserve dispatch policy)**.** *Assuming an uncorrelated net load process* $\mathbf{n}$*, the optimal operating reserve dispatch decision at time* k *constitute a threshold policy. Specifically,* $r_k{}^* = r_+^* - r_-^*$ *where:*

$$r_+^* = \begin{cases} 0 & \text{if } \tilde{z}_+ \ge e_k \\ e_k - \tilde{z}_+ & \text{if } \tilde{z}_+ \in (e_k - \overline{C},\, e_k) \\ \overline{C} & \text{if } \tilde{z}_+ \le e_k - \overline{C} \end{cases} \tag{4.24}$$

$$r_-^* = \begin{cases} \underline{C} & \text{if } \tilde{z}_- \ge e_k + \underline{C} \\ -e_k + \tilde{z}_- & \text{if } \tilde{z}_- \in (e_k,\, e_k + \underline{C}) \\ 0 & \text{if } \tilde{z}_- \le e_k, \end{cases} \tag{4.25}$$

$\tilde{z}_+$ and $\tilde{z}_-$ are unconstrained minima of optimization problems (4.26) and (4.27) respectively:

$$\tilde{z}_+ = \underset{y_+}{\operatorname{argmin}} \; -p_r^+ y_+ \; + \; V_{k+1}\big(y_+ - B + n_k\big) \tag{4.26}$$

$$\tilde{z}_- = \underset{y_-}{\operatorname{argmin}} \; p_r^- y_- \; + \; V_{k+1}\big(y_- - B + n_k\big), \tag{4.27}$$

$\underline{C}$ and $\overline{C}$ are thresholds computed at each time k:

$$\overline{C} = \min\big(C^+, m - B + n_k\big) \tag{4.28}$$

$$\underline{C} = \min\big(-C^-, B - n_k\big)\,. \tag{4.29}$$

Proof. By induction. Using the Bellman principle of optimality, we obtain a recursive relation between value functions at successive time-steps $V_k(e_k)$ and $V_{k+1}(e_{k+1})$:

$$V_k\big(e_k\big) = \min_{-\underline{C} \le r_k \le \overline{C}} p_r^+ \left[r_k\right]^+ + p_r^- \left[-r_k\right]^+ \; + \; V_{k+1}\big(e_k - (B - n_k + r_k)\big). \tag{4.30}$$

Expressing r_k as the difference between two positive numbers, r_+ and r_- ($r_k = r_+ - r_-$), we rewrite (4.30) as the lesser minimum of two separate optimization problems in r_+ and r_-.

$$V_k\big(e_k\big) = \min \begin{cases} \min\limits_{r_+ \in \left[0, \overline{C}\right]} p_r^+ r_+ \; + \; V_{k+1}\big(e_k - B + n_k - r_+\big) \\ \min\limits_{r_- \in [0, \underline{C}]} p_r^- r_- \; + \; V_{k+1}\big(e_k - B + n_k + r_-\big) \end{cases} \tag{4.31}$$

Performing a change of variables ($y_+ = e_k - r_+, y_- = e_k + r_-$) on (4.31) gives us:

$$V_k(e_k) = \min \begin{cases} p_r^+ e_k + \min\limits_{\substack{y_+ \le e_k \\ y_+ \ge e_k - \overline{C}}} - p_r^+ y_+ + V_{k+1}\big(y_+ - B + n_k\big) \\ -p_r^- e_k + \min\limits_{\substack{y_- \ge e_k \\ y_- \le e_k + \underline{C}}} p_r^- y_- + V_{k+1}\big(y_- - B + n_k\big) \end{cases} \tag{4.32}$$

The threshold policies (4.24) and (4.25) of the claim reflect the solutions to these optimization problems.

Finally, we examine the base case. Due to the terminal constraint ($e_{N+1} = 0$), the solution ${r_N}^*$ to the value function (4.23) at time N must be:

$$\max(\min(e_N - B + n_N, \overline{C}), -\underline{C}). \tag{4.33}$$

Notice that at time N, the solutions to (4.26) and (4.27) are the same ($\tilde{z}_+ = \tilde{z}_- = B - n_N$). Hence, the dispatch decision outlined by the threshold policy is identical to (4.33). This completes the proof by induction. □

Remark 4.1. *While the reserve dispatch policy described by (4.24) and (4.25) assumes constant operating reserve prices p_r^+ and p_r^-, this formulation readily admits time-varying prices for operating reserves. To incorporate deterministic, time-varying prices, we merely find the policy according to Theorem 4.1 while allowing the reserve prices p_r^+ and p_r^- to vary as functions of k.*

4.4 Optimal Procurement Without Deferrability

In this section, we focus on optimal policies for forward market procurement assuming *no deferrability*. The policies presented here will serve as base cases with which we can compute the benefit of deferrability. Moreover, they help develop intuition regarding the interplay between forward market decisions and grid operating costs - intuition that will aid our analysis in the case with coordinated aggregation.

Throughout this section, we develop this intuition by solving variants of the forward market optimization problem (4.20). We first focus on the *no capacity* case.

4.4.1 No Capacity Case

The key simplifications made in the no capacity case are:

1. No reserve capacity procurement in forward market,
2. Unconstrained reserve dispatch within the operating window.

In this scenario, the only decision to be made in the forward market is the bulk power purchase B. Fortunately, ignoring reserve capacity considerations makes finding the optimal forward market policies, unlike (4.20), an *unconstrained* optimization problem. Assuming the reserve dispatch policy when there are no deferrable loads (4.22), we can find the optimal bulk power procurement by solving the following optimization problem.

$$B^* \quad = \quad \underset{B}{\operatorname{argmin}} \;\; p_B B \; + \; \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [n_k - B]^+ + p_r^- \Delta t \sum_{k=1}^{N} [B - n_k]^+ \right] \tag{4.34}$$

We remark that this cost function is *linear* in B. The solution to this optimization problem is a *quantile* on prices.

Theorem 4.2. *(Optimal Bulk Power Procurement: No reserve capacity) The optimal bulk power (B^*) procurement policy for meeting load requirements on an operating window of length T is the following quantile of the time-averaged net load CDF.*

$$B^* \; = \; F^{-1}\left(\frac{p_r^+ T - p_B}{(p_r^+ + p_r^-)T} \right) \tag{4.35}$$

Moreover, this optimal bulk power purchase minimizes costs if and only if: $p_r^+ + p_r^- \geq 0$.

Proof. Notice that the cost function of (4.34) $J_B(B)$ can be rewritten as:

$$\begin{aligned} J_B(B) &= p_B B \; + \; \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [n_k - B]^+ + p_r^- \Delta t \sum_{k=1}^{N} [B - n_k]^+ \right] \\ &= p_B B \; + \; p_r^+ \Delta t \sum_{k=1}^{N} \int_B^{\infty} (n - B) f_k(n) dn \; + \; p_r^- \Delta t \sum_{k=1}^{N} \int_{-\infty}^{B} (B - n) f_k(n) dn \end{aligned}$$

Clearly, $J_B(B)$ is continuous in B on $\mathbb{R}$. With the technical assumption of continuous net load distributions $f_k(n)$, it follows that $J_B(B)$ is differentiable in B. By applying Leibniz integral rule, we obtain the first and second derivatives of $J_B(B)$.

$$\frac{\partial J_B(B)}{\partial B} = p_B + p_r^+ T \frac{\sum_{k=1}^{N} F_k(B)}{N} - p_r^+ T + p_r^- T \frac{\sum_{k=1}^{N} F_k(B)}{N} \tag{4.36}$$

$$\frac{\partial^2 J_B(B)}{\partial B^2} = \left(p_r^+ + p_r^-\right) T \frac{\sum_{k=1}^{N} f_k(B)}{N} \tag{4.37}$$

where we have expressed the time-step Δt in terms of the operating window length T and the number of balancing times N.

Examining the second derivative of $J_B(B)$, we see that the condition $p_r^+ + p_r^- > 0$ guarantees convexity of the cost function. Under this condition, the cost-minimizing bulk power procurement must satisfy the first-order stationarity condition:

$$B^* = \left\{ B \in \mathbb{R} : \frac{\partial J_B(B)}{\partial B} = 0 \right\} \tag{4.38}$$

Through straightforward manipulation of the first derivative of $J_B(B)$, we recover the claim. □

Remark 4.2. *(Inventory Control) The nature of this solution - a quantile on prices - is reminiscent of well-known classical results in inventory control from the fields of economics and operations research [104, 102]. This is unsurprising. In essence, a CM performs a very similar task to the decision makers in those classical problems; it procures a product for which there is random demand ahead of time, while trying to balance costs stemming from under- and over-procurement. As we shall see in Chapter 5, similar results [103, 18] have been shown for power bids in electricity markets for wind energy. In those studies, the decision maker must commit ahead of time to selling fixed quantities of an inherently variable resource with penalties for deviating from those ex-ante commitments.*

Remark 4.3. *(Symmetric Reserve Prices) In the case of positive symmetric reserve prices, that is $p_r^+ = p_r^- = p_r > 0$, the optimal bulk power purchase can be expressed succinctly as:*

$$B^* = F^{-1}\left(\frac{1}{2} - \frac{p_B}{2 p_r T}\right) \tag{4.39}$$

*This expression has a simple, intuitive interpretation. The optimal bulk power purchase is expressed as a deviation from the **median** of the time-averaged net load process. The magnitude of this deviation depends on the ratio of prices for bulk power price p_B and reserves p_r. Figure 4.2 graphically illustrates this relationship. If the cost of procuring bulk power is significant compared to that of reserve energy, less bulk power is purchased. In effect, the CM is willing to tolerate increased reserve costs to avoid the heightened cost of acquiring bulk power. Conversely, if the reserve energy price far exceeds that of bulk power (i.e.: $p_r >> p_B$), the appropriate bulk power purchase is the median of the time-averaged net load process.*

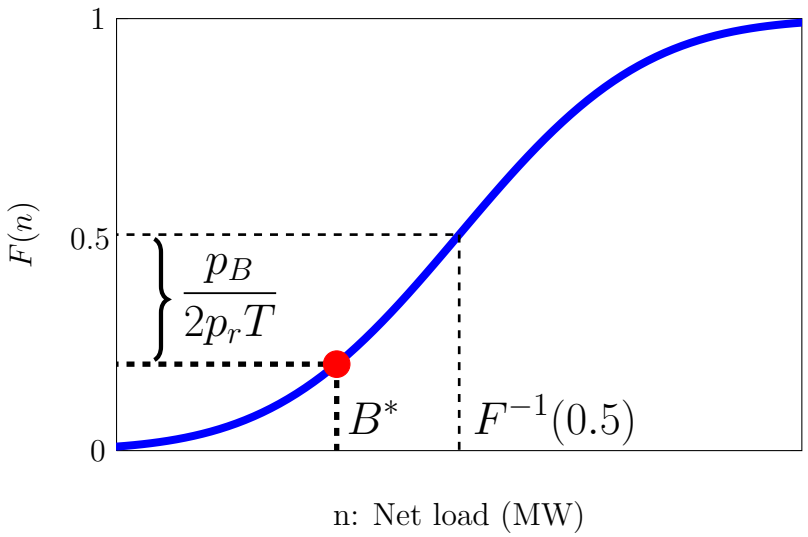


Figure 4.2: Graphical illustration of the optimal bulk power procurement policy in the cases of no reserve capacity and symmetric reserve energy costs

We remark that the bulk power procurement policy (4.39) minimizes cost if and only if $p_r > 0$. This is readily apparent from observing the structure of the second derivative of $J_B(B)$ (4.37), which for a convex function must be positive on $\mathbb{R}$.

4.4.2 Symmetric Capacities Case

Having characterized the optimal forward market procurement policy in the no reserve capacity case, we now focus on the *symmetric capacities* case. In this scenario, both up and down reserve capacities are identical ($C^+ = C^- = C$), as are their corresponding prices ($p_C^+ + p_C^- = p_C$). As a result, there are only two decisions made in the forward market: (1) bulk power B, and (2) a single reserve capacity C. Assuming the optimal reserve dispatch policy in the absence of deferrable loads (4.22), we can find the forward market procurement policies by the chance-constrained optimization problem (4.40)

$$\begin{aligned}(B^*, C^*) = \underset{B,C}{\text{argmin}} \quad & p_B B + p_C C + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [n_k - B]^+ + p_r^- \Delta t \sum_{k=1}^{N} [B - n_k]^+\right] \\ \text{subject to:} \quad & \mathbb{P}\left\{|n_k - B| \leq C\right\} \geq \eta, \forall k \\ & C \geq 0,\end{aligned} \tag{4.40}$$

This is a convex optimization problem if the underlying stochastic process, the net load process $\mathbf{n}$, has *log-concave* probability distributions [106]. We focus on solving this optimization problem assuming the net load process is *normally distributed*. Solving this particular

case, rather than one applicable for a broad class of log-concave probability distributions, offers intuition into the interplay between reserve dispatch and forward market decisions.

Lemma 4.1. *Assume the net load process* $\mathbf{n}$ *is a sequence of normally distributed random variables with time-varying means* m_k *and variances* ${\sigma_k}^2$*. The linear inequality constraints (4.41), (4.42), and (4.43) are a sufficient characterization of the constraint set of (4.40).*

$$B + C \geq \gamma_k^1 \tag{4.41}$$
$$B - C \leq \gamma_k^2 \tag{4.42}$$
$$C \geq 0 \tag{4.43}$$

where:

$$\gamma_k^1 = m_k + \phi^{-1}\left(\frac{1}{2} + \frac{\eta}{2}\right)\sigma_k \tag{4.44}$$
$$\gamma_k^2 = m_k - \phi^{-1}\left(\frac{1}{2} + \frac{\eta}{2}\right)\sigma_k \tag{4.45}$$

and $\phi^{-1}(y)$ *is the quantile function of a zero-mean, unit-variance normal distribution.*

Proof. Define three events A_1, A_2, and A_3:

A_1: $\{n_k \leq B - C\}$

A_2: $\{n_k \in (B - C, B + C]\}$

A_3: $\{n_k > B + C\}$

Clearly, $\mathbb{P}(A_2) \geq \eta$ is equivalent to $\mathbb{P}(A_1 \cup A_3) < 1 - \eta$. Using the union bound and the fact that A_1 and A_3 are disjoint events, we express $\mathbb{P}(A_1 \cup A_3)$ as $\mathbb{P}(A_1) + \mathbb{P}(A_3)$. Sufficient conditions to guarantee $\mathcal{P}(A_2) \geq \eta$ are:

$$\mathcal{P}(A_1) \leq \frac{1}{2} - \frac{\eta}{2} \Rightarrow \mathcal{P}\{n_k \leq B - C\} \leq \frac{1}{2} - \frac{\eta}{2}, \tag{4.46}$$
$$\mathcal{P}(A_3) < \frac{1}{2} - \frac{\eta}{2} \Rightarrow \mathcal{P}\{n_k \leq B + C\} \geq \frac{1}{2} + \frac{\eta}{2}. \tag{4.47}$$

We remark that (4.46), and (4.47), are *sufficient*, and not necessary, conditions to guarantee $\mathcal{P}(A_2) \geq \eta$. These constraints can be converted, using quantile functions, into the linear inequality constraints (4.41) and (4.42). □

This characterization of loss of load chance constraints enables an analytical solution of (4.40). Namely, the optimal bulk power and reserve capacity procurement policies can be expressed on a partition of prices $\left(p_B, p_C, p_r^+ T, p_r^- T\right) \in \mathbb{R}^4$ given by $\mathcal{P}_1 \cup \mathcal{P}_2 \cup \mathcal{P}_3 = \mathbb{R}^4$, where:

$$\begin{aligned}\mathcal{P}_1 =&\{\ (x, y, z_1, z_2)\ \ |g(x, z_1, z_2, \alpha) \leq -y\ \}\\ \mathcal{P}_2 =&\{\ (x, y, z_1, z_2)\ \ |\,|g(x, z_1, z_2, \alpha)| \leq y\ \}\\ \mathcal{P}_3 =&\{\ (x, y, z_1, z_2)\ \ |g(x, z_1, z_2, \alpha) \geq y\ \},\end{aligned}$$

and α, β, and $g(x, z, B)$ are:

$$\alpha = \frac{\max_k\ \gamma_k^1\ +\ \min_k\ \gamma_k^2}{2} \tag{4.48}$$

$$\beta = \frac{\max_k\ \gamma_k^1\ -\ \min_k\ \gamma_k^2}{2} \tag{4.49}$$

$$g(x, z_1, z_2, B) = x + (z_1 + z_2)F(B) - z_1 \tag{4.50}$$

Theorem 4.3 (Optimal forward market procurement policies)**.** *Assume the net load process* $\mathbf{n}$ *is a sequence of normally distributed random variables, the optimal forward market bulk power* (B^*) *and reserve capacity* (C^*) *procurement policies can be expressed as a function of the energy procurement prices* $\rho \in \mathbb{R}^4$*:*

$$B^* = \begin{cases} \inf\{B : g(p_B, p_r^+T, p_r^-T, B) \geq -p_C\}, & \text{if } \rho \in \mathcal{P}_1\\ \alpha, & \text{if } \rho \in \mathcal{P}_2\\ \sup\{B : g(p_B, p_r^+T, p_r^-T, B) \leq p_C\}, & \text{if } \rho \in \mathcal{P}_3 \end{cases} \tag{4.51}$$

$$C^* = \begin{cases} -\min\limits_k\ \gamma_k^2 + B^*, & \text{if } \rho \in \mathcal{P}_1\\ \beta, & \text{if } \rho \in \mathcal{P}_2\\ \max\limits_k\ \gamma_k^1 - B^*, & \text{if } \rho \in \mathcal{P}_3 \end{cases} \tag{4.52}$$

Moreover, this procurement policy minimizes costs if and only if: $p_r^+ + p_r^- \geq 0$.

