

The IMA Volumes in Mathematics and its Applications

Sean Meyn
Tariq Samad · Ian Hiskens
Jakob Stoustrup *Editors*

Energy Markets and Responsive Grids

Modeling, Control, and Optimization



The IMA Volumes in Mathematics and its Applications

Volume 162

Series editor

Daniel Spirn, *University of Minnesota, MN, USA*

Institute for Mathematics and its Applications (IMA)

The Institute for Mathematics and its Applications (IMA) was established in 1982 as a result of a National Science Foundation competition. The mission of the IMA is to connect scientists, engineers, and mathematicians in order to address scientific and technological challenges in a collaborative, engaging environment, developing transformative, new mathematics and exploring its applications, while training the next generation of researchers and educators. To this end the IMA organizes a wide variety of programs, ranging from short intense workshops in areas of exceptional interest and opportunity to extensive thematic programs lasting nine months. The IMA Volumes are used to disseminate results of these programs to the broader scientific community.

The full list of IMA books can be found at the Web site of the Institute for Mathematics and its Applications:

<http://www.ima.umn.edu/springer/volumes.html>.

Presentation materials from the IMA talks are available at

<http://www.ima.umn.edu/talks/>.

Video library is at

<http://www.ima.umn.edu/videos/>.

Daniel Spirn, Director of the IMA

More information about this series at <http://www.springer.com/series/811>

Sean Meyn • Tariq Samad • Ian Hiskens
Jakob Stoustrup
Editors

Energy Markets and Responsive Grids

Modeling, Control, and Optimization



Springer

Editors

Sean Meyn
Department of Electrical
and Computer Engineering
University of Florida
Gainesville, FL, USA

Ian Hiskens
Department of Electrical Engineering
and Computer Science
University of Michigan
Ann Arbor, MI, USA

Tariq Samad
Technological Leadership Institute
University of Minnesota
Minneapolis, MN, USA

Jakob Stoustrup
Department of Electronic Systems
Aalborg University
Aalborg, Denmark

ISSN 0940-6573

ISSN 2198-3224 (electronic)

The IMA Volumes in Mathematics and its Applications

ISBN 978-1-4939-7821-2

ISBN 978-1-4939-7822-9 (eBook)

<https://doi.org/10.1007/978-1-4939-7822-9>

Library of Congress Control Number: 2018942505

Mathematics Subject Classification: 46N10

© Springer Science+Business Media, LLC, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Foreword

This volume contains a selection of essays based on a workshop “Control at Large Scales: Energy Markets and Responsive Grids” held at the Institute for Mathematics and its Applications from May 9–13, 2016 and organized by Sonja Glavaski, Ian Hiskens, Sean Meyn, Tariq Samad, and Jakob Stoustrup. These papers provide a landscape of the mathematical, financial and policy challenges that are present with the design of an efficient, stable and resilient electrical grid. The workshop ran as part of an annual thematic year organized by Fariba Fahroo, Tryphon Georgiou, J.W. Helton, Anders Rantzer, Tariq Samad, Eduardo Sontag and Allen Tannenbaum on Control Theory and its Applications that ran at the IMA during the 2015–2016 academic year. We would like to especially thank volume editors Ian Hiskens, Sean Meyn, Tariq Samad and Jakob Stoustrup. Finally, we acknowledge the National Science Foundation for its support of the IMA.

Minneapolis, MN, USA

Daniel Spirn

Introduction

The electric power infrastructure in any large region amounts to a system of systems—dynamically interconnected domains with communication, computation, and control functions at multiple temporal and spatial scales. The control loops that regulate electricity exist alongside electricity markets that introduce their own dynamics as they encourage generators to come on-line, or take a break from operations. The grid today is remarkably reliable, given its inherent complexity and uncertainty.

However, a tremendous transformation of the power grid is under way across the globe. The movement towards a so-called smart grid has been driven by many different players in industry and by societal pressure—people are concerned about the future of the planet, and in particular the impact of global warming. A truly smart transformation of the grid will bring about many societal benefits, including a reduction in pollution and greenhouse gases, reduced capital and operational expenses, and improved energy security. To ensure that our electricity supply remains reliable requires careful consideration of control strategies, communications, and market design.

In the future, as is true today, the ultimate challenge is to control generation, transmission, distribution, storage, and consumption of electricity. Consumers, markets, and regulators are also participants and stakeholders, and the multiple roles and interrelationships may exacerbate the challenge in the absence of appropriate market rules and control designs. Quoting one of the closing statements of the first chapter: In order to sustain such a drastic and rapid change, new control paradigms have to be developed moving the grid to a flexible, cooperative structure providing survivability of the system. This cannot be achieved without revisiting traditional reliability criteria and adding such new concepts as resilience, robustness and flexibility.

The editors of this volume organized the IMA workshop on Control at Large Scales: Energy Markets and Responsive Grids in May, 2016, as part of the year-long IMA program on Control Theory and its Applications, held at the University of Minnesota. The goal of the workshop was to bring together experts and newcomers interested in all aspects of the challenges facing the creation of a more sustainable

electricity infrastructure. Included in the meeting were experts in distributed control, stochastic control, stability theory, economics, policy, and financial mathematics, as well as in all aspects of power system operation.

This monograph consists of selected essays by participants in the workshop on the challenges we face today and in the future, along with potential solutions. All contributions were subjected to a peer-review process, with significant revisions in many cases.

The chapters are loosely organized according to theme, beginning with a survey from three authors from ISO New England. The next few chapters consider several significant challenges in the domain of market design. A theme in these chapters is the question of incentives for innovation in markets with significant risk on many time scales, and where assets may cost billions of dollars. These chapters are followed by chapters on optimization and distributed control, and the book concludes with articles addressing resilience and vulnerability.

Large-scale renewable generation, distributed energy resources, integration of supply-side and demand-side management, and dynamic markets herald a revolutionary change in power systems. The associated challenges are daunting and will require multidisciplinary approaches. With the breadth and depth of expertise it encapsulates, we are hopeful that this volume will contribute towards the envisioned future for serving humanity's energy needs.

We are grateful to our authors for their patience with the review process and other, less excusable, delays. The workshop itself was a hive of discussion and debate and all participants deserve our thanks as well. As with all IMA workshops, the arrangements were excellent and allowed the organizers to dedicate their attention to the workshop technical program. We would like to thank Fadil Santosa, the IMA Director, in particular for his support and encouragement. Finally, it has been a pleasure to work with the Springer team: Achi Dosanjh, Nick Valente, and Danielle Walker.

Gainesville, FL, USA
Minneapolis, MN, USA
Ann Arbor, MI, USA
Aalborg, Denmark

Sean Meyn
Tariq Samad
Ian Hiskens
Jakob Stoustrup

Contents

How to Manage the Complexity of the Grid?	1
Eugene Litvinov, Feng Zhao, and Tongxin Zheng	
Naïve Electricity Markets	29
David B. Spence	
Capacity Markets: Rationale, Designs, and Trade-Offs	59
Alfredo Garcia	
Redesign of US Electricity Capacity Markets	73
Robert W. Moye and Sean P. Meyn	
A Swing-Contract Market Design for Flexible Service Provision in Electric Power Systems	105
Wanning Li and Leigh Tesfatsion	
A Dynamic Framework for Electricity Markets	129
Anuradha Annaswamy and Stefanos Baros	
Fast Market Clearing Algorithms	155
Arvind U. Raghunathan, Frank E. Curtis, Yusuke Takaguchi, and Hiroyuki Hashimoto	
Small Resource Integration Challenges for Large-Scale SCUC	177
Cuong Nguyen, Lei Wu, Muhammad Marwali, and Rana Mukerji	
Multi-Grid Schemes for Multi-Scale Coordination of Energy Systems	195
Sungho Shin and Victor M. Zavala	
Graphical Models and Belief Propagation Hierarchy for Physics-Constrained Network Flows.....	223
Michael Chertkov, Sidhant Misra, Marc Vuffray, Dwijotham Krishnamurthy, and Pascal Van Hentenryck	
Profit Maximizing Storage Integration in AC Power Networks	251
Anya Castillo and Dennice F. Gayme	

Virtual Inertia Placement in Electric Power Grids.....	281
Bala Kameshwar Poolla, Dominic Groß, Theodor Borsche, Saverio Bolognani, and Florian Dörfler	
A Hierarchy of Models for Inverter-Based Microgrids	307
Olaoluwapo Ajala, Alejandro D. Domínguez-García, and Peter W. Sauer	
Asynchronous Coordination of Distributed Energy Resources with Packetized Energy Management	333
Mads Almassalkhi, Luis Duffaut Espinosa, Paul D. H. Hines, Jeff Frolik, Sumit Paudyal, and Mahraz Amini	
Ensemble Control of Cycling Energy Loads: Markov Decision Approach.....	363
Michael Chertkov, Vladimir Y. Chernyak, and Deepjyoti Deka	
Distributed Control Design for Balancing the Grid Using Flexible Loads	383
Yue Chen, Md Umar Hashmi, Joel Mathias, Ana Bušić, and Sean Meyn	
Disaggregating Load by Type from Distribution System Measurements in Real Time	413
Gregory S. Ledva, Zhe Du, Laura Balzano, and Johanna L. Mathieu	
Risk-Aware Demand Management of Aggregators Participating in Energy Programs with Utilities	439
William D. Heavlin, Ana Radovanović, Varun Gupta, and Seungil You	
Toward Resilience-Aware Resource Allocation and Dispatch in Electricity Distribution Networks	461
Devendra Shelar, Saurabh Amin, and Ian Hiskens	
A Cautionary Tale: On the Effectiveness of Inertia-Emulating Load as a Cyber-Physical Attack Path	491
Hilary E. Brown and Christopher L. DeMarco	

How to Manage the Complexity of the Grid?



Eugene Litvinov, Feng Zhao, and Tongxin Zheng

“... complex systems are counterintuitive. That is, they give indications that suggest corrective action which will often be ineffective or even adverse in its results.”

Forrester, Jay Wright

Abstract Power industry is facing revolutionary changes. The direction of the US Government to low carbon footprint and, as a consequence, high penetration of renewable energy resources and smart grid technologies are completely transforming planning and operational patterns for electric grid. As more and more variable and demand response resources being integrated into the electric grid, the grid operation is experiencing increasing level of uncertainties. The decision-making process under such environment becomes more challenging. The grid architecture and control also become more and more decentralized requiring new control paradigms and reliability metrics to be investigated in order to achieve much higher level of flexibility and resilience. These changes are disruptive enough to cause even transformations in utility business dealing with completely unknown situations. On the other hand, the evolution in computing; generation, transmission, and distribution technologies; and mathematical methods creates opportunities for innovation in power system design and control. New mathematical models for power system analysis and operation are being developed to address above challenges. We will discuss the need for new power system control and electricity market design directions while managing grid complexity.

E. Litvinov (✉) · F. Zhao · T. Zheng
ISO New England Inc., Holyoke, MA, USA
e-mail: elitvinov@iso-ne.com; fzhao@iso-ne.com; tzheng@iso-ne.com

1 Electric Grid Architecture Evolution

Modern power systems are going through different stages of evolution driven by technical, economic, and regulatory events. They went from decentralized, very loosely coupled grid to highly interconnected and centrally controlled systems. The increased complexity and lack of ability to manage it led to major blackouts forcing significant changes in system planning and operation. The Great Northeast Blackout of 1965 led to the creation of the power pools with control centers running energy management systems (EMS) and centralized regional planning and control. Each pool linked together multiple neighboring transmission companies with much stronger ties among them (Figure 1). Besides local control centers, power pools created pool control centers. Not only did this help in increasing reliability and resilience by the ability to provide balancing assistance, but also created savings for the member companies by using less expensive generation to meet the regional load. The interties between the pools were still weak and only used for emergency help. With the inception of the markets in the late 1990s and the creation of ISOs/RTOs, market players started placing economic transactions across the pool boundaries, increasing the complexity of the grid operation. This led to the reinforcement of the transmission system and tighter integration of the interconnected systems. The complexity of such an architecture required new ways of system control. The economic dispatch (ED) being done in each market area independently created so-called seams issues – inefficient utilization of the interties. This, in turn, requires additional information technology and communication infrastructure to **coordinate** market operations across large geographic areas. The electric grid had become a very large complex cyber-physical system. All these changes and attempts to increase grid reliability have not lowered the risk of large blackouts. On the contrary, the number and frequency of blackouts are increasing, which is the property of a very large complex system that exhibits self-organized criticality [1]. The blackouts follow the power law.

Currently, the power industry is facing another revolutionary change. Government directives to lower the carbon footprint and, as a consequence, high penetration of renewable energy resources and smart grid technologies are completely transforming planning and operational patterns of the electric grid again.

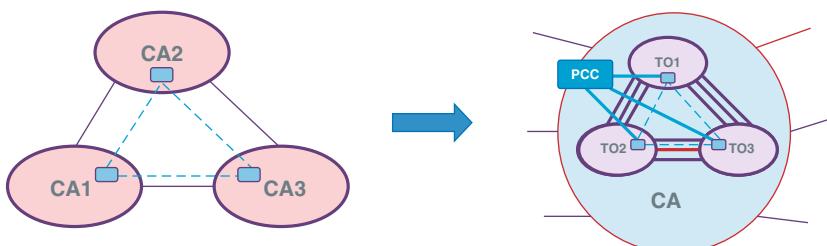


Fig. 1 Creation of power pools

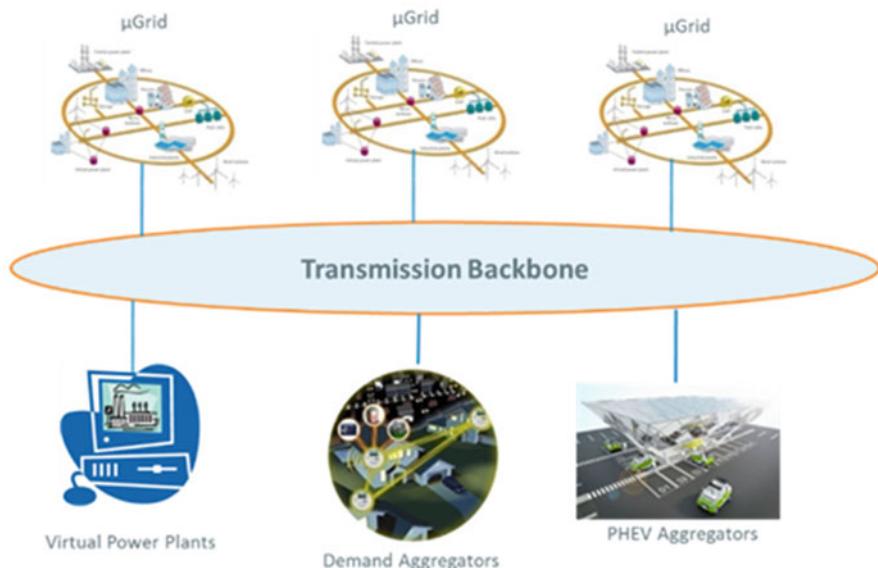


Fig. 2 Proliferation of DER

Distributed energy resources are being built deeply in the distribution networks, and the boundary between transmission, sub-transmission, and distribution is blurring. Traditionally, electric grid upgrades have been done centrally during transmission planning process. The process follows very strict reliability standards and requires large number of system studies, both in the steady state and transient regimes. Today, numerous changes to the grid are made ad hoc: distributed generation, microgrids, storage, etc. System operators lose control of the network perimeter. That topological uncertainty adds to the intermittent nature of the renewable resources. The architecture of the modern grid becomes more and more decentralized, while the control architecture is staying the same (Figure 2). Significant part of the generation resources is unobservable to the system operators. The unprecedented level of **uncertainty** is introduced not only in the location of distributed resources but their intermittent nature as well. The output of wind and PV generation can also swing significantly in time. The tribal knowledge of system operators is failing in dealing with completely different patterns of the system behavior. Even the concept of contingency is changing from being binary (the element of the grid is on or off) to continuous in time. The system load or generation can change by several gigawatts in a comparatively short period of time. This behavior, considered as abnormal or emergency, becomes part of the normal operation. This creates tremendous complexity in power system control.

In addition to DER proliferation, new “green” policies and low gas prices are causing retirements of coal, oil, and nuclear stations which leads to significant change in the generation mix and even capacity shortage. This as well makes

real-time operation decision-making process much more complicated and counterintuitive. Implementation of green and smart grid technologies is significantly increasing amount of power electronics connected to the transmission and distribution networks. Interactions of such a large number of interconnected controllers introduce another level of complexity and potential stability problems.

Another property of large cyber-physical systems is high interdependence of different infrastructures. Not only do we have to monitor electric grid contingencies but the failures in communication and information technology systems as well. The system **resilience** is getting much weaker, which requires new solutions for system planning and operation. Today, power systems are operated almost exclusively under the preventive paradigm. Every contingency is considered to be of probability 1, and the system is dispatched in such a way that no one failure would cause the violation of reliability criteria (N-1 standard). This approach, being quite expensive in the first place, becomes economically prohibitive in the new environment. More corrective actions must be introduced to make power system operation less expensive.

In order to understand the change in the power system operation, one can use Dy-Liacco's system state diagram [2] as shown in Figure 3. Each state is triggered by certain events and characterized by either getting very close to or violating specific constraints: physical, reliability, economic, etc. In the "alert" state system operator is facing a trade-off between preventive and corrective actions. By using preventive actions, the operator forces system away from the operating constraints increasing the margins. Alternatively, he/she may decide to defer actions until the system enters into the "emergency" state, especially if the process of moving from "alert" to "emergency" is comparatively slow. This is definitely a choice between reliability and economics. In the system with a reasonable level of uncertainty, the operator's actions are comparatively stable under wide range of conditions and situations. With the introduction of much higher level of uncertainty, the conditions that traditionally

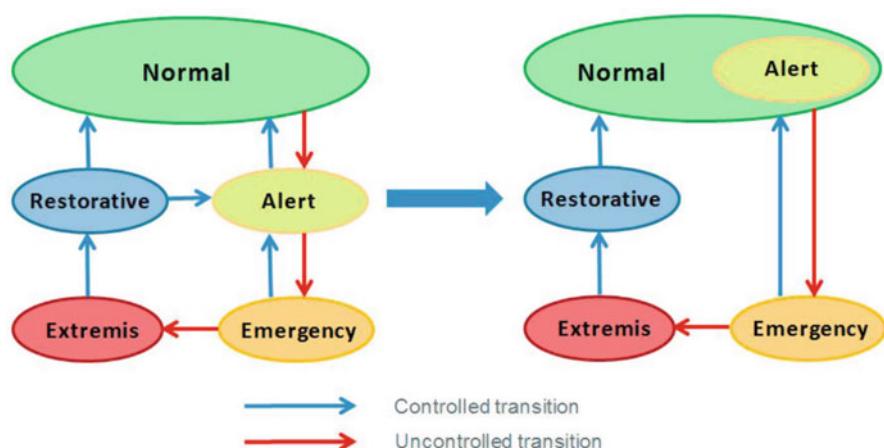


Fig. 3 New state transition diagram

are considered being “alert” become everyday “normal” phenomena, so we are observing the merging of these two states (Figure 3). Under new circumstances, the economics of the trade-off between preventive and corrective actions is changing. Corrective actions and remedial action systems (RAS) become more economic to use, which, in turn, forces the industry to review its control paradigm.

The **complexity induced by the large-scale distributed components, the lack of observability, and the uncertainty** in the future grid brings significant challenges in **modeling, decision-making, and control** of the system. To manage the above complexity by addressing these challenges, the industry needs **different control paradigm, new grid architecture, new algorithms, new models, and new reliability criteria**. The foundation for these new changes should be a more flexible grid architecture, e.g., a decentralized and distributed grid. Decision-making for the grid will have to be augmented by lowering the interdependence among different components and using robust solutions that are insensitive to external disturbances and economically efficient at the same time. The resulting robust components in turn will enable flexibility in distributed control structures and achieve the increasingly needed resilience of the grid. To efficiently design and implement such control architecture, we will need to formalize the new concepts of resilience and survivability and create metrics to be used to manage quality of the control.

In the following, we first discuss the general needs for **control architecture** (*Section 2*) and the likely **additional control components** needed for the existing control centers (*Section 3*). Then we explore some specific aspects of the new control architecture: **the corrective controls** (*Section 4*), **the uncertainty management** (*Section 5*), **the system flexibility** (*Section 6*), **the coordination algorithm** (*Section 7*), and **the new system resilience metrics** (*Section 8*). These aspects are by no means the complete list, but rather reflect what we have considered some major new pieces that will be needed for a future grid control.

2 New Power System Control Architecture

The new grid needs more flexibility to be able to operate with so much uncertainty. The flexibility is a very fuzzy concept and being used very loosely in the industry. It has to be formalized to be used in control and design algorithms. An attempt of such formalization is presented later in this chapter.

The industry is also very imprecise about the control architecture of the grid. Many different definitions of the control architecture being used: centralized/decentralized, hierarchical, coordinated, hierarchical-coordinated, distributed, collaborative, cooperative, etc. All these terms are not clearly defined even in the control theory literature and, in our opinion, require special attention from the control community. Today’s control seems to be strictly hierarchical and centralized. Such system is very rigid and has very little room for flexibility. With the increasing complexity, such an approach is insufficient to maintain system reliability and resilience.

Changing only the grid architecture to provide more flexibility while maintaining reliability is not sufficient. In order to reduce complexity, we have to make control system flexible as well, with the ability to adapt to different system states. This is impossible without some degree of distributed decision-making and decentralized control adapting to the unknown and dynamic environment. Additionally, decentralized systems are more resilient to disturbances or faults. These new qualities could be achieved by implementing distributed cooperative control paradigm with the capability of assembling temporary control entities collaborating in addressing specific events. Such a capability would allow decomposing a very complicated control problem into smaller, more manageable tasks. Large percentage of the system events are developing slowly enough so the corrective control would be capable of addressing large number of events. A new generation of state monitoring systems should be developed to take advantage of new information available from different devices and sensors. Decentralized control also requires careful design of the standard communication and control protocols and interfaces to enable interaction among heterogeneous components while cooperating in solving a common problem.

The increase of the computational capabilities and new IT architectures create opportunities for implementation of innovative control algorithms and infrastructure. Rapidly evolving cloud technology introduces unprecedented capabilities for online cooperation and collaboration. Being accessible from geographically wide area and capable of high-performance computing, cloud could serve as a medium for decentralized and distributed decision-making and control. The tremendous flexibility of this computing infrastructure will very quickly transition from very simple to highly complex control problems as needed. A simple example of such problem is resolving anticipated imbalance caused by a major contingency with the help of neighboring systems:

- Assembling model on the fly.
- Communicating coordination constraints (max imbalance allowed by participating entities), etc.
- Once resolved, the temporary collaborator is dropped.

Another benefit is ability to capture, accumulate, and use the patterns of the best control actions and strategies making it available during future events – stigmergy [3]. The system of such complexity also requires a different approach to reliability. Being under stress most of the time, power grid has to develop a survivability property, which is more general than just reliability. In addition, new reliability criteria together with resilience have to be investigated and implemented in order to formalize the objective of the power system control and required constraints.

3 Introducing New System Components to Control Center

The majority of power systems in the US are operated in an organized market environment or controlled by the RTO/ISO. In general, RTO/ISO performs two major functions: maintaining reliable system operation and managing wholesale electricity markets. Both functions can be considered as centralized control.

Modern power system operation deals with the physical aspect of the electric grid, and it is a challenging task. It involves many interacting processes. These processes can start from planning the system operating mode, coordinating generation and transmission outages with market participants and local control centers, forecasting system conditions, committing units for the real-time operation, scheduling generator outputs and interchanges with external control areas to meet the varying demand, collecting real-time system operating information through the supervisory control and data acquisition (SCADA) system, monitoring and alleviating static and dynamic security violations in the transmission system, and maintaining system voltages and frequency through the automatic generation control to taking emergency actions such as demand response, load shedding, emergency purchases, as well as conducting system restoration after a blackout. Some of these processes are automatic, and some of them require operators' manual actions.

Market operations, on the other hand, deal with the financial aspect of the electric system. Depending upon the structure of each regional market, each RTO/ISO may have different market operation procedures. However, broadly speaking, it includes clearing and settling the day-ahead energy, real-time energy, ancillary service markets, financial transmission rights (FTR), and forward capacity markets, monitoring and mitigating market power, and assessing the financial risk of market participants. Market operation and system operation are interconnected and affect each other. This is especially true for the real-time market and due to the fact that financial markets consider physical limitations of the transmission system.

The current RTO control system can be divided into two subsystems, the market system (MS) and the EMS, as shown in the dotted region of Figure 4. The market system performs all the market operation functions as described above, and the EMS facilitates the execution of all system operation processes. With the increasing penetration of renewables, distributed energy resources, demand response, and grid level smart devices, system operators are facing a much more complex system that contains a large number of controllable transmission and generation resources, various control models, fast-changing operating conditions, a high degree of uncertainty and is vulnerable to the changes in such external systems as the fuel delivery system, the regulatory regime, and commodity and financial markets. The existing control structure needs to be enhanced to facilitate the management of ever increasing complexity. In Figure 4, three new subsystems are introduced: dynamic decision support system (DDSS), risk management system (RMS), and market analysis, training, and simulation system (MATSS).

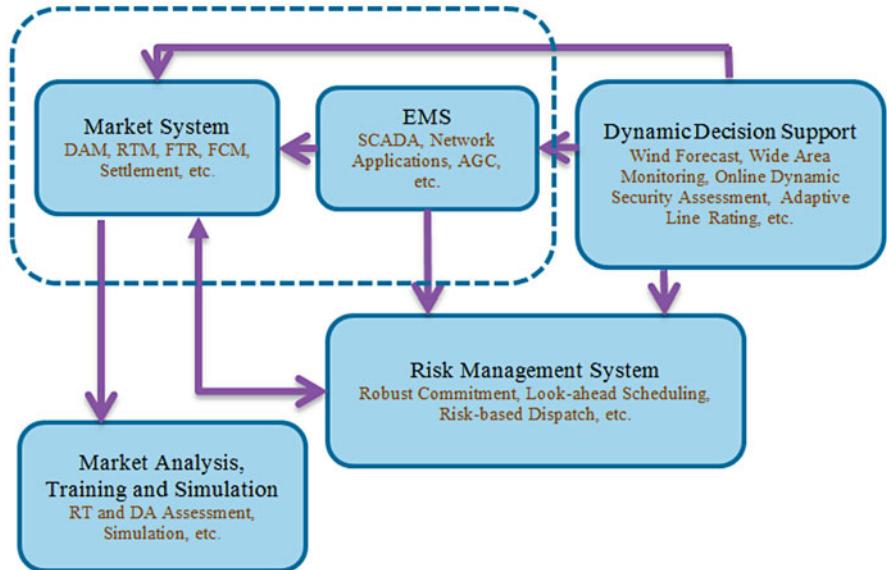


Fig. 4 System components for the future RTO

DDSS is a system that provides valuable control parameters to the system and market operation. The system is dynamic in the sense that it utilizes the latest available information in producing operational parameters. DDSS may have many functions and utilize different technologies depending on the task at hand. It should have the capability to perform the day-ahead and real-time renewable forecast including wind, solar, and DERs. It provides system operator with the most recent state of the system. Wide area monitoring using the phasor measurement unit (PMU) technology is a perfect fit to this task. Online dynamic security analysis or cascading event analysis will help the system operator define the secure region of the current system and provide possible corrective action plans. Online interface limit calculation and adaptive line rating [4] are also key functions of DDSS.

RMS is a system that deals with the increasing level of uncertainty faced by RTOs. It contains three major functions: collecting statistical information, assessing the system risk, and mitigating risk. Historical data, such as area control errors, load, wind production, solar generation, interchange level, transmission and generation failures, gas pipeline capacity reductions, etc., can be collected for statistical analysis. The system risk can then be assessed based on the statistical model established using historical data. Different risk indices, such as operational flexibility index [5], static security severity index [6], short-term loss of load expectation, etc., can be computed and displayed to the system operator. Different risk management techniques can be used to mitigate the system risk. They include, but not limited to, stochastic [7] and robust unit commitment [8], risk-based economic dispatch [9], dispatch with ramp constraints [10], etc.

MATSS performs an important function in assessing the efficiency of both market and system operations. As a recent trend, market operation is tightly integrated with the system operation. Actions taken in the system operation could have a large financial impact on the market participants. A comprehensive market simulator that is integrated with the traditional dispatcher training system is a very useful tool in simulating different system and market conditions, quantifying the financial impact of operator actions, and measuring the operational efficiency. In addition, such a simulation environment can be used to test future market designs, to assess the market competitiveness, and to perform the cost-benefit analysis of new market designs.

DDSS, RMS, and MATSS interact with MS and EMS directly and provide valuable information such as risk index, system security, cost of actions, and corrective action plans to the system operators. Introducing three new subsystems into the existing control scheme could help the system operator to better manage the increased complexity of the power system.

4 Exploring Corrective Controls

Under today's centralized control scheme, the risk associated with the power system uncertainty is mostly managed through *preventive* actions by the system operator. A typical example is the enforcement of contingency power flow limits. Namely, the power flow under any contingency will be within the safety limits, e.g., long-term emergency (LTE) limits, even without any remedial actions. However, in reality, a power line has different ratings such as short-term emergency (STE) and LTE, each associated with certain sustainable time based on thermal conditions. An STE rating associated with a short time period is higher than an LTE rating associated with a longer time period, indicating that the line can sustain a higher power level for a shorter time period. This feature could allow the contingency power flow to go above the current LTE limit without causing system reliability issues, provided that *corrective* actions such as unit redispatch can be taken to return the flow back to LTE within a certain time period. Consideration of such post-contingency corrective actions in the dispatch problem allows additional choices, thus providing more flexibility for the system control and lowering the dispatch cost [11–14]. With increasing penetration of renewable resources, such flexibility becomes more important because the conventional “preventive” control that requires covering every possible contingency scenario without factoring in the available corrective actions would become prohibitively expensive and may even lead to infeasibility. Below we present mathematical models of how to incorporate corrective actions into system operator's dispatch problem.

First consider a conventional security-constrained economic dispatch (SCED) problem:

$$\min_{\mathbf{p}} \mathbf{e}^T \cdot \mathbf{C}(\mathbf{p}) , \text{ s.t.} \quad (1)$$

$$\mathbf{h}(\mathbf{p}, \mathbf{d}) = \mathbf{0} , \quad (2)$$

$$\mathbf{f}(\mathbf{p}, \mathbf{d}) \leq \mathbf{f}_{\max} , \quad (3)$$

$$\mathbf{f}^c(\mathbf{p}, \mathbf{d}) \leq \mathbf{LTE} , \forall \text{Contingency } c \quad (4)$$

where \mathbf{p} is the vector of unit dispatch decisions, $\mathbf{C}(\mathbf{p})$ is the vector of unit dispatch costs, \mathbf{d} is the vector of load at different buses, $\mathbf{h}()$ is the power balance equation, $\mathbf{f}()$ is the vector of power flows in monitored lines, $\mathbf{f}^c()$ is the vector of power flows under Contingency c , and \mathbf{f}_{\max} is the vector of normal ratings of lines.

In the above SCED problem, dispatch decisions \mathbf{p} are made such that the power flow under any contingency would be retained within the safe limit of LTE (4). This is a very conservative control approach in the sense that the post-contingency flow could have been allowed to rise above LTE limits for a short time period (e.g., 15 minutes) without causing network reliability problems. As a result, the conventional SCED may unnecessarily use some expensive resources to contain a contingency flow to LTE, despite the chance of that contingency happening could be slim. With the increasing level of uncertainty in the system, the contingency definition must be expanded to cover a wide range of uncertainty spectrum, making the dispatch even more costly. Moreover, the risk of having no dispatch solution to cover a wide range of contingencies will increase. To address these problems, considering available corrective actions (e.g., unit redispatch) during contingency period becomes a natural choice to exploit system flexibility.

The SCED problem with **corrective actions** can be formulated as the following:

$$\min_{\mathbf{p}, \{\mathbf{p}^c\}} \mathbf{e}^T \cdot \mathbf{C}(\mathbf{p}) , \text{ s.t.} \quad (5)$$

$$\mathbf{h}(\mathbf{p}, \mathbf{d}) = \mathbf{0} , \quad (6)$$

$$\mathbf{f}(\mathbf{p}, \mathbf{d}) \leq \mathbf{f}_{\max} , \quad (7)$$

$$\mathbf{f}^c(\mathbf{p}, \mathbf{d}) \leq \mathbf{STE} , \forall \text{Contingency } c \quad (8)$$

$$\mathbf{f}^c(\mathbf{p}^c, \mathbf{d}) \leq \mathbf{LTE} , \forall \text{Contingency } c \quad (9)$$

$$|\mathbf{p} - \mathbf{p}^c| \leq \mathbf{R}_{15} , \forall \text{Contingency } c \quad (10)$$

where \mathbf{p}^c is the vector of unit redispatch under contingency c , $\mathbf{f}^c()$ is the vector of power flows under contingency c , and \mathbf{R}_{15} is the vector of units' 15-minute ramp capabilities. The corrective actions in the above formulation are the unit redispatch under each contingency c . The goal of the corrective actions is to retain the contingency power flow below LTE (9). The corrective actions are constrained by the unit's ramping capability (10). By considering the corrective redispatch actions \mathbf{p}^c , the power flow immediately after the contingency is relaxed from LTE in (4) to STE in (8), thus reducing the dispatch cost. From a mathematical perspective, the introduction of corrective actions \mathbf{p}^c in (5)–(10) allows a larger feasibility region for the dispatch decision \mathbf{p} than the original SCED formulation (1)–(4). This is due to

the fact that the corrective SCED will turn into the conventional SCED if one fixes the redispatch variables \mathbf{p}^c to \mathbf{p} .

Compared to the conventional SCED, the numbers of variables and constraints of the SCED with corrective redispatch increase dramatically by a factor of N (the number of contingencies). The solution of such a problem, in particular for real-time applications, is challenging. Decomposition techniques would have to be used together with parallel computing. Significant progress has been made on solving such problems [12, 13], and the latest reported results show that the problem can be tackled within several minutes for a large power system [14].

5 Modeling Uncertainty in Grid Operation

Uncertainty caused by the renewable integration is a key element of the system complexity. How to manage the system change caused by the sudden wind drop, cloud covering of solar panels, and high-speed wind cutout becomes an important field of study. Several methods exist today: deterministic method with increased operating margins such as additional reserve and ramp requirements, stochastic optimization, robust optimization, and chance-constrained optimization. The deterministic method is simple, but its efficiency is heavily dependent on the operating margin selected. Recent studies have shown that both stochastic and robust optimization techniques can achieve better efficiency in the uncertainty management. In this section, we first present the deterministic approach and then discuss two techniques in the process of making unit commitment (UC) decisions under uncertainty.