Proof. The feasible set $\mathcal{X}$ for problem (4.40) on the space of decision variables $x = (B, C) \in \mathbb{R}^2$ is the intersection of constraints (4.41), (4.42), and (4.43). Figure 4.3 is a graphical illustration of these constraints and the resulting set $\mathcal{X}$. Clearly, $\mathcal{X}$ is fully characterized by two inequality constraints:

1. $B + C \geq \max\limits_k \gamma_k^1$
2. $B - C \leq \min\limits_k \gamma_k^2$

The lines defining these constraints intersect at the point (α, β).

Let $J_{BC}(B, C)$ refer to the cost function for this optimization problem. With the assumption of a normally distributed net load process $\mathbf{n}$, $J_{BC}(B, C)$ is differentiable in B and

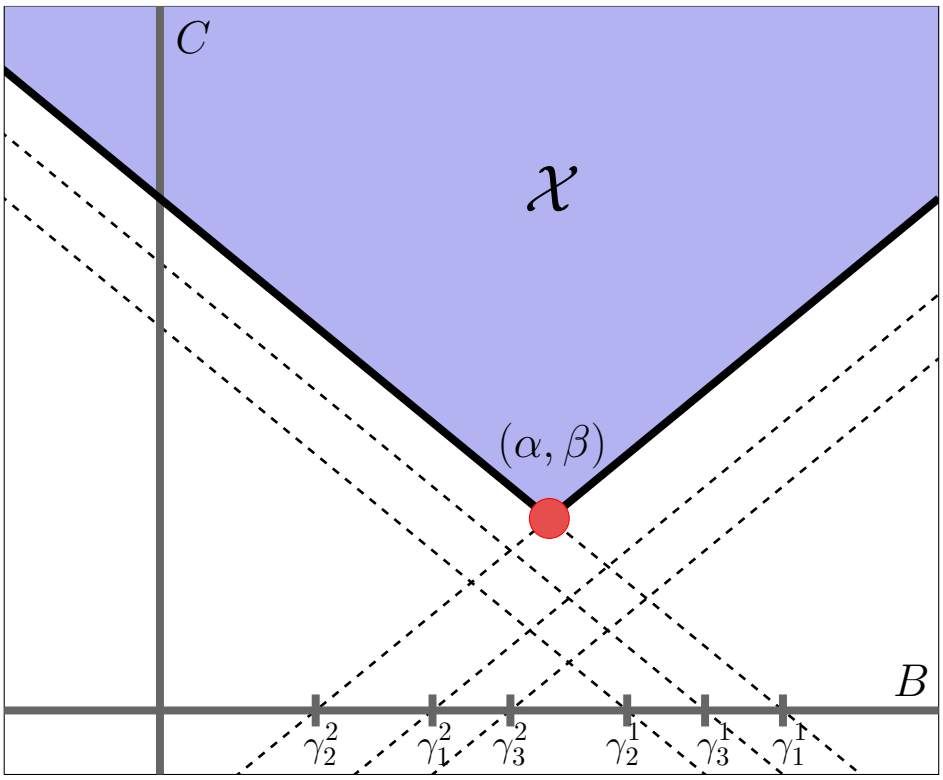


Figure 4.3: Illustration of the feasible set of DA procurement policies $\mathcal{X}$

C. By applying Leibniz integral rule, we obtain the first derivatives of $J_{BC}(B, C)$:

$$\frac{\partial J_{BC}}{\partial B} = p_B + \left(p_r^+ T + p_r^- T\right) \frac{\sum_{k=1}^{N} F_k(B)}{N} - p_r^+ T \tag{4.53}$$

$$\frac{\partial J_{BC}}{\partial C} = p_C \tag{4.54}$$

Notice that $\dfrac{\partial J_{BC}(B)}{\partial B}$ is a *non-decreasing* function of B as long as $p_r^+ + p_r^- \geq 0$. In fact, simple analysis of the Hessian of $J_{BC}(B, C)$ confirms that this condition guarantees convexity of the cost function.

The optimal procurement policy $x^* = (B^*, C^*)$ must satisfy:

$$\nabla J_{BC}(x^*)^T (x - x^*) \geq 0, \ \forall x \in \mathcal{X} \tag{4.55}$$

Clearly, as $p_C \geq 0$ (4.3), x^* must belong to the boundary of the feasible set, $\text{bn}(\mathcal{X})$. $\text{bn}(\mathcal{X})$ can be decomposed into:

1. (α, β)
2. $\{(B, \max_k \gamma_k^1 - B), \ \forall B < \alpha\}$

3. $\{(B, B - \min_k \gamma_k^2),\ \forall B > \alpha\}$

We first examine (α, β). In this case, the optimality condition (4.55) can be expressed as $\left|\frac{\partial J_{BC}(\alpha)}{\partial B}\right| \leq \frac{\partial J_{BC}}{\partial C}$. Comparing this condition to the price partition defined in (4.48), it is clear that this particular optimality condition condition is equivalent to $\left(p_B, p_C, p_r^+ T, p_r^- T\right) \in \mathcal{P}_2$. Moreover, notice that:

- $\frac{\partial J_{BC}(\alpha)}{\partial B} \leq -\frac{\partial J_{BC}}{\partial C}$ is equivalent to $\left(p_B, p_C, p_r^+ T, p_r^- T\right) \in \mathcal{P}_1$, and
- $\frac{\partial J_{BC}(\alpha)}{\partial B} \geq \frac{\partial J_{BC}}{\partial C}$ is equivalent to $\left(p_B, p_C, p_r^+ T, p_r^- T\right) \in \mathcal{P}_3$.

If $\left(p_B, p_C, p_r^+ T, p_r^- T\right) \in \mathcal{P}_1$, the optimal policy belongs to the set $\{(B, B - \min_k \gamma_k^2),\ \forall B > \alpha\}$ as $\frac{\partial J_{BC}(B)}{\partial B}$ is a non-decreasing function of B. On this set, the optimality condition (4.55) reduces to $\frac{\partial J_{BC}(B)}{\partial B} = -p_C$. It follows that the optimal procurement policy is $\left(B^*, B^* - \min_k \gamma_k^2\right)$ where B^* is the smallest power purchase satisfying the optimality condition $\left(\inf\left\{B : \frac{\partial J_{BC}(B)}{\partial B} \geq -p_C\right\}\right)$.

Alternatively, in the case when $\left(p_B, p_C, p_r^+ T, p_r^- T\right) \in \mathcal{M}_3$, the minimum is in the set $\{(B, \max_k \gamma_k^1 - B),\ \forall B < \alpha\}$. In this case, the optimality condition (4.55) reduces to $\frac{\partial J_{BC}(B)}{\partial B} = p_C$. Accordingly, the optimal procurement policy is $\left(B^*, \max_k \gamma_k^1 - B^*\right)$ where B^* is the largest power purchase that satisfies the optimality condition $\left(\sup\left\{B : \frac{\partial J_{BC}(B)}{\partial B} \leq p_C\right\}\right)$. As $x^* \in \mathrm{bn}(\mathcal{X})$, the choice of B^* completely characterizes C^*. □

Remark 4.4. *(IID Net Load Process) If the net load process* $\mathbf{n}$ *is a sequence of independent and identically-distributed (IID) normal random variables, the optimal DA procurement policy has a simple, intuitive interpretation. Let* m *be the mean and* σ^2 *the variance of this IID process. If the procurement prices satisfy:*

$$\left|p_B + \frac{(p_r^- - p_r^+)\, T}{2}\right| \leq p_C, \tag{4.56}$$

the optimal DA procurement policy, using (4.51) and (4.52), is:

$$B^* = m \tag{4.57}$$

$$C^* = \phi^{-1}\left(\frac{1}{2} + \frac{\eta}{2}\right)\sigma \tag{4.58}$$

where η, as in (4.40), describes the probability load is satisfied.

Clearly, the optimal bulk power purchase B^* is equal to the ***expected net load****. This is identical to existing DA practices employed by many system operators faced with scheduling generation to meet system load requirements [27, 41].*

The optimal policy for reserve capacity (4.58) suggests procuring an amount proportional to the ***standard deviation*** *of net load. Crucially, this proportionality constant is directly dependent on the loss of load probability η the CM can tolerate. Currently employed reserve capacity procurement levels, determined by federal and other regulatory mandates, do not explicitly depend on such metrics of renewable generation and load variability [95]. For instance, the operating reserve capacity requirement for system operators in the Western US is 5-7% of total load [95]. System operators employing (4.58) can explicitly incorporate load and renewable uncertainty estimates in their capacity procurement procedures. Moreover, this policy enables reserve capacity cost reductions for delivery intervals with low forecast uncertainty.*

Remark 4.5. *(Degenerate Procurement Price Conditions) Theorem 4.3 shows that the optimal procurement policy can be found for a given combination of procurement prices and statistics of the net load process. However, it is apparent that for certain combinations of prices, (4.51) and (4.52) suggest a degenerate policy of infinite bulk power or reserve capacity procurement.*

According to (4.51), infinite forward market procurements can only occur on the price partitions $\mathcal{P}_1$ and $\mathcal{P}_3$. To examine price conditions that result in degenerate procurements, we observe the following properties of $g(p_B, p_r^+T, p_r^-T, B)$. For compactness of notation, we refer to this function as $\hat{g}(B)$.

$$\max_B \hat{g}(B) \;=\; p_B + p_r^- T \tag{4.59}$$

$$\min_B \hat{g}(B) \;=\; p_B - p_r^+ T \tag{4.60}$$

Consider the price partition $\mathcal{P}_1$. Clearly, (4.51) suggests an infinite optimal procurement $(B^*, C^*) = (\infty, \infty)$ if $\max_B \hat{g}(B) < -p_C$. This condition can be expressed as:

$$-p_C - p_r^- T > p_B. \tag{4.61}$$

In effect, the cost of procuring bulk power is less than the reward possible by selling down reserves. Hence, the optimal policy amounts to one of arbitrage - procure bulk power ex-ante and supply said power back to the SO as down reserves during the operating interval.

Now, consider the price partition $\mathcal{P}_3$. In this case, a degenerate optimal procurement $(B^*, C^*) = (-\infty, \infty)$ occurs if $\min_B \hat{g}(B) > p_C$. This condition is equivalent to:

$$p_B > p_C + p_r^+ T. \tag{4.62}$$

In this case, the cost of reserving and dispatching up reserves is less than the price of bulk power. Therefore, the CM can engage in a different form of arbitrage - sell bulk power in the

forward market and meet these commitments using up reserves dispatched by the SO during the operating interval.

These special cases illustrate energy arbitrage opportunities for a CM participating in this type of two-stage market system. Importantly, they can help SOs design pricing structures for similar market systems that prevent such behavior. However, it is important to note that these degenerate cases also allude to the limitations of our modeling framework. In this analysis, we have assumed that the CM behaves as a price-taker in the forward market for bulk power. This assumption only holds if the CM's bulk power purchase is small relative to the total generation cleared through the market. In situations where the optimal policy amounts to one of infinite bulk power purchase, the price-taker assumption is no longer valid.

4.4.3 Asymmetric Capacities Case

Finally, we tackle the general case with separate reserve capacities for up (C^+) and down (C^-) reserves. In this case, the optimal procurement policy can be found by solving the optimization problem (4.20) with the optimal reserve scheduling policy r^* described by (4.22).

$$\begin{aligned}\left(B^*, C^{+*}, C^{-*}\right) = \underset{B,C^+,C^-}{\operatorname{argmin}} \quad & J_I(B, C^+, C^-) \\ \text{subject to:} \quad & \mathbb{P}\left\{(n_k - B) \in \left[C^-, C^+\right]\right\} \geq \eta, \forall k \\ & C^+ \geq 0, C^- \geq 0\end{aligned} \tag{4.63}$$

where the cost function $J_I(B, C^+, C^-)$ is described by:

$$\begin{aligned}J_I(B, C^+, C^-) \;=\; & p_B B + p_C^+ C^+ + p_C^- C^- \\ & + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [n_k - B]^+ + p_r^- \Delta t \sum_{k=1}^{N} [B - n_k]^+\right]\end{aligned} \tag{4.64}$$

As with the symmetric capacity case, this is a chance-constrained optimization problem that is convex if the net load process ($\mathbf{n}$) has *log-concave* probability distributions. Again, we focus on the case where the net load is normally distributed. Using the same technique described in Lemma 4.1, the constraints of this optimization problem can, with the same technique as , be sufficient characterized by a series of linear inequality constraints (4.65).

$$\begin{aligned}\underset{B,C^+,C^-}{\operatorname{argmin}} \quad & J_I(B, C^+, C^-) \\ \text{subject to:} \quad & B + C^+ \geq \gamma_k^1, \quad \forall k \\ & B - C^- \leq \gamma_k^2, \quad \forall k \\ & C^+ \geq 0, C^- \geq 0\end{aligned} \tag{4.65}$$

where the quantities γ_k^1 and γ_k^2 were defined by (4.44) and (4.45) respectively.

In this case, a precise analytical characterization of the optimal policy is not possible. We can however present the Karush-Kuhn-Tucker conditions for this problem. The KKT conditions, when paired with a suitable regularity condition, provide *necessary* first-order conditions for optimality [25].

Remark 4.6. *(KKT Optimality Conditions) A set of ex-ante procurements $x^* \in \mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+$ is optimal with respect to (4.65) if there exists a corresponding set of dual variables $\lambda^* \in \mathbb{R}^4$ such that the pair (x^*, λ^*) satisfies the following conditions:*

First-Order Stationarity

$$\begin{aligned} g(B^*) &= {\lambda_1}^* - {\lambda_2}^*, \\ p_C^+ &= {\lambda_1}^* + {\lambda_3}^*, \\ p_C^- &= {\lambda_2}^* + {\lambda_4}^*, \end{aligned} \tag{4.66}$$

Dual Feasibility

$$\lambda_i{}^* \geq 0, \quad \forall i \in \{1,2,3,4\}, \tag{4.67}$$

Primal Feasibility

$$\begin{aligned} {C^+}^* &\geq 0, \qquad B^* + {C^+}^* \geq \gamma^1, \\ {C^-}^* &\geq 0, \qquad B^* - {C^-}^* \leq \gamma^2, \end{aligned} \tag{4.68}$$

Complementary Slackness

$$\begin{aligned} {\lambda_3}^* {C^+}^* &= 0, \qquad {\lambda_1}^* \left(B^* + {C^+}^* - \gamma^1\right) = 0, \\ {\lambda_4}^* {C^-}^* &= 0, \qquad {\lambda_2}^* \left(B^* - {C^-}^* - \gamma^2\right) = 0, \end{aligned} \tag{4.69}$$

where

$$g(B) \;=\; p_B + \left(p_r^+ T + p_r^- T\right) F(B) - p_r^+ T, \tag{4.70}$$

$$\gamma^1 = \max_k \gamma_k^1 \,, \quad \gamma^2 = \min_k \gamma_k^2 \,; \tag{4.71}$$

and *the problem is convex, which is the case if and only if: $p_r^+ + p_r^- \geq 0$.*

Proof. We begin by observing that the feasible set for problem (4.65), on the space of decision variables $x = (B, C^+, C^-) \in R^3$, can be fully characterized by the 4 inequality constraints presented in (4.68). With this description of the feasible set, the conditions (4.66-4.69) are simple applications of the standard KKT conditions [25, Section 5.5.3]. These conditions signify optimality if the problem: (1) is convex, and (2) satisfies some *constraint qualification* condition.

First, we tackle convexity. Notice that the feasible set for this problem, defined by linear inequality constraints, is a convex set. Moreover, we observe that all second-order derivatives of $J_I(B, C^+, C^-)$ are 0 except for:

$$\frac{\partial^2 J_I}{\partial B^2} = \left(p_r^+ T + p_r^- T\right) \frac{\sum_{k=1}^N f_k(B)}{N} \tag{4.72}$$

The condition $p_r^+ + p_r^- \geq 0$, by ensuring $J_I(B, C^+, C^-)$ has a positive semi-definite Hessian, is a necessary and sufficient condition for convexity.

Second, we show that this problem satisfies *Slater's condition*, a well-known constraint qualification condition. For an optimization problem, Slater's condition holds if there is a *strictly feasible* point within its set of decision variables. For problem (4.65), this condition holds if there exists some point that satisfies the inequality constraints described by (4.68) with *strict* inequality. We submit that for some small $\epsilon > 0$, the point

$$B = \frac{\gamma^1 + \gamma^2}{2}, \quad C^+ = C^- = \frac{\gamma^1 - \gamma^2}{2} + \epsilon, \tag{4.73}$$

is one such feasible point that strictly satisfies all inequality constraints. As Slater's condition holds for this convex problem, the KKT conditions described are indeed optimality conditions. □

4.5 Optimal Procurement With Deferrability

In this section, we consider optimal procurement policies in the presence of deferrable loads. We will work within the framework of the *symmetric capacities case*. While we can apply many aspects of this analysis to the asymmetric capacities case, the critical arguments and observations that relay intuition about the benefit of deferrability are not as apparent as in the symmetric capacities case.

Consider the threshold policy for reserve dispatch outlined in Theorem 4.1. Clearly, this policy determines power delivery schedules to deferrable loads within the operating window. Notice that this policy is the optimal reserve dispatch policy with access to the operating interval information state ($\mathcal{I}_{II}$). We seek optimal forward market policies assuming this policy for reserve, and in turn deferrable load, scheduling. This is a difficult problem. To the best of our knowledge, this threshold policy for reserve dispatch prevents a simple analytical characterization of the optimal forward market policy (B^*, C^*).

We instead approximate the solution to the two-stage optimization problem by introducing additional decision variables, representing reserve scheduling in the presence of deferrable loads, when finding the optimal forward market procurement. Thus, we find the optimal forward market and load scheduling policies $(B^*, C^*, \mathbf{d}^*)$ by solving the following modified version of the optimization problem (4.20).

$$\begin{aligned} \underset{B,C,\mathbf{d}}{\text{argmin}} \quad & J_I\left(B, C, \mathbf{d}\right) \\ \text{subject to:} \quad & \mathbb{P}\left\{|d_k + n_k - B| \leq C\right\} \geq \eta, \forall k \\ & C \geq 0, \\ & \sum_{k=1}^{N} d_k \, \Delta t = L, \quad 0 \leq d_k \leq m \ \ \forall k. \end{aligned} \tag{4.74}$$

where the cost function $J_I(B, C, \mathbf{d})$ is described by:

$$\begin{aligned} J_I(B, C, \mathbf{d}) \;=\; & p_B B + p_C C \\ & + \mathbb{E}\left[p_r^+ \Delta t \sum_{k=1}^{N} [d_k + n_k - B]^+ + p_r^- \Delta t \sum_{k=1}^{N} [B - d_k - n_k]^+ \right] \end{aligned} \tag{4.75}$$

In effect, solving this relaxed problem (4.74) yields the optimal reserve dispatch policy with access to the forward market ($\mathcal{I}_{II}$), rather than the real-time ($\mathcal{I}_{II}$), information state. While this approach yields sub-optimal forward market procurement policies with respect to the original two-stage problem formulation, this approach offers insight into the impact of reserve scheduling on forward market decisions.

As described in Section 4.4.2, this problem is convex if the net load process $\mathbf{n}$ has log-concave probability distributions. Employing the same technique described in Lemma 4.1, we observe the following.

Lemma 4.2. *Assuming the net load process* $\mathbf{n}$ *is a sequence of normally distributed random variables with time varying means* m_k *and variances* ${\sigma_k}^2$*, the chance constraints in (4.74) can be sufficiently characterized by the following linear inequalities.*

$$B + C \geq \gamma_k^3 + d_k \tag{4.76}$$
$$B - C \leq \gamma_k^4 + d_k \tag{4.77}$$

where:

$$\gamma_k^3 = m_k + \phi^{-1}\left(\frac{1}{2} + \frac{\eta}{2}\right)\sigma_k \tag{4.78}$$
$$\gamma_k^4 = m_k - \phi^{-1}\left(\frac{1}{2} + \frac{\eta}{2}\right)\sigma_k \tag{4.79}$$

and $\phi^{-1}(y)$ *is the quantile function of a zero-mean, unit-variance normal distribution.*

Proof. The proof is identical to that for Lemma 4.1 with $\gamma_k^3 + d_k$ playing the role of γ_k^1 and $\gamma_k^4 + d_k$ playing the role of γ_k^2. □

Therefore, the optimization problem (4.74) only has linear inequality or equality constraints.

Remark 4.7. *(Constraints Comparison with No Deferrability Case) Clearly, the constraints modeling loss of load probabilities in the case with deferrable loads, (4.76) and (4.77), differ from their counterparts in the no deferrability case, (4.41) and (4.42), only through choice of* d_k*. Specifically, these constraints are equivalent if:*

$$\gamma_k^1 = \gamma_k^3 + d_k, \tag{4.80}$$
$$\gamma_k^2 = \gamma_k^4 + d_k. \tag{4.81}$$

In the absence of deferrable loads, these constraints are fixed and determined solely by the statistics of the net load process. In the presence of deferrable loads, judicious allocations to such loads, parametrized by d_k, enables the modulation of these constraints. The cost reductions on ex-ante procurements (B, C) afforded by this ability to modulate constraints is an estimate of the financial benefit afforded by deferrability.

We now analyze the impact of deferrable load scheduling decisions ($\mathbf{d}$) on the optimal forward market procurement policy $(B^*.C^*)$. Specifically, we show that, similar to the no deferrability case, the optimal procurement policies as functions of the deferrable load allocations $(B^*(\mathbf{d}), C^*(\mathbf{d}))$ can be expressed on a partition of procurement prices $\left(p_B, p_C, p_r^+T, p_r^-T\right) \in \mathbb{R}^4$ given by $\mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3 = \mathbb{R}^4$. Crucially, this partition *depends* on the load scheduling decisions $\mathbf{d}$.

$$\begin{aligned}
\mathcal{D}_1 =&\{\ (x, y, z_1, z_2)\ |g_{\mathbf{d}}(x, z_1, z_2, \alpha) \leq -y\ \} \\
\mathcal{D}_2 =&\{\ (x, y, z_1, z_2)\ |\,|g_{\mathbf{d}}(x, z_1, z_2, \alpha)| \leq y\ \} \\
\mathcal{D}_3 =&\{\ (x, y, z_1, z_2)\ |g_{\mathbf{d}}(x, z_1, z_2, \alpha) \geq y\ \},
\end{aligned}$$

and α, β, and $g(x, z, B)$ are:

$$\alpha = \frac{\max_k\ (\gamma_k^3 + d_k)\ +\ \min_k\ (\gamma_k^4 + d_k)}{2} \tag{4.82}$$

$$\beta = \frac{\max_k\ (\gamma_k^3 + d_k)\ -\ \min_k\ (\gamma_k^4 + d_k)}{2} \tag{4.83}$$

$$g_{\mathbf{d}}(x, z_1, z_2, B) = x + (z_1 + z_2)\frac{\sum_{k=1}^N \bar{F}_k(B)}{N} - z_1 \tag{4.84}$$

where $\bar{F}_k(x)$ is the CDF of a normal distribution with mean $d_k + m_k$ and variance ${\sigma_k}^2$.

Theorem 4.4. *(Optimal forward market procurement policy with deferrable loads) Assume the net load process* $\mathbf{n}$ *is sequence of normal random variables. The optimal forward market bulk power and reserve capacity procurement policies, as functions of the deferrable load allocations* $\mathbf{d}$*, can be expressed on a partition* $(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3)$ *of the space of procurement prices* $\left(p_B, p_C, p_r^+T, p_r^-T\right) \in \mathbb{R}^4$*.*

$$B^*(\mathbf{d}) = \begin{cases} \inf\{B : g_{\mathbf{d}}(p_B, p_r^+T, p_r^-T, B) \geq -p_C\}, & \text{if } \rho \in \mathcal{D}_1 \\ \alpha, & \text{if } \rho \in \mathcal{D}_2 \\ \sup\{B : g_{\mathbf{d}}(p_B, p_r^+T, p_r^-T, B) \leq p_C\}, & \text{if } \rho \in \mathcal{D}_3 \end{cases} \tag{4.85}$$

$$C^*(\mathbf{d}) = \begin{cases} -\min_k\ \left(\gamma_k^4 + d_k\right) + B^*, & \text{if } \rho \in \mathcal{D}_1 \\ \beta, & \text{if } \rho \in \mathcal{D}_2 \\ \max_k\ \left(\gamma_k^3 + d_k\right) - B^*, & \text{if } \rho \in \mathcal{D}_3 \end{cases} \tag{4.86}$$

Moreover, this procurement policy minimizes costs if and only if: $p_r^+ + p_r^- \geq 0$.

Proof. The proof technique is identical to that of Theorem 4.3. □

Remark 4.8. *(Load-Scheduling Heuristic) Given a particular choice of feasible deferrable load allocation* $\mathbf{d}$*, the optimal forward market procurement policy* (B^*, C^*) *has the same form (Theorem 4.4) as that in the no deferrability case (Theorem 4.3). Notably, the choice of* $\mathbf{d}$ *only affects the optimal procurement policy in two ways:*

1. *The chance constraints described by (4.76) and (4.77) and,*
2. *The rule for bulk power procurement through the function* $g_{\mathbf{d}}(p_B, p_r^+ T, p_r^- T, B)$ *as defined by (4.84).*

Armed with this characterization of bulk power and reserve capacity, we could formulate and numerically solve a stochastic optimization problem to determine the optimal deferrable load allocation $\mathbf{d}$*. However, we eschew this approach in favor of a simple load scheduling heuristic motivated by the structure of the optimal forward market policy.*

Consider the quantities α *(4.82) and* β *(4.83). Clearly, decreased values of* α *and* β *result in a feasible set for the optimization problem that accomodates reduced procurements of bulk power and reserve capacity. Minimal values of* α *and* β *are obtained by simultaneously:*

1. ***minimizing*** $\max_k \gamma_k^3 + d_k$ *while,*
2. ***maximizing*** $\min_k \gamma^4 + d_k$.

Intuitively, the choice of deferrable load allocations $\mathbf{d}$ *that reduces the difference between these two quantities will minimize ex-ante procurement costs. Accordingly, a load-scheduling heuristic can be found by solving (4.87).*

$$\begin{aligned} \mathbf{d}^* = \underset{\mathbf{d}}{\text{argmin}} \quad & \max_i \left(\gamma_i^3 + d_i\right) + \max_j \left(-\gamma_j^4 - d_j\right) \\ \text{subject to:} \quad & 0 \leq d_k \leq m, \ \forall k \\ & \sum_{k=1}^{N} d_k \Delta t = L \end{aligned} \tag{4.87}$$

4.6 Simulation Test Cases

In this section, we illustrate key aspects of the theory developed in this chapter through simulations. Specifically, we calculate the reductions in ex-ante procurements offered by deferrability. First, we demonstrate the concepts developed in this chapter on a synthetic example. We then apply this analysis framework to a more realistic test case based on load and solar generation data.

4.6.1 Synthetic Example

Parameters

For this example, we work under the assumption of symmetric capacities as described by (4.4). Hence, the CM procures two quantities, B and C, for each hour-long operating window. There are 4 balancing times within this operating window. At each balancing time k, the net load is normally distributed with the means (μ_k) and standard deviations (σ_k) listed in Table 4.1.

k	μ_k (kW)	σ_k (kW)
1	525	50
2	550	50
3	475	50
4	450	50

Table 4.1: Net load statistics for the synthetic test case

We choose a loss of load probability $\eta = 0.997$. This makes the load constraint a 'three-σ' rule - generation procurement must account for net load deviations within 3 standard deviations of the mean.

To facilitate analysis, we assume the procurement prices $(p_B, p_C, p_r^+T, p_r^-T)$ always belong to the set $\mathcal{D}_2$ in the partition of procurement prices described in Theorem 4.4, or the set $\mathcal{P}_2$ in the partition defined in Theorem 4.3. This ensures that the optimal procurement policy is described by the vertex (α, β) of the feasible set $\mathcal{X}$ (Figure 4.3).

To understand the value of deferrable loads, we vary both the energy need of the deferrable load component (L) as well as the servicing rate limit (m). The energy need is varied on the interval $[0, 75]$ (kWh). As the total load requirement over this operating window is 500 kWh, this corresponds to deferrable load penetrations of upto 15%. The servicing rate limit is varied from $\frac{L}{4\Delta t}$ (constant, inflexible deferrable load profile) to $\frac{L}{\Delta t}$ (no rate limits). For each of these deferrable load characterizations, we compute the ex-ante procurements based on the load-scheduling heuristic described in (4.87).

Results

Figure 4.4 compares the feasible sets for ex-ante procurements for cases with and without deferrability. The values γ_1 and γ_2 characterize this set when load is purely static. With deferrable loads, employing the load-scheduling heuristic (4.87) yields a set of feasible procurements described by γ_1^* and γ_2^*. Clearly, the load-scheduling heuristic reduces ex-ante procurement needs by modulating constraints on these decision variables. In this case, the optimal reserve capacity procurement decreases from 200 to 160 kW.

Figure 4.5 shows the improvements in reserve capacity procurements possible at various levels of deferrable energy need (L) and rate limit (m). It is apparent that greater reserve

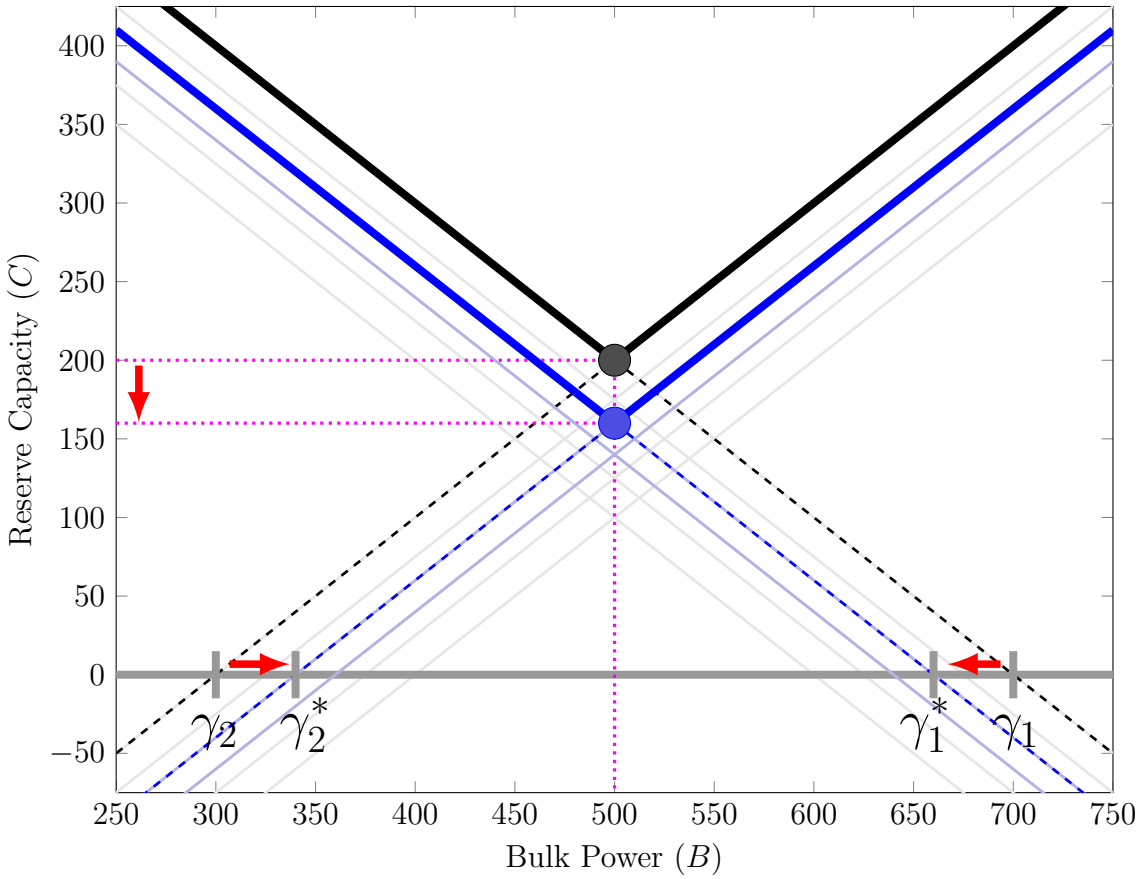


Figure 4.4: Illustration of how deferrable loads reduce optimal ex-ante procurements using this framework. The blue and black circles correspond to the optimal procurements with and without deferrability respectively. This particular case assumes deferrable load parameters: $L = 40$ kWh and $m = 80$ kW.

capacity reductions are possible with: (1) higher deferrable load penetrations, and (2) less stringent servicing rate limits. However, these results also suggest that there is an upper bound on the marginal benefit of additional deferrable loads. This occurs because there is a fundamental limit to the benefit resulting from modulating constraints. Further reductions in ex-ante procurements can only be realized through other means such as, for instance, improved forecasting of net load.

4.6.2 Solar Penetration Test Case

Parameters

This test case is based on solar generation data from the PV integrator Solar City and aggregate load data from the California ISO. These time series data are the same as those

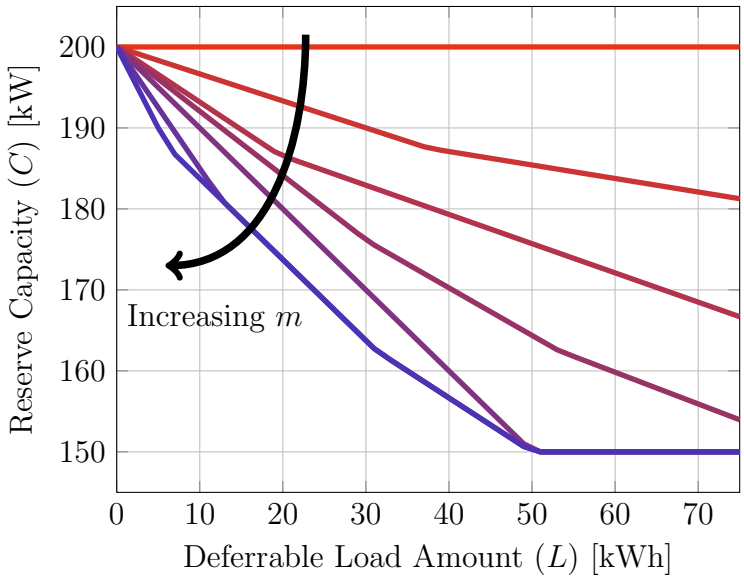


Figure 4.5: Optimal reserve capacity procurement as a function of total deferrable load energy need (L) and servicing rate limit (m). The total load requirement is 500 kWh.

used for analysis of real-time scheduling algorithms in Section 3.4.

As in the case with the synthetic example, we make the following assumptions in our analysis:

A1 Symmetric Capacities for up and down reserves as described by (4.4).

A2 Loss of load probability $\eta = 0.997$ in line with the 'three-σ' loss of load rule.

A3 Procurement prices $(p_B, p_C, p_r^+ T, p_r^- T)$ always belong to the set $\mathcal{D}_2$ in the partition of procurement prices described in Theorem 4.4, thereby ensuring that the optimal procurement policy is described by the vertex (α, β) of the feasible set $\mathcal{X}$ (Figure 4.3).

We compute ex-ante procurements separately for each hour of the day. There are 6 balancing times within each hour-long operating window ($N = 6$, $\Delta t = 10$ mins). We study the impact of deferrable load size on procurements by varying the proportion, α, of the total load energy requirement that is deferrable. Let $\mathbf{l}$ be the DA forecast of total load over an operating window. For a given value of α, we compute the deferrable load size and consequently, the static load profile $\mathbf{s} = \{s_k\}$ according to (4.88) and (4.89).

$$L = \alpha \sum_{k=1}^{N} l_k \Delta t \tag{4.88}$$

$$s_k = l_k - \frac{L}{N\Delta t}, \quad \forall k \in \{1, \ldots, N\} \tag{4.89}$$

For each balancing time k, we compute day-ahead forecasts of net load based on this data. To do this, we use variants of the methods outlined in Section 3.4.