5.1 Deterministic Unit Commitment

A unit commitment problem can be stated as the system operator finding the optimal schedules of resources over a short time period, typically 24 hours for a day-ahead market or 1–4 hours for the real-time operation under the ISO environment, based on a cost minimization principle. For a deterministic UC problem, the optimal solution must satisfy the physical characteristics of resources, a set of operating constraints, and the demand forecast. A generalized deterministic security-constrained UC (SCUC) problem can be formulated as the following compact matrix form:

$$\min_{\mathbf{x}, \mathbf{y}} \mathbf{c}^T \cdot \mathbf{x} + \mathbf{b}^T \cdot \mathbf{y}, \text{ s.t.} \quad (11)$$

$$\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \leq \mathbf{g}, \quad (12)$$

$$\mathbf{H}\mathbf{y} \leq \mathbf{h}, \quad (13)$$

$$\mathbf{I}_d\mathbf{y} = \bar{\mathbf{d}}, \quad (14)$$

$$\mathbf{F}\mathbf{x} \leq \mathbf{f}, \quad (15)$$

$$\mathbf{y} \geq 0, \mathbf{x} \text{ is binary}. \quad (16)$$

where \mathbf{x} is the vector of binary commitment-related decision variables that may include a unit's on/off status and start-up or shutdown variables. \mathbf{c} is the vector of the commitment costs that include the start-up cost and no-load cost. \mathbf{y} is the dispatch decision variable that includes energy dispatch and ancillary service dispatch from both generators and loads, and \mathbf{b} is the vector of the incremental energy and ancillary service costs. Equation (12) represents the coupling constraints between the commitment decisions and dispatch decisions, e.g., units' maximum and minimum operating limits and start-up and shutdown ramps. \mathbf{A} , \mathbf{B} and \mathbf{g} are the coefficient matrixes and parameter vectors associated with (12). Equation (13) represents the dispatch constraints, e.g., reserve requirements constraints, transmission constraints, units' ramp limits, energy and reserve capacity constraints, etc. The equality constraint (14) corresponds to the expected energy balance constraint. \mathbf{I}_d is an indicator matrix that selects the components of vector \mathbf{y} to meet the expected demand \mathbf{d} . (15) represents constraints related to the commitment decisions, e.g., units' minimum up and down constraints, start-up cost constraints, etc. \mathbf{F} and \mathbf{f} are the coefficient matrix and the limit vector for (15).

Deterministic UC problem is often formulated as a mixed integer linear programming problem, which can be solved efficiently by commercial MILP solvers or Lagrangian relaxation method.

5.2 Stochastic Unit Commitment

Different from the deterministic UC, which determines the commitment schedule to meet the expected system condition such as the expected system load and the expected renewable generation, the stochastic optimization approach explicitly incorporates the probability distribution of the uncertainty [15–17]. A general form of a two-stage stochastic UC problem with the consideration of random system demand can be represented as

$$\begin{aligned} & \min_{\mathbf{x}, \mathbf{y}} \mathbf{c}^T \cdot \mathbf{x} + E(\mathbf{b}^T \cdot \mathbf{y}(\omega)), \text{ s.t.} \\ & \mathbf{Ax} + \mathbf{By}(\omega) \leq \mathbf{g}, \\ & \mathbf{Hy}(\omega) \leq \mathbf{h}, \\ & \mathbf{Idy}(\omega) = \mathbf{d}(\omega), \\ & \mathbf{Fx} \leq \mathbf{f}, \\ & \mathbf{y}(\omega) \geq 0, \mathbf{x} \text{ is binary}. \end{aligned} \quad (17)$$

Compared to the deterministic UC, the objective function of the stochastic UC contains two parts: the first-stage commitment cost $\mathbf{c}^T \mathbf{x}$ and the expected second-stage dispatch cost $E(\mathbf{b}^T \mathbf{y})$. $E()$ is the expectation function over the random event ω . $\mathbf{y}(\omega)$ is the recourse action or the dispatch solution in event ω . The first-stage decision is the commitment variable \mathbf{x} , and the second stage decision is the dispatch solution $\mathbf{y}(\omega)$, which has to meet the random demand realization $\mathbf{d}(\omega)$.

Many methods exist in solving the stochastic UC problem. [18] adopted the progressive hedging method, [19] utilized the Lagrangian decomposition technique. The most common solution technique is the Benders decomposition, where the master problem and subproblems are solved iteratively until convergence. The major limitation of stochastic UC in applying to large-scale power systems is the need for probability distribution of random variables and the possible large number of scenarios that requires intensive computation.

5.3 Robust Unit Commitment

Robust optimization has recently gained substantial popularity as a modeling framework for optimization under uncertainty, led by the work in [20–26]. The approach is attractive in several aspects. First, it only requires moderate information about the underlying uncertainty, such as the mean and the range of the uncertain data; and the framework is flexible enough that the modeler can incorporate more probabilistic information such as the correlation to the uncertainty model, when such information is available. Second, the robust model constructs an optimal solution that immunizes against all realizations of the uncertain data within a deterministic uncertainty set. Hence, the concept of robust optimization is consistent with the risk-averse fashion in which the power systems are operated.

Following the decision-making process (UC decision before the operating day and the dispatch against the uncertainty realization), we extend the previous deterministic formulation and discuss a two-stage adaptive robust unit comment model that considers adaptive economic dispatch actions in the real-time operation and produces robust commitment solutions to account for the uncertainty in the individual load. In this model, demand is assumed to belong to a polyhedral uncertainty set, which can be represented in the following general form:

$$\mathcal{D} \equiv \{\mathbf{d} \mid \mathbf{M} \cdot \mathbf{d} \leq \mathbf{N}, \mathbf{d} \geq 0\}. \quad (18)$$

Therefore, we replace (14) in the deterministic model by the following equation: $y_{i,t} = d_{i,t}$, $\forall (i, t) \in \mathcal{L} \times \mathcal{J}$ where $d_{i,t}$ is uncertain demand level and $\mathbf{d} \in \mathcal{D}$. The two-stage adaptive robust UC model is formulated as follows:

$$\min_{\mathbf{x}} (\mathbf{c}^T \mathbf{x} + \max_{\mathbf{d} \in \mathcal{D}} \min_{\mathbf{y} \in \{\mathbf{y} \mid \mathbf{B}\mathbf{y} \leq \mathbf{g} - \mathbf{A}\mathbf{x}, \mathbf{H}\mathbf{y} \leq \mathbf{h}, \mathbf{I}_d\mathbf{y} = \mathbf{d}, \mathbf{y} \geq 0\}} \mathbf{b}^T \mathbf{y}), \text{ s.t.} \quad (19)$$

$$\mathbf{F}\mathbf{x} \leq \mathbf{f}, \mathbf{x} \text{ is binary}.$$

The first-stage decision variables are the binary decisions that are related to the unit commitment. The system operator implements the unit commitment (here-and-now) decision before the observation of the actual load values. The power outputs and reserves are the second-stage (wait-and-see) decision variables, which are chosen after the uncertainty is realized. The goal of the above adaptive UC model is to find a robust unit commitment decision that minimizes the sum of the commitment costs for first-stage decisions and the worst-case dispatch costs induced by the first-stage together with the second-stage decisions.

Uncertainty set is an important aspect of the robust optimization. Different characterization of uncertainty set can affect the conservativeness and thus the solution of a robust optimization problem. Uncertainty sets described by different norms and the concept of uncertainty budget are discussed in [27]. To reduce the conservativeness of the robust optimization, some researchers adopt the data-driven approach in constructing the uncertainty set, which could also incorporate the spatial and temporal correlation of uncertain parameters.

Compared to stochastic UC, robust UC does not require probabilistic information about the uncertainty and tries to minimize the worst dispatch cost rather than the expected dispatch cost. The computation effort is relatively small. Methods used in the stochastic UC can be used to solve the robust UC problem. These methods include Benders decomposition, column and constraint generation, and affine policy approximation of the adaptive actions.

6 Managing System Flexibility

As more variable resources are integrated into the electric power system, supply and demand uncertainty increases dramatically. This requires the system to have the ability to react to sudden changes and accommodate new status within acceptable time period and cost. Therefore, the notion of flexibility recently has been drawing extensive attention in the power industry.

Most of the flexibility definitions in the literature [28–33] and metrics proposed pertain to particular aspects of power systems. Many of the assumptions underlying some of the metrics make their field of application very narrow. A unified flexibility framework for power systems is needed and will allow flexibility to be explicitly considered in the design of the system from both short-term and long-term perspectives and in control algorithms. In this section, we identify four elements, response time window, uncertainty, course of action, and cost, that are common to the flexibility literature in power systems. These four crucial elements serve as a basis for constructing effective measures of flexibility that can be applied to a wide range of situations.

6.1 Definition of Flexibility

Flexibility at a particular state is the ability of the system to respond to a range of uncertain future states by taking an alternative course of action within acceptable cost threshold and time window. Flexibility is an inherent property of a system. The following four elements are identified as the determinants of the flexibility: *response time window (T)*, *set of corrective actions (A)*, *uncertainty (U)*, and *response cost (C)*. The first three elements are affected by the power system operating criteria while the last element is determined by the economic criteria. Next, we will describe each element in detail.

6.1.1 Response Time Window (T)

The response time window indicates how fast the system is expected to react to the state deviations and restore the system to its normal state. The time window can be seconds, minutes, hours, days, or months depending on the purpose of the study. Based on the selected response time, a system may have different flexibility levels. Shorter time windows focus on the short-term operational flexibility, which indicates a system's timely response to emergency in minutes or hours. Longer time windows focus on the long-term planning flexibility, which shows a system's ability to cope with changes such as generation mix, regulatory policy, and electricity consumption pattern changes in years. Therefore, the time horizon has to be determined when we compare and evaluate system flexibility.

6.1.2 Set of Corrective Actions (A)

The set of corrective actions A represents the corrective actions that can be taken within the response time window under certain operating procedure. Therefore, the corrective actions set depends on the response time window T , i.e., $A(T)$. For instance, if $T=1$ hr, the corrective action set may include actions such as voltage control, commitment of units, and interchange scheduling. The size of the available corrective action set reflects the diversity of corrective actions. The larger the set $A(T)$ is, the more options operators have to respond to unexpected events. In turn, the response cost can be reduced or more uncertainty can be accommodated. Operating procedure changes or technology improvement will affect the corrective action set.

6.1.3 Uncertainty (U)

Uncertainty is the lack of complete information of the state of the system in the future. There has always been uncertainty in power systems operations and planning. Uncertainty is traditionally associated with the likelihood of failure of

components, forecast errors, or strategic gaming behavior of market participants. In recent years, the increase in variable generation creates new sources of uncertainty in the system because its output cannot be perfectly foreseen. The magnitude of the uncertainty determines how much flexibility a system requires to handle uncertainty and how flexible a system is. For example, the uncertainty considered under the N-1 criterion, U_{N-1} , is the loss of any single transmission or generation elements whereas the uncertainty considered under the N-2 criterion, U_{N-2} , is any combinations of two random outages of transmission or generation elements. A system that is flexible with respect to U_{N-1} may not be flexible if U_{N-2} is considered. We call the variation range of uncertainty that the system aims to accommodate the target range. The target range implies the risk level which the flexibility is in relation to and is subjectively set by operation or planning criteria. The larger the target range is set, the more conservative the system is designed or operated to be.

6.1.4 Response Cost (C)

The response cost C depends on the corrective action $a (\in A)$. This implies that the cost is a function of a , i.e., $C(a)$. In some cases, there can be a response cost threshold \bar{C} , which sets an upper bound on the cost to cope with the uncertainty realization. In other words, $C(a) \leq \bar{C}$. As a result, the cost threshold puts restriction on the available corrective actions in addition to the physical limitation associated with the time scales as illustrated in Figure 5. If the cost threshold is infinitely large, then there is no restriction on corrective actions associated with the cost limitation. If the cost threshold is low, some corrective actions become uneconomical and will

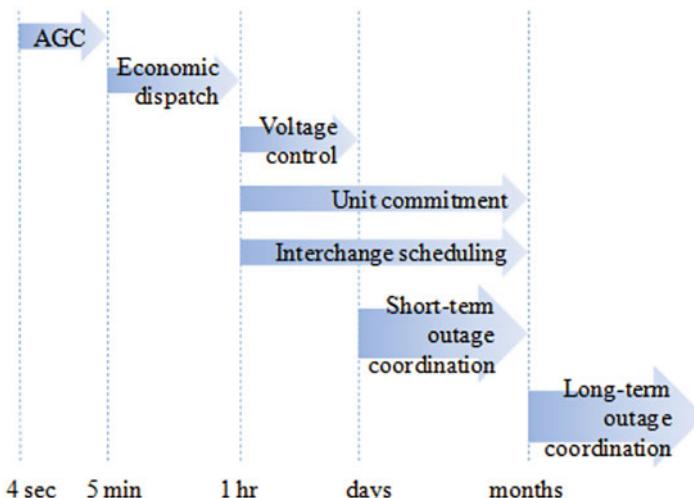


Fig. 5 Corrective actions in different time scales

not be taken into consideration. In some other cases, the objective of a decision-maker can be minimizing the response cost, i.e., $\min_{a \in A} C(a)$. Under this objective, the most economic corrective actions are sought in response to uncertainty.

6.2 Flexibility Metrics

With the 4-element flexibility concept, we can construct different flexibility metrics to serve the needs of system operation and planning. In particular, we first identify the largest variation range of uncertainty within which the system can remain feasible under given response time horizon and cost threshold. The flexibility metric is obtained by comparing the largest variation range with the target range to reflect excessive availability of the system relative to the target variation range.

Given a response time window T , the target variation range \bar{U}_T that decision-makers wish to accommodate at the time T can be characterized by a hypercube as follows:

$$\bar{U}_T = \left\{ u \mid \bar{u}^{LB} \leq u \leq \bar{u}^{UB} \right\},$$

where u is a n-dimensional vector, representing n uncertain sources in the system. The parameters \bar{u}^{LB} and \bar{u}^{UB} represent the lower and upper bounds. The largest variation range problem can be formulated in an abstract form as follows, for a given response time window T and a response cost threshold \bar{C} :

$$\max_{u^{LB}, u^{UB}, a(\cdot)} \|u^{UB} - u^{LB}\|, \text{s.t.,} \quad (20)$$

$$A \cdot a(u) + B \cdot u \leq b, \forall u \in [u^{LB}, u^{UB}], \quad (21)$$

$$c^T \cdot a(u) \leq \bar{C}, \forall u \in [u^{LB}, u^{UB}], \quad (22)$$

The objective function (20) of the above problem is to maximize the size of variation range of uncertainty, which is measured by norm $\|\cdot\|$. Equation (21) describes how system reacts to each uncertainty realization via the corrective actions $a(u)$. This constraint must hold for any uncertainty realized in the range $[u^{LB}, u^{UB}]$. Equation (22) indicates that the cost of the corrective actions must not exceed the cost threshold \bar{C} for any realization of uncertainty. The optimal solution (u^{*LB}, u^{*UB}) of the problem corresponds to lower and upper bounds of the largest range of uncertainty that the system can sustain within the response time window T and the cost threshold \bar{C} .

We define a flexibility metric by comparing the largest variation range with the target range. In an abstract form, the flexibility metric, denoted by F_T , is a function of the tuple $(u^{*LB}, u^{*UB}, \bar{u}^{LB}, \bar{u}^{UB})$, i.e.,

$$F_T = f(u^{*LB}, u^{*UB}, \bar{u}^{LB}, \bar{u}^{UB}), \quad (23)$$

Depending on the applications of interest, decision-makers can choose appropriate function f . For example, the metric can reflect the relative size of the largest variation range as compared to the target by letting $F_T = \|u^{*UB} - u^{*LB}\| / \|\bar{u}^{UB} - \bar{u}^{LB}\|$. It is straightforward to see that if the F_T is less than 1, it implies that the system cannot meet the target variation range.

Additionally, when the uncertainty materialized is beyond the largest variation range $[u^{*LB}, u^{*UB}]$, it means that the system is unable to accommodate such realization, hence at risk. Knowing what may potentially jeopardize system reliability is very important for designing an effective strategy to avoid such catastrophes.

7 New Coordination Algorithm

Under a centralized and hierarchical control scheme, the central entity (e.g., the system operator) at the top of hierarchy has access to the information of the entire system through the hierarchical path. While this allows the system operator to have the full control of the system, the communication burden is high, e.g., all information needs to be sent to the system operator through the sequential paths. Also, the hierarchical structure is vulnerable to communication attacks or errors since any disconnection on the sequential information path would cut the connection from the downstream entities. Thus the cost of maintaining such centralized control scheme could be high. Furthermore, the increasing penetration of distributed resources located in the distribution system makes the extension of transmission system operator's direct control to these resources an impossible task. As discussed in the previous sections, we envision a more decentralized control scheme for the future grid, e.g., balancing authorities' subsystems interact with each other through the transmission network. Also on the microgrid level, components within each microgrid are likely to act as autonomies (e.g., variable resources). For both situations, there is no central entity with access to all information in the system, e.g., each autonomy possesses its own private information, and the access to another autonomy's private information is dictated by a coordination protocol. As a result, the communication burden is distributed among autonomies. Also, multiple information paths exist between two autonomies, indicating a more resilient structure against communication failures or attacks.

The transformation of a centralized hierarchical control scheme into more decentralized schemes entails increased coordination among the subsystems, components, or autonomies since one subsystem can only make locally optimal decision without the critical information of other subsystems. A coordination scheme determines what information is exchanged between subsystems and how the information is used in each subsystem's decisions. General coordination schemes, e.g., Lagrange relaxation, Benders' decomposition, parametric optimization, etc., and their applications in power system have been well documented in the literature [34–41]. However, most of these decomposition algorithms suffer from parameter tuning, slow convergence, or infeasible solution before convergence. We have developed a

new general coordination scheme, i.e., marginal equivalent algorithm [42], that can be used for coordination between distributed subsystems. The algorithm works for any linear program problems such as

$$\begin{aligned} \min_{\mathbf{X}} \mathbf{C}^T \cdot \mathbf{X}, \text{ s.t.} \\ \mathbf{A} \cdot \mathbf{X} \leq \mathbf{B}, \\ \underline{\mathbf{X}} \leq \mathbf{X} \leq \bar{\mathbf{X}}. \end{aligned} \quad (24)$$

where \mathbf{X} is the vector of decision variables; $\underline{\mathbf{X}}, \bar{\mathbf{X}}$, respectively, are the vectors of lower and upper bounds of \mathbf{X} ; \mathbf{C} is the vector of coefficients in the objective; \mathbf{A} is the coefficient matrix in the linear constraints; and \mathbf{B} is the vector of constraint limits.

Each subproblem is formed by a subset of the original variables and a subset of the original constraints. During the iterative process, each subproblem is solved to identify the free variables (i.e., the variables that are not on its boundaries) and the binding constraints. Such information is shared among all subproblems, and each subproblem in the next iteration models the free variables and binding constraints from other subproblems. The algorithm is described in the following steps:

- Step 0: Initialize the free variable set and the binding constraints set;
- Step 1: Solve each subproblem with its own variables/constraints and the free variables/constraints of other subproblems;
- Step 2: If all subproblem solutions yield no change of free variables and binding constraints, then the algorithm converges; otherwise, go to Step 1.

The algorithm leads to the same solution as the centralized control scheme through the exchange of critical but not full information of the neighboring subsystems. The algorithm is proven to converge within a finite number of iterations, and feasible solutions can be obtained even before the convergence. A salient feature of the decomposition algorithm is that it does not rely on specific problem structures and does not require any parameter tuning. Also, the information exchanged between subproblems is not overwhelming. Furthermore, the convergence properties of the algorithm indicate a fast convergence rate similar to the simplex method. With the above features, the marginal equivalent algorithm could be an excellent method for the coordination among different subsystems in an increasingly decentralized power system. It can also be adapted to address today's coordination between system operators (i.e., the seams issue) where each area's system is formed as a subproblem with the marginal buses and binding constraints exchanged between areas.

8 Toward a Resilient Power System

Conventional power system reliability criteria were built for a centralized system. In an increasingly decentralized power system, the conventional system reliability model is insufficient to effectively evaluate and plan for the systems as the structure of these systems evolve over subsequent years. To address that deficiency, concepts such as resilience, robustness, sustainability, and survivability, which are a small subset of the terminology that has been used in other fields such as ecology and network analysis to describe the well-being of systems, may prove valuable in extending traditional reliability theory for power systems. However, one problem with using these terms is that their meaning has become confused, interchangeable, and often varies between and within disciplines [43]. However, the common thread through this maelstrom of terminology is a set of core concepts applicable to power system well-being analysis. After presenting these concepts and providing a terminological framework, a mathematical foundation is proposed from which an individual power system's performance may be measured.

8.1 Core Concepts

In order to address this broader spectrum of issues surrounding system well-being, five core characteristics of power systems are presented [44–48], indicating whether or not the system is in a state of well-being, and/or it will remain in a state of well-being as time progresses, and in the face of disturbances. Each concept is then associated with a specific word for the purpose of providing cohesion to the concepts within this section, not to add additional knots to the terminology entanglement. The core concepts are as follows:

- *Reduction of the number and severity of disturbances to the system and operating the system far from critical points.* This concept relates to system protective actions taken to decrease the impact (which can result from few disturbances or smaller disturbances) of those disturbances which are able to be controlled or protected against. For example, decreasing the forced outage rate of a unit and improving the lightning protection on transmission lines would both help protect the system from undesirable contingencies. This concept may be defined through **stability** during normal operation with regard to endogenous disturbances and **robustness** with regard to exogenous disturbances.
- *Acceptable quality of service, minimized value loss, and maximized speed of recovery during and after the system are subject to endogenous disturbances (or the absence of disturbances - normal operation).* This is the field analyzed in conventional reliability theory: given that system components may fail, how often do such failures occur and what is the impact on the system, its customers, and on power delivery. A major focus of this area is on maximizing the speed of system recovery. This also encompasses normal operation in the sense that

the system should provide an acceptable quality of service when there are no contingencies, so there are no inherent flaws in the system design. This concept may be defined as **reliability**.

- *Acceptable quality of service, minimized value loss, and maximized speed of recovery during and after the system are subject to **exogenous** disturbances.* The response of the system to external challenges is considered in this characteristic. For example, how will the system respond to a directed attack on the most critical infrastructure or perhaps from a high-impact natural disaster? Again, the time to recovery of the system should be as short as possible. This concept may be defined as **resilience**.
- *Reactive adaptation in the medium term to better handle disturbances and improve the quality of service.* In response to disturbances, a system is able to react on the time scale of days to months to better protect itself from the effects of unexpected contingencies. The system's aptitude to be able to do this is addressed in this characteristic. This concept may be defined as **survivability**.
- *Proactive evolution in the long term to better handle disturbances and improve the quality of service and allows for enhanced functionality.* This characteristic deals with the proclivity of the system to make long-term changes (on the order of years to decades) that will anticipate future challenges and add enhanced functionality. This includes the ability to integrate smart grid concepts, while controlling or at least understanding the complexity, so as to only elicit beneficial autonomous behavior and self-organization. This concept may be defined as **sustainability**.

These concepts were compiled so that they spanned the space of system well-being. The relationship between these concepts is shown in Figure 6.

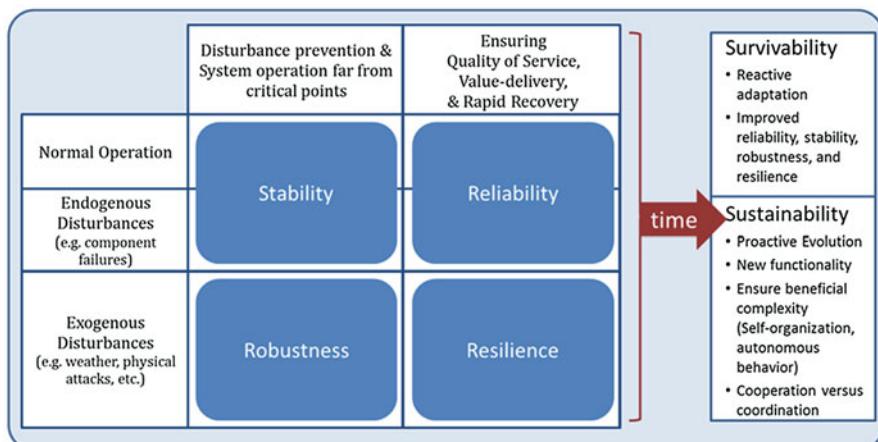


Fig. 6 System well-being characteristics and how they relate to one another

8.2 System Metrics

The ensemble of concepts presented in the previous section can be further solidified by quantitative metrics to evaluate many of these concepts. To that end, Figure 7 shows a hypothetical system disturbance where $f(t)$ could be an indicator of system health including frequency or voltage. The ensemble of concepts presented in the previous section can be further solidified by quantitative metrics to evaluate many of these concepts. To that end, Figure 7 shows a hypothetical system disturbance where $f(t)$ could be an indicator of system health including frequency or voltage.

Satisfactory response of the system in the face of exogenous disturbances is a critical component of system well-being. In the example disturbance shown in Figure 7, a system which is least affected by the disturbance will be preferable. To measure how much the system is impacted, a number of potential metrics are introduced. First, the average change in f during a disturbance, averaged over all events:

$$\overline{\Delta f} = \frac{1}{N} \sum_{i=1}^N \Delta f_i .$$

This quantity provides a measure of how significant the average effect of the disturbance is. A second metric for resilience is the rate of change of f just after the onset of the i th event:

$$\left. \frac{df}{dt} \right|_{t=t_{e_i}^+} .$$

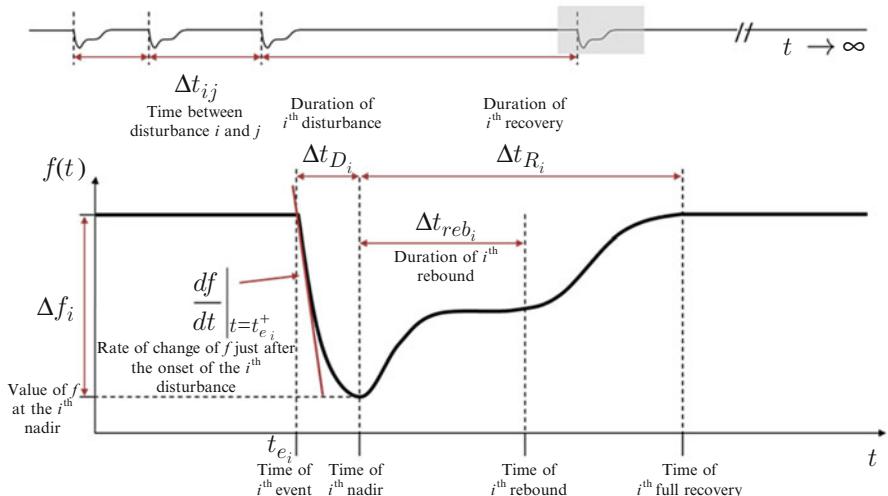


Fig. 7 Terminology related to an event in a power system [48]

Systems with greater inertia will not change as rapidly as those with less inertia, and thereby changing less rapidly in the face of disturbances allows system operators to prevent greater damage and reductions in value delivery. Along that idea, the longer the system is in a degraded operational state, the greater the potential for further disturbances and damage and the greater the loss of value delivery. Therefore, a natural metric is the average duration of the recovery, averaged over all events:

$$\overline{\Delta t_R} = \frac{1}{N} \sum_{i=1}^N \Delta t_{R_i} .$$

Stability and robustness metrics. The ability of a system to operate as far as economically possible from critical points is essential for the well-being of the system. In this context critical points are the threshold between low and high probability of system disturbances, either during normal operation, or in the presence of exogenous disturbances. As discussed in [45], as a system increases its level of loading, it may reach a value at which the probability of large cascading failures rapidly increases. This is demonstrated graphically in Figure 8, where, as

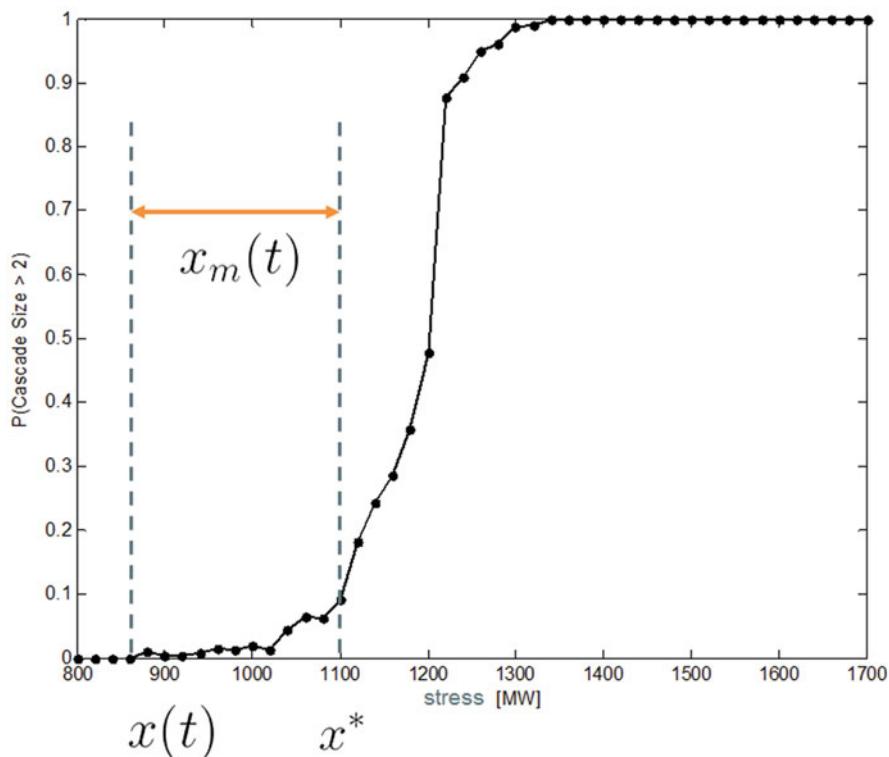


Fig. 8 Measuring system stress and the system's operational distance from criticality

in [45], system stress is measured as constant coefficient multiplying each system load. If we assume that the system is operating at stress level $x(t)$ at time t and that the critical stress level is x^* , then the operating margin between the two is a measure of how stable the system is, e.g.,

$$x_m(t) = x^* - x(t).$$

Similarly, another measure of how robust the system is would be how rapidly that phase transition occurs from low probability of cascading failure to almost certain cascading failure. Suppose that Δp_c is the change of the probability of cascading and Δx is the change of the system stress level. Then this metric can be defined as follows:

$$\varepsilon = \frac{\Delta p_c}{\Delta x}.$$

Reliability metrics. The conventional reliability theory is, of course, an important part of an overall system well-being analysis. However, even here other metrics may be added to the standard loss of load expectation (LOLE), including expected unserved energy (EUE) and equivalent load carrying capacity (ELCC) for stochastic generation.

9 Conclusion

The transformation of the power industry led to significant changes in the electric grid business. The complexity of the grid is growing exponentially, especially on the periphery due to proliferation of distributed energy resources and power electronic devices. Increasing uncertainty requires novel approaches to the grid planning, operation, and control. Traditionally centralized grid architecture is transforming to more decentralized and constantly being stressed by the volatility of the wind and solar sources of energy. In order to sustain such a drastic and rapid change, new control paradigms have to be developed moving the grid to flexible, cooperative structure providing survivability of the system. This cannot be achieved without revisiting traditional reliability criteria adding such new concepts as resilience, robustness, and flexibility. These concepts, in turn, require formalizing their definition and creating metrics to be able to use them in system design and operation. New grid also needs much more sophisticated real-time decision support system providing new ways of dealing with stochastic nature of the grid behavior. Probabilistic approaches and stochastic and robust optimization methods are being developed to make usually very computationally complex algorithm tractable for solving real-size problems of electric grid planning and operation. New information technologies and more powerful computers create new opportunities for the real-

time use of traditionally intractable computational methods. Such technology as cloud computing creates natural environment for the cooperative control enabling fast and reliable communication of the distributed control systems.