1. **Static Loads**: We use day-ahead load forecasts, corresponding to the load time-series data, from the California ISO. These hourly demand forecasts are normalized. Let $\widetilde{s}$ refer to the forecasted static load for a given hour.

2. **Renewable Generation**: For each hour of the operating window, we calculate the average generation over the past 5 days. Let $\widetilde{w}$ refer to the forecasted renewable generation for a given hour.

Let $\widetilde{n}$ refer to the net load forecast for this hour. Clearly, we have $\widetilde{n} = \widetilde{s} - \widetilde{w}$.

For each balancing time k, we compute a mean scaling factor to model typical sub-hourly trends present in the data. These factors, denoted $\{f_k\}_{k=1}^{N}$, are computed independently for each hour of the day using the net load profile.

Consider a particular hour of the day. Let n_k^j refer to net load at balancing time k within this hour on day j. Moreover, let N_d be the total number of days present in the data set and $\bar{n}^j$ be the average net load across all balancing times for this hour on day j. The factors can be computed according to (4.90).

$$f_k = \frac{\sum_{j=1}^{N_d} \frac{n_k^j - \bar{n}^j}{\bar{n}^j}}{N_d} \tag{4.90}$$

Given these factors, the mean net load forecast at each balancing time k is given by:

$$\mu_k = (1 + f_k)\,\widetilde{n} \tag{4.91}$$

where $\widetilde{n}$ is the net load forecast for the hour under consideration.

For each hour, we assume constant net load variance σ^2 across all balancing times. We use the mean squared error in the net load forecasts as an estimate of this variance. For a particular hour, let $\widetilde{n}^j$ be the net load forecast and $\bar{n}^j$ be the hourly net load average on day j respectively. The net load variance is given by:

$$\sigma^2 = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\widetilde{n}^j - \bar{n}^j\right)^2 \tag{4.92}$$

To investigate the marginal benefit of deferrability under this framework, we vary both the proportion by energy of the deferrable load component (α) as well as the servicing rate limit (m). The deferrable load proportion is varied on the interval $[0, 0.1]$. The servicing rate limit is varied from $\dfrac{L}{6\Delta t}$ (constant, inflexible deferrable load profile) to $\dfrac{L}{\Delta t}$ (no rate limits). For each of these deferrable load characterizations, we compute the ex-ante procurements based on the load-scheduling heuristic described in (4.87).

Results

Figure 4.6 shows ex-ante procurements for a typical day with and without deferrable loads. Two trends are immediately evident. First, the bulk power purchase is lowest at midday (10:00-14:00) and highest in the evening (17:00-21:00). Second, the capacity procurement is highest during the late morning and early afternoon hours. Both these trends agree with typical solar power output and load patterns. Solar generation is highest during the mid-day hours, which explains both the reduced bulk power requirements as well as the increased capacity requirements. While peak load requirements typically occur in the afternoon, peak net load only occurs later in the evening once solar power is no longer available.

Figure 4.7 shows the reductions possible in reserve capacity procurements possible at various levels of deferrable load penetration (α) and rate limit (m). These results suggest that higher deferrable load proportions and less stringent rate limits offer reserve capacity reductions. Moreover, the vast majority of these reductions can even be achieved at low levels of deferrable load penetration. In fact, we observe that, similar to the synthetic test case, there is a bound on the capacity reductions possible through addition load scheduling.

Both figures 4.6 and 4.7 suggest that deferrable loads only offer *small* reductions ($< 5\%$) in capacity requirements. These relatively modest benefits are explained by the deferrable load model used in this analysis. Note that we performed independent analyses for each hour-long operating window. As a result, we effectively limit our analysis to deferrable loads that have fixed hourly energy needs and do not model loads with deferrability windows longer than an hour. A modeling framework that can accommodate deferrable loads over longer windows, or involves multiple ex-ante procurements for intervals within the operating window, would indicate greater benefits of deferrability to ex-ante operations. However, such a formulation may not be amenable to simple analytical characterizations like those presented in this chapter. More precise quantification of the benefit of deferrability would require detailed deferrable load models.

4.7 Conclusions & Possible Extensions

In this chapter, we have investigated the effect of coordinated deferrable load scheduling on ex-ante forward market decisions. We first performed this analysis in the case of a fully static load under various sets of working assumptions. Assuming no reserve capacity procurement, we expressed the optimal bulk power procurement as a quantile on prices. Assuming symmetric reserve capacities, we analytically characterized the optimal bulk power and reserve capacity decisions on a partition of their forward market prices and the energy price of operating reserves. Using an aggregate model for deferrable loads, we derived a threshold policy for optimal reserve scheduling in the face of uncertain supply. We also offered a mathematical framework to analyze the effect of load scheduling on forward market decisions, and subsequently quantified the operational cost reductions offered by deferrability over a series of synthetic test cases.

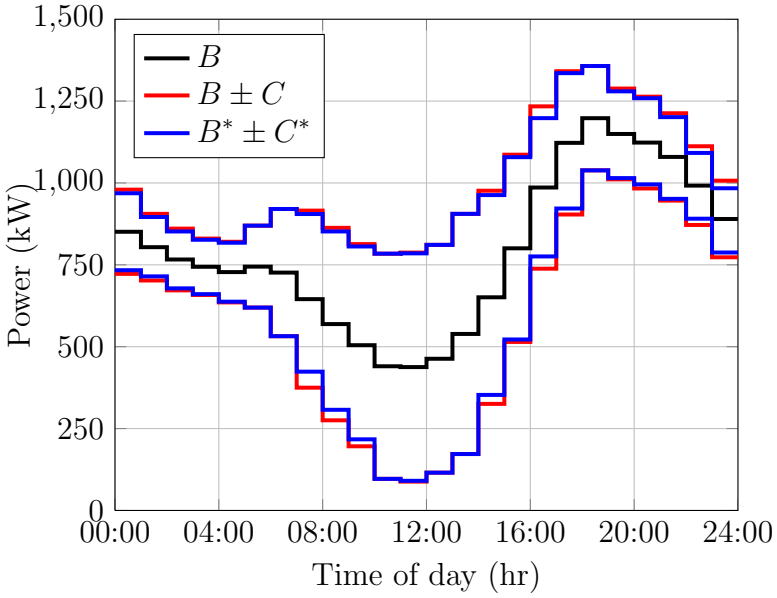


Figure 4.6: Variation in ex-ante procurements over a typical day with $(B \pm C)$ and without $(B^* \pm C^*)$ deferrable loads.

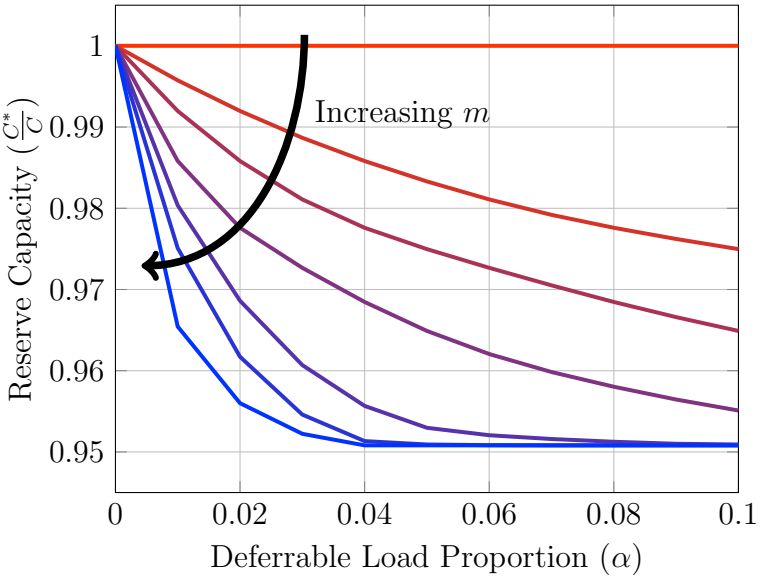


Figure 4.7: Optimal reserve capacity procurement as a function of deferrable load proportion (α) and servicing rate limit (m). All reserve capacities C^* are expressed as fractions of the initial reserve capacity requirement C. Results are averaged over 50 days.

There are several practical aspects of integrating deferrable loads that require further exploration. In this study, we work exclusively with an aggregate deferrability model assuming a single deferrable load. Extensions of these formulations to situations with multiple loads having varying degrees of flexibility will enable more accurate estimates of grid operation costs. Moreover, the optimal reserve scheduling policy computed in this chapter assumes an *uncorrelated* renewable generation process. This is a strong assumption and future work will center around finding appropriate optimal policies for a correlated process.

In this chapter, we do not address various power network constraints that affect generation procurement. Incorporating network considerations in our formulations will further improve the accuracy of our operating cost estimates. Finally, we also want to develop fair compensation mechanisms for flexible loads that are cognizant of the value they offer grid operations.

Chapter 5

The Benefit of Wind Curtailment

In Chapters 3 and 4, we demonstated the reserve cost reductions achievable by exploiting the flexibility present in loads to better match renewable generation. In this chapter, we explore another approach to facilitating renewable integration focused on the sources of variability – the renewable power producers. Specifically, we investigate mechanisms that allow wind power producers (WPPs) to better manage variability in their output. Curtailing WPP output depending on system conditions is a key component of such schemes. To this end, we study individual WPP decision-making in response to market signals.

5.1 Motivation

Wind power will play a central role in achieving renewable penetration targets in many countries. In fact, certain countries have set specific targets for wind generation alone. The US Department of Energy (DOE) has set a goal of reaching 20% wind penetration by 2030 [52], while Denmark has set a more ambitious target of wind power generation accounting for 50% of national energy consumption by 2025 [126]. To incentivize construction of utility-scale wind farms, many countries and states have provided preferential treatment for wind power in the form of guaranteed grid access. However, such support cannot continue indefinitely. The impact of variability in renewables at high levels of wind energy penetration on reserve costs will become unacceptable. Indeed, certain countries have already begun to modify rules governing wind power producer (WPP) participation in wholesale electricity markets [2, 109, 55].

Motivated by such transitions, we consider the scenario in which WPPs offer power in conventional two-settlement electricity markets. Specifically, the WPP agrees to produce a contracted amount ex-ante and is penalized for deviations (up and down) from this contracted position. In this setting, a WPP with the ability to *curtail* wind production can avoid imbalance penalties by adjusting production levels in real-time.

In this context, curtailment is a voluntary and rational action of a profit-maximizing WPP and is not the result of ISO-mandated dispatch adjustments. Currently, certain ISOs

curtail wind generation through manual dispatch to alleviate congestion in the transmission network [110]. There are two issues with this form of generation re-dispatch. First, this requires fair schemes for compensating WPPs for their curtailed output. To the best of our knowledge, such schemes do not exist - an assertion supported by the sizable variation in payment practices employed by different ISOs [110]. Second, the ISO assumes a degree of legal risk under this paradigm as WPPs may allege that certain curtailment requests were unnecessary. This issue came to light in a 2011 petition to the Federal Energy Regulatory Commission (FERC), in which WPPs in the Bonneville Power Administration's (BPA) balancing area challenged BPA's curtailment practices. FERC, ruling in favor of the WPPs, called the BPA's mandatory curtailment policy under system overproduction conditions 'unduly discriminatory' [57]. In line with this recent FERC ruling, we focus exclusively on *market-based* mechanisms inducing voluntary curtailment.

Much of the recent literature outlines mechanisms, such as retail dynamic pricing tariffs [114] and energy storage [58], that mitigate energy spillage resulting from wind power curtailment. Alternatively, other studies have attempted to quantify the benefit accruing from wind curtailment. Evans [54] quantifies the economic gain of coordinated wind power curtailment with a hydro-electric storage system – enabling price arbitrage to increase profit margins. Ela [50] argues that curtailment of wind power in response to LMPs may relieve transmission congestion and simultaneously provide economic benefits to the WPP. To the best of our knowledge, no attempt has been made to explore the strategic self-induced curtailment of wind power in response to asymmetric imbalance prices – an effect we refer to as *market induced curtailment*.

In this chapter, we quantify the financial benefit to a WPP of having *full* curtailment capability. We begin by developing a stochastic model for wind power production and present a stylized model of a competitive two-settlement electricity market. We then formulate and solve a two-stage stochastic optimization problem yielding the profit-maximizing contract and curtailment policies. Using these results, we quantify the increase in expected profit attainable through curtailment. We show that the optimal contract offer and the resulting expected profit can be expressed analytically on a partition of the space of expected imbalance prices. We also show that curtailment *always* results in an increase in the expected profit. Moreover, this curtailment benefit depends solely on expected values of imbalance prices and wind power. Finally, we conduct several empirical studies, using time series data from the NYISO and MISO, to demonstrate the financial benefit achievable through curtailment.

Implicit in the decision to curtail is the problem of optimizing forward contract offerings in the face of uncertain supply. This is a well studied problem. Morales et al. [93] and Bathurst et al. [17] used stochastic programming approaches to numerically compute DA contract offers that maximize expected profit in a two-settlement market setting. Recently, Botterud et al. [24] studied optimal bidding strategies through nonlinear programming in the case where a WPP has the choice of bidding into either the DA or RT market. These studies are computational in nature. Other studies have focused on analytical characterizations of the optimal contract offerings. Bitar et al. [18] and Pinson et al. [103] showed the optimal

DA contract offer is a quantile on prices – a well established result in inventory theory [102]. With the inclusion of curtailment capability, the results obtained in this chapter are a natural extension of those presented in [18]. Indeed, we compare our results to those of [18] to ascertain the impact of curtailment on optimal WPP behavior.

The remainder of this chapter is organized as follows. In Section 5.2, we present market and wind power production models. Sections 5.3 and 5.4 contain our main results on optimal curtailment and forward contract policies, respectively. In Section 5.5, we empirically compute the financial benefit afforded by curtailment using wind and price time series data. We conclude and suggest possible directions for future research in Section 5.6.

5.2 Problem Formulation

5.2.1 Wind Power Model

Let $w(t)$ refer to the wind power available to a WPP at time t. To accommodate the variable nature of wind speed and direction, we model $w(t)$ as a scalar-valued stochastic process. We normalize this value by the farm's nameplate capacity, thus $w(t) \in [0, 1]$. Let $\mathbf{w}$ refer to this stochastic process over a specific delivery window $[t_0, t_F]$:

$$\mathbf{w} = \{w(t) | t \in [t_0, t_F]\}. \tag{5.1}$$

For fixed $t \in \mathbb{R}$, $w(t)$ is a random variable with probability density function $\phi(w; t)$. Let $f_w(w)$ and $F_w(w)$ be time-averaged probability density and cumulative distribution functions on the delivery window $[t_0, t_F]$ of width $T = t_F - t_0$. These quantities can be readily computed using the formulae (5.2) and (5.3).

$$f_w(w) = \frac{1}{T} \int_{t_0}^{t_F} \phi(w; t) dt. \tag{5.2}$$

$$F_w(w) = \int_0^w f_w(x) dx. \tag{5.3}$$

In addition, we define the *quantile function* $F_w^{-1} : [0, 1] \to [0, 1]$ corresponding to the cumulative distribution function (CDF):

$$F_w^{-1}(y) = \inf \{x \in [0, 1] : F_w(x) \geq y\}. \tag{5.4}$$

This function plays a key role in characterizing the optimal contract offerings.

5.2.2 Curtailment Model

The curtailment decision at time t is expressed as a scalar $\alpha(t)$, where $\alpha(t)w(t)$ denotes the amount of power produced at time t. We make two simplifying assumptions to facilitate analysis.

A1. WPP output can be fully curtailed at any time within the delivery window. As a result, the curtailment factor $\alpha(t)$ takes values on the closed interval $[0, 1]$.

A2. There are no inter-temporal constraints limiting the curtailment rate. In other words, for all times $t_1, t_2 \in [t_0, t_F]$, the curtailment decisions $\alpha(t_1)$ and $\alpha(t_2)$ are *independent*.

Remark 5.1 (Curtailment in practice)**.** *Physically, wind curtailment is achieved by pitching the angle of turbine blades with respect to wind direction. This alters the angular velocity of the blades and consequently, the amount of power generated. In practice, the curtailment rate is limited as sudden decreases in plant output result in real power imbalances that adversely affect system frequency. As a result, while fast curtailment rates are possible, typical ramp rate limits for WPPs in practice are on the order of 3–5 MW/min [91].*

Let α refer to the continuous set of curtailment decisions on the time interval $[t_0, t_F]$:

$$\alpha = \{\alpha(t) | t \in [t_0\ t_F]\}. \tag{5.5}$$

We highlight the fact that these decisions are made *within* the delivery interval.

5.2.3 Market Model

Model Description

We assume the wind power producer (WPP) participates in a competitive two-settlement market system. This system consists of an *ex-ante* day-ahead (DA), or forward, market and an *ex-post* imbalance settlement mechanism to penalize uninstructed deviations from contracts scheduled in the DA market. In this framework, the WPP agrees in the DA market to supply a level of generation C over the delivery window. The WPP is obligated to deliver this level of generation and is penalized for deviations from this contracted position. Figure 5.1 is a graphical illustration of this model for WPP participation in electricity markets.

Working within this framework, the WPP faces three market prices:

1. p: ($/MWhr) contract settlement price in the DA market or *DA price*,
2. q: ($/MWhr) imbalance price for negative deviations ($w(t) < C$) or *shortfall price*.
3. λ: ($/MWhr) imbalance price for positive deviations ($w(t) > C$) or *surplus price*.

This pricing scheme for penalizing contract deviations reflects both the energy imbalance in the control area and the price of balancing energy at times between the DA market and the delivery window. We assume that the imbalance prices (q, λ) are unknown during the DA market and are not revealed until the delivery window. These imbalance prices can be either positive or negative, and play the role of market signals from the ISO indicating system generation conditions. For instance, a negative shortfall price q can occur in the event of systemic over-production, thereby inducing WPPs to deliberately supply less than their contracted positions.

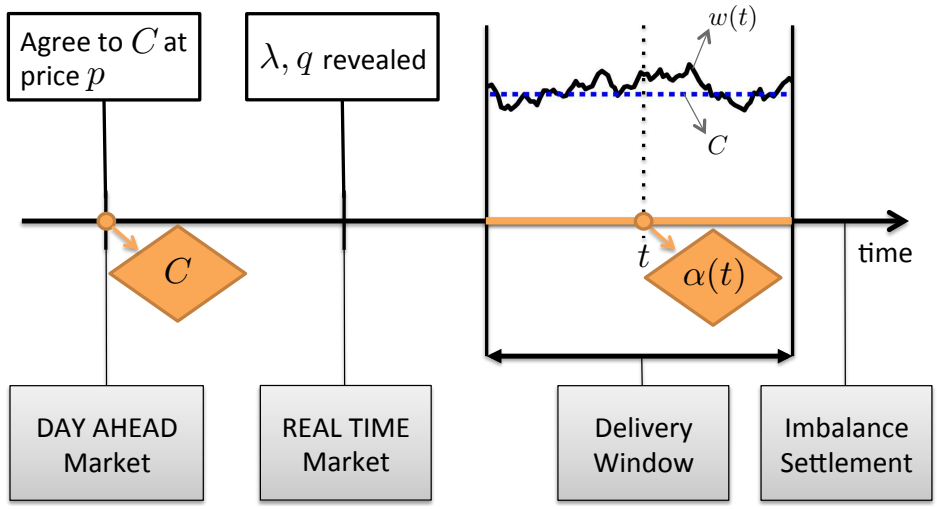


Figure 5.1: Timeline illustrating the two-settlement market model assumed for WPP participation.

Assumptions

We make the following assumptions regarding prices and production costs.

A3. The WPP is assumed to have *zero marginal cost of production*.

A4. As the capacity of a single WPP is small relative to the entire market, the effect of individual offers on the forward market clearing price p is minimal. Thus, we assume the WPP behaves as a *price taker* in the forward market. This market price, p, is assumed *fixed* and *known*.

A5. The imbalance prices (q, λ) are constant over contract intervals and are revealed 5-15 minutes before the delivery window.

A6. As imbalance prices are generally difficult to forecast in advance of the delivery window, we model q and λ as random variables with a joint probability distribution function $f_{q,\lambda}(q, \lambda)$ assumed known at the forward market.

A7. Again, as the capacity of a single WPP is small relative to the entire market, associated deviations have negligible effect on the imbalance prices. Thus, we assume q and λ to be statistically independent of the corresponding wind power process, $\mathbf{w}$.

A8. The scenario in which the WPP is compensated by the ISO for *both* generation surpluses and shortfalls ($q < 0$ and $\lambda < 0$) is not practical in any envisaged market. Accordingly, the following constraint is imposed on the joint probability distribution function:

$$\mathbb{P}\{q < 0, \lambda < 0\} = 0. \tag{5.6}$$

5.2.4 Metrics

Working under these assumptions, the profit acquired by a WPP over time interval $[t_0, t_F]$ is given by:

$$\begin{aligned}\Pi(C, \alpha, \mathbf{w}, \lambda, q) \;=\;& pCT \\ & - q \int_{t_o}^{t_F} [C - \alpha(t) w(t)]^+ \, dt \\ & - \lambda \int_{t_o}^{t_F} [\alpha(t) w(t) - C]^+ \, dt. \end{aligned} \tag{5.7}$$

where $[x]^+ = \max(x, 0)$.

The first term refers to revenue obtained from agreeing to the contract C signed at the time of the DA market. The second and third terms represent imbalance costs for generation shortfall and surplus respectively.

This quantity is inherently random, as it depends on the realization of wind power $\mathbf{w}$ and imbalance prices (q, λ). Consequently, we focus on the metric of expected profit:

$$J(C, \alpha) = \mathbb{E}\ \left[\Pi(C, \alpha, \mathbf{w}, q, \lambda)\right], \tag{5.8}$$

where expectation is taken with respect to the imbalance prices (q, λ) and the wind power process $\mathbf{w}$.

Using this model, we formulate a two-stage stochastic program consisting of an initial forward contract offering C – scheduled to be delivered continuously over a time interval $[t_0, t_F]$ – followed by a sequence of curtailment decisions α upon realization of imbalance prices (q, λ) and wind power $\mathbf{w}$.

5.2.5 Optimization Problem

Our objective is to identify a forward contract C and curtailment policy α that maximize expected profit (5.8). Specifically,

$$(C^*, \alpha^*) = \underset{C, \alpha}{\operatorname{argmin}}\, J(C, \alpha). \tag{5.9}$$

Because of the significant lead time on delivery of the ex-ante commitment C, problem (5.9) can be decomposed into a two-stage stochastic optimization problem, where the set of decisions (C, α) are assumed to have a causal information structure. Specifically, the contract C is agreed to in the forward market (Stage I) well in advance of the actual delivery interval, while the curtailment factor α is chosen at time of delivery (Stage II). While the forward contract offering C is decided upon in the face of uncertainty in wind and imbalance prices, the decision to curtail $\alpha(t)$ at a time t is made with full knowledge of the wind realization $w(t)$, contract C, and imbalance prices (q, λ). It follows that (5.9) can be decomposed into the following two-stage optimization problem.

$$\text{Stage I: } C^* = \underset{C}{\operatorname{argmin}} \quad \mathbb{E}\left[\max_{\alpha} \Pi(C, \alpha, \mathbf{w}, q, \lambda)\right] \tag{5.10}$$

$$\text{Stage II: } \alpha^* = \underset{\alpha}{\operatorname{argmin}} \quad \Pi(C, \alpha, \mathbf{w}, q, \lambda) \tag{5.11}$$

Effectively, probability distributions for the wind process $\mathbf{w}$, and the imbalance prices (q, λ) are only necessary in determining the optimal contract. Our approach in this chapter is as follows. First, we find the optimal choice of curtailment factors α^* assuming an *arbitrary* contract C (Stage II). We then compute the optimal contract C^* (Stage I) given the *optimal* curtailment policy, α^*. The resulting solution is optimal with respect to the original optimization problem (5.9).

5.3 Optimal Curtailment Policy

As we have assumed no inter-temporal constraints limiting the rate of curtailment $\alpha(t)$, the curtailment decisions decouple across time. It follows that the optimal curtailment decision $\alpha^*(t)$ at each time t depends only on the available information at that time - the wind power available $w(t)$ and the imbalance prices (q, λ). Specifically, the curtailment decision depends on the following properties of the available information:

1. The wind power available $w(t)$ compared to the forward contract C,
2. The signs of the imbalance prices (q, λ).

To simplify the characterization of the optimal curtailment policy, we partition the space of possible available generation into two half-spaces representing generation *shortfall*, $\Omega_-(C)$, and *surplus*, $\Omega_+(C)$.

$$\begin{aligned} \text{Shortfall:} \quad & \Omega_-(C) = \{w(t) < C\} \\ \text{Surplus:} \quad & \Omega_+(C) = \{w(t) \geq C\} \end{aligned}$$

Clearly, $\Omega_-(C) \cup \Omega_+(C) = [0, 1]$, the support for available wind power $w(t)$. We now describe the optimal curtailment policy, $\alpha^*_-(t)$ and $\alpha^*_+(t)$, on each half-space $\Omega_-(C)$ and $\Omega_+(C)$ respectively.

5.3.1 Shortfall ($\Omega_-(C)$)

In the case of generation shortfall, the surplus price λ does not affect the realized profit for all possible curtailment decisions ($\alpha(t) \in [0, 1]$). Accordingly, the choice of $\alpha(t)$ depends solely on the sign of the shortfall price q. Clearly, for a positive shortfall price ($q > 0$), the WPP, attempting to avoid under–producing as best it can, delivers the entire quantity –

yielding $\alpha^*(t) = 1$. Alternatively, for a negative shortfall price ($q \leq 0$), the WPP profits from under-producing, yielding optimality of full curtailment $\alpha^*(t) = 0$.

Combining these characterizations, we see that the optimal curtailment policy, during realized generation shortfalls, is piece-wise constant on the space of imbalance prices (q, λ). In summary, we have:

$$\alpha_-^*(t) = \begin{cases} 1, & q > 0, \lambda \in \mathbb{R} \\ 0, & q \leq 0, \lambda \in \mathbb{R} \end{cases}. \tag{5.12}$$

5.3.2 Surplus ($\Omega_+(C)$)

In the case of generation surplus, the profit depends on the signs of both surplus λ and shortfall q imbalance prices. The optimal curtailment policy is different on each *quadrant* of the space of imbalance prices $(q, \lambda) \in \mathbb{R}^2$. This dependence is as follows:

1. $\{(q, \lambda) \in \mathbb{R}^2 : q > 0, \lambda > 0\}$
   This is the case where both surpluses and shortfalls are penalized. Here, the obvious policy is to curtail down to the ex-ante commitment C, thereby avoiding any imbalance. The corresponding curtailment factor $\alpha^*(t) = C/w(t)$.

2. $\{(q, \lambda) \in \mathbb{R}^2 : q > 0, \lambda \leq 0\}$
   In this case, surpluses are rewarded and shortfalls penalized. The optimal policy is one of full delivery, which maximizes revenue gained from surplus while minimizing the chance of paying a shortfall penalty – i.e.: $\alpha^*(t) = 1$.

3. $\{(q, \lambda) \in \mathbb{R}^2 : q \leq 0, \lambda > 0\}$
   In this case, shortfalls are rewarded and surpluses penalized. the optimal policy is one of full curtailment, which maximizes revenue gained from shortfall while avoiding any penalty for over-production – i.e.: $\alpha^*(t) = 0$.

4. $\{(q, \lambda) \in \mathbb{R}^2 : q \leq 0, \lambda \leq 0\}$
   This case is not considered for reasons outlined in Assumption A8, Section 5.2.

Combining these characterizations, the optimal curtailment can be explicitly characterized as a piece-wise affine function on the space of imbalance prices. More specifically, the optimal policy is:

$$\alpha_+^*(t) = \begin{cases} C/w(t), & q > 0, \lambda > 0 \\ 1, & q > 0, \lambda \leq 0 \\ 0, & q \leq 0, \lambda > 0 \\ -, & q \leq 0, \lambda \leq 0 \end{cases}. \tag{5.13}$$

Combining (5.12) and (5.13), we have the optimal curtailment policy defined on the entire space of available information $(w(t), q, \lambda)$:

$$\alpha^*(t) \;=\; \alpha^*_-(t)\,\mathbb{I}\{w(t) < C\} \;+\; \alpha^*_+(t)\,\mathbb{I}\{w(t) \geq C\}. \tag{5.14}$$

5.4 Optimal Contract Offer

Given the optimal curtailment policy outlined in Section 5.3, we now compute the expected-profit maximizing contract offering C^* in the DA forward market, defined as:

$$C^* = \underset{C\in[0,1]}{\operatorname{argmin}}\, J(C, \alpha^*). \tag{5.15}$$

We now define four weighted conditional expectations that will play a central role in characterizing the optimal contract.

Definition 5.1. *Average* ***shortfall penalty****:* μ_q^+

$$\mu_q^+ = \mathbb{E}\,[\,q\,|\,q > 0\,]\;\;\mathbb{P}\{q > 0\} \tag{5.16}$$

Definition 5.2. *Average* ***shortfall reward****:* μ_q^-

$$\mu_q^- = \mathbb{E}\,[\,q\,|\,q < 0\,]\;\;\mathbb{P}\{q < 0\} \tag{5.17}$$

Definition 5.3. *Average* ***surplus penalty****:* μ_λ^+

$$\mu_\lambda^+ = \mathbb{E}\,[\,\lambda\,|\,\lambda > 0\,]\;\;\mathbb{P}\{\lambda > 0\} \tag{5.18}$$

Definition 5.4. *Average* ***surplus reward****:* μ_λ^-

$$\mu_\lambda^- = \mathbb{E}\,[\,\lambda\,|\,\lambda < 0\,]\;\;\mathbb{P}\{\lambda < 0\} \tag{5.19}$$

These weighted conditional expectations depend solely on the probability distribution of imbalance prices $f_{q,\lambda}(q, \lambda)$. Notice that the law of total expectation yields:

$$\begin{aligned}\mu_q^+ + \mu_q^- &=: \mu_q = \mathbb{E}\,[q]\\ \mu_\lambda^+ + \mu_\lambda^- &=: \mu_\lambda = \mathbb{E}\,[\lambda]\end{aligned}$$

In an analogous manner to [18], we now show that the optimal contract policy C^* can be analytically expressed on a partition of the space of expected imbalance prices $(\mu_q, \mu_\lambda) \in \mathbb{R}^2$ given by $\mathcal{P}_1 \cup \mathcal{P}_2 \cup \mathcal{P}_3 = \mathbb{R}^2$, where:

$$\begin{aligned}\mathcal{P}_1 &= \{\,(x,y) \in \mathbb{R}^2 \mid y \leq -p + \mu_q^- + \mu_\lambda^+\,,\\ &\qquad\qquad \mu_w y + (\mu_w - 1)\,x \leq -p + \left(\mu_q^- + \mu_\lambda^+\right)\mu_w\,\}\\ \mathcal{P}_2 &= \{\,(x,y) \in \mathbb{R}^2 \mid x > p\,,\; y > -p + \mu_q^- + \mu_\lambda^+\}\\ \mathcal{P}_3 &= \{\,(x,y) \in \mathbb{R}^2 \mid x \leq p\,,\\ &\qquad\qquad \mu_w y + (\mu_w - 1)\,x > -p + \left(\mu_q^- + \mu_\lambda^+\right)\mu_w\,\}.\end{aligned} \tag{5.20}$$

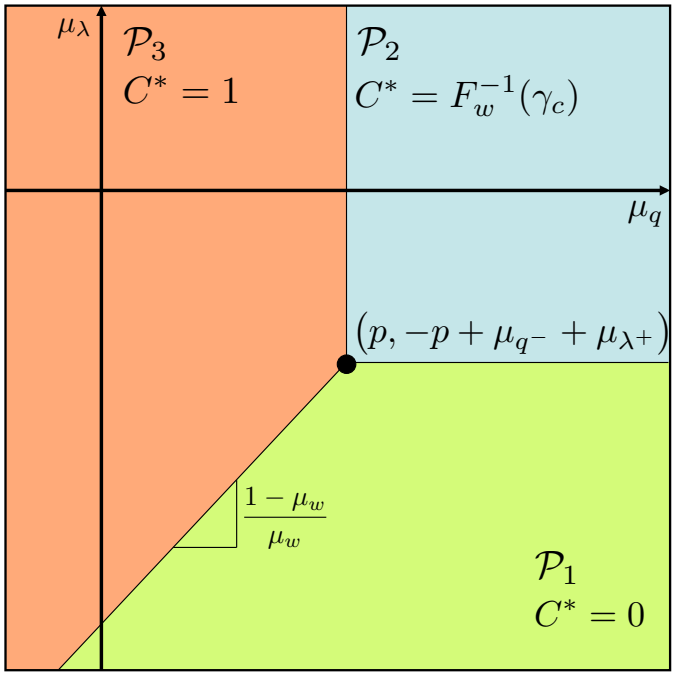


Figure 5.2: Graphical illustration of the optimal contract policy on the partition $(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3)$ of expected imbalance prices (μ_q, μ_λ)

Here, μ_w denotes the expected value of the time-averaged distribution of the wind power process on $[t_0, t_F]$. This partition is graphically illustrated in Figure 5.2.

We are now ready to state the main result of this chapter - the optimal contract policy that maximizes expected profit for a WPP with full curtailment capability.

Theorem 5.1 (*Optimal Contract*)**.** *For a WPP, exercising the curtailment policy α^* given by (5.14), that faces the pair of expected imbalance prices $\xi = (\mu_q, \mu_\lambda) \in \mathbb{R}^2$,*

(a) the optimal contract offering C^* is given by:

$$C^* = \begin{cases} 0, & \text{if } \xi \in \mathcal{P}_1 \\ F_w^{-1}(\gamma_c), & \text{if } \xi \in \mathcal{P}_2 \\ 1, & \text{if } \xi \in \mathcal{P}_3. \end{cases}, \ \gamma_c = \frac{p - \mu_q^- + \mu_\lambda^-}{\mu_q^+ + \mu_\lambda^-} \tag{5.21}$$

(b) the optimal expected profit $J(C^*, \alpha^*)$ is given by:

$$\frac{J(C^*, \alpha^*)}{T} = \frac{J^*}{T} = \begin{cases} -\mu_\lambda^- \mu_w, & \text{if } \xi \in \mathcal{P}_1 \\ \mu_q^+ \int_0^{\gamma_c} F_w^{-1}(\theta)\, d\theta - \mu_\lambda^- \int_{\gamma_c}^1 F_w^{-1}(\theta)\, d\theta, & \text{if } \xi \in \mathcal{P}_2 \\ \left(p - \mu_q^-\right) - \mu_q^+ (1 - \mu_w), & \text{if } \xi \in \mathcal{P}_3. \end{cases} \tag{5.22}$$

Proof. **(a)**: Recall that the time-averaged probability density function for the wind is denoted $f_w(w)$, and the joint probability density for the imbalance prices, $f_{q,\lambda}(q,\lambda)$. The expected profit function for a WPP over a time interval of length $T = t_F - t_0$ can be expressed as:

$$\frac{J(C,\alpha)}{T} = pC - \mathbb{E}_{w,q,\lambda}\left[q\left[C-\alpha w\right]^+ - \lambda\left[\alpha w - C\right]^+\right] \tag{5.23}$$

Assuming the WPP employs the optimal curtailment factor α^* outlined in (5.14), the imbalance costs encountered by the WPP differ depending on the particular combination of w, q and λ:

$$\begin{aligned}\frac{J(C)}{T} = \frac{J\left(C,\alpha^*\right)}{T} \;=\; & pC - \mathbb{E}_{w,q,\lambda}\Big[\mathbb{I}\{w<C, q<0\}\; qC + \mathbb{I}\{w<C, q>0\}\; q\left(C-w\right) \\ & + \mathbb{I}\{w>C, \lambda>0, q>0\}\;\left(0\right) + \mathbb{I}\{w>C, \lambda<0, q>0\}\;\lambda\left(w-C\right) \\ & + \mathbb{I}\{w>C, \lambda>0, q<0\}\; qC\Big]\end{aligned} \tag{5.24}$$

Notice that we have invoked the assumption that at least one of the imbalance prices is non-negative (Assumption A8, Section 5.2) and excluded the appropriate cases from (5.24).