References

1. Jensen HJ (1998) Self-organized criticality: emergent complex behavior in physical and biological systems. Cambridge lecture notes in physics. Cambridge University Press, Cambridge
2. DyLiacco T (1968) Control of power systems via the multi-level concept. Technical report SRC-68-19, Case Western Reserve University, Systems Research Center
3. Bonabeau E (1999) Editor's introduction: stigmergy. Special issue of *Artif Life Stigmergy* 5(2):95–96
4. Maslennikov S, Litvinov E (2009) Adaptive emergency transmission rates in power system and market operation. *IEEE Trans Power Syst* 24(2):923–929
5. Zhao J, Zheng T, Litvinov E (2016) A unified framework for defining and measuring flexibility in power system. *IEEE Trans Power Syst* 31(1):339–347
6. Wang Q, McCalley J, Zheng T, Litvinov E (2013) A computational strategy to solve preventive risk-based security constrained OPF. *IEEE Trans Power Syst* 28(2):1666–1675
7. Carpentier P, Gohen G, Culoli JC, Renaud A (1996) Stochastic optimization of unit commitment: a new decomposition framework. *IEEE Trans Power Syst* 11:1067–1073
8. Bertsimas D, Litvinov E, Sun XA, Zhao J, Zheng T (2014) Closure to the discussion of “adaptive robust optimization for the security constrained unit commitment problem”. *IEEE Trans Power Syst* 29(2):996–997
9. Wang Q, Yang A, Wen F, Li J (2013) Risk-based security-constrained economic dispatch in power systems. *J Mod Power Syst Clean Energy* 1(2):142–149
10. Nivad N, Rosenwald G, Chatterjee D (2011) Ramp capability for load following in MISO markets. Available <https://www.misoenergy.org/layouts/MISO/ECM/Redirect.aspx?ID=112806>
11. Monticelli A, Pereira MVF, Granville S (1987) Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Trans Power Syst* 2(1):175–180
12. Capitanescu F, Wehenkel L (2008) A new iterative approach to the corrective security-constrained optimal power flow problem. *IEEE Trans Power Syst* 23(4):1533–1541
13. Pinto H, Stott B (2010) Security constrained economic dispatch with post-contingency corrective rescheduling. In: FERC conference, Washington, DC, June 23–24
14. Liu Y, Ferris MC, Zhao F (2014) Computational study of security constrained economic dispatch with multi-stage rescheduling. *IEEE Trans Power Syst* 30(2):920–929
15. Takriti S, Birge JR, Long E (1996) A stochastic model for the unit commitment problem. *IEEE Trans Power Syst* 11:1497–1508
16. Ozturk UA, Mazumdar M, Norman BA (2005) A solution to the stochastic unit commitment problem using chance constrained programming. *IEEE Trans Power Syst* 19(3):1589–1598
17. Wu L, Shahidehpour M, Li T (2007) Stochastic security-constrained unit commitment. *IEEE Trans Power Syst* 22:800–811
18. Ryan SM, Silva Monroy C, Watson JP, Wets RJB, Woodruff DL (2013) Toward scalable, parallel progressive hedging for stochastic unit commitment. In: Proceedings of the IEEE 2013 IEEE power and energy society general meeting
19. Papavasiliou A, Oren SS, Rountree B (2015) Applying high performance computing to transmission-constrained stochastic unit commitment for renewable energy integration. *IEEE Trans Power Syst* 30(3):1109–1120
20. Ben-Tal A, Nemirovski A (1998) Robust convex optimization. *Math Oper Res* 23:769–805
21. Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. *Oper Res Lett* 25(1):1–13

22. Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Math Program* 88:411–421
23. Ghaoui LE, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM J Matrix Anal Appl* 18:1035–1064
24. Ghaoui LE, Oustry F, Lebret H (1998) Robust solutions to uncertain semidefinite programs. *SIAM J Optim* 9(1):33–52
25. Bertsimas D, Sim M (2003) Robust discrete optimization and network flows. *Math Program* 98:48–71
26. Bertsimas D, Sim M (2004) The price of robustness. *Oper Res* 52:35–53
27. Bertsimas D, Brown DB (2009) Constructing uncertainty sets for robust linear optimization. *Oper Res* v57(6):1483–1495
28. Lannoye E, Flynn D, O’Malley M (2013) Evaluation of power system flexibility. *IEEE Trans Power Syst* 27(2):922–931
29. Ma J, Kirschen D, Silva V, Belhomme R (2011) Optimizing the flexibility of a portfolio of generating plants to deal with wind generation. In: IEEE power and energy society general meeting, San Diego, CA
30. Menemenlis N, Huneault M, Robitaille A (2011) Thoughts on power system flexibility quantification for the short-term horizon. In: IEEE power and energy society general meeting, San Diego, CA
31. Bouffard F, Ortega-Vazquez M (2011) The value of operational flexibility in power systems with significant wind power generation. In: Power and energy society general meeting, San Diego, CA
32. Studarus K, Christie R (2013) A deterministic metric of stochastic operational flexibility. In: IEEE power and energy society general meeting, Vancouver, BC
33. Ulbig A, Andersson G (2012) On operational flexibility in power systems. In: IEEE power and energy society general meeting, San Diego, CA
34. Conejo AJ, Aguado JA (1998) Multi-area coordinated decentralized DC optimal power flow. *IEEE Trans Power Syst* 13(4):1272–1278
35. Echeverri MG, Lezama JML, Roberto J, Mantovani S (2010) Decentralized AC power flow for multi-area power systems using a decomposition approach based on Lagrangian relaxation. *Rev Fac Ing Univ Antioquia* 53:225–235
36. Bakirtzis AG, Biskas PN (2003) A decentralized solution to the DC-OPF of interconnected power systems. *IEEE Trans Power Syst* 18(3):1007–1013
37. Arnold M, Knopfli S, Andersson G. Improvement of OPF decomposition methods applied to multi-area power systems. Available <http://www.eeh.ee.ethz.ch/uploads/txthpublications/ArnoldImprovementOfOPFDecompositionMethods.pdf>
38. Cadwalader MD, Harvey SM, Hogan WW, Pope SL. Coordinating congestion relief across multiple regions. Available www.hks.harvard.edu/fs/whogan/isoc1099r.pdf
39. Ilic M, Lang J. Methods of selecting the desired net interchange between New York and New England, and between New York and Ontario. Available at http://www.ferc.gov/EventCalendar/Files/20120627090023-Wednesday_SessionA_Ilic.pdf
40. Baldick R, Chatterjee D (2010) Final phase I report on coordinated regional dispatch framework. Available at www.midwestiso.org
41. Phulpin Y (2009) Coordination of reactive power scheduling in a multi-area power system operated by independent utilities. Ph.D. thesis, Georgia Institute of Technology
42. Zhao F, Zheng T, Litvinov E (2014) A marginal equivalent decomposition method and its application to multi-area optimal power flow problems. *IEEE Trans Power Syst* 29(1):53–61
43. Grimm V, Wissel C (1997) Babel or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. *Oecologia* 109:323–334
44. NERC (2007) Definition of adequate level of reliability. Available at https://www.nerc.com/comm/Other/Adequate%20Level%20of%20Reliability%20Task%20Force%20%20ALRTF%20DL/Final%20Documents%20Posted%20for%20Stakeholders%20and%20Board%20of%20Trustee%20Review/2013_03_26_ALR_Definition_clean.pdf

45. Liao H, Apt J, Talukdar S (2004) Phase transitions in the probability of cascading failures. In: Electricity transmission in deregulated markets
46. Richards M et al (2009) Metrics for evaluating survivability in dynamic multi-attribute tradespace exploration. *J Spacecr Rockets* 46:1049–1064
47. Sterbenz J et al (2010) Resilience and survivability in communication networks: strategies, principles, and survey of disciplines. *Comput Netw* 54:1245–1265
48. Richards MG, Hastings DE, Rhodes DH, Ross AM, Weigel AL (2009) Design for survivability: concept generation and evaluation in dynamic tradespace exploration. In: Second international symposium on engineering systems, MIT, Cambridge, MA

Naïve Electricity Markets



David B. Spence

Abstract The push toward competition, market pricing, and less regulation in the electricity industry embraces the logic and elegance of markets. It means that participants are exposed to more price risk than in the past, and it represents a narrowing of both the notion of the public interest and the government's role in protecting that interest. But electricity markets can never resemble the idealized markets of economic theory that have become so popular in conservative policy discourse. This chapter explores why that is. More specifically, it (i) reviews the work of economic thinkers whose work shapes the conservative challenge to regulation and the push for further deregulation, (ii) explores why the economist's goal of allocative efficiency does not subsume elements of fairness and risk management that are important to voters and policymakers and why economic models continue to have trouble incorporating important lessons from behavioral research, and (iii) explains why these lessons are important to understanding the operation of electricity markets and to an understanding of the problem of ensuring a reliable, reasonably priced energy supply.

1 Economic Theory and Electricity Regulation

In the nineteenth century, electricity titan Samuel Insull sought price stability through government regulation of the electricity industry and is credited with creating the first modern electric utility, Commonwealth Edison [155]. Since at least

This chapter is adapted from a broader paper presented at the IAMS conference in Minneapolis in 2016 and published in the Notre Dame Law Review in 2017.

D. B. Spence (✉)

University of Texas at Austin, School of Law and McCombs School of Business, Austin, TX,
USA

e-mail: david.spence@mccombs.utexas.edu

the early twentieth century, governments have used ex ante regulation (public utility law) to achieve price and supply stability in the electricity industry. Reasoning that the electricity sector is a natural monopoly, state regulators opted for administrative price setting and monopoly service in retail markets; the Federal Power Act of 1935 imposes a fairness requirement on wholesale electric prices.

In the last few decades, however, federal (and some state) regulators began to introduce competition and market pricing into electricity markets; over that same time period, government policy has favored greener, and more decentralized, electric generation sources. The trends toward more competition and market pricing of energy and toward a greener, and distributed, energy mix are not the product of some broad national consensus. Rather, they represent political victories won (and defended) in an increasingly contentious political environment. Some states embrace competitive markets; others oppose them with equal resolve. Similarly, the battle over whether and how to green the energy mix is a continuous, multifront battle. In the last few years, the US Supreme Court has twice addressed jurisdictional disputes over electricity market regulatory authority between the Federal Energy Regulatory Commission (FERC) and states, and policy fights over such issues as demand response, capacity markets, net metering, the EPA's Clean Power Plan, renewable portfolio standards, and more clog the dockets of state legislatures, regulatory commissions, and courts across the country. These are all fights over attempts to use regulation to alter the market allocation of costs and benefits.

In today's ideologically and politically polarized environment, it has become increasingly popular among conservatives to cite economic theory in support of deregulatory positions. Beyond general appeals to the wisdom of the market and the failures of government, more conservatives are appealing to specific economic thinkers, such as Austrian economist Friedrich Hayek's arguments in favor of the market's ability to promote innovation, and against certain types of economic regulation as "serfdom" [66]. Indeed, appeals to "Austrian economics" have been particularly popular among Republican politicians, including Ron Paul, Rick Perry, Michele Bachmann, and Paul Ryan. These appeals serve not only to buttress candidates' conservative bona fides with Republican primary voters but also as evidence that the scholarly economic critique of regulation has penetrated public debates over regulation, including the regulation of electricity markets, more than ever before.

That critique draws in part on a stylized notion of the public welfare. Welfare economists seek allocative efficiency, a distribution of costs and benefits that maximizes social net benefit [151]. The neoclassical model of perfect competition yields this optimal allocation, as Adam Smith foreordained more than two centuries ago, if individuals are free to exchange goods and services and to enter and exit markets. That way, freely floating prices attached to their exchanges will allocate capital and labor to their highest uses, thereby maximizing social net benefits. This is the "invisible hand" of the market [133]. In the 20th century, scholars and policymakers began to argue that existing regulatory regimes were smothering these largely beneficial market forces and pushed for deregulation of the airline, banking, telecommunications, and energy sectors, among others.

In the electricity industry, after a century of regulated prices and service, policymakers at the Federal Energy Regulatory Commission (FERC) and some (but not all) state utility commissions ordered the “unbundling” of electricity sales from electricity delivery, the introduction of competition into the power sales segment of the industry, and the opening of the (still-regulated) delivery network to all on equal terms [129]. It has been well documented that the move to competition and market pricing has not been without its bumps. California’s newly competitive electricity markets failed spectacularly in 2000–2001,¹ due to a combination of poor market design, bad luck, and illegal manipulation of the market by sellers. However, diagnoses of the California market failure by Hayek’s disciples did not blame the sellers who were subsequently fined for manipulating those markets; rather, they blamed the regulation (e.g., [18]). Others were shocked by the crisis, and it slowed the transition to competition in many parts of the United States and the world. Nevertheless, competitive markets survived in many parts of the United States, and market overseers (like FERC and so-called independent system operators (“ISOs”) and regional transmission organizations (“RTOs”)) responded to the crisis by establishing market monitors to guard against market manipulation in electricity markets. The process of tweaking market rules to prevent market failure has been ongoing since the California crisis.

In addition to the notion that free markets beget efficiency, economic thought also addressed the “government versus markets” problem in another way, namely, by applying the tools of economic analysis to government policymaking. The period from the 1940s through the 1970s, in particular, saw the publication of seminal economic critiques of government decision-making and regulation. These analyses, which gained influence in the American policy debate in the ensuing decades, almost invariably suggested flaws in the regulatory process. The Coase theorem, for example, challenged the notion that externality problems (e.g., pollution) necessitated a regulatory response. Coase demonstrated that the most efficient policy response to a pollution problem is not command and control regulation or even a pollution tax but rather the establishment of property rights that will enable the holders of those rights to bargain to an efficient solution [29]. Arrow’s theorem offered another example of the use of formal logic to challenge the capacity of government to address market failures by demonstrating mathematically that no social choice mechanism—legislative or otherwise—could produce choices that satisfy certain basic democratic principles [3]. Arrow’s analysis became a pillar of so-called “public choice” economics, by supporting the inference that government cannot serve any “public interest” because no such interest exists. Subsequent public choice analyses complemented Arrow by characterizing regulation as the product of “rent-

¹Daily average prices on the California wholesale market soared to more than 20 times historical averages, triggering the bankruptcy of one major utility and the near bankruptcy of another, an Enron-centered market manipulation scandal, and more.

seeking”² by industry rather than attempts to address market failure and regulators as prone to “capture” by the very industries they oversee (e.g., [106, 111, 141]).

It is difficult to overstate the enormous influence economic and public choice analyses have exerted over scholarship within the field of energy law. Public choice scholarship fed the deregulatory impulse that produced the restructuring of American natural gas and electricity markets in the 1980s and 90s. That is, that transition was inspired both by the perceived failures of regulation (e.g., [76, 113, 114])³ and by economic analyses suggesting that competition and market pricing would benefit consumers (e.g., [20, 34, 142, 157]). And it is not uncommon for scholars to dismiss the notion of the public interest, to dismiss regulation as mere rent-seeking by private interests, and to look more benignly on the effects of monopoly and oligopoly, reasoning that markets are often self-correcting, because barriers to entry are lower, and economies of scale more common than traditional antitrust analysis assumed [76].

Friedrich Hayek’s writings provided much of the intellectual foundation for public choice scholarship. Writing two centuries after Adam Smith, Hayek argued that “it is essential that the entry into the different trades should be open to all on equal terms,” and that “[a]ny attempt to control prices or quantities of particular commodities deprives competition of its power” to promote efficiency. Hayek questioned the ability of regulators to have the foresight to regulate wisely and explored the market as a kind of complex adaptive system likely to allocate benefits and costs on its own better than regulators ever could [66, 67].⁴ This is because knowledge in this kind of system is not centralized: rather, it is diffuse and unevenly distributed among economic agents. The price signal transmits knowledge from agent to agent over time, unleashing a process not of equilibrium but of constant adaptation to constant change from which order emerges spontaneously. Crucially, this process produces better outcomes, said Hayek, than will government planning. Hayek did not explicitly apply his framework to electricity markets, but others have,

²Economic “rents” are returns that exceed the competitive return. Public choice economists used the term “rent-seeking” to refer to political action designed to allow firms or other economic actors to capture economic benefits not otherwise available but for the regulation.

³As Professor Richard Pierce [113] notes, the FERC was ordered to regulate natural gas wellhead sales by the Supreme Court, a decision that ultimately resulted in massive natural gas shortages in the 1970s. Later, scholars began to question the wisdom of rate regulation of wholesale sales in electricity markets [76].

⁴Hayek’s basic critique of regulators is laid out in *The Road to Serfdom*: “[G]overnment in all its actions is bound by rules fixed and announced beforehand—rules which make it possible to foresee with fair certainty how the authority will use its coercive powers in given circumstances and to plan one’s individual affairs on the basis of this knowledge. Though this ideal can never be perfectly achieved, since legislators as well as those to whom the administration of the law is entrusted are fallible men, the essential point that the discretion left to the executive organs wielding coercive power should be reduced as much as possible is clear enough”. Hayek denies that government can be impartial or reflect the public interest, noting that any collectivist state “must, of necessity, take sides,” thereby becoming a “‘moral’ institution . . . [that] imposes on its members its views on all moral questions” [66].

challenging the notion that public utility regulation, in particular, can “get prices right” or otherwise create conditions that mimic textbook competition (e.g., [89]).

Energy economists and their allies within the legal academy cite this logic to advocate the completion of the deregulatory project within electricity markets. They celebrate the fact that governments no longer seek to regulate price or supply in oil markets and wholesale electricity and gas markets (for the most part); but they lament vestigial government regulation of competition and retail prices in electricity and natural gas markets, as well as regulation in the markets for energy derivatives,⁵ as an impediment to the innovation and efficiency the unfettered market would bring, if given the chance (e.g., [18, 67, 82, 89]). Proponents of free electricity markets lament the market distortions created by continued regulation of retail natural gas and electricity prices in many states and caps on wholesale natural gas and electricity prices. They oppose regulatory incentives for particular energy investments (such as renewables) on similar grounds and licensing regimes which they see as barriers to entry [16, 69, 89]. These arguments are rooted, often explicitly, in the distrust of government and faith in markets popularly associated with Austrian economics and Adam Smith [18, 89].

Why, then, does the deregulatory project remain incomplete? Why do governments erect unequal barriers to entry for different kinds of energy projects? Why have modern electricity markets stopped short of the free market ideal to date? The answer lies, at least partly, in the realization that the economist’s highly stylized view of human nature is incomplete: it is mostly limited to that which can be deduced from the idealized abstraction that is *homo economicus*. It mostly ignores humans’ social side, *homo politicus*.

2 The Political Economy of Energy Law

While economics seeks allocative efficiency, voters and their agents in Congress care not only about what is efficient but also what is just or fair. Consequently, voters and policymakers sometimes use collective action to seek a more just distribution or to organize collective responses to risk [17, 37, 125, 134, 137].⁶ The distinction between efficiency and fairness, in turn, implicates a set of long-debated issues in welfare economics, political science, and philosophy. This section traces the reasons why the traditional tools of economic analysis have failed to account for considerations that are important to understanding energy regulation.

⁵Energy derivatives are financial contracts through which parties can secure a guarantee to purchase or sell energy or transmission rights at guaranteed future prices. Energy derivatives are regulated by the Commodity Futures Trading Commission.

⁶One cannot explain the broad body of public utility regulation or environmental regulation as mere rent-seeking or capture, at least not very persuasively [134, 137]. It has been the product of mass movements as well ([17, 37, 125], pp. 1635–1675).

Economics aspires to be a positive science, like physics, positing assumptions and deducing conclusions from those assumptions, often using mathematical or formal logic. Ideally, this system of logic yields testable conclusions about the real world, which are then subjected to rigorous empirical tests [28]. Of course, economists' first principles begin with the idea that individuals are rational maximizers of their self-interest (utility maximizers) and will behave in purposeful, sometimes strategic, ways in pursuit of that goal. In this way modern economics has formalized Adam Smith's argument that perfectly competitive markets produce Pareto efficient outcomes [47].⁷ As with physicists' models of the natural world, to many economic theorists, it does not matter that the assumptions on which the theory of perfect competition is built rarely exist in the real world. The theory is useful as a starting point, from which we can begin to understand how real markets work by comparing them to the competitive ideal. In positive science, they say, the value of the theory lies not in the realism of its assumptions but in its ability to illuminate that which is logical or to yield accurate predictions of aggregate behavior [19, 57, 93].

Arrow's theorem and the Coase theorem are examples of theories built in this way, though the former uses much more formal logic than the latter. From the assumption of purposeful rationality, both Arrow and Coase used logical deduction to reach conclusions about political and legal problems, respectively. Arrow's theorem posits Pareto efficiency as one of the several necessary characteristics of any democratic collective choice mechanism ([3], pp. 329–330). Coase concludes that under the conditions he posits, private bargaining is more likely to approach a Pareto-efficient solution to externality problems than government regulation or taxes. He contends that regulation of externality problems is bound to err by sometimes permitting activities with a negative net benefit and prohibiting activities with a positive net benefit and that when property rights are well defined and bargaining costless, those errors can be avoided. That is, under those circumstances, bargaining will result in a Pareto-superior distribution of costs and benefits, compared to regulation [29].

Since those conclusions implicate law and policy, several generations of legal scholars and political scientists have engaged both of these theorems in ways that illustrate that the conclusions of each are dependent upon disputable assumptions [44, 60, 99, 101, 109]. For example, both theorems posit the desirability of Pareto efficiency. But if politics is (axiomatically) a zero-sum game, economists' embrace of the Pareto criterion to evaluate policy or policymaking processes seems to expel political trade-offs from their discipline's domain [25, 117]. Why would welfare economists circumscribe their analyses in this way?

⁷In another emulation of physics, this conclusion is sometimes called the first theorem of welfare economics. This labelling echoes physicists' fundamental laws of thermodynamics. Pareto efficiency refers to distributions that cannot be changed without making at least one person worse off. It is distinguished from Kaldor-Hicks efficiency, which refers to distributions that maximize collective happiness even if they make some worse off, as long as it is possible for the winners to compensate the losers.

There is an historical reason for this. Traditional utilitarianism (of the kind associated with philosopher Jeremy Bentham) embraced the goal of maximizing utility, not simply maximizing the number of happy people [12]. Welfare economics rejected the normative questions at the heart of utilitarian philosophy sometime prior to the mid-twentieth century [126]. It did so by rejecting as “unscientific” interpersonal utility comparisons (and the idea that we can aggregate utility across individuals), based upon the premise that we cannot observe or measure individual utility; rather, we can only measure individual choices, from which we can infer individual preferences [63, 121]. It was this so-called “ordinal revolution” that elevated Pareto efficiency as the dominant goal of welfare economics [26, 63].⁸ It should be obvious, however, that the Pareto criterion is not value-free. To the contrary, it is by definition a rejection not only of redistribution but also of the intuition that disparities in wealth influence the amount of utility different individuals derive from a given quantity of goods or income. It is a “Trojan horse smuggling ethical commitments into the theoretical citadel of positive mainstream economics” ([65], pp. 67–68).

If the only legitimate inferences about welfare are those we can make from individual market decisions, it is little wonder that modern welfare economic analyses favor market solutions over government regulation: by that logic, only through individual voluntary exchange can welfare ever be maximized.⁹ In this way, welfare economics implicitly endorses as *normatively* best a narrow, unrealistic view of social efficiency and does so by (i) concluding that Pareto efficiency is the only *scientifically* justified decision criterion and (ii) employing it as the touchstone of “efficiency” across a wide spectrum of policy problems.

Of course, Pareto efficiency seems a limited and inadequate decision criterion to scholars concerned with the distributional impacts of policies or who recognize the ubiquity of zero-sum decisions in policymaking [54]. Hence, some political economy scholars reject Pareto efficiency as the only defensible criterion by which to judge policy choices. Judge (and legal scholar) Richard Posner concludes that Pareto efficiency is of limited value as a measure of social good because it depends upon “the distribution of wealth—willingness to pay, and hence value, being a function of that distribution” [117]. Legal scholar Michael Dorff echoes Posner when he observes that “there is general agreement that the Pareto principle is largely irrelevant in policymaking because it is almost never true that a change in policy will make everyone better-off” [36]. Economist Amartya Sen devoted a

⁸The term “ordinal revolution” comes from the notion that *cardinal* utility cannot be measured. Rather, we can only measure *ordinal* utility, that is, we can infer preference rankings from choice behavior. It should be noted that there are dissenters from the view that only choice can reveal preferences within the economic profession (e.g., [95]).

⁹Some of the more purist strains of the Austrian economic school have gone further, embracing a version of the Pareto criterion as a normative basis for opposing most government action as tyrannical or illegitimate. Murray Rothbard, an American associated with the Austrian school, advocated a society based on a series of voluntary private exchanges and characterized most government regulation as a form of violent coercion (e.g., [122]).

good portion of his 1998 Nobel address to a plea for welfare economics to move beyond the Pareto criterion and embrace interpersonal utility comparisons in order to make a more meaningful contribution to discussion of important policy problems [128]. And similar concerns dominate seminal works in political science (e.g., [33]). Nonetheless, the goal of Pareto efficiency retains its perch atop welfare economics and is responsible for a kind of disconnect between economic theory and political reality, at times.

For example, consider the problem of monopoly pricing, which loomed so large in the history of public utility law. In neoclassical economics, monopoly pricing is inefficient *not* because it enables the (monopoly) firm to capture more (and consumers fewer) benefits than under pure competition but because it produces a so-called deadweight loss representing potential benefits captured neither by firms nor consumers [72, 151]. However, “it seems certain that Congress never thought in terms of [deadweight loss] when it passed the antitrust laws ([156], pp. 1104–1105), nor did it intend public utility laws to rid the market of deadweight losses, but rather to ensure that prices that were ‘just and reasonable’” for firms and consumers alike [160, 169, 170]. Similarly, American environmental law eschews reliance on Coasean solutions, partly because those solutions are practically unworkable [64] and partly because they frequently offend most voters’ sense of fairness [136]. As Coase acknowledged, his argument assumes away collective action problems: where multiple parties are affected by pollution from a single firm, each injured party has an incentive to “free ride” off of the efforts of other injured parties, which can lead to inefficient results. Coase’s analysis also ignores some important dimensions of fairness, such as the question of whether the polluter came to the injured party or the injured party came to the polluter or the effects of distribution of wealth on willingness to pay and more.

Thus, in these ways economics’ aspirations to positive science tilt the discipline’s conclusions toward disregard of collective notions of fairness and the influence of wealth disparities on utility and toward greater skepticism about regulatory solutions to important distributional problems [36, 86]. Scholars in behavioral economics and behavioral game theory have been working for decades to address this defect in modern welfare economics and have broadened our understanding of human decision-making in the process [25, 145]. However, a large segment of mainstream economics continues to resist those lessons or to deny their usefulness or to pay them no more than lip service by way of oversimplified nods toward “bounded rationality.” Much of this resistance is traceable to the ordinal revolution and the belief that economic models (ought to) “make no assumptions and draw no conclusions about the physiology of the brain” or that theorizing about behavioral departures from rationality is ad hoc ([25], p. 11; [46, 61]). Nevertheless, less doctrinaire social scientists, philosophers, and legal scholars seek to understand how “normal people” or “people with emotions and cognitive limits, . . . behave” and to grapple with the distributional issues at the center of policymaking and law [25]. This explains the greater influence of behavioral research within philosophy, political science, and law than in economics.

This is not surprising. The behavioral revolution challenges economists' assumptions about individual rationality directly, reflecting a skepticism that was probably always present among a minority inside the discipline and a majority outside it [40, 94, 127, 144]. Herbert Simon was skeptical "about substituting *a priori* postulates about rationality for factual knowledge of human behavior" ([131], p. 297), and the subsequent work of Tversky and Kahneman and others within the fields of psychology, sociology, anthropology, and neurobiology has illustrated myriad ways in which human motivation and human action deviates from the assumptions underlying the neoclassical rational choice model [83, 147, 148]. This is apparently true for *homo economicus* but is especially true for *homo politicus* as well [124].¹⁰

While the behavioral literature is far too large to summarize here, a few of its fundamental lessons have particular significance for current debates over electricity markets and their regulation. First, behavioral models emphasize *the importance of emotion* in motivating choice behavior. The sense that a particular outcome (energy prices, for example, or the allocation of pollution risk) is unfair is partly an emotional reaction, and emotion can dominate reason. Second, human choice is influenced by a *concern for the welfare of the group*. Experimental psychologists have repeatedly demonstrated the importance of social forces in explaining (seemingly irrational) behavior, including our impulses to conform to the norms of the group,¹¹ to cooperate,¹² and to treat each other fairly.¹³ In other words, *homo politicus* cares about others and less about her absolute wealth and more about her position relative to others. Third, experimental research supports the conclusion that our brains' emotional circuitry is also built to help us *avoid risk or danger*. Indeed, one of the early heuristics identified by Kahneman and Tversky was our heightened sensitivity to the risk of loss [85], that is, we experience a smaller increase in utility from a gain of \$X than the decrease in utility we experience from

¹⁰In his critique of the Pareto criterion and the application of rational actor models to political questions, philosopher Mark Sagoff put it this way: "[N]ot all of us think of ourselves simply as consumers. Many of us regard ourselves as citizens as well. We act as consumers to get what we want for ourselves. We act as citizens to achieve what we think is right or best for the community" ([124], p. 1286).

¹¹The famous Asch experiment demonstrated that a surprising percentage of subjects would provide an obviously incorrect answer to a simple question once it had become the apparent dominant view within the group [4]. Irving Janis' notion of "groupthink" emphasized this same point, though Janis used ex post analysis of high-profile group decisions rather than experiments [74]. More recently, Dan Kahan's experiments at the Yale cultural cognition project have shown how people's beliefs, and evaluation of empirical evidence, is biased by their need to be consistent with the views of those with whom they share a political ideology and cultural identity [80, 81].

¹²Prisoner's dilemma game experiments demonstrate that cooperative norms can arise within the context of the game, even though the payoff structure suggests that noncooperation is the behavioral equilibrium [7, 55]. Elinor Ostrom's research offers empirical support for the same conclusion [110].

¹³According to Ernst Fehr, "a large body of experimental [laboratory] evidence in economics and psychology . . . indicat[es] that a substantial percentage of people . . . [have social] preferences and that neither concerns for the well-being of others nor for fairness and reciprocity can be ignored in social interactions" (Fehr, 2009).

losing \$ X . We experience more pain, for example, from an unexpected spike in energy prices than the pleasure we derive from a price drop. Loss aversion may help explain voters' willingness to support policies that socialize risk.

None of these findings should surprise students of politics and regulation, nor would they have surprised classical political economists working before the ordinal revolution. Indeed in the words of Daniel Kahneman, “the definition of rationality in . . . [modern] economic theory is so outlandish that it is not a major achievement to find objections to it” [84]. Significantly, these behavioral lessons can explain the persistence of energy regulatory regimes that the economic critique deems “inefficient,” as explained in the next section.

3 Risk, Uncertainty, and Externalities in Electricity Markets

Electricity markets offer an ideal illustration of why the law continues to resist the vision of self-regulating and self-correcting markets that enjoys such strong support in the conservative policy community. How will society manage risk and uncertainty (about energy supply and energy prices) in electricity markets? How will it manage the distribution of external costs and benefits not captured by market prices? These are not only questions of efficiency; they are also political questions to which voters, firms, and interest groups bring their interests and ideologies to bear. Like the market, the political process by which these questions are answered is imperfect, but it seems to reflect at least a generalized collective preference for regulatory interventions in electricity markets [17, 75]. Some of these interventions aim directly at distributional questions; others address voters' and regulators' dissatisfaction with market failures. The following discussion looks specifically at regulation aimed at managing the shortcomings of competitive electricity markets and how that regulation responds to problems unlikely to be addressed satisfactorily by free markets.

3.1 Managing Risk and Uncertainty

In competitive electricity markets, a fundamental problem centers on the role that governments (or other planners) ought to play in helping market participants manage price and supply risk and uncertainty [92].¹⁴ Electricity markets are no exception to the rule that market participants value the risk of losses more highly than the equivalent risk of gains [85]; they also avoid situations characterized by uncertainty, where the risk cannot be estimated with sufficient precision [24, 39, 130].

¹⁴Frank Knight is often credited with first articulating the distinction between “risk,” an uncertain future event to which a quantitative probability can be attached, and “uncertainty,” an uncertain future event for which no probability can be assigned.

3.1.1 The Supply Side

On the production side of the market, public utility law has long focused on the question of whether price signals alone can attract sufficient private capital investment in energy supply to ensure a reliable, reasonably priced supply of energy when it is needed. For certain kinds of highly capital-intensive, long-lived, fixed-asset investments, investors are risk- or loss-averse, doubly so because of the tremendous amount of uncertainty in electricity markets. For the prospective investor in an expensive, 40-year asset, it is next to impossible to estimate the probability that the competitively priced energy produced by the asset will produce a sufficient return over its lifetime (compared to existing or yet-to-be-invented alternatives) or whether the asset will be rendered obsolete or uncompetitive by new regulation. Economists characterize this “asset specificity” problem as a rational reaction to the possibility of strategic behavior by counterparties or to uncertainty about the opportunity cost of investing [91]. However, to most other scholars, investor reticence is better explained in behavioral terms, as a form of risk or loss aversion [103], or an emotional reaction to uncertainty. Indeed, Judge Posner explains the latter phenomenon this way:

One response to uncertainty that is common to most economic actors, whether producers or consumers, is to freeze. The impulse is natural . . . By freezing, one tries to preserve the status quo in the hope that time will bring information, enabling the correct response to be determined . . .

Freezing may be sensible, but it is not a product of calculation. What actuates freezing is fear, specifically fear of the unknown ([118], pp. 1345–1346).

This behavior is consistent not with any expected value calculation but with behavioral experiments on loss aversion [10, 11].

Nuclear power plants, coal-fired power plants, and other large central station technologies trigger this investment dynamic. In traditionally regulated states, state regulators guarantee a fair return on that investment, thereby providing ample incentive to invest. Critics of traditional regulation argue that that guarantee creates unnecessarily high rates for ratepayers, windfalls to shareholders, and unnecessary capital investments [6, 30]. In competitive electricity markets, owners of plants have no such guarantee. They must make investment decisions based upon revenue projections in uncertain competitive markets over the life of the plant. This is problematic because it is difficult to project how much electricity will be needed in the future or whether any particular plant’s electricity will be competitively priced in the future [17, 77, 135]. Nor can plant owners always solve this problem by signing long-term contracts with prospective buyers. In states like Texas, New York, and Pennsylvania, which are characterized by retail competition, retailers are the buyers on wholesale power markets. Because retailers typically sign contracts with their customers for no more than 12 months in duration, it is difficult for retailers to commit to power purchases over decades—the length of time necessary to secure financing for large power plants [35, 52]. Indeed, a recent study by the

American Public Power Association found that almost all new capacity in 2013 was constructed under a long-term contract or ownership and that only 2.4 percent was built for sale into competitive markets [2].

Uncertainty (and the consequent disincentive to invest) is further exacerbated by the way electric power is dispatched on the grid. Because electricity cannot be stored in commercial quantities economically, the grid must be kept in balance—at any given point in time, the amount of electricity being dispatched to the grid by generators must equal the amount being taken off the grid by consumers—in order to avoid outages. When the grid operator dispatches power from individual electric generating facilities to the grid, it does so from the available generating facility that is willing to provide the power at the lowest marginal cost, subject to the caveat that the security of the grid must be maintained. This so-called “security-constrained economic dispatch” (SCED) rule governs most dispatch decisions. This rule protects ratepayers from paying unnecessarily high (unjust and unreasonable) rates and applies both in traditionally regulated systems and in competitive wholesale markets [48]. For buyers, this dispatch rule means that spot market prices face continuous downward price pressure, particularly in an era of inexpensive natural gas and as more zero-marginal cost power from wind and solar generators enters the system, increasing the opportunity cost (or decreasing the option value) of locking into a fixed-price long-term supply contract. For plant owners, this rule means that they cannot always or easily predict when their plants will actually be sending power to the grid.

This additional uncertainty has led overseers and regulators of competitive electricity markets to intervene in those markets in a variety of ways to try to promote reliability of supply.¹⁵ Grid operators in every competitive market employ a variety of mandatory and contractual arrangements to ensure that specified plants are available to provide short-term power to the market in order to balance loads and avoid outages [140]. For example, grid operators may use so-called “reliability must run” or “RMR” contracts with plant owners under which plants are obligated to supply power when called upon to do so. In some organized wholesale power markets, RTOs/ISOs operate capacity markets, which use auctions to pay owners of generating capacity in order to ensure that an adequate amount of generating resources will be available at some future date [78]. The Texas grid operator has eschewed capacity markets in favor of letting wholesale prices float freely as a way of rewarding investment in new capacity [41], but Texas regulators have explored intervening in ancillary services markets to increase payments to providers of short-term reserves [119]. The Texas initiative is essentially a very short-term capacity market [62, 87, 154]. This same sense that wholesale markets are undercompensating providers of reliable electric service is behind a recent FERC initiative requiring RTOs/ISOs to change their settlement procedures in wholesale

¹⁵In the absence of regulatory interventions designed to ensure an adequate supply, pivotal suppliers can acquire and abuse market power in competitive markets [138].

spot markets [51]. Some of these market interventions are intended as attempts to “get prices right” and represent rejections of the unfettered market allocation of costs, benefits, and risk.

Nor do these interventions necessarily address all of the reliability attributes voters and regulators might wish for from a diversified fuel mix, attributes that may not be reflected in the way electricity is priced in spot markets [62, 140]. For example, intermittent sources like wind and solar are less reliable than fossil-fueled plants, because the former can offer power to the grid only when the wind is blowing and the sun is shining, respectively. And coal-fired and nuclear power plants are more reliable than gas-fired plants because they do not depend on real-time (and, therefore, interruptible) supply of fuel from a pipeline. Uniquely among electric generation sources, nuclear power combines very high fuel reliability with zero-emission generation, which may account for the efforts of states in competitive markets to ensure that existing nuclear plants do not exit the market [97, 105, 153]. On the other hand, combustion turbines (usually gas fired) can ramp much more quickly and efficiently than coal-fired or nuclear generators [14, 102] and can be efficient providers of short-term reserves. For all of these reasons, policymakers may intervene to ensure fuel diversity in the electric generation mix in order to ensure reliability of supply [104]. Central planners can plan for a diverse fuel mix, whereas the free market has difficulty pricing these reliability attributes of the generation mix. They simply do not appear through a bottom-up Hayekian process of spontaneous order; rather, they are provided from the top-down, by a combination of grid operator decisions and reliability planning mandates.