As we consider the random variable w to be independent of the imbalance prices q and λ (Assumption A7, Section 5.2), the expected profit can be expressed as:

$$\begin{aligned}\frac{J\left(C\right)}{T} = pC &- \int_{-\infty}^{0} q \int_{-\infty}^{\infty} f_{q,\lambda}(q,\lambda)\; d\lambda\, dq \quad C\int_0^C f_w(w)dw \\ &- \int_{0}^{\infty} q \int_{-\infty}^{\infty} f_{q,\lambda}(q,\lambda)\; d\lambda\, dq \quad \int_0^C \left(C-w\right) f_w(w)dw \\ &- \int_{-\infty}^{0} \lambda \int_{0}^{\infty} f_{q,\lambda}(q,\lambda)\; dq\, d\lambda \quad \int_C^1 \left(w-C\right) f_w(w)dw \\ &- \int_{-\infty}^{0} q \int_{0}^{\infty} f_{q,\lambda}(q,\lambda)\; d\lambda\, dq \quad C\int_C^1 f_w(w)dw\end{aligned} \tag{5.25}$$

Assumption A8 also yields the following equalities:

$$\int_{-\infty}^{0} \lambda \int_{0}^{\infty} f_{\lambda,q}(\lambda,q)\; dq\, d\lambda = \int_{-\infty}^{0} \lambda \int_{-\infty}^{\infty} f_{\lambda,q}(\lambda,q)\; dq\, d\lambda \tag{5.26}$$

$$\int_{-\infty}^{0} q \int_{0}^{\infty} f_{\lambda,q}(\lambda,q)\; d\lambda\, dq = \int_{-\infty}^{0} q \int_{-\infty}^{\infty} f_{\lambda,q}(\lambda,q)\; d\lambda\, dq \tag{5.27}$$

Using (5.26) and (5.27) and the terminology defined in (5.16) - (5.19), we can express the expected profit as:

$$\frac{J\left(C\right)}{T} \;=\; pC \;-\; \mu_q^- C \;-\; \mu_q^+ \int_0^C \left(C-w\right) f_w(w)dw \;-\; \mu_\lambda^- \int_C^1 \left(w-C\right) f_w(w)dw \tag{5.28}$$

Assuming $f_w(w)$ is a continuous distribution on $[0,1]$, it follows that $J(C)$ is also differentiable on $[0,1]$. We apply the Leibniz integral rule to obtain the first and second derivatives of $J(C)$:

$$\frac{dJ}{dC} = T\left(p - \mu_q^- + \mu_\lambda^- - \left(\mu_q^+ + \mu_\lambda^-\right) F\left(C\right)\right) \tag{5.29}$$

$$\frac{d^2J}{dC^2} = -Tf(C)\left(\mu_q^+ + \mu_\lambda^-\right) \tag{5.30}$$

As $T > 0$ and $f(C) \geq 0\ \forall C \in [0\,,\,1]$, $J(C, \alpha^*)$ is concave $\Leftrightarrow \mu_q^+ + \mu_\lambda^- \geq 0$ and convex $\Leftrightarrow \mu_q^+ + \mu_\lambda^- \leq 0$. In the space of expected imbalance prices (μ_q, μ_λ), the concave half-space corresponds to $\mu_q + \mu_\lambda \geq \mu_q^- + \mu_\lambda^+$ and the convex half-space to $\mu_q + \mu_\lambda \leq \mu_q^- + \mu_\lambda^+$.

We first find the optimal policy, C^*, in the concave half-space. In this region $(\mu_q + \mu_\lambda \geq \mu_q^- + \mu_\lambda^+)$, $C^* \in [0,1]$ is a maximum $\Leftrightarrow$

$$(x - C^*)\left.\frac{dJ}{dC}\right|_{C=C^*} \leq 0,\ \forall x \in [0,1]$$

This can be interpreted as:

- $C^* = 1$ when $J(C)$ is a non-decreasing function of C.
- $C^* = 0$ when $J(C)$ is a non-increasing function of C.
- $C^* = \beta$ when $\dfrac{dJ}{dC}\left(C = \beta\right) = 0$. (A stationary point)

Consider the following partition of the concave half-space,:

$$\begin{aligned}
\mathcal{G}_1 &= \{\mu_\lambda \leq -p + \mu_q^- + \mu_\lambda^+\} \cap \{\mu_q + \mu_\lambda \geq \mu_q^- + \mu_\lambda^+\} \\
\mathcal{G}_2 &= \{\mu_q > p\} \cap \{\mu_\lambda > -p + \mu_q^- + \mu_\lambda^+\} \\
\mathcal{G}_3 &= \{\mu_q \leq p\} \cap \{\mu_q + \mu_\lambda \geq \mu_q^- + \mu_\lambda^+\}
\end{aligned}$$

It is easy to verify that when $(\mu_q, \mu_\lambda) \in \mathcal{G}_1$, $\dfrac{dJ}{dC} \leq 0\ \forall C \in [0,1]$ and so the optimal contract $C^* = 0$. Similarly, $C^* = 1$ when $(\mu_q, \mu_\lambda) \in \mathcal{G}_3$ as $\dfrac{dJ}{dC} \geq 0\ \forall C \in [0,1]$. Finally, when $(\mu_q, \mu_\lambda) \in \mathcal{G}_2$, the concave function $J(C)$ has a stationary point at $F_w^{-1}\left(\gamma_c\right)$ where $\gamma_c = \dfrac{p - \mu_q^- + \mu_\lambda^-}{\mu_q^+ + \mu_\lambda^-}$. Thus, the profit maximizing contract is $C^* = F_w^{-1}\left(\gamma_c\right)$.

We now analyze optimality on the half-space $\mu_q + \mu_\lambda \leq \mu_q^- + \mu_\lambda^+$. As $J(C)$ is convex in this region, the maximum must occur on the endpoints of the set of feasible contracts [0, 1]. The maximization problem is then a simple comparison of two values: $C^* = 0$ when

$J(0) \geq J(1)$ and $C^* = 1$ otherwise. The difference between the two candidate maxima is given by:

$$J(0) - J(1) = T\left(-p + \mu_q^+(1-\mu_w) + \mu_q^- - \mu_\lambda^-\mu_w\right)$$

Therefore, in the convex half-space, C^* is:

$$C^* = \begin{cases} 0, & -p + \mu_q^- \geq \mu_\lambda^-\mu_w + \mu_q^+(\mu_w - 1) \\ 1, & -p + \mu_q^- < \mu_\lambda^-\mu_w + \mu_q^+(\mu_w - 1) \end{cases}$$

In the (μ_q, μ_λ) space, the corresponding condition is:

$$C^* = \begin{cases} 0, & \mu_w\mu_\lambda + (\mu_w - 1)\,\mu_q \leq -p + (\mu_{q^-} + \mu_{\lambda^+})\,\mu_w \\ 1, & \mu_w\mu_\lambda + (\mu_w - 1)\,\mu_q > -p + (\mu_{q^-} + \mu_{\lambda^+})\,\mu_w \end{cases} \tag{5.31}$$

Combining the results obtained for the convex and concave half-spaces, one recovers Theorem 1(a).

(b): The values of the optimal expected profit when $(\mu_q, \mu_\lambda) \in \mathcal{P}_1$ and $\mathcal{P}_3$ can be computed easily by substitution of the appropriate optimal profit into (5.28). When $(\mu_q, \mu_\lambda) \in \mathcal{P}_2$, we perform a change of variables $\theta = F_w(w)$ to obtain:

$$\begin{aligned} \frac{J(C^*, \alpha^*)}{T} &= (p - \mu_q^-)\,C^* \;-\; \mu_q^+ \int_0^{C^*} (C^* - w)\, f_w(w)dw \;-\; \mu_\lambda^- \int_{C^*}^1 (w - C^*)\, f_w(w)dw \\ &= (p - \mu_q^-)\,C^* \;-\; \mu_q^+ \int_0^{\gamma_c} \left(C^* - F_w^{-1}(\theta)\right) d\theta \;-\; \mu_\lambda^- \int_{\gamma_c}^1 \left(F_w^{-1}(\theta) - C^*\right) d\theta \\ &= \left(p - \mu_q^- + \mu_\lambda^- - \left(\mu_q^+ + \mu_\lambda^-\right)\gamma_c\right) C^* \;+\; \mu_q^+ \int_0^{\gamma_c} F_w^{-1}(\theta)d\theta \;-\; \mu_\lambda^- \int_{\gamma_c}^1 F_w^{-1}(\theta)d\theta \\ &= \mu_q^+ \int_0^{\gamma_c} F_w^{-1}(\theta)d\theta \;-\; \mu_\lambda^- \int_{\gamma_c}^1 F_w^{-1}(\theta)d\theta \end{aligned} \tag{5.32}$$

by the definition of γ_c. Thus, we completely recover the claim. □

Remark 5.2 (Inverse supply function)**.** *Because of our price-taker assumption (Assumption A4, Section 5.2), the optimal contract offer C^* can be interpreted as the WPP's inverse supply function. Notice that C^* is an explicit function of the day-ahead clearing price p. Consequently, in the DA market, the WPP simply provides the ISO with the supply schedule $C(p)$ characterized by (5.21). Working under the assumption that $\mu_q^+ + \mu_\lambda^- > 0$, this supply function simplifies to:*

$$C(p) = \begin{cases} 0, & p \leq \mu_q^- - \mu_\lambda^- \\ F_w^{-1}(\gamma_c), & \mu_q^- - \mu_\lambda^- < p < \mu_q \\ 1, & p \geq \mu_q. \end{cases} \quad , \; \gamma_c = \frac{p - \mu_q^- + \mu_\lambda^-}{\mu_q^+ + \mu_\lambda^-} \tag{5.33}$$

In order to quantify the increase in expected profit attainable through curtailment, we compare our results to the baseline case where the WPP has *no curtailment capability*. For this baseline, explicit characterizations of the optimal contract offering and corresponding expected profit were given by [18]. That result is stated below to facilitate exposition.

In the absence of curtailment capability ($\forall t, \alpha^*(t) = 1$), the optimal contract offering can be analytically expressed on the following partition of the space of expected imbalance prices, $\mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3 = \mathbb{R}^2$:

$$\begin{aligned} \mathcal{M}_1 =& \{\ (x,y) \in \mathbb{R}^2 \mid y \leq -p\,, \mu_w y + (\mu_w - 1)\, x \leq -p\ \} \\ \mathcal{M}_2 =& \{\ (x,y) \in \mathbb{R}^2 \mid x \geq p\,,\ y \geq -p\ \} \\ \mathcal{M}_3 =& \{\ (x,y) \in \mathbb{R}^2 \mid \mu_w y + (\mu_w - 1)\, x > -p\ , x < p\}. \end{aligned} \tag{5.34}$$

For a WPP facing a pair of expected imbalance prices $\xi \in (\mu_q, \mu_\lambda)$,

(a) the optimal contract offering C_o^* is given by:

$$C_o^* = \begin{cases} 0, & \text{if } \xi \in \mathcal{M}_1 \\ F_w^{-1}(\gamma), & \text{if } \xi \in \mathcal{M}_2 \\ 1, & \text{if } \xi \in \mathcal{M}_3 \end{cases}, \gamma = \frac{p + \mu_\lambda}{\mu_q + \mu_\lambda} \tag{5.35}$$

(b) the optimal expected profit $J(C_o^*, 1)$ is given by:

$$J(C_o^*, 1) = J_o^* = T \begin{cases} -\mu_\lambda \mu_w, & \text{if } \xi \in \mathcal{M}_1 \\ \mu_q \displaystyle\int_0^\gamma F_w^{-1}(\theta)\, d\theta - \mu_\lambda \int_\gamma^1 F_w^{-1}(\theta)\, d\theta, & \text{if } \xi \in \mathcal{M}_2 \\ p - \mu_q\,(1 - \mu_w) & \text{if } \xi \in \mathcal{M}_3. \end{cases} \tag{5.36}$$

Examining the structure of the expected profit functions, we can make the following claim about the benefit of curtailment.

Theorem 5.2. *Curtailment only* ***increases*** *the optimal expected profit.*

$$\Delta J := J^* - J_o^* \geq 0 \tag{5.37}$$

Proof. $J_o^* = J\,(C_o^*, 1) \leq J\,(C_o^*, \alpha^*)$ as α^* was the optimal curtailment policy for *any* contract C. $J^* = \max_C J\,(C, \alpha^*) \geq J\,(C_o^*, \alpha^*) \geq J_o^*$. Hence, $\Delta J \geq 0$. □

Remark 5.3 (Quantiles)**.** *The quantile, γ_c , obtained in the case with curtailment (5.21) can be re-expressed as:*

$$\gamma_c = \frac{p + \mu_\lambda - \left(\mu_q^- + \mu_\lambda^+\right)}{\mu_q + \mu_\lambda - \left(\mu_q^- + \mu_\lambda^+\right)}, \tag{5.38}$$

which is ***identical*** *to γ, the quantile in the no curtailment result (5.35)* ***if and only if****:*

$$\mu_q^- + \mu_\lambda^+ = 0. \tag{5.39}$$

It is apparent from (5.38) that the main difference between optimal contracts in the cases with and without curtailment depend on surplus penalties (μ_λ^+) and shortfall rewards (μ_q^-). This agrees with intuition as curtailment offers financial benefits in both these cases.

Remark 5.4 (Surplus Penalties)**.** *In the event of a surplus penalty $\{\lambda > 0\}$, a WPP avoids financial penalties by curtailing any excess generation when generation $w(t)$ exceeds the contract C. Consider the scenario where a WPP offers no contract in the DA market ($C = 0$) but supplies power in the hope that it will be rewarded for surplus generation. In this case, a WPP without curtailment capability realizes an expected profit of $-\mu_\lambda \mu_w$. On the other hand, a WPP that can curtail its production, by avoiding surplus penalties, can increase its expected profit to $-\mu_\lambda^- \mu_w$.*

The ability to avoid these penalties clearly affects the forward contract offering (5.38). In response to high values of μ_λ^+, WPPs that cannot curtail their output offer larger contracts than those that can, to avoid surplus production scenarios.

Remark 5.5 (Shortfall Rewards)**.** *In the event of a shortfall reward $\{q < 0\}$, a WPP can curtail production to increase profit. Consider the scenario where a WPP offers a full capacity contract in the DA market ($C = 1$). Here, a WPP is subject to shortfall penalties irrespective of its ability to curtail generation. However, only a WPP with curtailment capability can fully benefit from shortfall rewards. Comparing the maximum expected profits realized with (5.22) and without (5.36) curtailment confirms our inference as the curtailment benefit is $-\mu_q^- \mu_w$.*

Clearly, the ability to realize shortfall rewards through curtailment impacts the forward contract offering (5.38). In response to highly negative values of μ_q^-, WPPs that can curtail production offer larger contracts, hoping to induce shortfall scenario, than those that cannot.

We now explore conditions under which the optimal contract offerings for WPPs with and without curtailment are the same.

Remark 5.6 (Contract Policy Equivalence)**.** *The condition described in Remark 5.3 is also a necessary and sufficient condition for the two partitions of the space of expected imbalance prices, $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$ and $\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$ to be the same. This is summarized by the following corollary.*

Corollary 5.1. *The optimal contract policies for the curtailment and no curtailment cases are **identical**, (i.e.: $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\} \equiv \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$ and $\gamma_c = \gamma$) $\Leftrightarrow \mu_q^- + \mu_\lambda^+ = 0$.*

This condition can be interpreted by examining (5.38). As average shortfall rewards (μ_q^-) increase in magnitude, WPPs with curtailment capability offer larger contracts than WPPs that cannot curtail production. Conversely, as average surplus penalties (μ_λ^+) increase, WPPs with curtailment capability offer smaller contracts than WPPs that cannot curtail production. The condition $\mu_q^- + \mu_\lambda^+ = 0$ corresponds to an equilibrium between these incentives to increase or decrease the contract offer compared to the offer in the no curtailment case.

We remark that the equivalence condition (5.39) holds true for imbalance pricing mechanisms where $q = -\lambda$. This symmetric penalty structure is currently applied to imbalances for conventional thermal generation in many balancing areas [96, 27, 28].

Remark 5.7 (Curtailment Benefit)**.** *When $\mu_q^- + \mu_\lambda^+ = 0$, that is when curtailment does not affect the optimal contract policy, the financial benefit (ΔJ) resulting from curtailment is* ***identical*** *across the entire space of expected imbalance penalties.*

Corollary 5.2.

$$\Delta J = J^* - J_o^* = -\mu_q^- \mu_w T = \mu_\lambda^+ \mu_w T. \tag{5.40}$$

Proof. According to Corollary 5.1, the partitions on which the optimal contract policy are defined are identical $\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\} \equiv \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$ if $\mu_q^- + \mu_\lambda^+ = 0$. We then calculate ΔJ on each set of the partition.