3.1.2 The Demand Side

Does economic theory do a better job of predicting demand behavior? If freely floating wholesale and retail energy prices do not always provide a sufficient incentive to invest in supply, might prices be used to influence demand decisions more efficiently? When prices are high in oil markets, we drive less and convert home heating systems from heating oil to gas or electricity. Proponents of freer markets argue that electricity market price caps disrupt this dynamic: if wholesale and retail power prices floated freely in ways that reflected the full cost of delivering electricity to each location on the grid over time, price signals could cure the capacity assurance problem more efficiently than market interventions (such as capacity markets), in part by influencing (reducing) demand. At grid locations where prices are consistently high, not only will new capacity be built; at those same locations, consumers will reduce demand, obviating the need for peaking capacity in the first place [79, 82, 89]. Or, if consumers wish to avoid outages, they will pay more for electricity or find their own alternative sources of supply. If consumers are not willing to pay rates that sustain the amount of generating capacity necessary to prevent outages, we can infer, therefore, that consumers do not really want that higher level of reliability. Instead, they have revealed their true preferences for more frequent outages [89].

This sort of real-time or dynamic retail pricing would elicit from consumers their true willingness to pay to ensure a reliable supply (and avoid outages), in much the same way that Coasean bargaining ought to reveal the parties' true willingness to pay to resolve pollution problems. Dynamic pricing is technically possible in the era of smart meters [50] and commonly used in organized wholesale markets [150]; yet it is largely absent from both competitive and regulated retail markets, where customers pay mostly fixed rates [49].¹⁶ This is inefficient in that it leads the market to undervalue generating capacity, a problem electricity economists call "the missing money problem" [31, 70, 78].

Pilot experiments indicate that consumers respond to dynamic pricing by altering their consumption patterns in response to price signals (saving money in the process) [45, 89]. If dynamic retail pricing is efficient and technically possible, why is it so rare? It may be that for most residential consumers, the stakes (savings of a few dollars per month) may not be worth the bother of responding manually to price signals or of purchasing and programming a device to do so. Or it may be that consumers, like investors, prefer to avoid downside price risk and may be willing to pay a premium to avoid it in the form of higher-but-predictable rates [79]. Moreover, the subjects of dynamic pricing experiments may not be a representative sample of ratepayers: most were not selected randomly, and they may be more responsive to price signals than the average ratepayer. Alternatively, the Hawthorne effect may be at work in some of these experiments, making the results unrepresentative of these same participants' behavior outside the experimental context. Finally, in some of these pilot programs, participants were insulated against downside risk as a condition of their participation in the experimental program, which also can distort results [45]. If consumers really do not want dynamic retail pricing, are they being irrational in forgoing the ultimate savings available from dynamic rates? Perhaps, but this behavior seems perfectly consistent with the loss aversion heuristic in the behavioral literature.

Retailers may yet coax consumers into acceptance of dynamic rates, since retailers face dynamic prices as buyers on wholesale markets. A few retailers, many of them traditionally regulated utilities in competitive wholesale markets, are trying to entice their customers to embrace dynamic pricing by offering risk-free trial periods during which the utility guarantees that the customer's rate will not increase regardless of consumption patterns [9]. After the price ceiling guarantee expires, risk-averse retail customers could conceivably purchase financial hedges, thereby reducing their exposure to price risk. However, financial hedges make more sense for high-volume market participants (like retailers or generators) than for individual residential consumers for whom the stakes are small and the transaction costs relatively high. Alternatively, there has arisen a niche market of demand-side aggregators, who sign up retail customers to contracts in which the customer pledges to reduce demand (or to allow the aggregator to do so) during peak demand periods;

¹⁶According to a 2011 FERC survey, less than one percent of households pay rates that vary according to time of use [49].

the aggregator and the customer then share the resulting savings [149]. Even in the absence of dynamic retail pricing, aggregated demand response (DR) could theoretically bid into wholesale markets just as generators do, offering to provide X MW of DR at specified times, for a price. Indeed, the FERC encourages DR participation in wholesale markets.¹⁷

Another alternative to dynamic pricing are behavioral “nudges,” policies that might reduce demand peaks with fewer transaction costs for consumers [146]. Nudges usually take the form of informational appeals to users to reduce consumption during peak periods, for varied reasons. The appeal can be to assist in the achievement of a policy goal, such as environmental protection or avoiding health-based costs of power generation [5], or to the individual’s sense of peer or community norms [1, 8, 71]. These sorts of appeals aim to activate individuals’ sense of social responsibility or desire to conform to social norms [13]. Companies like Opower manage these sorts of nudge programs for an increasing number of retailers [108].

Economists tend to see behavioral nudges as inferior to dynamic pricing because they induce consumers to bear a cost (forgoing consumption at a convenient time) and provide uncompensated benefits (shaving system peaks) to others; dynamic retail pricing, on the other hand, allows consumers to sell that benefit to the retailer. Thus, dynamic pricing represents a Pareto improvement: each party gains from the trade, or they would not make the trade. Nudges may not represent a Pareto improvement, because consumers forgo benefits of uncertain value. However, one can argue that nudges represent Pareto improvements. The consumer is not compensated monetarily for her inconvenience; but it may be that the consumer derives utility from contributing to the achievement of a social goal or from conforming to social norms. After all, nudges induce behavior; they do not compel it. On the other hand, the nudge may induce disutility by alerting the customer that she is conflicting with social norms; her change in behavior represents a desire to remove that disutility. It is not clear whether that is a welfare-enhancing outcome. In any case, nudges are a form of regulatory intervention in the market, one whose relative success (compared to dynamic retail pricing) seems to be a function of its embrace of the behavioral (rather than the rational actor) model.¹⁸

¹⁷The Supreme Court recently endorsed the FERC’s efforts to encourage DR in *Federal Energy Regulatory Commission v. Electric Power Supply Association*, 136 S. Ct. 760 (2016), which overturned a lower court decision finding DR participation in wholesale markets inconsistent with the Federal Power Act.

¹⁸There is a growing literature in finance on “market microstructure” that examines ways in which financial markets fail to resemble textbook markets in important ways (e.g., [107]). Debates over the wisdom or effects of trying to price ancillary services, for example, are analogous to the microstructure literature in financial markets.

3.2 Managing (Negative and Positive) Externalities in Electricity Markets

Nor has economics' prescription for externalities—namely, to “get prices right” through taxes, subsidies, or assigning property rights to public goods—prevailed in the law. Part of the reason is that getting prices right in this context is very difficult; and for reasons suggested by the behavioral literature, voters may not consider pricing externalities a sufficient solution to the problem. The economics literature on negative (environmental) externalities is rich, well-developed, and tends to favor pollution taxes over command-and-control permitting regimes. It tends to view permitting regimes as unnecessarily costly barriers to entry in electricity markets [115]. Some dedicated Coaseans prefer privately negotiated solutions to externality problems, even over environmental taxes [15, 38], and most conservative scholars agree that permitting regimes impede efficiency (e.g., [98]). Nonetheless, permitting and licensing continue to dominate American environmental regulation, despite decades-long challenges from economic theory and the ideological right. Their abolition seems unlikely primarily because they enjoy public support, support we might infer is rooted in the sense of security that comes from the existence of a regulator preventing firms from shifting too many environmental costs to the rest of us [42, 116].

Economics also struggles with how to “get prices right” in the supply of network infrastructure—oil and gas pipelines and electricity transmission and distribution lines; these networks produce their own kind of missing money problem, one that is also in need of a regulatory fix. This problem is one of the *positive* externalities. Not only does the presence of the network stimulate economic benefits in the form of transactions over the network that otherwise would not have happened; the beneficiaries of individual segments of the network include nonusers of those segments. Absent some system for spreading the costs of the system to those non-customer beneficiaries, prospective investors do not anticipate being fully compensated for the benefits their investment creates, suggesting a role for government in this market.

Economists struggle to fit energy delivery networks neatly into the public or private goods category. Access to the network is excludable (like a private good) but for the common carriage obligation; consumption of space on the network is non-rivalrous (like a public good), but only up to the point of congestion [59, 89, 140]. However, the benefits of a robust network extend beyond paying users, both geographically and temporally. For example, all of the New Englanders who use natural gas to heat their homes (or natural gas-fired electricity) would benefit immediately from investment in additional pipeline capacity into New England, in the form of lower gas and electric prices and fewer gas or electric supply interruptions. Likewise, all electricity users benefit from regional investment in high-voltage transmission lines because the investment brings improved system reliability (fewer outages and less congestion on the system). Furthermore, if transmission is built to serve renewable generation, which is typically located far from load, then the beneficiaries are even farther flung and include those who

would otherwise have been the downwind receptors of pollution from fossil fuel plants displaced by the renewable generation made possible and/or economic by the transmission line.

Traditionally, jurisprudence under the Federal Power Act follows the cost causation principle, which historically has limited cost allocation to customers of the transmission line. This is an artifact of the statutory requirement that rates be just, reasonable, and nondiscriminatory. The problem of incentivizing investment in electricity transmission infrastructure is acute, because the need for new investment is widely perceived to be urgent and because the Federal Power Act (unlike the Natural Gas Act) does not grant the FERC the power to site lines or line owners the power of eminent domain, creating additional (state and local) barriers to entry. Applying the cost causation principle is especially difficult in the electricity context because electricity follows its own path (of least resistance) across the interstate grid, spreading the impacts of transmission investment more widely throughout the network [21]. In the first of three transmission cost allocation opinions involving challenges to FERC orders by the Illinois Commerce Commission (ICC), the Seventh Circuit overturned a FERC order authorizing a transmission tariff that would spread the cost of a new high-voltage transmission line among all of the utilities (and their customers) within the PJM¹⁹ regional wholesale market, on a pro rata basis. The court found the order to be inconsistent with the cost causation principle, because FERC had not met the burden of demonstrating that costs were being allocated in a way that was at least “roughly commensurate” with benefits, though it acknowledged in principle that costs could be spread more widely than the transmission customer base ([163], pp. 473–477).

The FERC has since tried to encourage transmission tariffs allocating costs to non-customer beneficiaries who reap the reliability or clean energy benefits of new transmission lines. The Seventh Circuit has approved a transmission tariff for MISO²⁰ regional wholesale market that spreads costs of new lines broadly across the MISO region [164] but rejected a second attempt by PJM to spread the costs of a new high-voltage line across its region in 2014 [165]. This cost allocation problem prompted a debate between Judge Posner and the late Judge Cudahy over the leeway that market regulators ought to be afforded in managing the market for transmission. Judge Posner’s majority opinions in these cases reflect his belief that it ought to be possible to identify the distribution of the benefits of new transmission among existing ratepayers, and to apportion the costs accordingly. Judge Cudahy disagreed:

However theoretically attractive may be the principle of “beneficiary pays,” an unbending devotion to this rule in every instance can only . . . discourage construction while the nation suffers from inadequate and unreliable transmission. Unsurprisingly, it is not possible to realistically determine for each utility . . . the precise value of not having to cover the costs

¹⁹PJM is an RTO whose territory comprises most of the Middle Atlantic states and parts of the Midwest.

²⁰MISO is an RTO whose territory extends from Minnesota south through the central portion of the country and includes parts of the upper Midwest.

of power failures and of not paying costs associated with congestions, and all this *over the next forty to fifty years.* ([163], p. 479)

Judge Cudahy noted that the positive externalities are partly temporal, making the kind of accounting sought by Posner impossible in his view, in part because many of the beneficiaries of improvements to this network cannot be identified. We can be almost certain that sometime in the next four or five decades people on the network (who do not directly use the new line) will benefit from its presence in predictable ways, but we cannot come close to identifying who those people are right now. To Judge Cudahy, the problem was one of “incommensurable forces and conditions” and therefore required deference to agency discretion ([165], pp. 565–566) [166].

When network infrastructure produces positive externalities not easily compensated by the market, there is a role for regulators to assist in spreading the costs more broadly to ensure the provision of sufficient supply. In the words of Brett Frischmann, “[t]he societal need for nondiscriminatory community access to infrastructure and the generation of substantial spillovers each appears to independently constitute grounds for identifying a potential market failure and for supporting some role for government” ([59], p. 6). Positive externalities lead markets to undersupply network infrastructure, as do ill-designed rules (like the beneficiary pays rule) that attempt to mimic that same flawed market. Furthermore, for delivery networks, the cost causation rule poses an equity problem if access to the network is essential in order to participate in economic life. If we build networks only to locations where the users have the ability to pay, the rich will have access to the network, and the poor will not.²¹ In the post-ordinal revolution framework of neoclassical economics, that fact does not necessarily imply a problem, because we cannot assume that those who are unable to pay would derive as much utility from access to the network as those who are able to pay; to many others, however, the problems associated with relying on willingness to pay measures as the best measure of utility in that instance are obvious.

This is the same problem that provoked a government solution in the form of the Rural Electrification Act in the 1930s [27, 123]. Similarly, in the 1950s, neither President Eisenhower nor the Congress justified government funding of interstate highways by identifying and taxing only those people likely to use each segment of the interstate highway system. Nor could they have done so, which may be partly why American taxpayers shared that burden. Interestingly, electric transmission lines are being approved and built in Texas with relative speed and ease, where much of the grid lies beyond the jurisdiction of the Federal Power Act’s cost causation rule [53, 143].²² This may be because the state has chosen to emulate the financing of the

²¹The cost causation principle produces a level of network investment that maximizes net benefits only if one subscribes to the view that willingness to pay is the best available measure of utility and that we cannot make inferences about the relative amounts of utility different individuals derive from a good or service.

²²The Electric Reliability Council of Texas (“ERCOT”) is an RTO that manages a grid that is functionally separate from the remainder of the American power grid and comprises most of the

federal highway system by spreading the cost of the new lines to all ratepayers.²³ In other words, these governments have seen fit to address market failure in the market for network investment, and they do so by spreading the costs more widely than rigid adherence to a (simulated) willingness-to-pay regime would.

4 The New (Old) Political Economy of Regulation

There is a contradiction at the heart of capitalist democracy, one that government regulation attempts to manage. We want an economy that incentivizes innovation and offers the social benefits of efficiency *and* a polity that protects us from the various harms associated with market failure. In electricity markets those harms include sudden price spikes, harmful pollution, and the undersupply of energy infrastructure. Americans seem willing to support policies that reduce our exposure to these harms and to ensure that energy prices and competition in electricity markets are “fair.” Since its inception more than a century ago, modern American energy law—public utility law and environmental law—has sought to reconcile these conflicting impulses. Certainly, regulation sometimes produces distributions that economists suspect are suboptimal. When voters and policymakers choose these policies anyway, it is tempting to ascribe to them a misunderstanding of markets or of what is best for society. But it may be that voters and policymakers believe they are choosing between two imperfect systems and reject the pure forms of both; it may very well be that regulation is an informed choice.

American electricity markets are shaped by bottom-up innovation that responds to market incentives and by top-down regulation that aims to minimize the dangers of market failure. It seems extremely unlikely that American energy policy will veer sharply toward central planning or toward eliminating regulation of electricity markets altogether and for good reason. Because the Pareto criterion is both practically and politically an unrealistic goal and because we often fail to behave like *homo economicus*, regulators intervene in electricity markets to incentivize investment and to manage the distribution of the externalities of energy production. Economic models of politics may conceive of these interventions as rent-seeking likely to distort markets, but this explanation is convenient and unpersuasive,

grid within the State of Texas. The lack of an interstate connection means that the Federal Power Act requirement that transmission rates be just, reasonable, and nondiscriminatory does not apply to the ERCOT grid.

²³The state offered financial incentives for investment in renewable power within the CREZ zones and decided to “socialize” the costs of building transmission from the CREZ zones eastward to San Antonio, Houston, and the remainder of central and east Texas. The presence of this new transmission, in turn, has sparked the development of more generation in Texas than any other state.

because it is the product of the a priori assumptions economists employ [158].²⁴ Rather, regulatory interventions are better explained as the product of Americans' revealed preferences for *some* regulation of electricity markets.

Ironically, while Smith and Hayek condemned governments' failure to understand the motives of market participants and the sometimes harmful consequences of regulation, neither man sought to vindicate the kind of elegant, mathematical expression of human behavior found in modern economic theory. Rather, Smith and Hayek each wrote in response to the specific, problematic forms of government interference in the economy they observed during their lifetimes. Smith wrote at a time when guilds controlled access to most professions under the guise of protecting the public; Hayek wrote in the shadow of Nazi and Soviet totalitarianism. Their writings should be understood in those contexts. It is a sizeable leap from their criticism of the misguided regulation they witnessed to the kind of idealized free electricity markets being advocated by some conservatives today, markets which Judge Cudahy long ago accurately described as "folklore" [32]. To the contrary, one could argue that Smith and Hayek would endorse the kind of electricity markets we see now: markets into which regulators have introduced competition and market pricing cautiously and iteratively, coupled with regulatory experimentation to ensure an adequate supply of infrastructure and to internalize the externalities of energy production [68].

Both Smith and Hayek recognized a role for government in addressing public goods and externality problems and in incentivizing investment where markets fail to supply enough of any good that society needs. Here is Hayek (quoting Smith) on the importance of "intelligently designed and continuously adjusted" legal institutions in an efficient market:

To create conditions in which competition will be as effective as possible, to supplement it where it cannot be made effective, to provide the services which, in the words of Adam Smith, "though they may be in the highest degree advantageous to a great society, are, however, of such a nature, that the profit could never repay the expense to any individual or small number of individuals", these tasks provide indeed a *wide and unquestioned field for state activity*. In no system that could be rationally defended would the state just do nothing. An effective competitive system needs an intelligently designed and continuously adjusted legal framework as much as any other. Even the most essential prerequisite of its proper functioning, the prevention of fraud and deception (including exploitation of ignorance) provides a great and by no means yet fully accomplished object of legislative activity. [66]

Hayek also endorsed health and safety regulation and regulation that mandates the provision of information that "can never be adequately provided by private enterprise" [66]. For his part, Adam Smith envisioned for government "the duty of protecting ... every member of the society from the injustice or oppression of

²⁴Nobel laureate Oliver Williamson chastised this kind of slavish adherence to economic theory in the wake of the California electricity crisis, arguing that designers of the California market applied theory "naively" without regard to "the realities of the political and regulatory process" ([158], p. 384).

every other member . . . [and] of erecting and maintaining certain public works and certain public institutions” which the market will not provide [133].

Nor would Smith or Hayek be comfortable with the mathematical version of modern neoclassical economics that was cause and consequence of the ordinal revolution: Smith because he would reject its narrow view of *homo economicus* and Hayek because he was skeptical of the ability of mathematical economists to capture the dynamics at work inside markets [90].²⁵ According to economist Alan Krueger, “Smith was a Rawlsian before . . . Rawls,” implying that Smith cared so much about distributional justice that he would have rejected Pareto optimality as a goal [96, 120]. Rather, Adam Smith’s [132] was the behavioral view of human nature, one that embraced social preferences: “How[ever] selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it” [132]. Unfortunately, too many modern economists have jettisoned classical economists’ broader understanding of human nature and of the interdependence of politics and markets in their attempts to make the discipline more scientific (and hence more rigorously logical and mathematical). Hayek’s contemporary and rival John Maynard Keynes wrote that “the master-economist” is not only a logician or mathematician but also a “historian, statesman, [and] philosopher” [88, 152]. Hayek disagreed with Keynes on most things, but not on this point. In an address at the London School of Economics, Hayek lamented narrow specialization within economics, noting that “if you know economics and nothing else, you will be a bane to mankind, good, perhaps, for writing articles for other economists to read, but for nothing else” ([139], quoting Hayek).

Some scholars trace the ancestry of the ordinal revolution back to Smith’s contemporary David Hume and his admonition that an “ought” cannot be derived from an “is” [63]. But Hume did not believe human nature was fully captured by *homo economicus* any more than Smith did; nor would Hume endorse the modern public choice view of the policy process. When Hume famously described reason as a “slave to” passion, he was making a descriptive statement about human nature that echoes modern behavioralists [73], one central to his (and James Madison’s) theory of government [100]. The American Founders were students of mathematical theories of collective choice that predated Arrow’s theorem, but the Madisonian theory of government was (and is) about more than mere preference aggregation. Rather, it is about structuring the *delegation* of decision authority by voters to a *deliberative* government. Despite their clear-eyed view of human ambition and selfishness, the Founders aimed to create a decision process that minimizes rent-seeking and favors deliberation and that pushes policy toward “the permanent . . .

²⁵In his 2014 best-selling book on wealth and income inequality, Thomas Piketty made a similar point: To put it bluntly, the discipline of economics has yet to get over its childish passion for mathematics and for purely theoretical and often highly ideological speculation, at the expense of historical research and collaboration with the other social sciences. . . . This obsession with mathematics is an easy way of acquiring the appearance of scientificity without having to answer the far more complex questions posed by the world we live in ([112], p. 32).

interests of the community” [100]. In that sense, Madison’s goal for government resembled that of his contemporary Edmund Burke: government should decide as the people would decide if the people could devote the resources and time necessary to understand the problem [22, 43].

The problem we face in today’s polarized American polity is that the meaning of the permanent interests of the community is particularly hotly contested. But that does not negate the worthiness of pursuing that goal. In American energy policy that contest seems to be between two visions of the good: a top-down vision of ever-greener electricity markets, on the one hand, and bottom-up vision of ever-freer electricity markets, on the other. Both visions can be naive, at times. Proponents of both visions lament the lack of “an energy policy” in line with their vision and the fact that American energy policy falls somewhere in between.

However, American energy policy is forever destined to lie in between, because it appears that that is what the well-informed median voter wants. Voters want a reliable, affordable, and clean energy supply. Energy and environmental regulators, working within constraints imposed by statutes and courts, have proven quite adept at the kind of cautious experimentation by which the permanent interests of the community can be identified and realized. Defying the caricature of the power-hungry central planner, American regulators have long balanced the benefits of markets against their dangers in ways that reflect the goal of serving the well-informed median voter [161, 168]. This has been particularly true in the modern era of congressional gridlock [56].

By contrast, it has been elected legislators, and sometimes even courts [167], who have been much more prone to clumsy interventions in markets. State *legislators* have tried to “correct” energy prices they perceive to be discriminatory against their citizens, from the earliest days of public utility regulation to the present day.²⁶ While today’s legislators must curry votes by paying verbal lip service to one or the other ideal visions of our energy future, public utility commissions and environmental agencies are free to do the hard work of reconciling markets with community needs in an industry that produces what is often described as “the lifeblood of the economy.”²⁷ Thus, in solidly Republican Texas, policymakers pursue a vision of free electricity markets but are willing to compromise that vision in order to ensure the security of energy supply or to promote wind development. And in solidly Democratic California, policymakers pursue a vision of green electricity markets but are willing to compromise that vision in order to ensure that prices do not get too high [23]. Despite a policy debate fought using the language of ideological archetypes, regulation is a collective project involving continual interaction between

²⁶In the early 2000s, New Jersey and Maryland grew dissatisfied with wholesale electricity prices in eastern PJM. Policymakers in both states concluded that the PJM capacity market was not inducing sufficient investment in new generation facilities in eastern PJM and undertook to subsidize construction of new natural-gas fired generation within their state borders. Reasoning that these subsidies would distort prices in the PJM market, the Supreme Court struck them down in *Hughes v. Talen Energy Mktg., LLC*, 136 S. Ct. 1288 (2016).

²⁷A Google search of this phrase reveals more than 400,000 results (last searched Nov. 8, 2016).

policy and markets [17]. As human beings, we participate in this project in two ways: *homo politicus* participates in the policy process in order to place limits on *homo economicus* in the market. We bring different concerns and motives to each role, and it is little wonder that the best tools we have to analyze markets provide such an incomplete picture of the policy process. We recognize the virtues of the market, but we do not entirely trust it to maximize social net benefit, and so we retain the option to regulate [58]. In this way American electricity policy is an ongoing, contested effort to define which costs and benefits will be allocated by the market and which will be allocated by law and policy.

Acknowledgments The author would like to acknowledge valuable comments received on earlier drafts of this article from William Boyd, Emily Hammond, Jody Freeman, Sean Meyn, Amy Stein, and Alex Klass.

References

1. Alcott H (2011) Social norms and energy conservation. *J Pub Econ* 95:1082–1059
2. Am. Pub. Power Ass'n (2014) Power plants are not built on spec: 2014 update. http://www.publicpower.org/files/PDFs/94_2014_Power_Plant_Study.pdf. Accessed 20 Jan 2017
3. Arrow K (1950) A difficulty in the concept of social welfare. *J Political Econ* 58:328–346
4. Asche SE (1951) Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow H (ed) Groups, leadership and men: research in human relations. Carnegie Press, Pittsburgh
5. Asensio OI, Delmas MA (2015) Nonprice incentives and energy conservation. *Proc Natl Acad Sci* 112:510–515
6. Averch H, Johnson LL (1962) Behavior of the firm under regulatory constraint. *Am Econ Rev* 52:1052–1069
7. Axelrod R (1984) The evolution of cooperation. Basic Books, New York
8. Ayres I et al (2012) Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *J Law Econ Org* 29:992–1022
9. Badtke-Berkow M (2015) A primer on time-variant electricity pricing. Environmental Defense Fund, New York
10. Bechara A et al (1997) Deciding advantageously before knowing the advantageous strategy. *Science* 275:1293–1295
11. Bechara A et al (2003) Role of the amygdala in decision-making. *Ann N Y Acad Sci* 985:356–369
12. Bentham J (1967) A fragment of government and an introduction to the principles of morals and legislation. Basil Blackwell, Oxford
13. Berkowitz L, Daniels LR (1964) Affecting the salience of the social responsibility norm: effects of paste help on the response to dependency relationships. *J Abnorm Soc Psychol* 68:275–281
14. Black & Veatch (2012) Cost and performance data for power generation technologies. <http://bv.com/docs/reports-studies/nrel-cost-report.pdf>. Accessed 5 Feb 2017
15. Black G et al (2014) The Coasean framework of the New York City watershed agreement. *Cato J* 34:1–32
16. Borenstein S, Bushnell J (2015) The U.S. electricity industry after 20 years of restructuring. Energy Institute at Haas. <https://ei.haas.berkeley.edu/research/papers/WP252.pdf>. Accessed 16 May 2018.
17. Boyd W (2014) Public utility and the low-carbon future. *UCLA Law Rev* 61:1614–1710

18. Bradley R Jr (2009) Capitalism at work: business, government and energy. M&M Scrivener Press, Salem, MA
19. Brams SJ (1976) Paradoxes in politics: an introduction to the nonobvious in political science. Free Press, New York
20. Breyer S (1979) Analyzing regulatory failure: mismatches, less restrictive alternatives, and reform. Harv Law Rev 92:547–609
21. Brown MH, Sedano RP (2004) National Council of Energy Policy. Electricity transmission: a primer. <http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/primer.pdf>. Accessed 20 Jan 2016
22. Burke E (1790) Reflections on the revolution in France. Oxford University Press, Oxford
23. Cal. St. Senate (2016) California climate leadership: powering the new economy. <http://focus.senate.ca.gov/climate>. Accessed 9 Nov 2016
24. Camerer CF (2000) Prospect theory in the wild: evidence from the field. In: Kahneman D, Tversky A (eds) Choice, values, and frames. Cambridge University Press, New York
25. Camerer CF (2003) Behavioral game theory. Princeton University Press, Princeton
26. Caplin A, Glimcher PW (2014) Basic methods from neoclassical economics. In: Glimcher PW, Fehr E (eds) Neuroeconomics: Decisions making and the brain. Academic Press, San Diego
27. Caro RA (1982) The years of Lyndon Johnson: the path to power. Random House, New York
28. Clarke KA, Primo DM (2012) Overcoming ‘physics envy’. New York Times. http://www.nytimes.com/2012/04/01/opinion/sunday/the-social-sciences-physics-envy.html?_r=0. Accessed 21 Jan 2017
29. Coase RH (1960) The problem of social cost. J Law Econ 56:837–877
30. Courville L (1974) Regulation and efficiency in the electric utility industry. Bell J Econ Manag Sci 5:53–74
31. Cramton and Stoof (2006) The convergence of market designs for adequate generating capacity with special attention to the CA ISO’s resource adequacy problem, White Paper <http://www.cramton.umd.edu/papers2005-2009/cramton-stoof-market-design-for-resource-adequacy.pdf>. Accessed 4 Feb 2017
32. Cudahy RD (1998) The folklore of deregulation (with apologies to Thurman Arnold). Yale J Reg 15:427–442
33. Dahl RA (1956) A preface to democratic theory. University of Chicago Press, Chicago
34. Demsetz H (1968) Why regulate utilities? J Law Econ 11:55–65
35. Direct Energy (2016) <https://www.directenergy.com/ny/electricity-plans>. Accessed 31 Oct 2016
36. Dorff MB (2002) Why welfare depends on fairness: a reply to Kaplow and Shavell. South Calif Law Rev 75:847–900
37. Downs A (1972) Up and down with ecology-the “issue-attention cycle”. Public Interest 28:38–50
38. Ellickson RC (1989) The case for Coase and against “Coaseanism”. Yale Law J 99:611–630
39. Ellsberg D (1961) Risk, ambiguity, and the savage axioms. Q J Econ 75:643–669
40. Elster J (1983) Sour grapes: studies in the subversion of rationality. Cambridge University Press, New York
41. ERCOT (2013) System-wide offer cap and scarcity pricing mechanism methodology. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiPwqSn44rbAhUBP6wKHf_8DpUQFgnMAA&url=http%3A%2F%2Fwww.ercot.com%2Fcontent%2Fmktrules%2Fissues%2Fnpr%2F451-475%2F468%2Fkeydocs%2F468NPRR-03 _System-Wide_Offer_Cap_and_Scarcity_Pricing_Mechan.doc&usg=AOvVaw0Hk9FbwCuQ_QtNxT84cMKc. Accessed 16 May 2018
42. Farber DA (1992) Politics and procedure in environmental law. J Law Econ Org 8:59–81
43. Farber DA, Frickey PP (1991) Law and public choice: a critical introduction. University of Chicago Press, Chicago
44. Farina CR (2000) Faith, hope, and rationality or public choice and the perils of Occam’s razor. Fla State Univ Law Rev 28:109–136

45. Faruqui A, Sergici S (2010) Household response to dynamic pricing of electricity: a survey of 15 experiments. *J Regul Econ* 38:193–225
46. Fehr E, Krajich I (2014) Social preferences and the brain. In: Glimcher PW, Fehr E (eds) *Neuroeconomics: decision making and the brain*. Academic Press, San Diego
47. Feldman AM (2008) Welfare economics. In: Durlauf SN, Blume LE (eds) *The new Palgrave dictionary of economics*. Palgrave Macmillan, London
48. FERC (2006) Security constrained economic dispatch: definition, practices, issues, and recommendations. <http://www.ferc.gov/industries/electric/indus-act/joint-boards/final-cong-rpt.pdf>. Accessed 20 Jan 2017
49. FERC (2011) Assessment of demand response and advanced metering, staff report 27. <http://www.ferc.gov/legal/staff-reports/2010-dr-report.pdf>. Accessed 20 Jan 2017
50. FERC (2013) Assessment of demand response and advanced metering. <http://www.ferc.gov/legal/staff-reports/2013/oct-demand-response.pdf>. Accessed 20 Jan 2017
51. FERC (2016) Settlement intervals and shortage pricing in markets operated by regional transmission organizations and independent system operator (Order 25). <https://www.ferc.gov/whats-new/comm-meet/2016/061616/E-2.pdf>. Accessed 16 May 2018
52. First Choice Power (2016) <https://www.firstchoicepower.com/texas/electricity-plans>. Accessed 31 Oct 2016
53. Fleisher JM (2008) ERCOT's jurisdictional status: a legal history and contemporary appraisal. *Tex J Oil Gas Energy Law* 3:4–21
54. Fleurbaey M (2012) Economics and economic justice. In: Stanford Encyclopedia of Philosophy. <http://plato.stanford.edu/archives/win2014/entries/economic-justice/>. Accessed 5 Feb 2017
55. Frank RH (1993) Does studying economics inhibit cooperation? *J Econ Perspect* 7:159–171
56. Freeman J, Spence DB (2014) Old statutes, new problems. *Univ Pa Law Rev* 163:1–94
57. Friedman M (1953) The methodology of positive economics. In: *Essays in positive economics*. University of Chicago Press, Chicago
58. Friedman D (2008) *Morals and markets: an evolutionary account of the modern world*. Palgrave Macmillan, New York
59. Frischmann BM (2012) *Infrastructure: the social value of shared resources*. Oxford University Press, New York
60. Green DP, Shapiro I (1994) *Pathologies of rational choice theory*. Yale University Press, New Haven
61. Gul F, Pesendorfer W (2005) The case for mindless economics. Princeton University Working Paper
62. Hammond E, Spence DB (2016) The regulatory contract in the marketplace. *Vanderbilt Law Rev* 69:141–216
63. Hands DW (2012) The positive-normative dichotomy and economics. *Handbook of the Philosophy of Science* 13:219–239
64. Hardin R (1982) *Collective action*. The Johns Hopkins University Press, Baltimore
65. Hausman DM, McPherson MS (2006) *Economic analysis, moral philosophy, and public policy*. Cambridge University Press, New York
66. Hayek FA (1944) *The road to serfdom*. University of Chicago Press, Chicago
67. Hayek FA (1948) *Individualism and economic order*. University of Chicago Press, Chicago
68. Hayek FA (1960) *The constitution of liberty*. The University of Chicago Press, Chicago
69. Hogan WW (2001) Statement before the committee on governmental affairs. United States Senate, Washington DC
70. Hogan WW (2005) On an “energy only” electricity market design for resource adequacy. Center for Business and Government John F. Kennedy School of Government Harvard University. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.4422&rep=rep1&type=pdf>. Accessed 5 Feb 2017
71. Houde S et al (2013) Real-time feedback and electricity consumption: a field experiment assessing the potential for savings and persistence. *Energy J* 34:87–102

72. Hovenkamp H (1994) Federal antitrust policy: the law of competition and its practice. West Publishing, St. Paul
73. Hume D (1739) A treatise of human nature. Clarendon Press, Oxford
74. Janis IL (1972) Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes. Houghton Mifflin, Boston
75. Joskow PL (1974) Inflation and environmental concern: structural change in the process of public utility price regulation. *J Law Econ* 17:291–327
76. Joskow PL (1989) Regulatory failure, regulatory reform, and structural change in the electrical power industry. *Brook Pap Econ Act Microecon* 1989:125–208
77. Joskow PL (1991) Asset specificity and the structure of vertical relationships: empirical evidence. In: Williamson OE, Winter SG (eds) *The nature of the firm: origins, evolution, and development*. Oxford University Press, New York
78. Joskow PL (2008) Capacity payments in imperfect electricity marked: need and design. *Util Policy* 16:159–170
79. Joskow PL, Wolfram CD (2012) Dynamic pricing of electricity. *Am Econ Rev* 102:381–385
80. Kahan DM, Braman D (2006) Cultural cognition and public policy. *Yale Law Policy Rev* 24:147–172
81. Kahan DM et al (2007) The second national risk and culture study: making sense of – and making progress in – the American culture ware of fact. The Cultural Cognition Project at Yale Law School. Available via http://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=1271&context=faculty_publications. Accessed 5 Feb 2017
82. Kahn A (1970) *The economics of regulation: principles and institutions*. MIT Press, Cambridge
83. Kahneman D (2011) *Thinking, fast and slow*. Farrar, Straus and Giroux, New York
84. Kahneman D (2013) Behavioral economics and investor protection: keynote address. *Loyola Univ Chic Law J* 44:1333–1340
85. Kahneman D, Tversky A (1984) Choices, values, and frames. *Am Psychol* 39:341–350
86. Kaplow L, Shavell S (2001) Fairness versus welfare. *Harv Law Rev* 114:961–1389
87. Kempton W, Tomic J (2005) Vehicle-to-grid power fundamentals: calculating capacity and net revenue. *J Power Sources* 144:268–279
88. Keynes JM (1924) Alfred Marshall, 1842–1924. *Econ J* 34:311–372
89. Kiesling L (2008) *Deregulation, innovation, and market liberalization: electricity restructuring in a constantly evolving environment*. Routledge, New York
90. Kilpatrick HE Jr (2001) Complexity, spontaneous order, and Friedrich Hayek: are spontaneous order and complexity essentially the same thing? *Complexity* 6:16–20
91. Klein B et al (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J Law Econ* 21:297–326
92. Knight FH (1921) *Risk, uncertainty and profit*. Signalman Publishing, Kissimmee
93. Korobkin R (2013) Daniel Kahneman's influence on legal theory. *Loyola Univ Chic Law J* 44:1349–1356
94. Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. *Calif Law Rev* 88:1051–1144
95. Kőszegi B, Rabin M (2008) Choices, situations, and happiness. *J Public Econ* 92:1821–1832
96. Kreuger AB (2001) Economic scene; the many faces of Adam Smith: rediscover 'The Wealth of Nations'. *New York Times*. <http://www.nytimes.com/2001/08/16/business/economic-scene-the-many-faces-of-adam-smith-rediscovering-the-wealth-of-nations.html>. Accessed 5 Feb 2017
97. Larson A (2016) Exelon gets its Christmas wish – Illinois legislation will save nuclear plants. Power. <http://www.powermag.com/exelon-gets-its-christmas-wish-illinois-legislation-will-save-nuclear-plants>. Accessed 2 Dec 2016
98. Loris ND (2014) Free markets supply affordable energy and a clean environment. Heritage Foundation Backgrounder No. 2966. <http://www.heritage.org/research/reports/2014/10/free-markets-supply-affordable-energy-and-a-clean-environment>. Accessed 5 Feb 2017

99. Lowi TJ (1979) *The end of liberalism: the second republic of the United States*. W. W. Norton & Company, New York
100. Madison J (1787) *The Federalist No. 10*. In: Shapiro I (ed) *The Federalist papers (rethinking the western tradition)*. Yale University Press, New Haven
101. Medema SG, Zerbe RO Jr (1999) The Coase Theorem. In: Encyclopedia of Law and Economics. <http://encyclo.findlaw.com/0730book.pdf>. Accessed 5 Feb 2017
102. MIT Energy Initiative (2011) Managing large-scale penetration of intermittent renewables. <http://mitei.mit.edu/system/files/intermittent-renewables-full.pdf>. Accessed 5 Feb 2017
103. Montier J (2007) *Behavioral investing: a practitioner's guide to applying behavioral finance*. John Wiley & Sons, West Sussex
104. Morrison J (2016) Capacity markets: a path back to resource adequacy. *Energy Law J* 37:1–60
105. Nearing B (2016) Multi-billion dollar state nuclear payout draws environmental challengers, defenders. Albany Times-Union. <http://www.timesunion.com/tuplus-business/article/Multi-billion-dollar-state-nuclear-payout-draws-10688616.php>. Accessed 3 Dec 2016
106. Niskanen WA (1971) *Bureaucracy and representative government*. Aldine-Atherton, Chicago
107. O'Hara M (1995) *Market microstructure theory*. Blackwell, Oxford
108. Opower (2016) Company overview. <http://vignet.opower.com/company>. Accessed 2 Nov 2016
109. Ordeshook PC (1986) *Game theory and political theory*. Press Syndicate of the University of Cambridge, New York
110. Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, New York
111. Peltzman (1976) Toward a more general theory of regulation. *J Law Econ* 19:211–240
112. Picketty (2014) *Capital in the twenty-first century*. Harvard University Press, Cambridge
113. Pierce R Jr (1986) A proposal to deregulate the market for bulk power. *Va Law Rev* 72:1183–1235
114. Pierce R Jr (1988) Reconstituting the natural gas industry from wellhead to burnertip. *Energy Law J* 9:1–58
115. Pierce R Jr (2005) Environmental regulation, energy, and market entry. *Duke Environ Law Policy Forum* 15:167–186
116. Pope JG (1990) Republican moments: the role of direct popular power in the American constitutional order. *Univ Pa Law Rev* 139:287–454
117. Posner RA (1998) *Economic analysis of law*. Wolters Kluwer, New York
118. Posner RA (2013) Behavioral finance before Kahneman. *Loyola Univ Chic Law J* 44:1341–1348
119. Pub. Util. Comm'n of Texas, 40000: Commission proceeding to ensure resource adequacy in Texas. <http://www.puc.texas.gov/industry/projects/electric/40000/40000.aspx>. Accessed 5 Feb 2017
120. Rawls J (1971) *A theory of justice*. Belknap Press, Cambridge
121. Robbins L (1932) *An essay on the nature and significance of economic science*. Macmillan & Co., London
122. Rothbard MN (1962) *Man, economy, and state: a treatise on economic principles*. The Ludwig von Mises Institute, Auburn
123. Rural Electrification Act of 1936, Pub. L. No. 74-605, 49 Stat. 1363 (1936)
124. Sagoff M (1981) At the shrine of Our Lady of Fatima or why political questions are not all economic. *Ariz Law Rev* 23:1283–1298
125. Schroeder CH (1998) Ration choice versus republican moment – explanations for environmental laws, 1969–1973. *Duke Environ Law Policy Forum* 9:19–53
126. Sen A (1970) The impossibility of a Paretian liberal. *J Polit Econ* 78:152–157
127. Sen AK (1977) Rational fools: a critique of the behavioral foundations of economic theory. *Philos Public Aff* 6:317–344
128. Sen A (1999) Nobel lecture: the possibility of social choice. *Am Econ Rev* 89:349–378
129. Shapiro SA, Toman JP (2003) *Regulatory law and policy: cases and materials*. LexisNexis, New York

130. Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. *J Finance* 19:425–442
131. Simon HA (1984) Human nature in politics: the dialogue of psychology with political science. *Am Polit Sci Rev* 79:293–304
132. Smith A (1759) Theory of moral sentiments. A. Millar, London
133. Smith A (1789) An inquiry into the nature and causes of the wealth of nations. Methuen & Co, London
134. Spence DB (2001) A public choice progressivism continued. *Cornell Law Rev* 87:397. 419–443
135. Spence DB (2008) Can law manage competitive electricity markets? *Cornell Law Rev* 93:765–818
136. Spence DB (2014) The political economy of local vetoes. *Tex Law Rev* 93:395–413
137. Spence DB, Cross F (2000) A public choice case for the administrative state. *Georgetown Law J* 89:97–142
138. Spence DB, Prentice R (2012) The transformation of American energy markets and the problem of market power. *Boston Coll Law Rev* 53:131–202
139. Steele GR, Hayek F (2008) The complete economist. *Econ Aff* 28:67–69
140. Stein AL (2015) Regulating reliability. *Houst Law Rev* 54 (forthcoming issue 5)
141. Stigler GJ (1971) The theory of economic regulation. *Bell J Econ Manag Sci* 2:3–21
142. Stigler GJ, Friedland C (1962) What can regulators regulate? The case of electricity. *J Law Econ* 5:1–16
143. Sustainablebusiness.com News (2013) Texas to double wind capacity, deliver to major cities. <http://www.sustainablebusiness.com/index.cfm/go/news.display/id/24725>. Accessed 20 Jan 2017
144. Taylor M (1995) Battering RAMs. *Crit Rev* 9:223–234
145. Thaler R (1980) Toward a positive theory of consumer choice. *J Econ Behav Org* 1:39–60
146. Thaler RH, Sunstein CR (2008) Nudge: improving decisions about wealth, health, and happiness. Penguin Group, New York
147. Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. *Science* 185:1124–1131
148. Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211:453–458
149. Tweed K (2010) The top 5 players in demand response. <http://www.greentechmedia.com/articles/read/top-5-demand-response>. Accessed 20 Jan 2017
150. U.S. Energy Info Admin (2011) Today in energy: Wholesale power price maps reflect real-time constraints on transmission of electricity. <http://www.eia.gov/todayinenergy/detail.cfm?id=3150>. Accessed 20 Jan 2016
151. Varian HA (1990) Intermediate microeconomics: a modern approach. W. W. Norton & Company, New York
152. von Neumann J (1947) The mathematician. In: Heywood RB (ed) *The works of the mind*. University of Chicago Press, Chicago
153. Walton R (2015) NY Gov. Cuomo: state will fight Entergy plans to shutter Fitzpatrick Nuclear Plant. <http://www.utilitydive.com/news/ny-gov-cuomo-state-will-fight-energy-plans-to-shutter-fitzpatrick-nuclea/408667/>. Accessed 20 Jan 2017
154. Walton R (2016) Texas grid operator continues debate over pricey reliability agreement. <http://www.utilitydive.com/news/texas-grid-operator-continues-debate-over-pricey-reliability-agreement/432545/>. Accessed 20 Jan 2017
155. Wasik JF (2006) The merchant of power: Sam Insull, Thomas Edison and the creator of the modern metropolis. St. Martin's Press, New York
156. Weiss LW (1979) The structure-conduct-performance paradigm and antitrust. *Univ Pa Law Rev* 127:1104–1140

157. Williamson OE (1976) Franchise bidding for natural monopolies – in general and with respect to CATV. *Bell Econ* 7:73–104
158. Williamson OE (2005) Why law, economics, and organization? *Ann Rev Law Soc Sci* 1:369–396

Cases

159. *Federal Energy Regulatory Commission v. Electric Power Supply Association*, 136 S. Ct. 760 (2016)
160. *FPC v. Hope Nat Gas Co*, 320 US. 591 (1944)
161. *FPC v. Sierra Pac Power Co*, 350 US. 348 (1956)
162. *Hughes v. Talen Energy Mktg., LLC*, 136 S Ct 1288 (2016)
163. *Ill. Commerce Comm'n v. FERC (ICC I)*, 576 F.3d 470 (7th Cir. 2009)
164. *Ill. Commerce Comm'n v. FERC (ICC II)*, 721 F.3d 764 (7th Cir. 2013)
165. *Ill. Commerce Comm'n v. FERC (ICC III)*, 756 F.3d 556 (7th Cir. 2014)
166. *Midwest ISO Transmission Owners v. FERC*, 373 F.3d 1361 (D.C. Cir. 2004)
167. *Phillips Petroleum Co. v. Wisconsin*, 347 U.S. 672 (1954)
168. *United Gas Pipe Line Co. v. Mobile Gas Serv. Corp.*, 350 U.S. 332 (1956)

Regulations and Statutes

169. 183 U.S.C. § 717c (2012)
170. 184 U.S.C § 824d (2012)

Capacity Markets: Rationale, Designs, and Trade-Offs



Alfredo Garcia

Abstract Many electricity markets around the world have implemented capacity markets (e.g., PJM, New England, New York, UK, Colombia), and many more are in the process of finalizing capacity market designs. In this chapter we will review the rationale behind capacity markets, the basic traits of the most popular designs and some outstanding design issues.

1 Introduction

In the last decade, many electricity markets have put into operation new “capacity markets” which aim to ensure an adequate level of investment in generation capacity expansion by supplementing generator revenues. Evidently, this regulatory intervention is motivated by a firmly held belief that expected energy payments do not promote the right levels of investment in new generation capacity. To provide context, consider the case of Colombia’s power system which largely relies on hydroelectric generation for supplying demand. In years with average to high precipitation levels, thermal plants with a high marginal cost have exceedingly low utilization factors. However, in a dry year (which occurs on average, every 5 years) these generation resources are badly needed to ensure reliable supply. In a market with only energy transactions, this type of generation technology only receives significant income 1 out of every 5 years when spot prices often equal the price cap (i.e., maximum allowable bid in the spot market) imposed by the regulator. Due to vertical disintegration between generation and retail, load-serving entities cannot cope with this risk by maintaining a portfolio of hydro and thermal generation assets. Therefore, in a market with energy-only payments and vertical disintegration, the risks related to volatile revenues due to cyclic precipitation patterns, price, and availability of fuels make private investment in peaking technologies very unlikely.

A. Garcia (✉)

Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX
e-mail: alfredo.garcia@tamu.edu

Why would a situation like this call for regulatory intervention? One could argue, for example, that a market for forward contracts (with duration longer than 5 years) could provide a solution to this conundrum: peaking technologies could price-in the risk hedging service, and load-serving entities could manage their exposure to high prices by procuring forward contracts from investors in peaking technologies. There are however significant caveats to this story. For instance, a thin or poorly subscribed forward contract market would be associated with insufficient investment in peaking technologies. Under intense price competition in the retail market, load-serving entities would have weak incentives to pay a risk premium in high to average precipitation years. It seems therefore that a forward contract solution would require some form of systemic risk regulation.

The structure of this document is as follows. In the first section we briefly review the different arguments that have been used to justify the creation of capacity markets. The second section describes the basic design structure of several capacity markets. In the final section, I offer some concluding thoughts.

2 Rationales for Capacity Markets

2.1 Price Caps and the Missing Money Problem

The most frequently cited reason to justify the introduction of capacity markets is known as the problem of *missing money*. To describe this problem, it is necessary to review the functioning of a perfectly competitive market with a spot price equal to the marginal cost of the marginal technology. A useful visual rendition is known as a screening curve (see Figure 1). The top describes the different technologies in terms of total cost per unit if installed capacity (e.g., MW) versus the total number of hours of operation in a year. Peak technologies have lower capital (fixed costs) but higher variable costs (fuel). Mid-merit and base technologies are more capital intensive but lower variable costs. Assuming the spot market operates efficiently (i.e., the spot price equals the marginal cost of the marginal technology), then one can examine which capacity configuration is equilibrium (i.e., capacity levels such that no incremental investment in any of the available technologies is profitable). The bottom figure describes the spot prices for the given capacity configuration. Note that there are realizations of demand (with a duration in $[0, H_0]$) for which there is not enough capacity and the spot price equals the marginal value of lost load (VOLL) (which is assumed constant here for ease of exposition). This is the maximum per unit amount consumers are willing to pay to avoid outage.

To see why the capacity levels described in Figure 1 are in equilibrium, consider, for example, the highest values of demand with durations in $[0, H_0]$. As we mentioned before, there is not enough capacity to meet these values of demand. Investors in peak technology earn a rent equal to difference between the VOLL and the marginal cost of the peaking technology times the duration H_0 . This rent equals

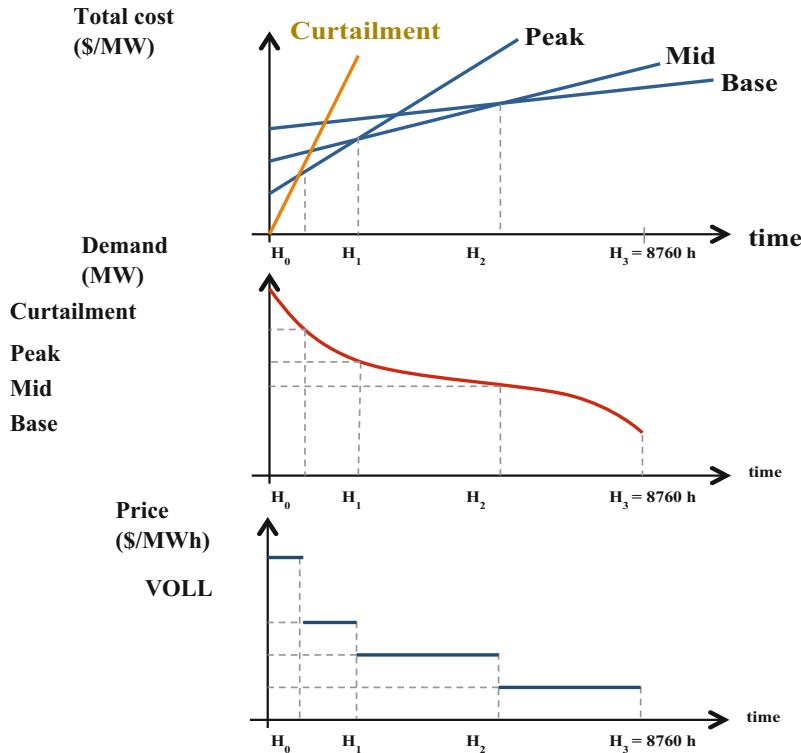


Fig. 1 Screening curves for capacity expansion under perfect competition

the average cost associated to this investment level. Any additional investment in this technology is not profitable. For demand values with durations in $[H_0, H_1)$ in equilibrium, the marginal technology is the peak technology (i.e., for these demand levels, the spot price equals the marginal cost of the peaking technology). As before, the mid-merit technology earns a rent that is proportional to the durations H_0 and H_1 , and there is no incentive for additional investment in this technology. For demand levels associated with durations greater than H_1 , the technology with the lowest average cost is the mid-merit. Hence, for demand values with durations in $[H_1, H_2)$, the mid-merit technology is at the margin. Finally, the base technology (lowest marginal cost, highest fixed cost) is marginal for load levels with durations exceeding H_2 . A similar argument can then be made to show that there is no incentive for additional investment in this technology.

For diverse reasons (e.g., political, market power mitigation), electricity regulators around the world have price caps for the spot price of electricity. In an electricity market where the price of spot energy is subject to a price cap (set at a value that exceeds the marginal cost of the peaking technology but that is lower than the (per unit) value of lost load $VOLL$), the aggregate levels of investment in

all technologies (in equilibrium) are distorted. In particular, the equilibrium levels of investment in the peaking technology are *reduced* inducing a market failure: the individual return of investors in peaking technology is lower than the social value of this type of investment (See [6]). In a similar fashion, the imposition of a price cap also distorts the equilibrium mix of mid-merit and base technologies.

The missing money argument crucially depends on the assumption that demand is inelastic. This assumption may have been well justified in the past, but it seems no longer tenable as novel technologies are enabling consumers to shift consumption over time. For example, a combination of distributed generation (e.g., solar) and storage capacity may enable consumers to respond to intraday price changes. If demand is sufficiently elastic, there is no need for a price cap to protect consumers. And without a price cap, there would not be a missing money problem.

Another key element in the missing money argument relates to the assumption of a perfectly competitive market. With market power the relationship between price caps and equilibrium spot prices is rather complicated. For example, a *lower* price cap may *increase* average spot prices when there is market power (see [2]). The evidence during the California market crisis (Summer 2000) seems to support this conclusion (see table below).

PriceCap\$/MWh	AvgSpotPrice\$/MWh	
750	106	(1)
500	126	
250	134	

2.1.1 A Counter Argument: Energy-Only Markets and High Price Caps

A simple counterargument to the “missing money” argument for capacity markets is to relax the price cap constraint. This is the case of the markets in New Zealand, Australia, and Texas in which the spot market is capped with relatively high values (e.g., 10000 US \$ / MWh in Texas and even higher in Australia and New Zealand). A relaxed price cap constraint implies that situations of scarcity induce significant rents on available capacity, thus allocating risk to consumers. Without any further regulatory intervention, consumers may be exposed to the risk of extremely high prices. Without proper incentives, load-serving entities may not actively hedge against high price events. To protect consumers, load-serving entities must be hedged via contractual commitments and subject to steep penalties if they fail to comply with this commitment. This induces a demand for long-term contracting that in turn drives investment in generation capacity and prudent management of fuel price risk.

2.2 *Systemic Risk*

Under vertical disintegration between generation and retail, load-serving companies have weak incentives to sign long-term forward contracts. The competitive pricing pressure in a business with low markups for commercial intermediation makes a relatively myopic procurement strategy fairly attractive. However, it is precisely long-term forward contracts that facilitate the financing of new generation plants. By failing to hedge against a potential but low probability scenario of scarcity and high prices, load-serving companies may be exposed to the risk of insolvency. A myopic procurement strategy induces a systemic risk: investment in new capacity is likely to be insufficient.

It is convenient here to make an analogy to the banking sector since it presents an analogous situation. Banks may have lending portfolios with risk levels that in aggregate pose a real threat to the viability of the system. In order to control this systemic risk, several schemes of prudential regulation have been implemented through, for example, stress tests that limit the levels of exposure to portfolio risk. The analogy is imperfect because there is no central bank in the electricity market that acts as a lender of last resort. Capacity markets in electricity aim to provide the market with a tool of last resort by forcing all load-serving entities to have some form of protection against critical system (high price) conditions.

2.3 *Market Power*

Another argument (which is perhaps less cited to justify a capacity market) relates to the incentives to limit entry by incumbent generation firms with market power. In an energy-only electricity market, there is an inverse relationship between the total amount of available generation capacity and the energy spot price: the higher the levels of excess capacity (i.e., the difference between available capacity and peak demand), the more competitive pressure is exerted in the short-term market leading to lower spot prices in the energy market. Low levels of excess capacity are typically associated with higher spot prices for energy and stronger incentives for entry. In this setting, spot market sales are the most important source of revenue for small or independent power plants. This is due to the fact that (*i*) vertically integrated utilities may self-procure electricity by having forward contracts between affiliates in generation and retail and (*ii*) large incumbent generators (with diverse assets) demand lower premiums for price-risk hedging via forward contracts. Therefore, in an oligopolistic market with entry costs, generation companies in the market can (indirectly) control entry by maintaining a level of excess capacity that is high enough to discourage entry by new investors. Thus, there may be a free-entry equilibrium in which oligopolists maintain a level of surplus capacity that is lower than that which would be socially optimal but high enough to deter entry (see [4, 5]). In these conditions there would also be a market failure: it is socially optimal to

invest in greater levels of excess capacity, but it is not profitable to do so either for an independent investor or for firms already installed in the market. The investment strategy mentioned by the oligopolists acts as a barrier to entry. Capacity market designs aim to eliminate this barrier to entry by subjecting all investment alternatives for new capacity to a competitive tender process.

3 Taxonomy of Capacity Market Designs

Before describing different regulatory approaches for capacity markets, it is important to outline the differences between *resource adequacy* and *security of supply*. Resource adequacy refers to the guarantee that there will be enough available generation and network capacity to meet forecasted demand. This long-term goal differs from *security of supply* which is the ability of the power system to deal with real-time disturbances in the short run. In the previous section we described arguments to support the claim that electricity markets subject to price caps may fail to ensure resource adequacy, hence the need for capacity markets. In what follows we will describe the different capacity designs which can be classified as depicted in Figure 2 below.

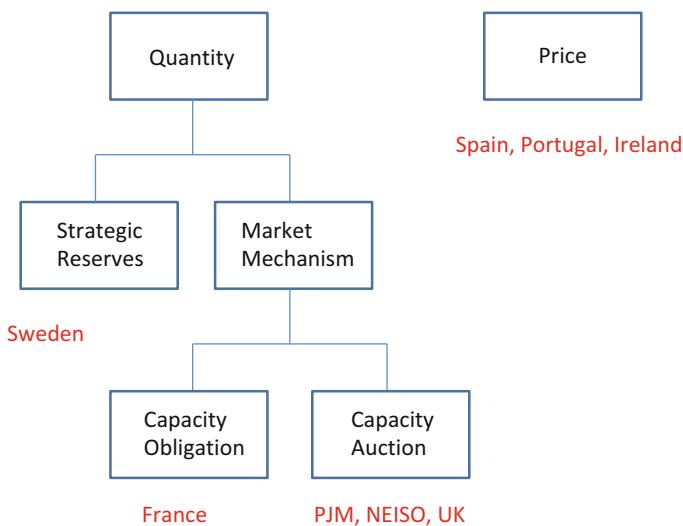


Fig. 2 Taxonomy of Capacity Market Designs

3.1 Quantity-Based Mechanisms: Capacity Obligation

A common approach to ensuring adequate investment in new generation capacity is quantity-based: load-serving utilities are obligated by regulation to procure an adequate level of generating capacity. For example, in the capacity obligation scheme to be implemented in France and the Southwest Power Pool (SPP), load-serving utilities are required to contract a certain level of capacity from certified generators or certified consumers (with the ability to reduce demand upon command) at a price negotiated between the parties. A load-serving utility that fails to comply with this obligation is subject to a fine that is proportional to the unmet requirement. This scheme includes checks and penalties in the event that the capacity guarantee is not honored by the issuer when required.

A capacity obligation induces an artificial demand for capacity guarantees. In a similar manner, a certification process (coupled with penalties for nonperformance) regulates the supply of capacity certificates. Market clearing for capacity certificates is decentralized though in certain designs, certificates can also be traded via periodic public auctions.

A recurrent criticism of this capacity market design is related to vertical integration between generation and retail by certain incumbents. Under decentralized market clearing, vertically integrated incumbents may self-supply, and this constitutes a barrier for more efficient entrants. A related criticism pertains to the duration of capacity certificates. A relatively short-duration certificate (e.g., 1 year) imposes a significant risk on new entrants thus increasing their cost of capital. In order to alleviate these concerns, the capacity design in France includes certificates with a 7-year duration for new capacity that are traded through a public auction. To prevent possible market manipulation (e.g., capacity withholding), the design includes active market monitoring of capacity certificates issuers with respect to historical benchmarks. In addition, large generation firms will be mandated to supply minimum amounts of capacity certificates to ensure liquidity in centrally run auctions.

3.2 Quantity-Based Mechanisms: Strategic Reserves

In this capacity market design, an independent system operator is in charge of determining the total amount of capacity reserves needed to meet a reliability target based on estimated demand and what the market would otherwise provide in terms of capacity without the mechanism.¹ Typically, the ISO is in charge of procuring via a competitive auction contracts for capacity that must be deployed upon command. These capacity resources are often exclusively dedicated for reserves purposes so that the strategic reserve acts as generator of last resort.

¹The notion of maintaining strategic reserves is not novel. For example, the USA has kept for several decades an inventory of oil reserves for national security purposes.

This mechanism is relatively easy to implement. However, it is not clear that it effectively addresses the resource adequacy problem since a relatively short contract duration may only attract interest from the owners of installed capacity. On the other hand, a long contract duration may render some ISO forecasts into a self-fulfilling prophecy: if large amounts of reserves are deemed necessary, then most new entry will likely be structured around reserves contracts thus reinforcing the need for large amounts of reserves.

In addition, the mechanism may create a nontrivial linkage between spot prices and the clearing price for reserves as incumbent generators with relatively large market shares may arbitrage between the spot and the reserves market. In this case, the mechanism induces a transfer of capacity from one market to another without necessarily providing incentives for new capacity.

3.3 *Quantity-Based Mechanisms: Reliability Options*

This scheme is centered around the procurement of *reliability* options by load-serving entities. Typically procurement is done in a centralized fashion by the ISO (on behalf of all load-serving entities) through a descending clock price auction. Under a reliability option contract, contracted capacity providers must make capacity available when the spot price exceeds a strike price (or *scarcity* price) which defined ex ante or pay the contracted capacity times the difference in prices. The duration of these contracts may vary between 1- and 5-year-long terms. This scheme has been adopted in New England and Colombia (see [1, 7]). Determining the aggregate demand for reliability options is an important aspect of this design. Since there is no demand-side bidding in the auction, a proxy for demand appropriately reflecting consumers' willingness to pay for reliability is needed. A starting point for computing this proxy demand curve is the engineering target of meeting a given loss of load expectation (LOLE) (e.g., for the New England Market, this target is 1 day in 10 years). A pivotal (quantity, price) two-tuple used in constructing a proxy for demand corresponds to the level of capacity needed to meet the given reliability target at the estimated cost of the marginal technology (this is sometimes referred to as cost of new entry). A linear function is then obtained by using a slope around this point that is again set by the regulator (see [8] for a more elaborate estimation of a proxy for the capacity demand curve).

This relatively complex design exhibits significant benefits: *(i)* it reduces long-term generator's risk for both incumbents and new entrants, *(ii)* consumers are hedged against high spot prices, and *(iii)* it mitigates market power that would otherwise emerge in times of scarcity. However, the implicit bundling of resource adequacy and high price concerns may result in a flawed instrument as we shall review in the ISO-NE capacity market performance during the Polar Vortex of 2014. (Figure 3).

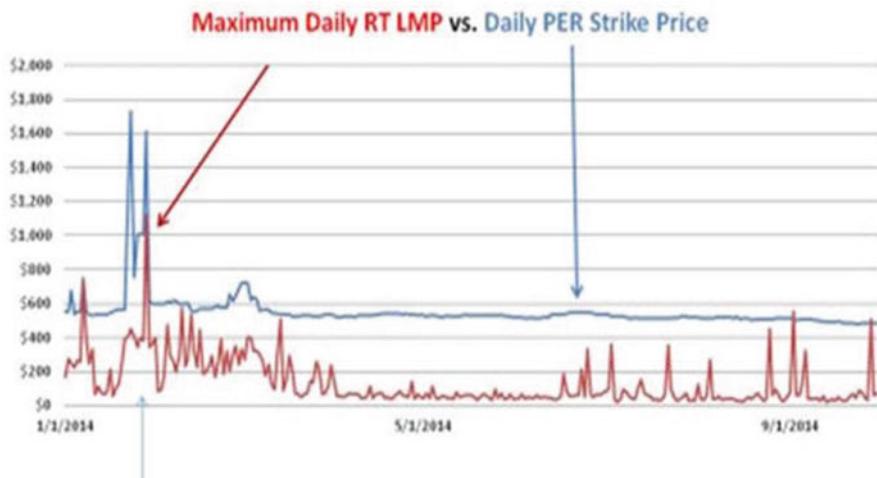


Fig. 3 ISO-NE spot prices vs strike price during polar vortex 2014

The Problem of Indexing the Strike (or Scarcity) Prices: Indexing or updating the scarcity price over time has proven to be quite problematic in both New England and Colombia. At the outset of the forward capacity market (FCM) in New England, the strike price was based on the marginal cost of a generator using natural gas (priced at the Henry Hub price index) and under a given (relatively high) heat rate. Due to the significant increase in shale gas production, natural gas prices dropped so that strike or scarcity price fell significantly. However, in early 2014 an extremely cold wave caused by a southward shift of the North Polar Vortex affected New England. Record-low temperatures implied an increased demand for natural gas for heating. Limited pipeline capacity implied extremely high gas prices and thus an extremely high strike price. The marginal plants in the market dispatch were burning liquid fuel at a cost that though relatively high was lower than the strike price. As a result of failing to properly capture the cost of the marginal technology, the reliability options were not exercised during this highly critical event. After a long process of consideration, in May 6, 2015, FERC approved the elimination of peak energy rent (the mechanism protecting consumers against high prices) by raising of the strike price to 1000 US \$ /MWh.

Another recurrent concern associated with this mechanism pertains to the uniform price auction design and the possibility of strategic behavior. With a single clearing price, there are incentives to withhold capacity from the auction in order to induce higher prices. This issue has become readily apparent to regulators and independent system operators alike. For example, in a recent ISO-NE report (Dec 2015), it is stated that:

...the current FCM rules do not address the potential for a capacity supplier to exercise market power by retiring a resource prematurely in order to decrease supply, artificially increase prices and benefit the remainder of the suppliers portfolio.

Evidently, more active demand-side participation in the auctions for reliability options can mitigate this risk. Finally, proponents of this market design argue the capacity market does not affect in any way the price formation in the spot market. The trouble with this argument is the fact that scarcity events always affect generators differently so that during critical event some generators are short of capacity to meet their reliability option obligation and some have excess capacity. Those with excess capacity can exercise market power by bidding excess capacity at the price cap. The resulting spot price is therefore affected by the capacity market as we illustrate in the following simple example:

Interaction Between Capacity and Spot Markets: To illustrate the interaction between capacity and spot markets, we will describe a simple hydro-thermal system with two periods. In the system there are two hydroelectric generators each with generation capacity of 100 and reservoir capacity of 150. In addition, there is one (price-taker) thermal plant with capacity 75. We assume there are no water inflows to the reservoirs in the first period, but reservoirs are filled up to capacity after the second period. Assuming initially, reservoir 1 has 150 units (energy) and reservoir 2 has 50. If demand is 150 per period, minimum cost dispatch is 100 units (hydro) and 50 thermal in each period. We also assume the marginal cost of the thermal plant is several orders of magnitude less than the strike price. Even if generators behave strategically, the equilibrium market dispatch is equal to the minimum cost dispatch, and the forward capacity obligations are *not exercised* since the spot price for energy is below the exercise or strike price (see Figure 4 (a)).

Consider now the case in which hydro-generators behave strategically and each one has a forward capacity obligation for 50 units. With the same initial conditions as above, in equilibrium, in the first period, reservoir 1 offers capacity (100 units) at the price cap while reservoir 2 offers its available capacity (only 50 units) at strike price. As a result of first period's market dispatch, reservoir 2 has no capacity left and reservoir 1 has monopoly over residual demand. In the second period, reservoir 1 offers capacity at a price slightly below scarcity price. In the first period, capacity obligations are exercised since spot price exceeds strike price. In the second period, the spot price is set by reservoir 1 at a price that is slightly below strike price and the capacity obligation is not exercised (see Figure 4 (b)). This strategy is an equilibrium provided the strike price is greater than half the price cap. The asymmetry of initial reservoir levels coupled with the forward obligation gives reservoir 1 significant market power.

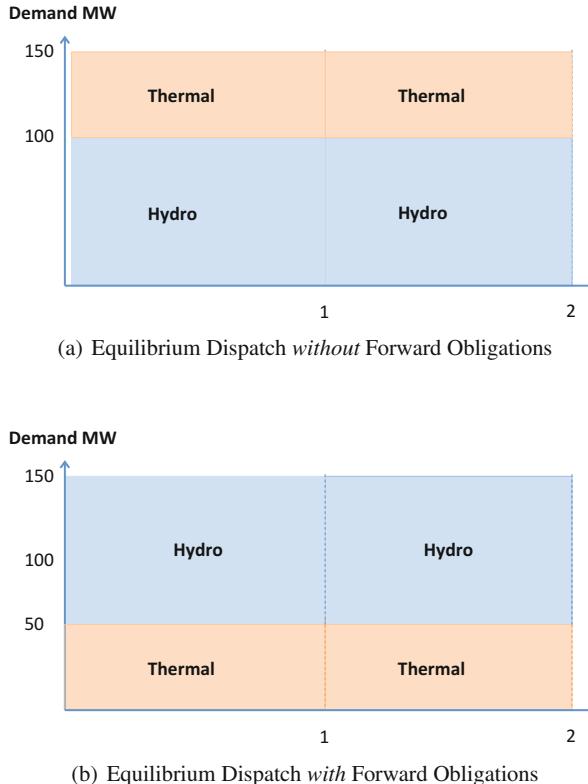


Fig. 4 Interaction between spot market and capacity market

3.4 Incentives for Performance and Moral Hazard

There are two approaches to limit the possibility that generators that are awarded reliability options fail to honor their contractual obligation. The first is preventive: constraints are placed on how much capacity each generator can bid into the capacity market. For example, initially gas-fired units could only participate in the capacity market when backed by a firm contract for natural gas supply (UK, PJM). This requirement is particularly as stringent on the underlying gas network. In countries with significant hydroelectric capacity (Colombia), hydro-plants with large reservoirs are subject to statistical analysis to determine which fraction of their energy could be reliably called on under a reliability option contract under scarcity scenarios.

Preventive schemes have not worked as well as expected. For example, in Colombia during the 2009–2010 drought (due to “El Niño”), hydropower plants continued to operate early in the dry year to meet their bilateral energy commitments, progressively emptying reservoirs. This was a clear indication that hydropower

companies preferred the risk of future underperformance on their reliability options obligations over the failure to honor their contractual obligations in the energy market.