It is straightforward to show that the claim, $\Delta J = \mu_\lambda^+ \mu_w T$, holds on $\mathcal{M}_1$ and $\mathcal{M}_3$. On $\mathcal{M}_2$, notice that the quantiles used in the optimal policies are identical ($\gamma_c = \gamma$) when $\mu_q^- + \mu_\lambda^+ = 0$ (Remark 5.3). As a result, the increase in expected profit can be expressed as: $\Delta J = \left(-\mu_q^- \int_0^\gamma F_w^{-1}(\theta)\, d\theta + \mu_\lambda^+ \int_\gamma^1 F_w^{-1}(\theta)\, d\theta \right) T = \mu_\lambda^+ \mu_w T.$ □

5.5 Empirical Results

In this section, we empirically compute the financial benefit afforded by curtailment using price data from the New York (NYISO) and Midwest (MISO) independent system operators. Currently, imbalances for conventional thermal generation are settled based on locational marginal prices (LMP) [96]. The LMP at a particular node represents the incremental system cost of serving an additional unit of demand at said node, subject to transmission losses and capacity constraints.

In our analysis, we assume WPPs participate in a two-settlement market (e.g.: Section 5.2), where the imbalance settlements are calculated based on LMPs. Specifically, a positive LMP represents a shortfall penalty and surplus reward, whereas a negative LMP represents a shortfall reward and surplus penalty. This scenario is equivalent to the two-sided imbalance penalty mechanism presented in Section 5.2 with the additional constraint

$$q = -\lambda \tag{5.41}$$

where q is equal to the LMP at the WPP bus.

Working under this pricing structure, the contracts offered by WPPs with and without curtailment capability are the same since $\mu_q^- = -\mu_\lambda^+$ (Corollary 5.1). It follows that the curtailment benefit can computed analytically using (5.40).

5.5.1 Data Set

In order to compute the benefit derived from curtailment capability, as characterized by Corollary 5.2, we require empirical estimates of the mean statistics $(\mu_w, \mu_q^-, \mu_\lambda^+)$. As typical DA contracts are constant on intervals of length one hour, we compute statistics independently for each hour of the day.

Wind generation

In order to estimate average wind power production, we work with time series data from the Bonneville Power Administration (BPA). This is time series data of measured wind power aggregated over the 14 wind power generation sites in the BPA control area. The wind power output is sampled every 5 minutes and covers the 2008 and 2009 calendar years. In order to account for additional wind power capacity coming online at various points in time over the 2-year horizon, we normalized all data by the aggregate nameplate wind power capacity as a function of time.

Due to the diurnal periodicity in wind speed, we assume that the daily wind power trajectories are drawn from distributions that are *identical* across days. To simplify analysis, we additionally assume these trajectories to be *independent* across days. Working under these assumptions, a consistent estimator for the mean wind power $\mu_w(t) = \mathbb{E}[w(t)]$ at time t is given by:

$$\hat{\mu}_w(t) = \frac{1}{N} \sum_{n=1}^{N} w_n(t), \tag{5.42}$$

where n indexes days. Figure 5.3 depicts the trajectory of the empirical mean averaged over contract intervals of length one hour. While the actual wind power trajectory varies with WPP location, this empirically obtained sample is a reasonable estimate for rough curtailment benefit calculations.

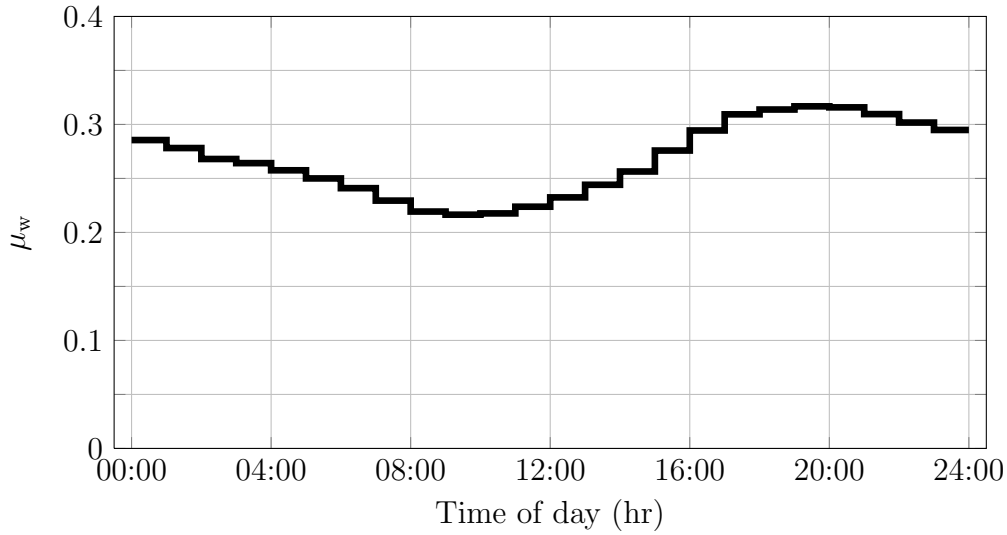


Figure 5.3: Empirical means of wind generation (μ_w) computed from BPA time series data. Empirical means are calculated for each hour of the day.

LMP

In order to compute estimates of the mean price statistics (μ_q^-, μ_λ^+), we work with LMP time series data from both the NYISO and MISO. For the NYISO 2008 calendar year, we work with real-time LMP time series sampled every 5 minutes from 4 different zones in the network. For the MISO, we work with real-time LMP time series from 2011 from 7 select nodes in this network. In both cases, this data is averaged over hour long intervals. Specifically, we estimate the weighted expected value of the LMP q condition on it being negative. This quantity, which is the *average shortfall reward*, is calculated for each hour $(t = 1, \cdots, 24)$ of the day according to:

$$\hat{\mu}_q^-(t) = -\hat{\mu}_\lambda^+(t) = \frac{1}{N}\sum_{n=1}^{N} q_n(t)\,\mathbb{I}\{q_n(t) \leq 0\}, \tag{5.43}$$

where n indexes days.

Figures 5.4 and 5.5 depict the average surplus penalties $\hat{\mu}_\lambda^+(t)$ for each hour of the day at select nodes in the NYISO and MISO balancing areas, respectively. From Figure 5.4, it is readily apparent that certain nodes (e.g.: ALTW.LKFLD.IPL) experience larger and more frequent negative excursions in their LMPs than other nodes (e.g.: OTP.BIGSTON1). This can be attributed to transmission capacity limitations and typical flow patterns within the network. Because of these LMP differentials, the financial benefit derived from wind curtailment will vary across nodes. It is also apparent from comparing figures 5.4 and 5.5 that there is greater location-based variation in MISO prices than in NYISO prices. This observation agrees with intuition as the MISO footprint is a far larger geographical area than that of NYISO. Transmission network considerations over this larger area cause greater LMP variation across the network.

Node	**State**	**Savings ($)**
OTP.BIGSTON1	SD	186,176
MEC.WSEC3	IA	245,971
NSP.NOBLESTR2	MN	687,058
ALTW.LKFLD.IPL	MN	935,414
NSP.WHEATO2	WI	73,540
OTP.COYOT1	ND	144,343
AMIL.CLINTO51	IL	361,061

Table 5.1: Annual curtailment benefit for a 50 MW WPP with full curtailment capability sited at 7 different locations within the MISO balancing area

5.5.2 Curtailment Benefit

Using the described wind generation and LMP data, we estimate the benefit derived from having full curtailment capability. We consider a representative WPP with a nameplate

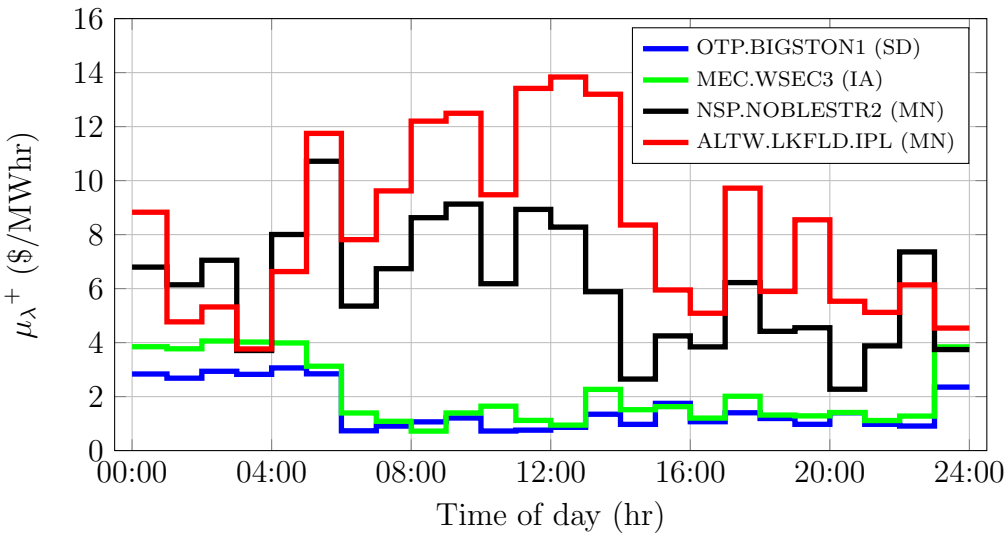


Figure 5.4: Average surplus penalties (μ_λ^+) computed using MISO LMP market data from January-June 2011. Empirical means are calculated for each hour of the day.

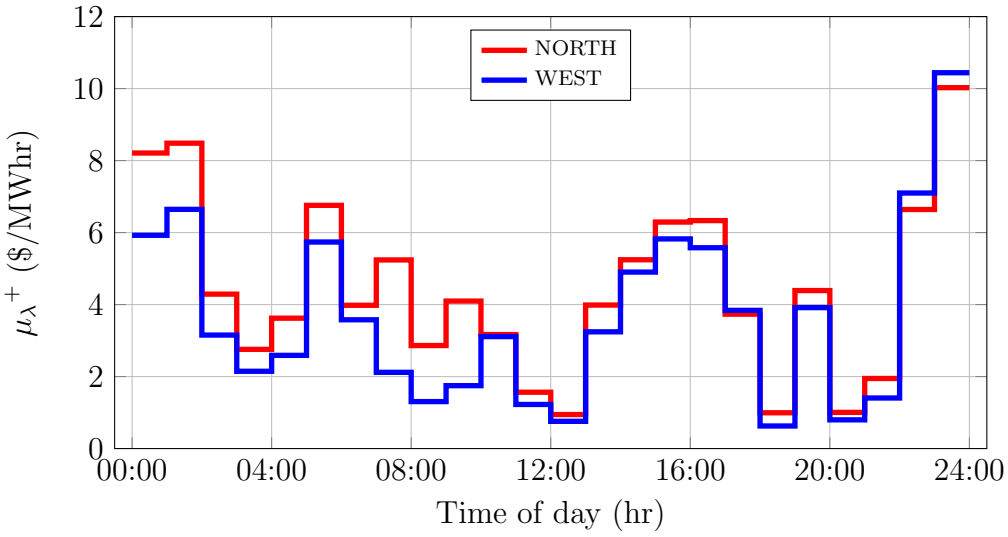


Figure 5.5: Average surplus penalties (μ_λ^+) computed using NYISO LMP market data from 2008. Empirical means are calculated for each hour of the day.

capacity of 50MW. The expected annual cost savings can be computed using:

$$\text{Savings} \quad = \quad \sum_{t=1}^{24} \mu_\lambda^+(t)\mu_w(t) \;\; (\text{Number of days}) \, (\text{WPP Capacity}) \tag{5.44}$$

Tables 5.1 and 5.2 list the expected annual savings realized by a 50 MW WPP at different

Node	Savings ($)
NORTH	525,514
WEST	438,443
GENESE	441,790
HQ	471,710

Table 5.2: Annual curtailment benefit for a 50 MW WPP with full curtailment capability sited at 4 different locations within the NYISO balancing area

nodes within the MISO and NYISO networks respectively. It is immediately apparent that bus location has a direct impact, through the LMPs, on the curtailment benefit. This geographical dependence, particularly evident in the MISO price data (Figure 5.4), is the only reason behind the substantial variation in annual savings ($50,000–1,000,000). The NYISO price data, owing to limited location-based LMP variation, exhibits far less variation in annual savings.

We remark that this calculation can be easily repeated for any WPP by using only historical time series data of wind power and market prices. In the case of a symmetric imbalance settlement mechanism, this calculation only requires mean statistics of wind generation, and imbalance prices (μ_w, μ_λ^+) at the WPP connection node.

5.6 Conclusions & Possible Extensions

In this chapter, we quantify how a wind power producer – participating in a two-settlement market – might leverage power curtailment capability to mitigate financial risk in the face of uncertain production and imbalance prices. Using a general stochastic model for wind power production and a stylized two-settlement market model, we explicitly characterize the expected profit-maximizing curtailment policy and forward contract offerings. Moreover, we demonstrate that power curtailment capability always leads to an expected profit higher than the profit achievable without curtailment. Using NYISO and MISO market data, we compute the curtailment benefit realized by a 50 MW WPP at various nodes exhibiting differing LMP mean statistics.

Using the tools presented in this chapter, a WPP can calculate an *upper bound* of their curtailment benefit *a-priori*, thus permitting a cost-benefit analysis of the installation of curtailment capability. As mentioned previously, this calculation can be readily performed using historical wind power and price data. Moreover, an ISO can use the results presented in this chapter to understand and predict how WPPs will respond to imbalance price signals.

There are several avenues for future work. The market model used in this chapter makes no mention of ancillary service (A/S) markets, markets used by ISOs to procure reserve generation capacity. While WPPs currently do not participate in A/S markets, such participation becomes more likely as wind penetration increases and WPPs, forced to curtail generation due to market conditions, are allowed to sell excess generation in these markets.

Alternatively, WPPs with storage capability could curtail their output, store this energy, and inject it into the grid during favorable market conditions. More nuanced analyses are required to compute the profit-maximizing contract for a WPP with curtailment capability in these situations.

Finally, this model can be used to inform the selection of imbalance prices. Using this model for WPP behavior, one can create mechanisms for setting these imbalance prices to induce power curtailment that helps manage: (1) sudden changes in renewable generation levels, and (2) transmission network congestion.

Bibliography

[1] A. Alarcon-Rodriguez, G. Ault, and S. Galloway. "Multi-objective planning of distributed energy resources: A review of the state-of-the-art". In: *Renewable and Sustainable Energy Reviews* 14.5 (2010), pp. 1353–1366.

[2] J.R. Abbad. "Electricity market participation of wind farms: the success story of Spanish pragmatism". In: *Energy Policy* 38.7 (2010).

[3] U.S. Energy Information Administration. *Annual Energy Outlook 2011*. Tech. rep. DOE/EIA-0383(2011). 2010.

[4] U.S. Energy Information Administration. *Annual Energy Outlook 2013*. Tech. rep. DOE/EIA-0383(2013). 2013.

[5] M.H. Albadi and E.F. El-Saadany. "Demand response in electricity markets: An overview". In: *IEEE Power Engineering Society General Meeting*. 2007, pp. 1–5.

[6] M. Alizadeh et al. "Information infrastructure for cellular load management in green power delivery systems". In: *Proceedings of the International Conference on Smart Grid Communication (SmartGridComm)*. 2011.

[7] O. Alsac et al. "The rights to fight price volatility". In: *IEEE Power and Energy Magazine* 2.4 (2004), pp. 47–57.

[8] 110$^{\text{th}}$ Congress of the United States of America. *Energy independence and security act of 2007*. Public Law 110-140.

[9] *Ancillary Service Market Transactions in the Day-Ahead and Real Time Adjustment Period*. Electric Reliability Council of Texas (ERCOT). 2011.

[10] W. Anderegg et al. "Expert credibility in climate change". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.27 (2010), pp. 12107–12109.

[11] *April 2013 - Monthly Energy Review*. U.S. Energy Information Administration, 2013.

[12] *Arizona Renewable Energy Standard and Tariff rules*. A.A.C. R14-2-1801:1815. 2006. URL: `http://www.azcc.gov/divisions/utilities/electric/res.pdf`.

[13] T.P. Baker. "An analysis of EDF schedulability on a multiprocessor". In: *IEEE Transactions on Parallel and Distributed Systems* 16.8 (2005), pp. 760–768.

[14] G. Barbose, C. Goldman, and B. Neenan. *A survey of utility experience with real time pricing*. Tech. rep. LBNL-54238. Lawrence Berkeley National Laboratory, 2004.

[15] S. Barker et al. "SmartCap: Flattening peak electricity demand in smart homes". In: *IEEE International Conference on Pervasive Computing and Communications*. 2012, pp. 67–75.

[16] Sanjoy Baruah and Joël Goossens. "Scheduling real-time tasks: Algorithms and complexity". In: Joseph Y-T. Leung. *Handbook of Scheduling: Algorithms, Models, and Performance analyses*. Boca Raton, FL: CRC Press, 2004. Chap. 28.

[17] G. Bathurst, J. Weatherill, and G. Strbac. "Trading wind generation in short term energy markets". In: *IEEE Transactions on Power Systems* 17.3 (2002), pp. 782–789.

[18] E. Bitar et al. "Bringing wind energy to market". In: *IEEE Transactions on Power Systems* 27.2 (2012), pp. 1225–1235.

[19] Eilyan Y. Bitar. "Selling Random Energy". PhD thesis. University of California at Berkeley, 2011.

[20] R. Bo and F. Li. "Probabilistic LMP forecasting considering load uncertainty". In: *IEEE Transactions on Power Systemss* 24.3 (2009), pp. 1279–1289.

[21] S. Borenstein. "The long-run efficiency of real-time electricity pricing". In: *The Energy Journal* 26.3 (2005), pp. 93–116.

[22] S. Borenstein and J. Bushnell. "Electricity restructuring: deregulation or reregulation?" In: *Regulation* 23.2 (2 2000), pp. 46–52.

[23] S. Borenstein, M. Jaske, and A. Rosenfeld. *Dynamic pricing, advanced metering and demand response in electricity markets*. Tech. rep. CSEM WP 105. University of California, Berkeley, 2002.

[24] A. Botterud et al. "Wind power trading under uncertainty in LMP markets". In: *IEEE Transactions on Power Systems* 27.2 (2012), pp. 894–903.