Another approach involves performance incentives. For example, in PJM's reliability pricing model (RPM), a peak-hour period availability (PHPA) charge is used based on actual performances during 500 peak hours. The scheme may have proved inadequate: during the polar vortex of 2014 with record-breaking peak demand, PJM experienced an abnormally high forced outage rate with underperformance of committed capacity resources of roughly 40%. This led to a strong criticism of the scheme in place [3]:

...capacity resources rarely face financial consequences for failing to perform, and therefore have little incentive to make investments to ensure that they can reliably provide what the region needs: energy and reserves when supply is scarce.

As a result PJM is phasing in a capacity performance requirement. It is fair to say that finding the optimal design combining preventive with incentives is still a work in progress.

3.5 Price-Based Mechanisms

The main idea in a price-based mechanism is to provide an additional source of revenue for certain types of peak units which receive a payment per unit capacity usually pre-set, for all or a fraction of their available capacity. This capacity premium aims to encourage generators to invest. It is the regulator that sets the capacity price and the market that determines the amount of capacity. There are different ways to determine the capacity price. For example, in the mechanism that operated in the UK up until 2000, the capacity payment was computed as the product of the likelihood of system failure (calculated based on the supply-demand balance) and the difference between the value of lost load and the market spot price. The experience was not satisfactory: by withholding capacity, incumbents were able to manipulate the market by reducing their available capacity particularly during peak periods. Here again we find very weak incentives for performance. Unlike a capacity market based on reliability options in which generators must assume the difference between spot prices and strike price when failing to honor their reliability obligation, in this design the penalties for underperformance are relatively weak as they are proportional to the capacity payment per unit.

4 Future Issues

4.1 Renewable Technologies and Capacity Markets

Renewable technologies are becoming increasingly competitive in the spot (energy) market. From a *security of supply* perspective, this trend raises a number of issues. For example, there might be a situation in which cheap renewable resources may not be used in an economic dispatch constrained by network transmission capacity and other reliability considerations. Evidently this situation can be addressed by having enough reserves (e.g., fast-ramping gas-fired turbines or demand response). This is an ancillary service, the cost of which must be taken into account in spot market clearing in order to identify efficient dispatch.

Though it may defy intuition, renewable resources may also contribute to *resource adequacy* because a certain level of capacity can be expected to be available with high enough probability. However, the contribution to resource adequacy by renewable resources may only be assessed by using sophisticated system-wide models of stochastic dispatch in the medium and long-run. Evidently, there is a risk involved in crediting renewable capacity as contributing to resource adequacy. This risk depends upon the specific characteristics of each power system and the overall renewable capacity configuration. Hence any capacity market that incorporates renewable resources must be price discriminatory. Such prices would vary greatly between energy-constrained (e.g., Colombia) and capacity-constrained systems (e.g., ISO-NE, PJM). Designing a capacity market that can incorporate renewable technologies is an open research question.

4.2 Interaction Between Capacity Market Designs in Interconnected Markets

The integration of interconnected electricity markets which face capacity constraints for cross-border trade has emerged as an important source of efficiency gains for all market participants. This integration has been achieved by means of market coupling protocols (e.g., France-Belgium-Germany). However, if the integration is limited to the short term (day-ahead to real-time trades), the failure to coordinate capacity market designs may induce inefficiencies that may be much greater in magnitude. However, the coordination between capacity markets is far from trivial. To illustrate consider the hypothetical case of two markets with coupled trading. Suppose one market has a capacity design in place and the other has no capacity market but a relatively high price cap. In this setting, a “leakage” of capacity may occur: under scarcity it may be best for a generator to fail to honor a reliability option in its home market in order to sell its capacity in the neighboring market. In conclusion, the interactions between capacity markets (or lack of) may create incentives for inefficient operation and/or location decisions. The harmonization of capacity markets for coupled markets is an important goal for regulatory agencies.

References

1. Cramton P, Stoft S (2005) A capacity market that makes sense. *Electr J* 18:43–54
2. Earle R, Schmedders K, Tatur T (2007) On price caps under uncertainty. *Rev Econ Stud* 74:93–111
3. FERC, Federal Energy Regulatory Commission (United States) Order on Tariff Filing and Instituting Section 206 Proceeding. Docket No. ER14-1050-000, May 30, 2014
4. Garcia A, Shen J (2010) Equilibrium capacity expansion under stochastic demand growth. *Oper Res* 58(1):30–42
5. Garcia A, Stacchetti E (2011) Investment dynamics in electricity markets. *Econ Theory* 46(2):149–187
6. Joskow P (2008) Capacity payments in imperfect electricity markets: need and design. *Util Policy* 16(3):159–170
7. Oren S (2005) Generation adequacy via call option obligations: safe passage to the promised land. *Electr J* 18(9):28–42
8. Zhao F, Zheng T, Litvinov E (2016) Decomposition and optimization in constructing forward capacity market demand curves. Available at <http://www.optimization-online.org>

Redesign of US Electricity Capacity Markets



Robert W. Moye and Sean P. Meyn

Abstract This paper surveys the different approaches in use today to ensure grid reliability and incentivize new resources. Market challenges are surveyed, as well as empirical findings that suggest that current market approaches do not provide proper incentives. It is argued that the primary problem is that organized capacity markets today do not consider risks and uncertainty over the proper time frame – decades instead of months or years. Because of this, the analyses ignore risks and other factors that are key to making optimal investment decisions. Solutions are proposed based in part on concepts from traditional resource planning.

1 Introduction

Much of the social history of the Western world over the past three decades has involved replacing what worked with what sounded good. – Thomas Sowell

Driven by claims during the last quarter of the twentieth century of anticompetitive behavior by electric utilities, frustrations by consumers having to bear much of the risk of large electric generation investments, and a desire by many to create a more economically efficient market for the electric power industry, the United States began to deregulate (or “reregulate”) certain aspects of the electric power industry. Starting with the *Public Utilities Regulatory Policies Act (PURPA)* in 1978 and continuing with energy policy acts in 1992 and 2005, the US wholesale power markets (and in some states, the retail markets) were opened to more competition. Much like the deregulation of AT&T in the 1980s, these actions were expected to spur innovation and reduce costs to consumers. And like the experience with AT&T, opinions on the results of the deregulation of the electric power industry are mixed.

R. W. Moye (✉) · S. P. Meyn

Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL, USA
e-mail: rmoye@ufl.edu; meyn@ece.ufl.edu

Today, the wholesale power markets in the United States represent a patchwork of policies and approaches that impact, if not outright dictate, how capacity,¹ energy, and related products are bought and sold. While some aspects of all markets appear to function well, there continues to be much debate about the efficacy of some. This is particularly true for the markets designed to ensure that sufficient electric generation is in place today and being planned and constructed for future use. These *capacity markets* currently take many and varied forms across the United States. Most markets, but not all, define a minimum *resource requirement* that entities directly serving customer load are required to meet. Some rely on bilateral markets to meet these requirements. Others have very structured processes (e.g., auctions) to facilitate the purchase and sale of generating capacity. At least one market relies entirely on energy price signals to provide the necessary incentives for investment in new generation.² And while there are many prominent individuals involved in the market structure debate today who believe long-term markets will self-optimize³ if only we can be patient, this issue is far from being settled.

There is little scientific basis to predict that long-term optimality will emerge from short-term decision-making by generation operators. One challenge to analysis is that there is no agreement on how to quantify risk to society or to an individual agent in the market. Another challenge is the enormous uncertainty over planning horizons of many decades. The risk to a generator operator is obvious in today's technological environment: the lifetime of an efficient gas turbine generator may be a half century, and its purchase price over one billion dollars. At the same time, revenue over this time horizon depends on uncertain energy prices and policy.

The importance of capacity planning cannot be overstated. Some of the most famous electric utility failures are associated with poor selection of resources, failure to adequately address potential future market scenarios that ultimately materialized, or simply poor management of the design and construction of resources [28, 35]. Furthermore, a primary argument of capacity market proponents, and market reregulation in general, was to get away from a rate-of-return approach to incentivize investment (wherein consumers bore essentially all capital and operational risks) and instead have the companies that build the generation bear these risks.

Few, if any, electric utilities have failed because of poor operating conditions. These can be rectified in short order. However, resource investments can long outlive the engineers, analysts, and managers that propose such projects. Perhaps the most famous failure regarding resource expansion was the Washington Public Power Supply System. In the 1970s, this group, owned by a consortium of municipal utilities committed to build five, large nuclear power plants in the Pacific Northwest.

¹Capacity represents the amount of power a resource is capable of delivering at a point in time, or over a period of time. Variations in this term are used in the markets to reflect availability, deliverability and other characteristics important in a particular setting.

²The Electric Reliability Council of Texas (ERCOT), the electricity market covering most of Texas, uses an energy-only market as a means of incentivizing resource investments.

³Throughout this chapter, the use of the terms optimize, optimum, or optimal follow the general definitions provided by the *Cambridge Dictionary* and others to mean “best” or “most effective in a particular situation.” In many cases used herein, it will simply mean “least cost.”

Cost overruns and delays, along with a significant drop in load growth, led to the cancellation of four of these units and the largest default on municipal bonds in history at the time (\$2.25 billion) and still the second largest default today [35]. Another famous failure was the construction of the Shoreham nuclear unit on Long Island, completed in 1984 at a cost of \$6 billion. Due to the lack of an approved evacuation plan, the plant never operated and was eventually sold for \$1 [28]. An even more recent failure is the impending cancellation of the Kemper County clean coal plant by Southern Company and Mississippi Power, with the utilities potentially realizing losses of over \$3 billion [19]. Failures like these are often pointed to when discussing the importance of market deregulation.

The remainder of this chapter is organized as follows: The evolution of regulation of the electric power industry is discussed in Section 2 to provide a backdrop for the state of the market today. Traditional power supply planning is the topic of Section 3. The methods and objectives employed in this approach are still in use today in several parts of the United States and provide a comparison to the capacity market solutions implemented in the organized markets. Section 4 contains a short history of marginal cost theory and its use in today's energy and capacity markets. This theory is based on the notion of *efficiency* and *competitive equilibria*, whose definitions are based on a hypothetical social planner's problem.

Current market structures are surveyed in Section 5, with emphasis on the elements of mechanisms used to incentivize investment in generating resources. It is here where we find potential gaps between the hoped-for optimal social planner's solution and the outcomes of markets in a real-world setting. Some of these shortcomings are discussed in Section 6, and potential solutions are presented in Section 7.

Because this chapter requires history of "traditional" power system planning and more recent organized markets, it may be useful for the reader to consult the glossary [16] for terminology.

2 Evolution of the Electric Power Industry

Today, electricity is so basic to the world economy that certain electricity indices are used to express a country's economic standing (consumption or production of electricity per capita) and the standard of living enjoyed by consumers (per capita electricity consumption in the domestic sector) [18]. As such, the availability and cost of electricity is fundamental to the economic well-being and prosperity of a society.

Primarily as the result of competitive market forces, the electric power industry has evolved significantly over time. Generating resources have become more reliable and efficient. High-voltage transmission networks, nonexistent at the birth of the industry, are now extremely reliable and highly efficient.

Practices and procedures, both for system operation and for long-term planning, have also improved greatly and now contribute to the overall value and efficiency of the industry. While many improvements were realized in the early days through trial and error, today's systems benefit from the extensive use of computers to optimize

both short-term operation and longer-term system expansion. These tools have been particularly helpful in enhancing short- and long-term planning techniques and practices.

The regulatory paradigm has also changed significantly over the history of the electric power industry – starting first with regulation by municipalities through the granting of franchises. This was followed by the creation of public service commissions in each state and eventually regulation of wholesale market activities at the federal level. In general, these changes were made to protect electric power consumers from anticompetitive behavior.

In addition to the changes in the manner in which utilities were regulated, several federal laws have had significant impact on the structure of the industry. In the first half of the twentieth century, antitrust laws forced the breakup of larger investor-owned utilities and the creation of many smaller utilities. During this period, other regional electric utilities were created to serve areas not covered by the investor-owned utilities. These included electric power cooperatives that were formed to serve rural areas and federal power marketing agencies like the Tennessee Valley Authority and the Bonneville Power Administration to harness the energy from large hydroelectric projects at federal dams.

In the second half of the twentieth century, the industry was again changed to promote more competition. The Public Utilities Regulatory Policy Act of 1978 for the first time allowed companies other than regulated utilities to sell electricity into the wholesale power market (limited to renewable energy and cogeneration⁴ resources).

The passage of the Energy Policy Act of 1992 marked a significant evolution of the industry. Following development of rules by the Federal Energy Regulatory Commission (FERC), the high-voltage transmission systems that interconnect the utilities in the United States began providing open access to all existing utilities and wholesale generators, and nonutilities would be allowed to own and operate electric generation for sale into the wholesale market. In addition, entities called power marketers (in the 1990s these were typically affiliates of utilities and investment banks) could participate freely in the market by purchasing electricity from one entity and selling to another.

The rules implemented by FERC also provided for regional organizations (albeit on a voluntary basis) to operate the high-voltage transmission systems on a statewide or multistate basis and to implement electricity markets for the purchase and sale of electricity products. While not all regions of the United States elected to create these organizations, those that have continue to evolve.

3 Traditional Electric Power Supply Planning

In the early days of the electric power industry, little formal analysis was done to evaluate existing systems and even less was done to evaluate system expansion [8].

⁴Cogeneration is a generating system where a single fuel source is used to simultaneously produce two or more forms of energy output – typically electricity and steam.

Without access to sophisticated computer models, design and system operation decisions were based on trial and error and heuristics developed over time. The systems were not “optimally” designed or operated in the sense that we think of today. Even if the means were available at that time to take a more analytical approach to planning, the pace at which technology was changing would have likely made any plan developed meaningless in a very short period of time. Load was growing exponentially, and new generators made old generators obsolete well before the end of their otherwise useful lives.

Prior to 1960, engineers and operators used judgment and primitive tools to evaluate the system. In addition to specialized slide rules, systems were evaluated using analog network analyzers. These were working-scale models of the major components of a power system and allowed engineers to test operating scenarios, study system contingencies, and evaluate system expansion plans. Questions like what is the best location for a new generator, a new transmission line, or even a reconfiguration of the entire network could be only partially answered.

These simulators were expensive to build and maintain and often difficult to use. Like many other scientific areas, it took the invention and use of the computer and the development of computer models to make a significant upgrade to the quality of power system evaluation and design. In addition to significantly improving the analysis of existing systems through network reconfigurations and operating practices, computers made possible the longer-term analysis of system expansion. *Power flow*⁵ (or *load flow*) models helped operations but also quickly lent themselves to longer-term optimization. *Production costing* programs⁶ also began to be used to estimate short-term operating costs of the generation fleet. This expanded to long-term power system analysis and the development of tools to deal with uncertainty as the power and use of computers expanded.

3.1 Resource Investments in “Traditional” Markets

While the focus of this chapter is more directed to the methods used in *organized markets*, many parts of the country continue to operate under a “rate-of-return” regime. That is, a utility determines an optimal generating expansion plan to meet future load requirements and secures approval of this plan by the Public Service

⁵A *power flow model* is used to simulate the operation of a high-voltage transmission system. Given assumptions for the topology of the transmission system, the complex impedances of all significant transmission elements, estimates of the real and reactive power loads at each node, and the real and reactive power output of each generator, the model estimates real and reactive power flows through each element of the system modeled and estimates voltage at every bus (magnitude and angle).

⁶A *production costing program* uses assumptions for loads and generation cost characteristics to simulate various operating scenarios (e.g., optimum hourly resource commitment and dispatch over the study period) in order to estimate individual resource and total system operating costs over the study period. Longer-term models also incorporate cost assumptions for resource additions, retirements, and repowering.

Commission that regulates their service territory. Once approved, these utilities are allowed to incorporate the cost of new resources in their *rate base*,⁷ and stock holders earn a return on equity invested over the entire *useful life*⁸ of the facility. This rate-of-return method of incentivizing investments is what today's *organized markets* are attempting to replace.

The traditional approach to developing generation expansion planning is to determine the combination of resources that, given the assumptions used, will result in the least total system cost over an extended period of time (e.g., 30 years) and over a range of scenarios. However, it will not necessarily represent the absolute least cost plan given the most likely scenario, but will represent a plan that is deemed "best" based on the evaluation criteria (both quantitative and qualitative) and reflects some robustness around a range of uncertain future outcomes.

Current-day power system planners consider a wide range of factors when developing and eventually selecting a final generation expansion plan. These factors include economies of scale attainable from larger resources, economies attainable from interconnection to other systems, progress associated with newer technologies, substitution among input factors, resource replacement, the risk-reward nature of investment decisions, and expected future operating conditions [34]. The process employed typically involves the following [1]:

1. Forecast of the system electric load for *30 or more years into the future*.
2. Evaluation of the energy resources (generation and demand-side resources) presently available and expected to be available in the future.
3. Forecast and evaluation of the economic and technical characteristics of the existing system and of potential expansion scenarios. These characteristics include capital investment costs, fuel costs, operation and maintenance costs, efficiencies, and construction times.
4. Determination of technical and cost characteristics of the resources available for expansion.
5. Determination of the economic, financial, and technical parameters affecting decisions. These include reliability standards, borrowing rates, required return on equity estimates, and financial discount rates.
6. Choice of a procedure to determine the optimal expansion strategy within the imposed constraints.⁹
7. Qualitative review of the results to estimate the viability of the proposed solution.

This analysis takes into account the present and future economic and technical environments within which the electric sector is expected to operate. Thus, available resources and fuel prices are related to the energy policy of the country. Economic development policies, existing and foreseeable, are considered in the demand forecast.

⁷*Rate base* represents the total value of facilities on which a public utility is permitted to earn a specified rate of return, in accordance with rules set by a regulatory agency.

⁸*Useful life* is the estimated lifespan of a depreciable fixed asset, during which it can be expected to contribute to utility operations.

⁹In this case "optimal" does not necessarily mean absolute "least cost" because other strategic considerations are included in the planning process.

Traditionally, the demand by consumers for electricity was assumed to be exogenous, and the objective was always cost minimization. The approach used in the past to address the interactions between supply and demand was usually an iterative one, where future demand was typically based on an initial pricing policy, the planning engineers developed least-cost solutions based on these demand estimates, and then the demand forecasts were revised based on marginal costs and prices. The search for an investment program was therefore one that satisfied engineering and economic criteria in an iterative, multidisciplinary process [1].

3.2 Long-Term Planning Models

Traditional methods for determining resource investments involve optimization of an objective function for a power system planning model that includes forecasts over many years [11]. A brief survey is presented here on the optimization methods for short-, intermediate- and long-term analysis of power systems because it is germane to the discussion that follows regarding “energy-only” markets and their potential to provide incentives for resource investment.

As will be made clear, optimization models used for long-term optimization take into consideration the long-term capital requirements associated with the addition and retirement of resources needed to provide a reliable supply of energy to consumers. An effective model also takes into account elements of short-term optimization: the same hour-by-hour simulation present in today’s organized markets (i.e., used to define *locational marginal prices* (LMPs)) is an important component of the traditional planning methodology.

One way to frame the nature of power system costs, and therefore the optimization problem to be addressed, is to consider these costs over differing time periods. Why certain costs are relevant in optimization analysis and other costs are not will be made clear in the remainder of this section.

- **Short-term** periods (covering the next few minutes to the next few hours), where only fuel and variable operation and maintenance (O&M)¹⁰ costs are relevant.
- **Intermediate-term** time periods (a few hours to several days or even months), where, in addition to fuel and variable O&M costs, unit startup costs and no-load energy costs must be considered.
- **Long-term** periods (periods from several months to many years), where all costs must be considered. These include fixed operating costs that can be avoided if a resource is shut down for an extended period, costs to overhaul or repower existing resources, and costs of new resources that can be brought online to replace or supplement existing resources.

¹⁰Variable O&M refers to nonfuel resource costs that vary with resource operation. While the determination of some components of variable O&M is very subjective, some components are easily quantified. Examples include lubricating oil and make-up water for cooling towers – each of which can be significant for small generators. Utilities can directly tie these costs to operating hours or MWh production. For simplicity, they are often reflected entirely in units of \$/MWh.

For the period we have described as short-term (ST), the optimum cost (OC) function used to develop an optimal solution for a system of J resources over a time period T can be represented as follows:

$$STOC = \min_G \left\{ \sum_{t=1}^T \sum_{j=1}^J V_j(t) G_j(t) I_j(t) \right\}, \quad (1)$$

subject to numerous constraints, such as generation matching demand and resources being dispatched within their allowable operating ranges. In this minimization, $V_j(t)$ represents the *variable* operating costs (fuel plus variable O&M costs, in \$/MWh) for resource j during time t ,¹¹ $G_j(t)$ represents the average output of this resource during the same period (expressed in MWh), and $I_j(t) \in [0, 1]$ indicates whether resource j is online during time t . Because the other costs associated with the ownership and operation of a system cannot be avoided during shorter time periods, they are considered “fixed” or “sunk” in this formulation.

For the period we have described as intermediate-term (IT), the cost function used to develop an optimal solution should take into consideration other costs (namely, those that can be avoided during such periods). The extent to which such additional costs are taken into consideration depends on the length of the time period being evaluated and the characteristics of each resource. Modifying (1) to incorporate these costs yields:

$$ITOC = \min_{F,G} \left\{ \sum_{t=1}^T \sum_{j=1}^J [V_j(t) G_j(t) + F_j(t)] I_j(t) \right\}, \quad (2)$$

subject to the same constraints described for Equation 1 but also other constraints such as minimum start times, minimum run times, maximum run times, and fuel limitations. In this equation, $F_j(t)$ represents startup costs, no-load energy costs,¹² and other avoidable operating and maintenance costs of resource j during the period being evaluated (T), and $I_j(t)$ now indicates if a resource is available to be committed and dispatched during the period being considered, given the above constraints. While these costs are often expressed in units of \$/kW-mo or \$/MW-day, for our use here we have assumed that they are reflected on the same timescale as t .

Note that when transmission constraints are taken into consideration, Equation (1) is essentially the form used in organized markets today for determining LMPs in real-time markets. When the startup times, ramping constraints, and other operational and reliability considerations are taken into consideration, Equation (2) is essentially the form used in day-ahead markets.

¹¹As used here, t represents a period of 5 minutes, 15 minutes, or 1 hour – all commonly used in power system analysis.

¹²No-load energy costs represent the fuel and variable O&M costs to operate a resource at 0 MW of output.

For analysis over time periods of years or even decades, the cost function used becomes even more complex. Over such long-term (LT) periods, additional avoidable operating and maintenance costs must be taken into account, but also costs associated with overhauling, repowering, or replacing existing resources must be considered:

$$LTOC = \min_{C,F,G} \left\{ \sum_{t=1}^T \sum_{j=1}^J \beta_t \left[V_j(t)G_j(t) + F_j(t) + C_j(t) \right] I_j(t) \right\}, \quad (3)$$

subject to the constraints described for Equations 1 and 2, but also subject to permitting and construction times limitations, limits on the ability to finance projects, and other related constraints. For this minimization, $C_j(t)$ is the amortized capital cost of future resource j (or the amortized costs of capital improvements made to existing resource j), β is a discount factor used to transform the results into the present-value costs more typically used in long-term analyses, and $I_j(t)$ indicates if a resource can be installed and made available for operation during the period under consideration.

It should be clear from (3), and an understanding of the size of the investment required by most generating technologies (that determines $C_j(t)$), that the choice of T can greatly influence the optimal solution. Because capital investments can represent a significant portion of the total cost for a resource over its lifetime (nearly 100% for solar and wind resources), the time period over which an investment must recover these costs is critical. For resources with service lives of several decades, the analysis must cover such periods to ensure the solution to the optimization problem is meaningful.

Traditional methods for determining resource investments employ the optimization formulation, albeit in more complex terms, shown in equation (3). However, traditional methods also incorporate probabilistic techniques to reflect the uncertainty and risks (to both the market and investors) that are associated with many assumptions and input variables used in the analysis. These include load uncertainty tied to future weather and economic conditions, resource availability, fuel availability and prices, technology changes, capital costs, regulatory uncertainty, and many others. To the extent such factors can be quantified, they are included in the numerical analyses performed. To the extent they cannot be quantified with sufficient accuracy, they are typically incorporated in a high-level subjective assessment. Often, multiple hypothetical scenarios are run to better understand the impact of key events before arriving at a final plan.

It is troubling that the methods used in today's organized markets ignore much of the above complexity and instead assume that a very short-term, so-called competitive market will realize results as good or better than from the use of traditional long-term optimization methods.

4 Marginal Analysis and Efficiency

[E]very tub must stand on its own bottom, and that therefore the products of every industry must be sold at prices so high as to cover not only marginal costs but also all the fixed costs, including interest on irrevocable and often hypothetical investments... Hotelling [21, pg. 242].

Since, when average costs are decreasing, marginal costs are less than average costs, the total amount paid for the product will fall short of total costs. Coase [10, pg. 169].

An hourly spot price (in dollars per kilowatt hours) reflects the operating and capital costs of generating, transmitting and distributing electric energy. [emphasis added] Scheweppe, et al [37, pg. xvii].

Short- and long-term optimization of resources in today's organized markets lean heavily on marginal cost theory and the concepts of economic efficiency. A major weakness we have identified is the reliance on short-run marginal costs to provide long-run investment signals. A review of the research regarding the use of marginal costs to set the price for factors of production reveals that some other means of addressing the fixed cost of assets is needed; this fact was recognized by commonly cited authors in this field, such as Coase and Scheweppe.

4.1 Marginal Costs

The discussion of the use of marginal cost pricing for public utility projects began with a French engineer in the 1800s. Jules Dupuit introduced the concept of marginal utility in an 1844 article concerned with the optimum toll for a bridge [12]. This theory was further formalized by Alfred Marshall in 1890 when he combined the ideas of supply and demand, marginal utility, and costs of production [26]. Marshall was also the first to introduce the concept of market equilibrium (used frequently in discussions regarding the design of today's electricity markets) and the ideas of consumer and producer surpluses.

In 1937, Harold Hotelling presented an update to the work of Dupuit (and used the supply and demand curves of Marshall) to argue, among other things, that the use of tolls on bridges in New Jersey was resulting in less-than-optimal use [21]. Hotelling argued that because the amount of the toll was above the marginal cost to allow people to use the bridge (which was essentially \$0), it prevented some from utilizing the bridge that would otherwise benefit from such use (because their marginal value was above \$0, but less than the amount of the toll). Because of this fact, Hotelling argued that the overall welfare of the potential users of the bridges, and indirectly the welfare of the community as a whole, was not being optimized.

In 1946, R. H. Coase addressed the issues presented by Hotelling and others and specifically focused on the “conditions of decreasing costs” [21, 22]; see also [25, 29]. Coase agreed that the amount paid for goods and services should equal the marginal cost to produce or provide the goods and services. However, he pointed out that whenever marginal costs are less than average costs, the total amount paid

for a product will fall short of total costs. In the case of power systems, average total costs are well above average marginal costs. For bridges, water systems, and other public infrastructure investments (that were the primary focus of Dupuit and Hotelling), the solution to this shortfall was to use taxation or some other means to make up the difference. What then is the solution for electric power systems?

Marginal analysis was first applied to investments in electric power supply by the Electricité de France (EDF) in the late 1940s and in the 1950s. While most efforts in the United States were focused on the theoretical aspects of marginal pricing, EDF was concerned more with the practical implementation [32, 42]. This work led EDF to implement a transmission tariff in 1957 that utilized marginal cost pricing and incorporated these same concepts into long-term investments. Marcel Boiteux (during this time an engineer at EDF and later its chairman) studied the relation between short- and long-run marginal cost pricing. Boiteux states in [13] that “provided there is an optimal investment policy, short-term pricing is also long-term pricing and there is no longer any contradiction between the two” [13, pg 70]. While this appears to support the position taken by some regarding the design of US Organized Markets today, it is contradicted by the following statement by Turvey: “If a price equal only to marginal operating costs creates excess demand... efficient short-run resource allocation requires a price higher than this [42, pg 429].” The solution provided by Boiteux et al. was to increase the price beyond marginal costs. Unfortunately, this leads to less than an optimal allocation of resources as argued by Dupuit et al.¹³

In the 1970s, work in this area continued by Baumol and Bradford [2] and Feldstein [15], where Ramsey-Boiteux pricing¹⁴ was used to derive how prices should be increased above marginal cost in order to meet “social revenue requirements.” In 1971, Vickrey introduced the concepts of “real-time pricing”¹⁵ for a product, albeit for telephone service pricing [2]. However, it wasn’t until the 1980s when work by Schweppe et al. focused specifically on electricity [5].

This work, along with other work done by his co-authors [3, 4, 40], led up to the book that many today point to as the basis for the use of marginal cost pricing in organized markets – *Spot Pricing of Electricity* [37].

It is a crucial fact that all of the prior research and analysis into the use of marginal costs from Dupuit to Schweppe et al. are consistent with the idea that while prices for electricity at marginal cost optimize the general welfare in the short-run,

¹³Any amount added to marginal cost will, in theory and in practice, lead to lower consumption and will therefore not maximize social welfare (producers will have surplus energy at a cost that is less than what customers value).

¹⁴Ramsey-Boiteux pricing is a policy concerning what price a monopolist should set, to maximize social welfare, subject to a constraint on profit.

¹⁵In general, real-time pricing refers to the price for energy over a relatively short period of time – typically between 5 minutes and 1 hour. In the organized RTO markets, real-time pricing refers to the LMPs calculated by the market for energy bought and sold at a specific location and for a set period of time (e.g., 5 minutes).

basing revenues entirely on short-run marginal costs is not sufficient to recover fixed costs and therefore insufficient to incentivize investment in generation.

4.2 Social Planner's Problem on Engineering Timescales

Economic systems are said to be Pareto optimal if there is no alternative way to “organize the production and distribution of goods that makes some consumer better off without making some other consumer worse off” [27].

From a power supply perspective, a power system is said to be operating under optimal conditions if there is no alternative way to lower short-run¹⁶ costs by redispatching or modifying the commitment of available generating resources. However, as discussed above, over the long run, a system can be said to be optimal only if investment decisions are also incorporated as shown in equation (3). That is, a long-term power supply plan can be said to realize Pareto optimality only if there is no other combination of existing and potential resources, along with the optimal commitment and dispatch once given these resources, over the useful life of the resources.

The primary challenge with incentivizing investments in today’s electricity markets centers around the time frame covered by our decisions. Operating decisions are short term, from a few minutes to a few years. Investment decisions are long term, from a few years to several decades.

In neoclassical welfare economics, the *social planner* is a hypothetical “benevolent dictator” who endeavors to achieve the best result for both producers and consumers. The optimal solution of the social planner is the maximization of a *social welfare function* – defined as the sum of the welfare of the suppliers and the consumers. By definition, this solution is Pareto optimal: no one’s economic status can be improved without worsening someone else’s. A market that achieves this optimum is called *efficient*.

While there is theory to support the emergence of efficiency as the result of short-term optimization by selfish agents in the market, this theory is not likely to be predictive on the timescales of interest in this chapter. ***We believe that a long-term investment scenario that is consistent with Pareto optimality can be achieved only with a certain level of long-term planning.***

5 Organized Markets in the United States

The electricity markets in the United States today can be viewed as falling into one of two paradigms. There continue to be “bilateral markets” in which buyers and sellers negotiate the purchase and sale of energy and capacity directly with each

¹⁶“Short-run” in this context refers to the period from the next five minutes through the next few years (i.e., as limited by the time it takes to install additional generating resources).

other.¹⁷ These transactions can range in timescale from the next hour up to several decades, and the characteristics (e.g., firmness, delivery location, and of course price) can be different for every transaction. And while under current regulations, any entity can participate in these transactions, it takes a certain set of knowledge and skills to be effective in this market.

Outside of the bilateral markets, and covering most of the United States, organized markets have been established to provide for the buying and selling of energy, ancillary services, and, in some cases, capacity, via a central clearing mechanism. The primary purpose of these markets is to separate generation and retail electric service from the natural monopoly functions of transmission and distribution.

The primary agents in these models are generation companies that supply the electric power and the *load-serving entities* (LSEs) that are responsible for providing electric service to retail customers [16]. Examples include investor-owned electric utilities such as Pacific Gas and Electric and not-for-profit community-choice aggregators (CCAs) such as Marin Clean Energy. The term “customer” is reserved for the end consumer of electricity – either residential or commercial.

5.1 Marginal Pricing in RTOs

Both Independent System Operators (ISOs) and Regional Transmission Organizations (RTOs) are organizations formed with the approval of the FERC to coordinate, control, and monitor the use of the electric transmission system by utilities, generators, and marketers. More specifically, an ISO, as specified in FERC Order 888, is a nonprofit organization that is designed to provide nondiscriminatory service to all market participants and is independent of the transmission owners and the customers who use its system.

RTOs, defined in FERC Order 2000,¹⁸ also provide nondiscriminatory access to the transmission network but have some additional responsibilities dealing with transmission planning and expansion for the entire region served by the RTO. One key distinction is that an RTO structure is used when the footprint of the organization covers more than one state.¹⁹

Today there are nine ISOs/RTOs operating in North America. They manage the systems that serve two thirds of the customers in the United States and over half the population of Canada. Over time, the distinction between ISOs and RTOs in the United States has become insignificant. Both organizations provide similar transmission services under a single tariff at a single rate, and they operate energy

¹⁷Bilateral markets for all capacity and energy products continue to operate in the southeast and parts of the western United States. Most of the country continues to utilize bilateral markets for capacity – at least for meeting part of the markets’ needs.

¹⁸While the functions of RTOs are similar to those of ISOs, FERC chose to use a new name in Order 2000 for its desired form of transmission organizations in the United States.

¹⁹All organized markets can be structured as an ISO (and most are), but only multistate organized markets can be structured as an RTO.

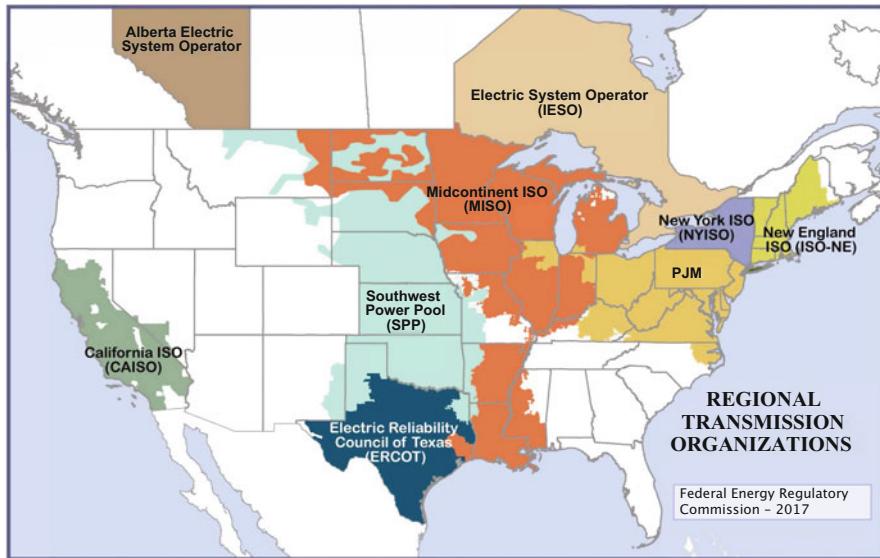


Fig. 1 ISO/RTOs in the United States and similar systems in Canada

markets within their footprints. For brevity, we refer to either ISOs or RTOs, or collectively organized markets, simply as “RTOs.” See Figure 1 for a depiction of the existing RTOs in North America (excluding Mexico) and note that the white regions represent areas in the United States and Canada where the utilities are not currently organized as an RTO (i.e., they remain entirely bilateral markets).

The *locational marginal price* (LMP) used in RTOs is intended to be the cost of supplying, at least cost, the next increment of electric demand at a specific location (node) on the electric power network, taking into account both supply (generation/import) offers and demand (load/export) bids and the physical aspects of the transmission system including transmission and other operational constraints [41]. By design, when the lowest-priced electricity can be delivered to all locations in the market footprint (i.e., there are no transmission constraints), and ignoring electrical losses, prices are the same across the entire RTO. However, when power flows over the transmission system reach limits designed to ensure reliable operation, the lowest-priced energy cannot flow freely to some locations, and more expensive generation is required to serve the load in the constrained regions. Under this scenario, LMPs are subsequently higher in those locations.

A key element of the structure of energy markets within all RTOs is that resource owners and LSEs submit offers to sell and bids to buy hourly blocks of energy for all 24 hours of the next operating day. The RTO takes these offers and bids and determines the least cost, security-constrained commitment, and dispatch of

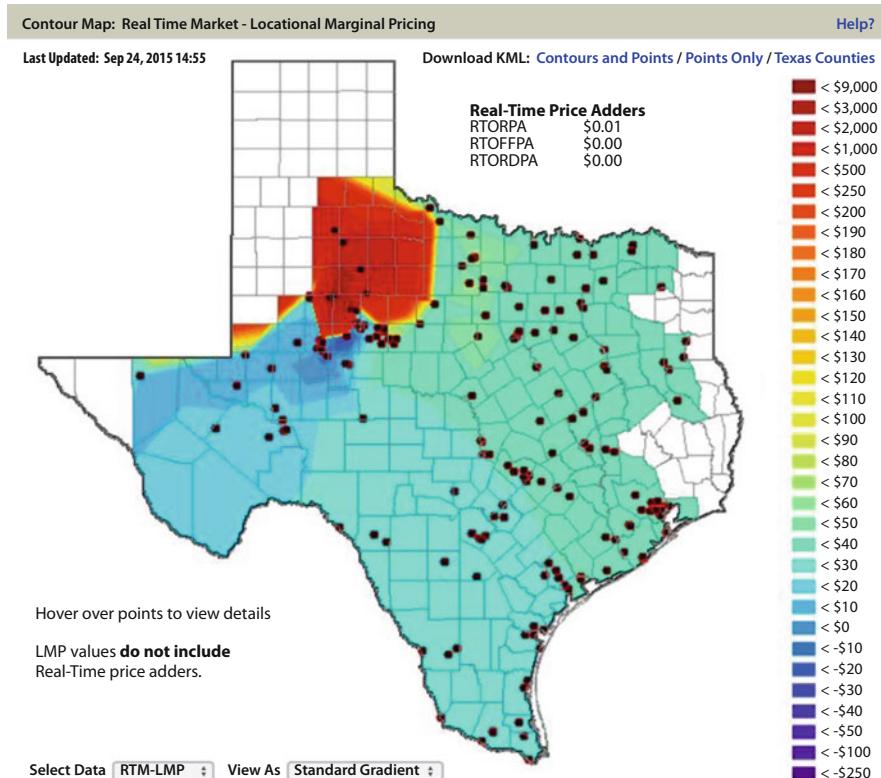


Fig. 2 Real-time prices at ERCOT on 2015-09-24 15.55.56. Source: Electric Reliability Council of Texas, with permission

resources to serve the LSEs for the next operating day. Out of this process, *day-ahead* LMPs are created from the prices offered and bid by the participants.²⁰

In addition to this *day-ahead* markets, RTOs also operate *real-time* energy markets. In most RTOs, “real-time” LMPs are calculated every 5 minutes and represent the price that LSEs will pay or generation will be paid for the next 5-minute period. This enables market participants to factor the information into their operational decision-making.²¹

When subject to transmission constraints or ramping constraints on generation, prices in excess of marginal cost and even *negative prices* are consistent with economic efficiency [9, 43]. Figure 2 illustrates one example of real-time energy

²⁰Note that this day-ahead market activity falls into the “intermediate” optimization problem we describe in Section 3.2. However, it covers only a portion of this time period. Balancing Authorities outside of RTOs typically optimize resource commitment over a rolling, seven-day period.

²¹For example, in a typical RTO operation, an LMP posted at 5 mins after the top of the hour reflects the price for all energy consumed or produced during the period from 10 minutes after the hour until 15 minutes after the hour.

LMPs that can occur in RTOs. This represents a 5-minute interval beginning September 24, 2015, at 2:55 PM Central Time. While the blue- and green-shaded areas represent relatively low LMPs (generally below \$50/MWh, but some less than \$0/MWh), the darker red areas represent prices that exceeded \$1,000/MWh.

5.2 “Administrative Actions” Adversely Affect LMPs

Many entities involved in the RTOs (RTO staff, market participants, market monitors,²² regulators, and market advisors) believe that an “economically efficient” market is one in which the only compensation paid to a generator is tied to a markets’ short-term LMP [7, 20]. Whether or not this is true, they universally recognize that problems with the markets’ design keep the markets from operating efficiently. They believe these problems include:

1. A lack of direct participation by the customers within the same timescales as the generators (e.g., hourly).
2. The use of price caps to limit the maximum price an LMP can rise to and thus limit the potential revenue a generator can receive.
3. The use of “administrative actions” by system operators (e.g., committing or dispatching otherwise uneconomical units) to ensure reliable system operations.

While by design, LMPs are not subject to manipulation by market participants, in practice, system operators have substantial discretion over LMP results through the ability to classify units as running in “out-of-merit dispatch.”²³ When this occurs, these units are excluded from the LMP calculation which often results in depressing market prices.²⁴ In most systems, units that are dispatched to provide reactive power to support transmission grids are declared to be “out-of-merit.” System operators also normally bring units online to hold as “spinning reserves” to protect against sudden outages or unexpectedly rapid ramps in demand and declare them “out-of-merit.” The result is often a substantial reduction in clearing prices at a time when increasing demand would otherwise result in escalating prices.

The *missing money problem* refers to a class of failures in organized markets: expected net revenues from sales of energy and ancillary services provide inadequate incentives for investors in new generating capacity (or equivalent demand-side resources) to invest in sufficient new capacity to match administrative reliability criteria [23]. The consequence is that prices paid to generators in the energy and ancillary service markets are substantially below the levels required to stimulate new entry.

²²Market monitors are independent entities hired by the RTOs to monitor market operations.

²³This indicates that one or more resources being dispatched are done so for reasons other than economics.

²⁴What happens under these circumstances is that more expensive generation is brought online but not allowed to set the market LMP. What is worse is that other generators are then required to reduce output so as to maintain the required instantaneous power balance, thus lowering the overall market’s LMP.

Organized markets have therefore been useful in bringing efficiencies to short-term system operations and dispatch but, in the opinion of some, have been a failure in what was advertised as a principal benefit: stimulating suitable new investment where it is needed and when it is needed. Some blame this lack of sufficient revenues to incentivize investments on features of organized markets, such as energy price caps, “out-of-merit” dispatch decisions, and the use of techniques such as voltage reductions during scarcity periods with no corresponding scarcity price signal.

5.3 Scarcity Pricing

Scarcity occurs when available generation is insufficient to cover the expected energy *and* operating reserves required for reliable operation. *Scarcity pricing* provides for an increase in the LMP during defined scarcity conditions – such conditions being tied to the level of reserves (regulating, spinning, standby, etc.) available to be called upon if needed. As touted by the supporters of this approach, this is a means to stimulate a more competitive market and to better provide incentives for investments in supply-side and demand-side resources.

Some of the RTOs have implemented versions of scarcity pricing. The design of pricing mechanisms are based on two concepts from traditional system planning: 1. the *value of lost load* (VOLL), in units of \$/MWh, that is intended to represent the cost to the ultimate electricity consumers when load is interrupted and 2. *loss of load probability* (LOLP), defined as the probability that the entire load cannot be served. Electric systems have traditionally been planned so that the probability of having insufficient capacity to meet their daily peak load is less than “1 day in 10 years.”

As an example of one scarcity pricing design, ERCOT utilizes an *operating reserve demand curve* (ORDC) which adds an additional price (the scarcity price) to the LMP during any defined periods of scarcity. Figure 3 illustrates the basic structure of the ORDC of the type used in the ERCOT market.

The primary components of an ORDC include (i) a price, assumed equal to the VOLL, to be paid to all resources participating in the real-time market when operating reserves fall below a set level (assumed equal to the market’s minimum operating reserve level) and (ii) a price to be paid to all resources participating in the real-time market as operating reserves approach the minimum designated level. For example, for any given settlement period:

- When operating reserves exceed 7%, the real-time locational marginal price (LMP) is not adjusted and remains equal to the actual LMP calculated for that period.
- When operating reserves fall to 3% or below, the real-time LMP is set equal to the higher of the actual real-time LMP calculated for that period or the VOLL.
- When operating reserves fall to a level below the initial threshold (e.g., 7%) but are above the minimum level (e.g., 3%), the LMP equals the higher of the actual real-time LMP or the product of a “real-time” LOLP and the VOLL.

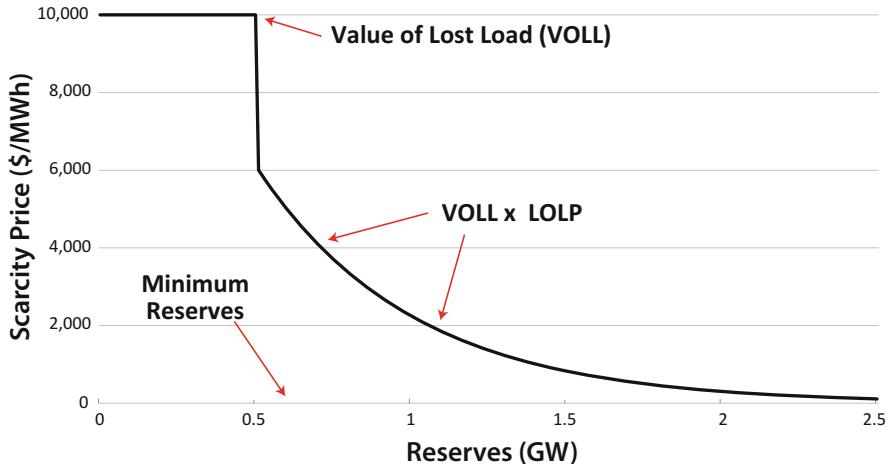


Fig. 3 Operating reserve demand curve

Though the ERCOT report is not clear, it is assumed that the market operator would determine a real-time LOLP that corresponds to the settlement period in question. The VOLL is assumed to be applicable to all customer classes and independent of time. Research has indicated a significant range of VOLL in practice, with values depending on the time of day, day of week, customer class, and also duration [24, 36]. Consequently, the real-time LOLP and stationary VOLL are parameters that the markets are currently not equipped to precisely define or calculate.

Proponents of scarcity pricing believe this market mechanism will i) improve system reliability, ii) be easily implemented within the framework of the current economic dispatch approach, iii) be fully compatible with other market-oriented policies, and, iv) through better scarcity pricing signals, contribute to long-term resource adequacy [20].

5.4 Incentivizing Investments in Organized Markets

The methods used in the RTOs in the United States to incentivize investment in new generation fall into three categories:

- **Energy-only:** An approach wherein revenues from the energy markets (and ancillary services markets) are expected to provide sufficient compensation and price signals to optimize resource investments.
- **Energy + capacity markets:** The above energy-only pricing approach, plus formal capacity markets, which together are expected to provide sufficient compensation and price signals to optimize resource investments.

- **Traditional:** Least-cost, long-term resource planning methods discussed in Section 3 have been used by utilities for decades as “incentives” for investment decisions. For utilities whose generation investments are still regulated by state utility commissions, this approach involves demonstrating that a proposed expansion plan is “best” (e.g., least cost or close thereto) of the plans evaluated. For such entities, the revenues to the utility will be based on a traditional rate-of-return methodology. For utilities and other power supply entities not under the jurisdiction of state utility commissions, long-term *bilateral* supply agreements are expected to support investment decisions.

Outside of the RTOs, the traditional method is being employed and with good success. Over the past 15 years, for example, the largest investor-owned utility in the state of Florida (Florida Power & Light or FPL) has invested in some of the most efficient and cost-effective combined-cycle facilities in the country [17].²⁵ And while FPL still operates under a cost-of-service, rate-of-return paradigm, even with this significant expansion of its generation base, it has seen very little increase in retail rates over this same period.

Within the RTOs in the United States, each of the above methods of incentivizing investment is being employed. Of the seven RTOs:

- One RTO (ERCOT) utilizes the energy-only approach
- One RTO (the Southwest Power Pool (SPP)²⁶) establishes a resource requirement for each LSE to meet and expects all market participants to secure the capacity needed via traditional, bilateral means.
- Four RTOs operate formal capacity markets but with significant differences among the group with regard to market structure and overall approach.

Figure 4 provides a high-level summary of the characteristics of each of the RTOs in the United States. As this table indicates, there are many similarities among the markets:

- Most markets have a Resource Adequacy Requirement. (ERCOT is the sole exception.)
- All markets require that capacity be physical and not financial.²⁷
- Most markets have price caps linked to a “demand curve” that is based on each market’s estimate of the *cost of new entry* (CONE²⁸). See Figure 4.²⁹

²⁵Since 2002, FPL has added over 15,000 MW of highly efficient, natural gas-fired generation.

²⁶SPP serves all of the states of Kansas and Oklahoma and portions of New Mexico, Texas, Arkansas, Louisiana, Missouri, Mississippi, and Nebraska.

²⁷*Physical capacity* is capacity provided by an actual operating generating resource. *Financial capacity* only provides a guarantee to make the purchaser of such capacity financially whole for any market losses. It does not ensure the actual delivery of electricity.

²⁸The *Cost of New Entry* is an estimate of the cost to build the least-cost resource in each market.

²⁹Figure 4 is taken from PJM’s 2014 Triennial Review of their “demand curve,” called the Variable Resource Requirements (VRR) curve in the PJM market.

	U.S. Market:	CAISO	ERCOT	ISO-NE	MISO	NYISO	PJM	SPP
Capacity Market Structure								
Formal Market	No ¹	No	Yes	Limited ²	Yes	Yes	Yes	No
Name			Forward Capacity Auction (FCA)	Planning Resource Auction (PRA)	Installed Capacity Market (ICAP)	Base Residual Auction (BRA)		
Resource Requirement	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Reserve Margin (%)	15%	"Target" of 13.75%	15 ³	14.7	17.0	15.9 ³	12.0	
Bilateral Market	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Market Participation			Mandatory	Voluntary	Mand. & Vol.	Mandatory		
Scarcity Pricing (Energy Market)	Yes							
Energy Market Price Cap ⁴	\$1,000/MWh	\$9,000/MWh	\$850/MWh	\$3,500/MWh	\$500/MWh	\$550/MWh	\$1,100/MWh	
Capacity Performance	Yes	no	no	no	no	Yes		
Market Timing								
Product Term								
New Resources			7 years	1 year	6 mo/1 mo	1 year		
Existing Resources			1 year	1 year	6 mo/1 mo	1 year		
Auction Timing								
Initial			3 years prior	30 days prior	3 mo prior	3 years prior		
Incremental 1			1 year prior		1 mo prior	1 year prior		
Incremental 2			3 mo prior		days prior	3 mo prior		
Incremental 3								
Price Caps								
Maximum			1.6 x CONE	1.4 x CONE	2.0 x CONE	1.5 x CONE		
Demand Curve Structure								

Notes: 1. Imposes a capacity requirement on load-serving entities (LSEs), and has a standardized capacity procurement mechanism but, at present, has no formal capacity market.

2. Limited capacity mechanism in the form of capacity auctions.

3. Estimated. Resource Adequacy set by LOLE analysis.

4. All RTOs approach Scarcity or "Shortage" pricing differently, and have prices for varying amounts of the actual shortage, and type of shortage (e.g., Spinning Res. vs. Non-Spinning Res.).

Fig. 4 Summary of RTO capacity markets

Figure 5 shows capacity supply and demand curves at PJM. The *system supply curve* represents the capacity supply offers submitted by market participants: Point *a* represents a point on the curve where PJM's reserves are 3% below the "1 day in 10 years" LOLP criterion. The capacity market price is capped at 1.5 x Net CONE at this point. Point *b* represents a point on the curve where PJM's reserves are 1% above the "1 day in 10 years" criterion. The capacity market price is capped at 100% of the Net CONE at this point.

However, there are also significant differences among these markets:

- In some markets, participation is mandatory (PJM and ISO-NE). In others, participation is voluntary (MISO and, to a lesser degree, NYISO).
- Some markets have strict performance metrics, wherein penalties are incurred if the resources that cleared the market and receiving capacity payments are not available at the time of critical market conditions (PJM & ISO-NE).³⁰ Some markets have no performance metrics at all and there are little to no consequences if capacity is unavailable when needed.³¹

³⁰The consequence of nonperformance may require that the defaulting party pay for any energy not provided up to \$5,000/MWh.

³¹Typically, the only consequence is that if the unavailability is repeated, future capacity payments will be reduced until capacity availability can be demonstrated.

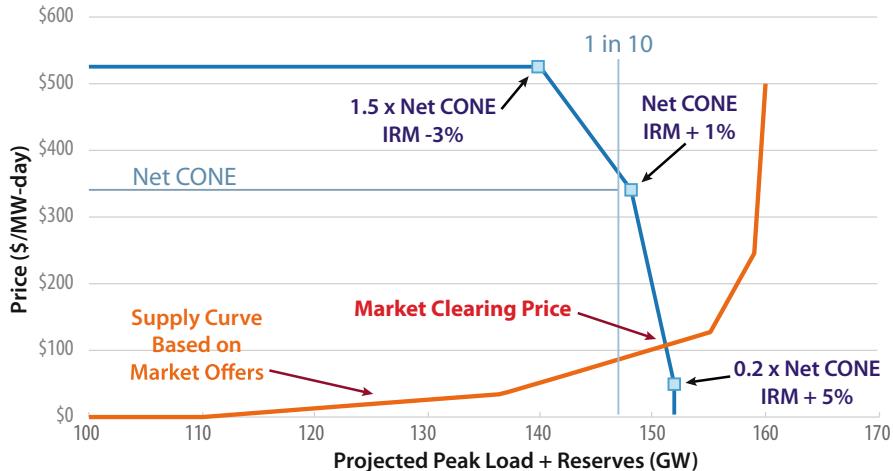


Fig. 5 PJM capacity market demand curve

- The term of the capacity commitments in the capacity markets varies from as short as 6 months (NYISO) to as much as 7 years (ISO-NE, for new capacity).³²
- The market price caps (both for resource offers and maximum allowable LMPs) vary significantly among the markets.
- *Scarcity pricing*, in place to some degree in all markets, also has significantly different thresholds and price caps.

5.5 Market Results

Figure 6 depicts the results (in terms of the capacity price that cleared) for major load regions in the PJM, ISO-NE, and MISO markets. As this figure indicates, capacity prices have been volatile over the first decade of market operation, at least for PJM and ISO-NE. With the *Cost of New Entry* estimated to be roughly \$230/MW-day, it is unclear if these markets are providing sufficient revenues to incentivize new resources. This will take additional analysis.³³ MISO, a next-month only capacity market, may not be comparable to PJM's and ISO-NE's markets and also requires additional analysis.

³²As a point of reference, capacity markets in the United Kingdom cover a term of 15 years.

³³Resources are expected to realize margins from the sale of energy and ancillary services and from Scarcity Pricing in some markets. These margins offset the need to otherwise be fully compensated via the capacity markets.

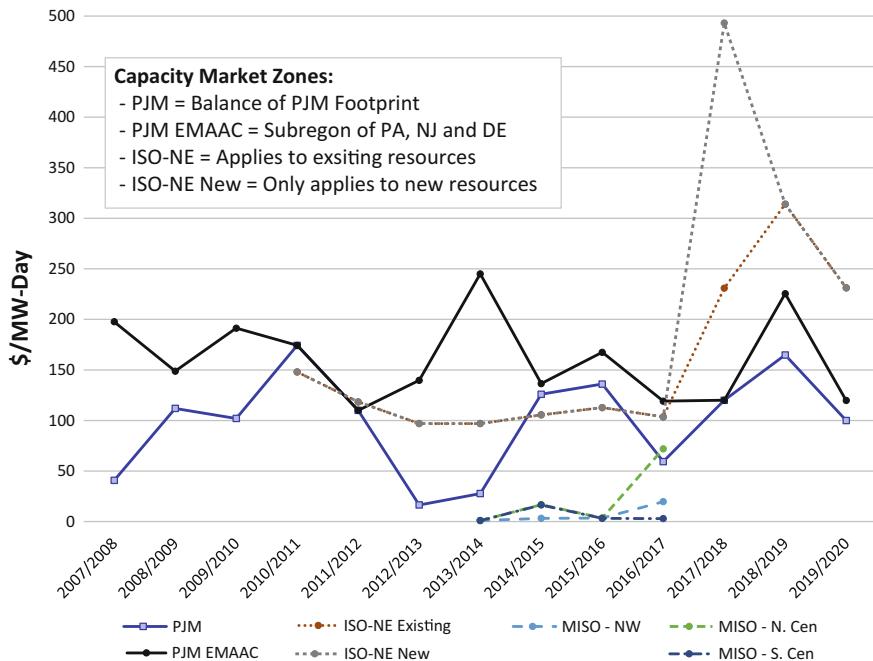


Fig. 6 Capacity market clearing results

6 Analysis and Critique of Methods Used to Incentivize Resource Investments in RTOs

Because the RTOs in the United States cover a significant percentage of all generation and load in the country and because most of these have implemented capacity markets, the economic impact of these markets, and the importance of their design, is significant. The value of the capacity bought and sold each year in the US capacity markets is several, if not tens of billions of dollars.³⁴

The fundamental assumption behind today's RTOs is that if competitive market forces are allowed to take place in the daily and hourly energy and ancillary services markets, the resulting prices (LMPs) will provide all of the incentives needed for supply side resources. That is, these markets alone can optimize the social welfare. The theory supporting this assumption is flawed, and moreover there is no empirical evidence that short-term markets will lead to long-term optimality.

³⁴During the last decade (2001–2010), over 265,000 MW of generating capacity was installed in the United States for an estimated cost of \$199 billion [14].

6.1 Critique of Energy-Only Market Designs

For the energy-only market theory to work, fundamental changes need to occur in the market design.

1. **Eliminate reserve margin criteria** – The reserve margin criteria in place in the markets, wherein defined levels of generating capacity in excess of what is necessary to serve the forecasted peak load, must be relaxed, if not eliminated. Without this change, the markets will rarely be without sufficient installed capacity to serve peak loads.³⁵ However, considering the economic and political ramifications of customers in the United States experiencing frequent and possibly prolonged outages, it is unlikely this criterion can be relaxed.
2. **Provide for direct retail participation** – Retail consumers of electricity must be given the ability, and have the desire, to participate directly in the market. It is claimed that they must in fact respond in real time to market price signals (see Borenstein's survey [6] and Wolak's testimony to the California state government [44].) While significant technological advances have been made (e.g., with *smart meters*), it is still difficult to see how price signals can be of value to customers or of value to the grid operator [30, 31].
3. **Eliminate administrative actions** – Administrative actions (market price caps and reliability-based, out-of-merit dispatches of resources) would have to be eliminated. Both of these have been pointed to as reasons the energy-only market approach is not working properly.

In the unlikely event that any of these changes could be made, we would still be challenged by other issues. Even if regulators go along with the plan, will customers accept the reliability construct required to create the price spikes needed to incentivize generation? Will the trigger prices set by retail customers for limiting service equal or exceed those required by resources to be adequately incentivized? Will demand-side solutions crush any price spikes expected as a result of lower reliability standards? *Like the scarcity pricing construct, without crisis there will be no opportunity!*

Because an energy-only structure would likely operate in a manner similar to that of today's scarcity pricing approaches, it is worthwhile to point out some significant shortcomings in that pricing scheme.

- Scarcity pricing in ERCOT and other markets is linked directly to conditions related to a lack of *operating reserves*³⁶ and not *planning reserves*.³⁷ Therefore,

³⁵Current reserve margin requirements in place across most of the United States (12% to 20% of projected annual peak load) ensure that generation is available to serve load 99.97% of the time.

³⁶*Operating reserves* represent resource capability above firm system demand required to provide for regulation, load forecasting error, equipment forced and scheduled outages, and local area protection.

³⁷*Planning reserves* represent installed capacity above the forecasted peak-hour firm system demand for a defined period in the future.

scarcity conditions occur when generation or transmission resources become unavailable or limited during the *operating conditions*. This rarely relates to the total capacity of resources installed in the market. In theory, the market could have many thousands of megawatts of capacity that are available, but offline and not able to be brought online in sufficient time to resolve the operating reserve deficiency. So, how is such an approach a proper investment signal?

- In the current market approach, scarcity pricing is only applied to generation and loads that deviate from the amount cleared in the day-ahead market. Therefore, any resources that submit offers in the day-ahead energy market, clear this market, exactly generate the amount during the scarcity event that they cleared in that market, and which have no surplus capacity beyond what cleared have no ability to receive the scarcity price for energy. They therefore receive none of these incentive revenues. This is an odd construct given that in most markets resources are expected, if not required, to submit offers into the day-ahead market. Shouldn't the objective be to incentivize all resources (generation, LSEs, and customers) to respond if possible during a scarcity event?
- By design, scarcity pricing is tied to one price (or percent of one price) that is assumed to represent the value to consumers for reliable service. The use of one value for the VOLL, regardless of the time of day, time of year, class of customer, and duration of outage, is inconsistent with a reasonable understanding of this parameter [24].
- Up until recently, LOLP has been used as an annual, long-term planning metric. The application in ERCOT and other markets to operating timescales and conditions is misguided. Therefore, unless the markets develop a mechanism to determine an LOLP-type metric given the exact operating conditions in place during the scarcity conditions, there is no foundation for its use in today's markets.
- Finally, the reliance on the energy markets to provide investment signals is inconsistent with the fundamental marginal cost theory as developed by Dupuit, Hotelling, Coase, et al. Marginal costs cannot be used to incentivize investments when such costs are lower than average costs.

6.2 Critique of Capacity Markets

The capacity markets in use in PJM, ISO-NE, NYISO, and, to a limited degree, MISO also suffer from poor design concepts. These include:

- **Short-term market horizon**— Thirty- to fifty-year investments cannot be optimized in a market that only provides for 1–7-year contracts. The consequence is that potential investors demand higher returns on equity due to the uncertain long-term economics of the arrangements (i.e., they shift the cost of these risks to the consumers), and they will naturally be biased toward resources that have

lower capital costs and higher variable costs (favoring peaking resources over base load resources) [33, 38].

- **Transmission investment coordination**— Current market designs provide no explicit co-optimization with potential transmission improvements nor much, if any, implicit co-optimization. Investors look only to existing and potential transmission topologies to decide on resource locations and have little to no control or influence over what the transmission owners may or may not do in the future. In fact, the value that a generating project believes it can realize through congestion revenues,³⁸ can be completely eliminated by a future transmission project, thus significantly impacting the generating project's long-term economics.
- **Natural gas infrastructure coordination** – Like with transmission, the lack of coordination with natural gas investments can significantly hamper the market from realizing the ultimate economically efficient solution.
- **Reliance on historical energy prices to set capacity price caps** – The demand curves developed by each capacity market and based on the CONE for that market nets out the expected value of energy revenues that such a hypothetical resource would realize in the market. However, these estimates are based on historical prices (PJM, e.g., uses an average of the past 3 years). Because the demand curve covers a period 3 years in the future (for PJM), this means there is a 6-year difference. Such a difference could mean that future energy market prices could be significantly higher or lower than those assumed for the demand curve. These changes could be due to natural gas prices, which could easily double in such a relatively long period of time. This could be either a windfall or a disaster for the investors.

The historical methods used to plan for and optimize resource expansion are not without faults. Projects were sometimes planned and built to serve load that never materialized. Some projects experienced significant cost overruns, with these costs typically passed on to consumers under the cost-of-service, rate-of-return paradigm. Some of these cost overruns were due to changes in regulations during the development and construction of projects, but some were due to poor management, or worse.

The primary complaint about the traditional, rate-of-rate method dealt with the assignment of risk – consumers bore the risk of uneconomical decisions made by utilities. If this can be solved, is this all we need? If this was the only complaint, are there better solutions to incentivizing investments in generation than the myriad approaches in place or proposed in RTOs today?

³⁸Congestion revenues are revenues realized in the energy markets that are associated with occasional or frequent transmission system congestion.

7 Resource Investment Solutions

The solution to the problems identified with today's capacity markets lie in one or more hybrid approaches that borrow from the methods perfected over time by traditional resource planners but which utilize open competitive markets for those aspects that are best suited for such activity.

Three specific options are presented below that implement this proposed construct.

- **Laissez-faire** – Allow LSEs to secure sufficient capacity to meet their capacity requirements (defined as serving their load, plus a level of reserves prescribed of the RTO) by any means they deem acceptable, as long as the capacity acquired meets criteria established by the RTO.
- **Long-term capacity markets** – Require each LSE to have a portion of their capacity requirements (perhaps the majority of their requirements) secured for a longer-term period (e.g., 20+ years instead of 1 year).
- **RTO as the system planner** – Have the RTO plan for the entire market footprint, competitively bid for generation to meet requirements, and allocate costs to market participants.

7.1 *Laissez-Faire*

This approach³⁹ would require that any LSE within the market footprint secure sufficient capacity to meet defined obligations, either through self-build of generating resources or via a bilateral market of willing buyers and sellers. The RTO's role would be to:

- Establish minimum resource requirements – This could be for any period the RTO deems prudent (e.g., the next year, the next 3 years, different percentages of projected peak loads for different future periods, etc.).
- Require that each LSE demonstrate capacity performance via periodic testing and demonstration that the required capacity will be available at the time of the RTO's annual peak demand.
- Validate reserve margin compliance through independent calculations based upon data submitted by LSEs.
- Validate that demand values submitted for reserve margin requirements are consistent with prior year actual data and forecasts.
- Validate that resources are accredited in accordance with the market criteria
- Establish and manage a performance program to ensure reliability criteria are met. If necessary, impose penalties to market participants that are not in compliance.

³⁹The approach described here is similar to that currently in use by SPP, wherein LSEs have specific reserve requirements, but the RTO operates no formal capacity market [39].

7.2 Long-Term Capacity Markets

Under this approach, many of the mechanisms and methods from the current capacity markets would be retained, but LSEs would be required to secure capacity for longer periods of time (e.g., 15–30 years) – if not for all of their projected loads, than at least a percentage of their load. For example: Require that all LSEs secure at least 50% of their forecasted requirements for the next 20 years, at least 75% of their requirements for the next 10 years, at least 90% of their requirements for the next 5 years, and 100% of their requirements for the next 3 years.

The market under this design should also provide for the reselling of capacity as the load forecasts change over time (and potentially less capacity is required), or other capacity resources become available to market participants. This should accommodate the needs of the smaller market participants.

7.3 RTO as System Planner

Under this solution, the RTO will act as a *system planner* and develop an optimal resource expansion plan for the entire market footprint. A competitive process would then follow that would determine who would provide the desired resources and what the final prices for these resources would be.

Like the approach still utilized in the “non-organized” markets in the Western and Southeastern United States, this approach would involve a process where the following are determined through a market-wide planning process that is conducted by the RTO itself.

- Reliability requirements will be established and tracked.
- Future load requirements are forecasted.
- Planned generation and transmission assets are identified and incorporated.
- Long-term analyses are performed by the RTO that identify:
 - the amount of capacity needed,
 - the desired location of the capacity considering existing and potential transmission and natural gas (or other fuel supply) infrastructure, and
 - the desired technology of resources used to provide the capacity (supply or demand side), with proper assessment of risks associated with newer technologies.
- Following development and agreement on a plan for the market, competitive auctions will be held for suppliers to build the desired resources.
- The RTO will contract with the successful bidders for the purchase of capacity and associated energy under long-term (e.g., 20-year+) agreements.
- LSEs will be allocated the cost of capacity required to reliably serve the market based on their load-ratio share.

While we believe the above approaches are all superior to those currently adopted in the capacity markets in the United States today,⁴⁰ it is the last approach that RTOs should implement. Under this best-of-both-worlds approach, the most valuable features of the traditional expansion planning approach would be retained, while competition for the development, ownership, and operation of the resources as seen in today's RTO capacity markets would be preserved.

A key element of this approach parallels what was used by RTOs in the development and current operation of the energy and ancillary services markets. Those markets borrowed heavily from the methods and processes to optimally commit and dispatch resources that were developed and refined in the electric power industry for the 100+ years prior to deregulation of the electric power industry. In many ways, the RTOs took these tried-and-true methods and simply applied them to a larger footprint.⁴¹ And while it is understood that the RTOs have implemented more advanced analysis methods (e.g., involving *security-constrained unit commitment* and *security-constrained economic dispatch*), these are due more to advances in computer hardware and software and the fact that some of the methods only have value over a larger footprint,⁴² than they are due to a better market structure or a new regulatory regime.

With the RTO as the overall system planner, it can incorporate all the “good” from traditional planning experience and take advantage of economies of scale to develop market-wide: (i) load and fuel price forecasts, (ii) technology assessments, (iii) transmission and fuel supply infrastructure studies, and (iv) assessments of political, legal, and regulatory frameworks within which that markets may operate in the future. And like the energy markets, these can all be done using best-in-class systems, models, and methods.

Critics of the traditional approach to expansion of power systems should also be satisfied. While such plans will be developed by the RTO, construction and performance risks will be borne by the independent developers and owners of the resources. Cost overruns and performance penalties will therefore not be directly passed through to eventual customers.

8 Conclusions

The organized markets in the United States continue to develop and implement solutions to incentivize market investments. In many cases, they struggle to find the best solutions as evidenced by the market results and the frequent changes to

⁴⁰SPP is excluded, because it is essentially the laissez-faire approach and not a “formal” capacity market per se.

⁴¹One notable exception is that the traditional energy optimization approach used at least a 7-day optimization period and was therefore not limited to just a day-ahead market.

⁴²For example, co-optimizing transmission use with generation commitment and dispatch.

market rules. Even if the economic theories used to support the market rules were sound, the significant departure in reality from key assumptions appears to make the approaches untenable. The active participation of the great majority of the retail electric consumers necessary to support an energy-only market approach seems elusive and may never materialize. Political pressure to protect consumers have also led to the use of “administrative actions” that have distorted market prices and therefore the signals to resources in the market. Because of these shortcomings (and because of weaknesses in the marginal cost theories used in general), the markets have resorted to the use of proxies to simulate demand-side response. Scarcity pricing is utilized in some form in both the energy markets and capacity markets for several of the RTOs. However, its design and use is less than ideal at best and significantly flawed at worst.