[25] Steven Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[26] D. Brooks et al. *2010 Resource Adequacy Report*. Tech. rep. California Public Utilities Commission, 2011.

[27] *Business Practice Manual for Market Operations*. California Independent System Operator (CAISO), Version 32. 2013.

[28] *Business Practices Manual: Energy and Operating Reserve Markets, BPM-002 r12*. Midwest Independent System Operator (MISO).

[29] California Independent System Operator (CAISO). *CAISO Markets: Day Ahead Markets Overview – Video Tutorial*. 2009. URL: `http://content.caiso.com/training/Day-Ahead_Market_Overview/index.html?pg=000`.

[30] California Independent System Operator (CAISO). *Integration of renewable resources: Operational Requirements and generation fleet capability at 20% RPS*. 2010. URL: http://www.caiso.com/2804/2804d036401f0.pdf.

[31] California Independent System Operator (CAISO). *Integration of renewable resources: Transmission and operating issues and recommendations for integrating renewable resources on the California ISO-controlled grid*. 2007.

[32] *California Renewable Portfolio Standard (RPS)*. California Public Utilities Commission (CPUC). URL: http://www.cpuc.ca.gov/PUC/energy/Renewables/.

[33] D. Callaway and I. Hiskens. "Achieving controllability of electric loads". In: *Proceedings of the IEEE* 99.1 (2011), pp. 184–199.

[34] Damian Carrington. *Wind power capacity grew 20% globally in 2012, figures show*. The Guardian. 11. URL: http://www.guardian.co.uk/environment/2013/feb/11/wind-power-capacity-grew-2012.

[35] S. Chen, T. He, and L. Tong. "Optimal deadline scheduling with commitment". In: *49th Allerton Conference on Communication, Control, and Computing*. 2011.

[36] S. Chen, P. Sinha, and N.B. Shroff. "Scheduling heterogenous delay tolreant tasks in smart grid with renewable energy". In: *Proceedings of the IEEE Conference on Decision and Control*. 2012, pp. 1130–1135.

[37] Federal Energy Regulatory Commission. *Energy primer: A handbook of energy market basics*. 2012.

[38] A.J. Conejo, J.M. Morales, and L. Baringo. "Real-Time Demand Response Model". In: *IEEE Transactions on Smart Grid* 1.3 (2010), pp. 236–242.

[39] S. Dahliwal and M. Guay. "A set-based estimation of heat loads for energy management in building systems". In: *Proceedings of the IEEE Conference on Decision and Control*. 2012, pp. 6957–6962.

[40] B. Daryanian, R.E. Bohn, and R.D. Tabors. "Optimal demand-side response to electricity spot prices for storage-type customers". In: *IEEE Transactions on Power Systems* 4.3 (1989), pp. 897–903.

[41] *Day-Ahead Scheduling Manual, Manual 11, Version 4.0*. New York Independent System Operator (NYISO). 2013.

[42] "Demand response experience in Europe: Policies, programmes and implementation". In: *Energy* 35.4 (2010), pp. 1575–1583.

[43] "Demand response in U.S. electricity markets: Empirical evidence". In: *Energy* 35.4 (2010), pp. 1526–1535.

[44] "Demand response models with correlated price data: A robust optimization approach". In: *Applied Energy* 96.0 (2012), pp. 133–149.

[45] Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education. *Renewable energy target*. Australian Goverment. URL: http://www.climatechange.gov.au/reducing-carbon/renewable-energy/renewable-energy-target.

[46] M.L. Dertouzos and A.K. Mok. "Multiprocessor online scheduling of hard real-time tasks". In: *IEEE Transactions on Software Engineering* 15.12 (1989), pp. 1497–1506.

[47] M.W. Doyle et al. "Dam removal in the United States: emerging needs for science and policy". In: *Eos, Transations American Geophysical Union* 84.4 (2003), pp. 29–33.

[48] J. Driesen and F. Katiraei. "Design for distributed energy resources". In: *IEEE Power and Energy Magazine* 6.3 (2008), pp. 30–40.

[49] Brain K. Edwards. *The economics of hydroelectric power*. Edward Elgar Publishing, 2003.

[50] E. Ela. *Using economics to determine the efficient curtailment of wind energy*. Technical Report TP-550-45071. National Renewable Energy Laboratory (NREL), 2009.

[51] E. Ela, M. Milligan, and B. Kirby. *Operating reserves and variable generation*. Tech. rep. NREL/TP-5500-51978. National Renewable Energy Laboratory (NREL), 2011.

[52] U.S. Department of Energy. *20% Wind energy by 2030 - Increasing wind energy's contribution to U.S. electricity supply*. Tech. rep. DOE/GO-102008-2567. 2008.

[53] U.S. Department of Energy. *The potential benefit of Distributed Generation and rate-related issues that may impede their expansion: A study pursuant to Section 1817 of the Energy Policy act of 2005*. Tech. rep. US DOE, 2007.

[54] J.I. Evans. "Benefits of wind power curtailment in a hydro-dominated electric generation system". MA thesis. Civil Engineering, University of British Columbia, 2009.

[55] Inc. Exeter Associates and GE Energy. *PJM renewable integration study – Task Report: Review of industry practice and experience in the integration of wind and solar generation*. 2012.

[56] Federal Government of Germany Federal Ministry for the Environment, Nature Conservation and Nuclear Safety. *Energy concept for an environmentally sound, reliable and affordable energy supply*. 2010.

[57] Federal Energy Regulatory Commission (FERC). *Iberdrola Renewables Inc., 137 FERC ¶61,185*. 7 December 2011. URL: https://www.ferc.gov/EventCalendar/Files/20111207083529-EL11-44-000.pdf.

[58] E. Fertig and J. Apt. "Economics of compressed air energy storage to integrate wind power: A case study in ERCOT". In: *Energy Policy* 39.5 (2011), pp. 2330–2342.

[59] Russell Flannery. *2011 China investment guide: the world's best hydro returns*. Forbes. 2011. URL: http://www.forbes.com/sites/russellflannery/2011/02/08/2011-china-investment-guide-the-worlds-best-hydro-returns/.

[60] M. Galus, R. la Fauci, and G. Andersson. "Investigating PHEV wind balancing capabilities using heuristics and model predictive control". In: *IEEE Power & Energy Society General Meeting*. 2010.

[61] L. Gan, U. Topcu, and S. Low. "Stochastic distributed protocol for electric vehicle charging with discrete charging rate". In: *IEEE Power & Energy Society General Meeting*. 2012.

[62] C.E. García, D.M. Prett, and M. Morari. "Model predictive control: Theory and practice - A survey". In: *Automatica* 25.3 (1989), pp. 335–348.

[63] J. Goossens, S. Funk, and S. Baruah. "Priority-driven scheduling of periodic task systems on multiprocessors". In: *Real Time Systems* 25.2/3 (2003), pp. 187–205.

[64] J. Hansen, M. Sato, and R. Ruedy. "Perception of climate change". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.37 (2012), E2415–E2423.

[65] J. Hansen et al. "Global temperature change". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.39 (2006), pp. 14288–14293.

[66] E. Hausman, R. Hornby, and A. Smith. *Bilateral contracting in deregulated electricity markets*. Synapse Energy Economics, Inc., 2008.

[67] D. Hawkins, J. Blatchford, and Y. Makarov. "Wind integration issues and solutions in California". In: *IEEE Power and Energy General Society Meeting*. 2007.

[68] U. Helman. *Resource and transmission planning to achieve a 33% RPS in California ISO Modeling tools and planning framework*. FERC Technical Conference on Planning Models and Software, 2010.

[69] J. Hong, X. Tan, and D. Towsley. "A performance analysis of minimum laxity and earliest deadline scheduling in a real-time system". In: *IEEE Transactions on Computers* (1989).

[70] Y. Y. Hsu and C. C. Su. "Dispatch of direct load control using dynamic programming". In: *IEEE Transactions on Power Systems* 6.3 (1991), pp. 1056–1061.

[71] G. Hug-Glanzmann. "Coordination of intermittent generation with storage, demand control and conventional energy sources". In: *Proceedings of IREP 2010 - Bulk Power Systems Dynamics and Control -VIII*. 2010.

[72] L. Jiang and S. Low. "Multi-period optimal energy procurement and demand response in smart grid with uncertain supply". In: *Proceedings of the IEEE Conference on Decision and Control and European Control Conference*. 2011, pp. 4348–4353.

[73] J.-Y. Joo and M.D. Ilic. "A multi-layered adaptive load management (ALM) system: Information exchange between market participants for efficient and reliable energy use". In: *IEEE Power and Energy Society Transmission and Distribution Conference and Expo*. 2010.

[74] A.C. Kemp et al. "Climate related sea-level variations over the past two millenia". In: *Proceedings of the National Academy of Sciences of the United States of America* (2011).

[75] Atul A. Khasnis and Mary D. Nettleman. "Global warming and infectious disease". In: *Archives of Medical Research* 36.6 (2005), pp. 689–696.

[76] Z.W. Kundzewicz et al. "The implications of projected climate change for freshwater resources and their management". In: *Hydrological Sciences Journal* 53.1 (2008), pp. 3–10.

[77] T.F. Lee et al. "Optimization and implementation of a load control scheduler using relaxed dynamic programming for large air conditioner loads". In: *IEEE Transactions on Power Systems* 23.2 (2008), pp. 691–702.

[78] M. Lehtonen and S. Nye. "History of electricity network control and distributed generation in the UK and Western Denmark". In: *Energy Policy* 37.6 (2009), pp. 2338–2345.

[79] M. Lijesen. "The real-time price elasticity of electricity". In: *Energy Economics* 29.2 (2007), pp. 249–258.

[80] Chung Laung Liu and James W. Layland. "Scheduling algorithms for multiprogramming in a hard-real-time environment". In: *Journal of the ACM (JACM)* 20.1 (1973), pp. 46–61.

[81] R. Liu, L. Dow, and E. Liu. "A survey of PEV impacts on electric utilities". In: *IEEE Power and Energy Society (PES) Innovative Smart Grid Technologies (ISGT)*. 2011.

[82] D. Lüthi et al. "High-resolution carbon dioxide concentration record 650,000-800,000 years before present". In: *Nature* 453 (7193 2008), pp. 379–382.

[83] Z. Ma, I. Hiskens, and D. Callaway. "A decentralized MPC strategy for charging large populations of plug-in electric vehicles". In: *Proceedings of the 18th IFAC World Congress*. 2011.

[84] P. Mahat, Z. Chen, and B. Bak-Jensen. "Review of islanding detection methods for distributed generation". In: *Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies, 2008 (DRPT 2008)*. 2008, pp. 2743–2748.

[85] *Manual 11, Energy & Ancillary Services Market Operations, Revision: 60*. PJM. 2013.

[86] S. Mariéthoz and M. Morari. "Modelling and hierarchical hybrid optimal control of prosumers for improved integration of renewable energy sources into the grid". In: *American Control Conference (ACC)*. 2012, pp. 3114–3119.

[87] California Independent System Operator (CAISO) Department of Market Monitoring. *Potential impacts of lower bid price floor and contracts on dispatch flexibility from PIRP resources*. 2011.

[88] J. L. Mathieu, S. Koch, and D. Callaway. "State estimation and control of electric loads to manage real-time energy imbalance". In: *IEEE Transactions on Power Systems* (accepted, to be published).

[89] A.J. McMichael and A. Haines. "Global climate change: the potential effects on health". In: *British Medical Journal* 315.7111 (1997), pp. 805–809.

[90] K. Mets et al. "Optimizing smart energy control strategies for plug-in hybrid electric vehicle charging". In: *Proceedings of the IEEE/IFIP Network Operations and Management Symposium Workshops*. 2010, pp. 293–299.

[91] N. Miller. "Facts on grid friendly wind plants". In: *IEEE Power & Energy Society General Meeting*. 2010.

[92] A.K. Mok. "Fundamental design problems of distributed systems for the hard real-time enviroment". PhD thesis. Massachusetts Institute of Technology, 1983.

[93] J. Morales, A. Conejo, and J. Perez-Ruiz. "Short-term trading for a wind power producer". In: *IEEE Transactions on Power Systems* 25.1 (2010), pp. 554–564.

[94] K. Morrow, D. Karner, and J. Francfort. *Plug-in hybrid electric vehicle charging infrastructure review*. Tech. rep. INL/EXT-08-15058. U.S. Department of Energy - Vehicles Technologies Program, 2008.

[95] North American Electric Reliability Corporation (NERC). *Reliability standards for the bulk electric systems of North America*. 2008.

[96] New York Independent System Operator (NYISO). *Market Participant's User Guide, Guide 01, Version 2012.00*. 2012.

[97] Z. O'Neill, S. Nayaranan, and R. Brahme. "Model-based thermal load estimation in buildings". In: *Simbuild, 4^{th} National Conference of IBPSA - USA*. 2010.

[98] N. Oreskes. "The scientific consensus on climate change". In: *Science* 306.5702 (2004), p. 1686.

[99] A. Papavasiliou and S. Oren. "Supplying renewable energy to deferrable loads: Algorithms and economic analysis". In: *IEEE Power & Energy Society General Meeting*. 2010.

[100] S. Park. *Climate change and the risk of statelessness: the situation of low-lying island states*. Tech. rep. United Nations High Commissioner for Refugees, 2011.

[101] M.L. Parry et al. "Effects of climate change on global food production under {SRES} emissions and socio-economic scenarios". In: *Global Environmental Change* 14.1 (2004), pp. 53–67.

[102] N. Petruzzi and M. Dada. "Pricing and the newsvendor problem: a review with extensions". In: *Operations Research* 47.2 (1999), pp. 183–194.

[103] P. Pinson, C. Chevallier, and G. Kariniotakis. "Trading wind generation from short-term probabilistic forecasts of wind power". In: *IEEE Transactions on Power Systems* 22.3 (2007), pp. 1148–1156.

[104] Evan L. Porteus. *Foundations of Stochastic Inventory Theory*. Stanford University Press, 2002.

[105] J. Prada. *The value of reliability in power systems - pricing operating reserves*. Tech. rep. MIT EL 99-005 WP. Massachusetts Institute of Technology, 1999.

[106] A. Prékopa. *Stochastic Programming*. Kluwer Academic Publishers, 1995.

[107] Kejun Qian et al. "Modeling of load demand due to EV battery charging in distribution systems". In: *IEEE Transactions on Power Systems* 26.2 (2011), pp. 802–810.

[108] R. Rajagopal et al. "Risk-limiting dispatch for integrating renewable power". In: *International Journal of Electrical Power & Energy Systems* 44.1 (2013), pp. 615–628.

[109] MISO Press Release. *MISO furthers wind integration in the market*. 1 June 2011. URL: `https://www.midwestiso.org/AboutUs/MediaCenter/PressReleases/Pages/MISOFurthersIntegrationofWindResources.aspx`.

[110] J. Rogers, S. Fink, and K. Porter. *Examples of wind energy curtailment practices*. Subcontract Report SR-550-48737. National Renewable Energy Laboratory (NREL), 2010.

[111] M. Roozbehani, M. Dahleh, and S.K. Mitter. "On the stability of wholesale electricity markets under real-time pricing". In: *Proceedings of the IEEE Conference on Decision and Control and European Control Conference*. 2010, pp. 1911–1918.

[112] M. Roozbehani et al. "Load-shifting under perfect and partial information: Models, robust policies, and economic value". In: *Operations Research* (2012). (submitted).

[113] *SAE Electric Vehicle Conductive Charge Coupler, SAE J1772*. Prepared by the EV charging systems committee. Society of Automotive Engineers (SAE). 2001.

[114] R. Sioshansi and W. Short. "Evaluating the impacts of real-time pricing on the usage of wind generation". In: *IEEE Transactions on Power Systems* 24.2 (2009), pp. 516–524.

[115] K. Spees and L.B. Lave. "Demand response and electricity market efficiency". In: *The Electricity Journal* 20.3 (2007), pp. 69–85.

[116] P. Stevens. *The 'Shale Gas Revolution': Developments and changes*. Tech. rep. EERG BP 2012/04. Chatham House, 2012.

[117] C.L. Su and D. Kirschen. "Quantifying the effect of demand response on electricity markets". In: *IEEE Transactions on Power Systems* 24.3 (2009), pp. 1199–1207.

[118] A. Subramanian et al. "Real-time scheduling of deferrable electric loads". In: *Proceedings of the American Controls Conference (ACC)*. 2012, pp. 3643–3650.

[119] J. Taylor, D.S. Callaway, and K. Poolla. "Competitive energy storage in the presence of renewables". In: *IEEE Transactions on Power Systems* 28.2 (2013), pp. 985–996.

[120] D. Thornley. "Texas wind energy: past, present and future". In: *Environmental & Energy Law & Policy Journal* (2009), pp. 69–126.

[121] L. Vandezande et al. "Well-functioning balancing markets: A prerequisite for wind power integration". In: *Energy Policy* 38.7 (2010), pp. 3146–3154.

[122] P.P. Variaya, F.F. Wu, and J.W. Bialek. "Smart operation of the smart grid: Risk-limiting dispatch". In: *Proceedings of the IEEE* 99.1 (2011), pp. 40–57.

[123] H. Vu and J. Agee. *Western electricity coordination council (WECC) tutorial on speed governors*. 1998.

[124] Y.-H. Wan. *Analysis of wind power ramping behavior in ERCOT*. Tech. rep. NREL/TP-5500-49218. National Renewable Energy Laboratory (NREL), 2011.

[125] J. Withgott and S. Brennan. *Environment: the science behind the stories*. 4th. Prentice Hall, 2011. Chap. 15.

[126] Z. Xu et al. "Towards a Danish power system with 50% wind - Smart grids activities in Denmark". In: *IEEE Power & Energy Society General Meeting*. 2009.

[127] C. Zhou et al. "Modeling of the Cost of EV Battery Wear Due to V2G Application in Power Systems". In: *IEEE Transactions on Energy Conversion* 26.4 (2011), pp. 1041–1050.