Because, by design, the capacity market solutions used by four of the RTO markets in the United States fail to incorporate many of the engineering principles used in long-term resource acquisition, they are unable to provide adequate investment signals to existing and potential resource owners. This includes the lack of a contract term consistent with the engineering timescales associated with generation investments. They also ignore most, if not all, of the strategic aspects concomitant with long-term planning (impacts on transmission, fuel supply infrastructure, fuel diversity, etc.). The solutions provided address these shortcomings.

Negative aspects of capacity markets has been the focus of this chapter. While short-term energy and ancillary markets are not perfect, the outcome is largely as intended: the RTOs took long-developed engineering methods and approaches used by electric utilities to optimize power supply systems in the short term and simply applied them to larger systems. They have also taken advantage of improvements in enhanced computing power to simultaneously optimize energy, reserves, and transmission over these larger systems – something not possible until relatively recently.

However, unlike the energy and ancillary services markets, the RTOs, either intentionally or unintentionally, ignored most of the long-developed engineering methods and approaches to optimizing power system expansion plans that covered longer time periods. This error adversely impacts the investment incentives for all market participants. In essence, and again unlike the energy markets, the RTOs did not take a well-functioning system and make it better – they have made it worse.

Finally, when it comes to the creation of capacity markets, perhaps it is appropriate to borrow a quote from a popular movie related to the cloning of dinosaurs: policy-makers and power economists “were so preoccupied with whether or not they could, they didn’t stop to think if they should.”⁴³

Acknowledgements This research was supported by the National Science Foundation under grants CPS-1646229 and EPCN-1609131.

⁴³Jurassic Park, 1993

References

1. Anderson D (1972) Models for determining least-cost investments in electricity supply. *Bell J Econ Manag Sci* 3(1):267–299
2. Baumol WJ, Bradford DF (1970) Optimal departures from marginal cost pricing. *Am Econ Rev* 60(3):265–283
3. Bohn RE (1981) A theoretical analysis of customer response to rapidly changing electricity prices. MIT Energy Laboratory, Cambridge
4. Bohn RE (1982) Spot pricing of public utility services. MIT Energy Laboratory, Cambridge
5. Bohn RE, Caramanis M, Schweppe F (1980) Optimal spot pricing of electricity. MIT Energy Laboratory, Cambridge
6. Borenstein S (2009) Electricity pricing that reflects its real-time cost. Technical report, NBER reporter: research summary. <http://www.nber.org/reporter/2009number1/borenstein.html>
7. Bowring J (2013) Capacity markets in PJM. Technical report, Monitoring analytics
8. Bradley RL (2011) Edison to Enron: energy markets and political strategies. Wiley, New York
9. Cho IK, Meyn SP (2010) Efficiency and marginal cost pricing in dynamic competitive markets with friction. *Theor Econ* 5(2):215–239
10. Coase RH (1946) The marginal cost controversy. *Econometrica* 13(51):169–182
11. Covarrubias AJ (1979) Expansion planning for electric power systems. *IAES Bull* 21(2/3):55–64
12. Dupuit AJEJ (1844) De la mesure de l'utilite des travaux publics. *Annales des ponts et chaussees* 2(8)
13. Electricité de France (1965) L'étude a long-terme des investissements a l'aide d'un programme non linéaire: le modèle investissements 85. Technical report, Electricité de France
14. Energy Information Administration (2011) Form EIA-860 annual electric generator report, and form EIA-860M. Technical report, U.S. Energy Information Administration
15. Feldstein M (1972) Equity and efficiency in public sector pricing: the optimal two-part tariff. *Q J Econ* 86:175–187
16. FERC. Guide to Market Oversight: Glossary. <https://www.ferc.gov/market-oversight/guide/glossary.asp>
17. FPL (2016) Florida Power & Light Company's 2016 Ten Year Power Plant Site Plan. Technical report, Florida Power & Light
18. Geddes RR (1992) Historical perspective on electric utility regulation. *CATO Review of Business and Government*
19. Geuss M (2017) \$7.5 billion kemper power plant suspends coal gasification. arsTechnica. www.arsTechnica.com/business/2017/06/7-5-billion-kemper-power-plant-suspends-coal-gasification/
20. Hogan WW (2013) Electricity scarcity pricing through operating reserves. *Econ Energy Environ Policy* 2(2):65–86
21. Hotelling H (1937) The general welfare in relation to problems of taxation and of railway and utility rates. *Econometric Society*, Evanston, IL
22. Hotelling H (1939) The relation of prices to marginal costs in an optimum system. *Econometrica* 7(2):151–155
23. Joskow PL (2013) Symposium on ‘capacity markets’. *Econ Energy Environ Policy* 2(2):v–vi
24. Keane D (1995) Outage cost estimation guidebook: EPRI Research Project 2878-04 Final Report. Technical report, EPRI
25. Lerner AP (1944) The economics of control: principles of welfare economics. Macmillan. www.books.google.com/books?id=2kcNAQAAIAAJ
26. Marshall A (1890) Principals of economics. Macmillan, New York.
27. Mas-Colell A (1995) Microeconomic theory. Oxford University Press, Oxford
28. McCallion K (1995) Shoreham and the rise and fall of the nuclear power industry. Praeger. www.books.google.com/books?id=dHvLpbSYMacC
29. Meade JE (1944) Price and output policy of state enterprise. *Econ J* 215:321–339

30. Negrete-Pincetic M (2012) Intelligence by design in an entropic power grid. PhD thesis, University of Illinois at Urbana Champaign, University of Illinois, Urbana, IL. Meyn S, chair, Sauer PW, de Castro L, Dominguez-Garcia AD, Shanbhag VK
31. Negrete-Pincetic M, Meyn S (2012) Markets for differentiated electric power products in a Smart Grid environment. In: IEEE power and energy society general meeting, pp. 1–7
32. Nelson JR (ed) (1964) Marginal cost pricing in practice. Prentice-Hall International, Englewood Cliffs, NJ
33. Petitet M (2016) Effects of risk aversion on investment decisions in electricity generation: what consequences for market design? In: Proceedings of the 13th international conference on the European energy market
34. Pinchot G (1945) Long struggle for effective federal water power legislation. George Wash Law Rev 14:9
35. Pope D (2011) Nuclear implosions - the rise and fall of the Washington public power system. Cambridge University Press, Cambridge
36. Schroder T, Kuckshinrichs W (2015) Value of lost load: an efficient economic indicator for power supply security? A literature review. Front Energy Res 3:55
37. Scheppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. Kluwer Academic, Dordrecht
38. Spence DB (2017) Naïve electricity markets. In: IMA volume on the control of energy markets and grids. Springer, Berlin
39. Staff S (2015) Load responsible entity for reserve margin obligation (presentation). Technical report, Southwest Power Pool
40. Tabors RD, Kirkley JL (1981) Homeostatic control: the utility/customer marketplace for electric power. MIT Energy Laboratory, Cambridge
41. Treinen R (2004) Locational marginal pricing (LMP): basics of nodal price calculations www.caiso.com/docs/2004/02/13/200402131607358643.pdf. CAISO market operations presentation
42. Turvey R (1964) Marginal cost pricing in practice. Econometrica 31(124):426–432
43. Wang G, Negrete-Pincetic M, Kowli A, Shafeeoorfard E, Meyn S, Shanbhag UV (2012) Dynamic competitive equilibria in electricity markets. In: Chakrabortty A, Illic M (eds) Control and optimization methods for electric smart grids. Springer, Berlin, pp 35–62
44. Wolak F (2011) Capturing the benefits of California's energy infrastructure investments. Technical report, Stanford University. https://web.stanford.edu/group/fwolak/cgi-bin/sites/default/files/files/little_hoover_testimony_wolak_sept_2011.pdf. Accessed June 2017

A Swing-Contract Market Design for Flexible Service Provision in Electric Power Systems



Wanning Li and Leigh Tesfatsion

Abstract The need for flexible service provision in electric power systems has dramatically increased due to the growing penetration of variable energy resources, as has the need to ensure fair access and compensation for this provision. A swing contract facilitates flexible service provision with appropriate compensation because it permits multiple services to be offered together in bundled form with each service expressed as a range of possible values rather than as a single point value. This paper discusses a new swing-contract market design for electric power systems that permits swing contracts to be offered by any dispatchable resource. An analytical optimization formulation is developed for the clearing of a swing-contract day-ahead market that can be implemented using any standard mixed-integer linear programming solver. The practical feasibility of the optimization formulation is demonstrated by means of a numerical example.

1 Introduction

The increased penetration of variable energy resources in electric power markets has increased the volatility of *net load* (i.e., load minus non-dispatchable generation) as well as the frequency of strong ramp events. *Variable energy resources (VERs)* are renewable energy resources, such as wind and solar power, whose generation cannot be closely controlled to match changes in load or to meet other system requirements.

In consequence, flexibility in ancillary service provision has become increasingly important to maintain the reliability and efficiency of power system operations.

W. Li

Department of Electrical and Computer Engineering, Coover Hall, Iowa State University, Ames, IA 50011, USA

L. Tesfatsion (✉)

Department of Economics, Heady Hall 260, Iowa State University, 518 Farm House Lane, Ames, IA 50011-1054, USA

e-mail: tesfatsi@iastate.edu

This has encouraged power system operators to introduce new products and market processes designed to permit more flexibility in ancillary service provision, thus enhancing net load following capability [12].

Nevertheless, three important issues arising from increased VER penetration still need to be resolved. First, power and reserve products are variously defined and compensated across the different energy regions; see, e.g., [10]. This lack of standardization makes it difficult to compare and evaluate the reliability, efficiency, and fairness of system operations across these regions.

Second, product definitions are specified in broad rigid terms (e.g., capacity, energy, ramp rate, regulation, spinning reserve). These rigid categorizations do not permit resources to be further differentiated and compensated on the basis of additional valuable flexibility in service provision, such as an ability to ramp up and down between minimum and maximum values over very short time intervals.

Moreover, the valued services provided by energy resources in power systems largely arise from one source: generated power paths. Since the attributes of power paths are highly correlated, attempts to unbundle these attributes into separately defined and priced products are conceptually problematic. For example, how can “ramp rate” be properly valued apart from a consideration of other power path attributes, such as start time, duration, and power range?

Third, attempts to accommodate new products have led to the introduction of *out-of-market (OOM)* compensation processes. In 2011 the US *Federal Energy Regulatory Commission (FERC)* issued Order 755 to address OOM payment problems for one particular product category in US centrally managed wholesale power markets, namely, regulation with different abilities to follow electronic dispatch signals with high accuracy [11]. However, given its limited scope, Order 755 does not fully eliminate the need in these markets to resort to OOM processes. As stressed in [4], the additional complexity resulting from OOM compensation processes provides increased opportunities for market participants to gain unfair profit advantages through strategic behaviors.

A group of researchers has been working to develop a new swing-contract market design for electric power systems that permits greater flexibility in service provision while at the same time addressing the above three issues [14, 22]. This work builds on important earlier work [2, 3, 7, 19] that stresses the relevance of options and two-part pricing contracts for electricity transactions.

The *swing contract (SC)* proposed in [14, 22] permits a resource with dispatchable power to offer into an electric power market a collection of available power paths with a wide range of specified services, such as location, start time, power level, ramp rate, duration, and volt/VAR support. Each of these services can be offered as a range of values rather than as a point value, thus permitting greater flexibility in real-time implementation to meet both power and reserve needs. Moreover, permitting the resource to offer its services into the market in bundled form, as a collection of available power paths, helps to ensure that all of its valued services receive appropriate compensation.

Simple examples are used in [14, 22] to illustrate how the trading of SCs could be supported by a sequence of linked centrally managed forward markets in a

manner that permits efficient real-time balancing of net load subject to system and reserve requirement constraints. In comparison with existing wholesale electric power market designs, the following key policy implications of this SC market design are highlighted:

- permits full market-based compensation for availability and performance
- facilitates a level playing field for market participation
- facilitates co-optimization of power and reserve markets
- supports forward market trading of power and reserve
- permits service providers to offer flexible service availability
- provides system operators with real-time flexibility in service usage
- facilitates accurate load forecasting and following of dispatch signals
- permits resources to internally manage unit commitment and capacity constraints
- permits the robust-control management of uncertain net load
- eliminates the need for out-of-market payment adjustments
- reduces the complexity of market rules

Left unresolved in this previous conceptual work, however, is whether the determination of optimal market-clearing solutions for SCs can be reduced to a routine operation suitable for real-world application. The present study provides an affirmative answer to this question for a general SC day-ahead market design permitting swing contracts to be offered by any dispatchable resource.¹

Section 2 presents and motivates an illustrative form of SC permitting the flexible provision of power and reserve services in electric power markets. The basic operational features of existing US day-ahead and real-time market designs are outlined in Section 3. Section 4 discusses in broad terms a new market design for the support of SC trading, with a particular focus on a centrally managed SC day-ahead market design that permits SCs to be offered by any dispatchable resource. Key distinctions between this SC day-ahead market design and existing US day-ahead market designs are highlighted.

Section 5 then presents a new optimization formulation for the market clearing of SCs in the SC day-ahead market. This formulation constitutes a *mixed-integer linear programming (MILP)* problem that can be solved by means of the same MILP solution software currently in use for standard security-constrained unit commitment optimization formulations [5, 13, 20, 24]. A numerical example is provided in Section 6 to demonstrate the practical feasibility of this new optimization formulation.

Concluding remarks are given in Section 7. A nomenclature table listing symbols and symbol definitions, Table 5, is provided in an appendix.

¹The present study is a substantial extension of an earlier preliminary study [17] by the authors appearing in an electronic conference proceedings.

2 An Illustrative Swing Contract in Firm Form

Four types of contracts are proposed in [14] to facilitate power and reserve trading, namely, firm contracts and option contracts taking either a fixed or swing form. A *firm contract (FC)* imposes specific obligations on the buyer and seller regarding how the buyer will procure services from the seller in accordance with contractually specified terms. In contrast, an *option contract (OC)* gives the buyer the right, but not the obligation, to procure services from the seller in accordance with contractually specified terms. The right can be activated by exercise of the OC at a contractually permitted exercise time, at which point the contractual terms of the OC become firm.

An FC or OC is a *fixed contract* if each of its offered services is expressed as a single value. An FC or OC is a *swing contract (SC)* if at least one of its offered services is expressed as a set of possible values, thus permitting some degree of flexibility in its implementation.

For concreteness, this study focuses on SCs in firm form that offer a particular spectrum of services expressed in time-domain terms.² The form of these SCs is as follows:

$$\text{SC} = [b, t_s, t_e, \mathcal{P}, \mathcal{R}, \phi] \quad (1)$$

b = location where service delivery is to occur;

t_s = power delivery start time;

t_e = power delivery end time;

$\mathcal{P} = [P^{\min}, P^{\max}]$ = range of power levels p ;

$\mathcal{R} = [-R^D, R^U]$ = range of down/up ramp rates r ;

ϕ = Performance payment method for real-time services.

In (1), the location b would typically refer to a bus or node of a transmission grid. The times t_s and t_e denote specific calendar times expressed at the granularity of time periods of length Δt (e.g., 1 hour, 1 minute), with $t_s < t_e$. The power interval bounds $P^{\min} \leq P^{\max}$ can represent pure power injections (if $0 \leq P^{\min}$), pure power withdrawals or absorptions (if $P^{\max} \leq 0$), or bidirectional power capabilities (if $P^{\min} \leq 0 \leq P^{\max}$). The down/up limits $-R^D$ and R^U for the ramp rates r (MW/ Δt) are assumed to satisfy $-R^D \leq 0 \leq R^U$.

The location b , the start time t_s , and the end time t_e are all specified as single values in (1). However, the power levels p and the down/up ramp rates r are specified in swing form with associated ranges \mathcal{P} and \mathcal{R} .

The performance payment method ϕ designates the mode of ex post compensation to be paid to the seller of the SC if this seller is called upon to provide actual services. This performance payment method can take a wide variety of forms.

²As stressed in [1], the services extracted from resources can alternatively be expressed in terms of their frequency bandwidth characteristics. The general concept of a swing contract does not depend on the exact manner in which services are characterized.

For example, ϕ could be a flat-rate price (\$/MWh) to be applied to the total amount of energy (MWh) injected into the grid between t_s and t_e . Alternatively, ϕ could specify that the price (\$/MWh) to be paid for power (MW) injected into the grid between t_s and t_e is contingent on the realization of some future event, such as the spot price of fuel between t_s and t_e . Also, ϕ might include a metric for the compensation of ramping, such as some form of “mileage” metric based on the length of any delivered down/up power path.³ In addition, ϕ could include penalty or incentive payments to encourage accurate following of dispatch instructions between t_s and t_e , thus permitting a market-based determination of these payments.⁴

To understand the obligations of the seller and buyer of an SC (1), should it be cleared, a numerical example might be helpful. Consider the following SC offered for sale in an ISO-managed day-ahead market by a market participant m in return for a requested *availability price* $\alpha = \$100$,⁵ where $\Delta t = 1$ hour.

$$b = \text{bus } b;$$

$$t_s = 8:00\text{am};$$

$$t_e = 10:00\text{am};$$

$$\mathcal{P} = [P^{\min}, P^{\max}] = [10\text{MW}, 40\text{MW}];$$

$$\mathcal{R} = [-R^D, R^U] = [-38\text{MW/h}, 28\text{MW/h}];$$

$$\phi = \$35/\text{MWh}.$$

This SC implies that market participant m is offering to provide power at bus b from 8:00am to 10:00am on the following day. The power levels at which m is willing to be dispatched range from 10MW to 40MW, but the required down/up ramp rates r to achieve these power levels must satisfy $-38\text{MW/h} \leq r \leq 28\text{MW/h}$. The performance payment method ϕ designates that m is to be paid the price $\phi = \$35/\text{MWh}$ for each MWh of energy it delivers under this SC.

³For example, CAISO defines the *mileage* of a planned power path for a dispatchable resource to be the summation of the absolute changes in the successive *automated generation control* (AGC) set points that are used to communicate power dispatch instructions to this resource [15].

⁴Current penalties for failure to follow dispatch instructions are administratively determined. For example, CAISO uses a comparison of AGC set points to actual telemetry in order to judge the accuracy with which dispatch instructions have been followed. It then adjusts mileage payments when a resource fails to provide the power movements called for by dispatch instructions [15].

⁵The availability price α requested by the seller of an SC, i.e., the SC’s offer price, is not considered to be part of the SC itself. In economics, physical commodities (e.g., apples) are considered separately from their offer prices. Similarly, standardized financial contracts (e.g., bonds) are treated as commodities that can be purchased in various market settings at possibly varying offer prices. In principle, this separation between a commodity/contract and its offer price facilitates price competition among commodity/contract sellers, thus increasing the likelihood that prices will be driven to efficient levels. See [21] for further discussion of this point.

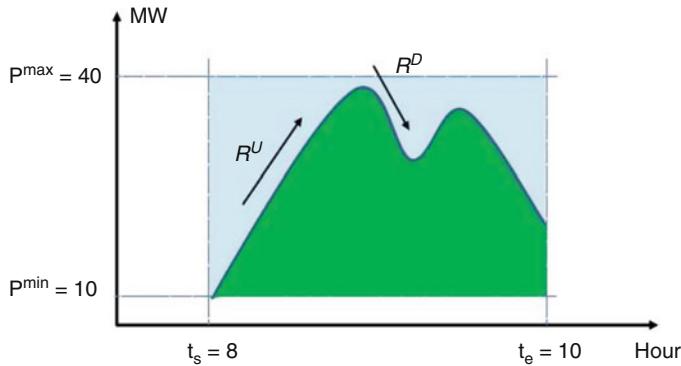


Fig. 1 A possible dispatched power path for the SC numerical example

Suppose the ISO announces that this SC has been cleared. The seller m is then immediately entitled to receive its availability price $\alpha = \$100$. In return for this payment, m is “committed” for next-day operations in the following sense: m is obligated to ensure it will be available to perform the services promised in its cleared SC if called upon to do so in next-day operations between 8:00am and 10:00am. In turn, the ISO is obligated to ensure that m is compensated fully, ex post, for any such service performance, in accordance with m ’s performance payment method ϕ .

Figure 1 depicts one possible power path that the ISO could dispatch in real-time operations, in accordance with the terms of this SC. The darker (green) area under this power path is the resulting energy (MWh) delivery, to be compensated ex post at the rate of \$35/MWh.

It is the responsibility of market participant m to ensure it is able to fulfill the terms of this offered SC. Two aspects must be considered: physical feasibility and financial feasibility. With regard to physical feasibility, the power delivery start time $t_s = 8:00\text{am}$ must precede the power delivery end time $t_e = 10:00\text{am}$, which is clearly the case. In addition, $[t_e - t_s] = 2h$ must be at least as great as m ’s minimum up time.⁶

With regard to financial feasibility, market participant m should make sure that all of its “*avoidable costs*” are covered. *Avoidable costs* are costs that can be avoided if an activity is not undertaken but that are incurred if it is undertaken.

Specifically, market participant m should make sure that its offered availability price $\alpha = \$100$ covers all of the avoidable costs that m would have to pay in order to guarantee service availability. Also, m should make sure that its offered performance payment price $\phi = \$35/\text{MWh}$ is sufficient to cover all avoidable costs that m would have to pay if called upon to perform actual services. Examples of avoidable service availability costs include avoidable *unit commitment (UC)* costs, such as start-up/shut-down and no-load costs, as well as *lost-opportunity*

⁶To help ensure the physical feasibility of offered SCs, an ISO might want to require all offered SCs to include in their performance payment methods some type of standardized failure-to-perform penalties. The severity of these penalties could be conditioned on the severity of past and current transgressions.

costs arising from m 's inability to receive revenues for its services in a next-best alternative use. Examples of avoidable service performance costs include avoidable costs for fuel and labor time.

3 Existing US Wholesale Power Market Designs

As depicted in Figure 2, seven US energy regions (CAISO, ERCOT, ISO-NE, MISO, NYISO, PJM, SPP) encompassing over 60% of US generation capacity currently have centrally managed wholesale power markets.⁷ Although specific market rules differ across these seven energy regions, particularly with regard to reserve procurement, their basic operational design can be roughly summarized as follows:

Private *generation companies* (*GenCos*) sell bulk power to other private companies called *load serving entities* (*LSEs*), who in turn resell this power to retail customers. The transactions between the *GenCos* and *LSEs* take place within a wholesale power market consisting of a *day-ahead market* (*DAM*) and a *real-time market* (*RTM*), operating in parallel, which are centrally managed by an *Independent System Operator* (*ISO*) or *Regional Transmission Organization* (*RTO*). Day-ahead generation schedules are determined in the *DAM* based on estimated next-day net

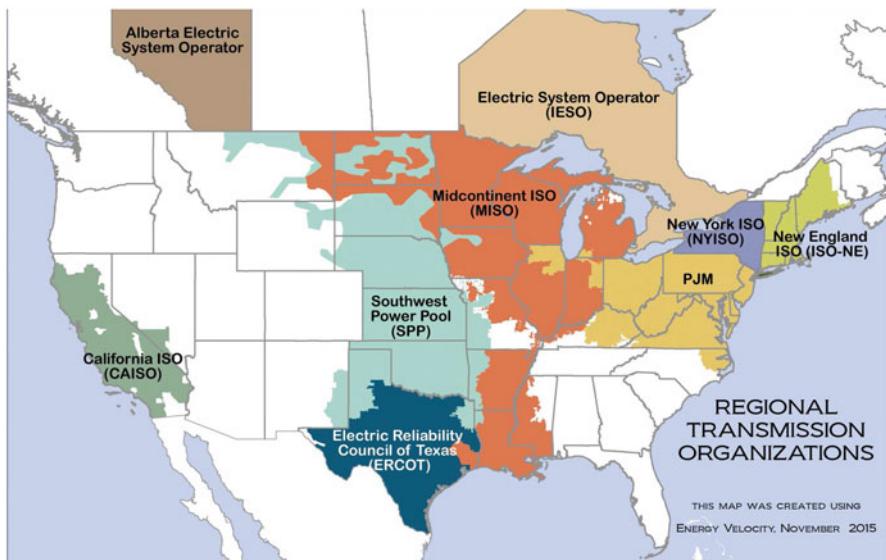


Fig. 2 Energy regions in North America that have ISO/RTO-managed wholesale power markets. Public domain source: [8]

⁷For background readings on current US wholesale power market operations pertinent for issues raised in the current study, see [6, 9, 16, 18], and [23].

loads. Any discrepancies that arise between DAM generation schedules for next-day operations and actual next-day needs for generation based on actual next-day net loads are handled in the RTM, which thus functions as a real-time balancing mechanism.⁸

The physical power flows underlying these transactions take place by means of a high-voltage transmission grid that remains centrally managed by the ISO/RTO in order to ensure open access at reasonable access rates. Transmission grid congestion is managed in the DAM and RTM by *locational marginal pricing (LMP)*.⁹

During the morning of each day d , the GenCos and LSEs submit into the DAM a collection of power supply offers and power demand bids, respectively, for all 24 hours h of day $d+1$. Given these offers and bids, the ISO/RTO solves *security-constrained unit commitment (SCUC)* and *security-constrained economic dispatch (SCED)* optimization problems subject to standard system constraints¹⁰ in order to determine the following planned outcomes at each transmission grid bus b for each hour h of day $d+1$: (i) GenCo unit commitments, (ii) scheduled dispatch levels (MW) for committed GenCos, and (iii) a locational marginal price $\pi^{DAM}(b, h, d)$ (\$/MWh). A committed GenCo located at bus b is paid $\pi^{DAM}(b, h, d)$ for each MW of power it is scheduled to inject at b during hour h of day $d+1$, and an LSE must pay $\pi^{DAM}(b, h, d)$ for each MW of power its retail customers are scheduled to withdraw at bus b during hour h of day $d+1$.

The ISO/RTO undertakes an RTM SCED optimization at least once every five minutes during each day d . At the start of an RTM SCED on any day d , immediately prior to some operating period t , the ISO/RTO forecasts the net load for t . The ISO/RTO then conducts the RTM SCED optimization to resolve any discrepancies between the dispatch schedule determined in the day-($d-1$) DAM for t on day d and the ISO/RTO's current forecasted net load for t on day d . Any dispatch adjustment and/or load curtailment needed to ensure load balancing at a particular bus b for operating period t on day d is settled at the LMP determined for bus b in the RTM SCED optimization conducted for operating period t on day d .

For later purposes, four key features of this existing wholesale power market design need to be stressed. First, the design does not provide for the coverage of UC costs through market-based processes. Rather, start-up/shut-down, no-load, and

⁸ A GenCo is an entity that produces (supplies) power for an electric power grid. The term *load* is used in two senses: (i) to refer to an entity that consumes (absorbs) power from an electric power grid and (ii) to refer to the power demands of such entities. The term *net load* is defined to be power demand net of non-dispatchable generation, such as wind or solar power. An LSE is an entity that secures power, transmission, and related services at the wholesale level in order to service the load (power demands) of its retail customers. An ISO/RTO is an organization charged with the primary responsibility of maintaining the security of an electric power system and often with system operation responsibilities as well. The ISO/RTO is required to be independent, meaning it cannot have a conflict of interest in carrying out these responsibilities, such as an ownership stake in generation or transmission facilities within the power system.

⁹ LMP is the pricing of electric power according to the timing and location of its withdrawal from, or injection into, an electric power grid.

¹⁰ These system constraints include power balance constraints, line and generation capacity limits, down/up ramping restrictions, minimum down/up-time requirements, and reserve requirements.

other forms of UC costs incurred by GenCos are compensated by various forms of *out-of-market (OOM)* payments, generally referred to as *uplift payments*.

Second, DAM/RTM settlements (including uplift payments) do not carefully distinguish between avoidable costs and unavoidable (sunk) costs. All of the avoidable costs incurred by DAM/RTM market participants due to their fulfillment of DAM/RTM service obligations should be compensated through DAM/RTM settlements. However, the unavoidable costs of these participants – i.e., the costs they would incur whether or not they participated in the DAM/RTM – should not be compensated through DAM/RTM settlements.

Third, settlement obligations for scheduled next-day service performance are incurred in the DAM in advance of actual service performance. These DAM settlement obligations are based on DAM net load estimates formed from LSE demand bids and from ISO/RTO forecasts for next-day non-dispatchable generation. Thus, subsequent RTM dispatch and settlement adjustments are typically needed in order to balance *actual* next-day net loads. Having multiple points in time (DAM, RTM) at which settlement obligations are incurred for the same operating period increases the chance that market inefficiency (deadweight loss) will arise.

Fourth, considered together, the above three features result in extremely complex market rules. This, in turn, opens up opportunities for market gaming.

4 The SC DAM Design: Overview

As discussed in [14, 22], swing-contract (SC) trading can be supported by a sequence of linked centrally managed forward markets whose planning horizons range from years to minutes. Forward markets with very long planning horizons can be used to encourage new capacity investment, while forward markets with very short planning horizons can be used to correct last-minute imbalances between available generation and forecasted real-time net loads.

In this study, for concreteness, we demonstrate how an ISO-managed *SC DAM* can be designed that permits SC trading by the set \mathbb{M} of all market participants (MPs) with dispatchable resources. The entities in \mathbb{M} can include GenCos, demand response resources (DRRs),¹¹ electric storage devices (ESDs), and dispatchable variable energy resources (VERs). Additional market participants include non-dispatchable VERs and LSEs with fixed (must-serve) loads.

To retain the ISO's nonprofit status, all costs incurred by the ISO for SC procurement must be passed through to market participants. This cost pass-through

¹¹An example of a DRR would be an entity that manages a collection of *distributed energy resources (DERs)*, such as household appliances. Even if individual DERs have relatively small amounts of down/up flexibility in their power usage due to local goals and constraints, a sufficiently large collection of these DERs could permit the extraction of down/up demand response services with substantial flexibility.

		Current DAM	Proposed SC DAM
Similarities		<ul style="list-style-type: none"> Conducted day-ahead to plan for next-day operations ISO-managed MPs can include GenCos, DRRs, ESDs, VERs, & LSEs Subject to same physical constraints: e.g. transmission, generation, ramping, & power-balance constraints 	
Differences	Optimization formulation	SCUC & SCED	Contract-clearing
	Settlement	Locational marginal pricing	Availability prices
	Payment	Payment for next-day service before actual performance	Payment for availability now & performance ex post
	Out-of-market payments	Uplift payments (e.g., for unit commitment costs)	No out-of-market payments
	Information given to MPs	Unit commitments, LMPs, & next-day dispatch schedule	Which contracts have been cleared

Fig. 3 Comparison of the SC DAM design with current DAM designs

could simply require all procurement costs to be allocated to the LSEs in proportion to their share of real-time loads. However, the presence of performance payment methods ϕ in offered SCs permits more sophisticated cost-sharing arrangements. For example, reserve requirement costs could arise in part due to the inability of some resources with cleared SCs to follow dispatch instructions with high accuracy. The ISO could require standardized failure-to-perform penalties to be included in the performance payment methods of SCs to help defray these costs.

Figure 3 provides a summary comparison of our proposed SC DAM design to current DAM designs. The basic features characterizing current DAM designs are explained in Section 3. To understand the similarities and differences highlighted in Figure 3, it is important to recall the key attributes of SCs discussed in Section 2. These key attributes are summarized below:

- (i) The swing in the contractual terms of SCs permits these contracts to function as both power and reserve products. This eliminates the need to provide separate pricing and settlement processes for power versus reserve services.
- (ii) The two-part pricing of SCs permits full separate market-based compensation for service availability and service performance. The availability price of an SC permits the seller to be compensated for all avoidable costs associated with service availability, while the performance payment method included among the terms of an SC permits the seller to be compensated ex post for all avoidable costs arising from actual real-time service provision.
- (iii) SCs require sellers to internally manage unit commitment and generation capacity constraints for their resources. By offering an SC into an SC DAM,

a seller is communicating to the ISO in charge of this SC DAM that it can feasibly perform the services represented in the SC if called upon to do so.

- (iv) The performance payment method ϕ included among the contractual terms of an SC can designate special incentives and/or penalties to assure the ISO that the seller of the SC will fulfill the terms of the SC if the SC is cleared.

5 The SC DAM Design: Analytical Formulation

5.1 SC DAM Analytical Formulation: Summary Description

As discussed in Section 3, current DAM designs rely on standard SCUC/SCED optimizations to determine unit commitment, economic dispatch, and pricing solutions. In a sharp break from this practice, we propose a new analytical optimization formulation for the SC DAM that permits the optimal clearing of SCs.

Figure 4 highlights key distinctions between our proposed optimization formulation for the SC DAM and traditional SCUC/SCED optimization formulations. Section 5.2 clarifies these distinctions by setting out our proposed SC DAM optimization formulation in concrete equation form.

		SCUC	SCED	Proposed SC DAM
Similarities		<ul style="list-style-type: none"> Both SCUC and the proposed SC DAM are solved as mixed integer linear programming (MILP) problems subject to physical constraints 		
Differences	• Objective	Min {Start-Up /Shut-Down Costs + No-Load Costs + Dispatch Costs + Reserve Costs}	Min {Dispatch Costs + Reserve Costs}	Min {Availability Cost + Expected Performance Cost}
	• Start-up & shut-down constraints	Yes	No	Start-up/shut-down constraints are implicit in submitted contracts
	• Primary decision variables	Unit commitments	Energy dispatch & reserve levels	Cleared contracts
	• Settlement	No	LMPs calculated as SCED dual variables	Availability prices paid for cleared contracts

Fig. 4 Comparison of the SC DAM optimization formulation with current SCUC/SCED DAM optimization formulations

5.2 SC DAM Analytical Formulation: Equations

Consider an ISO-managed SC DAM to be optimally cleared over a set $\mathbb{T} = \{1, \dots, T\}$ of successive next-day operating periods t with length Δt . For clarity of exposition, five assumptions are made.

First, it is assumed that all loads serviced by the LSEs are fixed (must-serve) loads that do not provide dispatchable services. Second, it is assumed that LSE demand bids have a simple block-energy form, i.e., an LSE's demand bid for any given period t consists of a power demand (MW) that is not responsive to price. Third, it is assumed that each market participant m with dispatchable resources, i.e., each $m \in \mathbb{M}$, offers a single swing contract SC_m into the SC DAM, where SC_m takes form (1).¹² Fourth, it is assumed that the performance payment method ϕ_m appearing within SC_m takes the form of a collection of flat-rate energy prices $\phi_m(t)$ (\$/MW Δt), one price for each $t \in \mathbb{T}$. Fifth, it is assumed that only system-wide down/up spinning reserve requirements are imposed; contingency reserve requirements for generator or line outages are not considered.¹³

Given these simplifications, the objective of the ISO managing the SC DAM reduces to the minimization of total cost (\$) over \mathbb{T} subject to system constraints. *Total cost* is the summation of SC availability cost plus expected performance cost arising from the need to balance expected net loads $\{NL_b(t) : b \in \mathbb{B}, t \in \mathbb{T}\}$ as determined by LSE demand bids and ISO-forecasted generation from non-dispatchable VERs. Total cost is expressible as follows:¹⁴

$$\sum_{m \in \mathbb{M}} \alpha_m c_m + \sum_{t \in \mathbb{T}} \sum_{m \in \mathbb{M}} \phi_m(t) |p_m(t)| \Delta t \quad (2)$$

The ISO minimizes (2) by appropriate selection of the following ISO decision variables:

- *Market participant contract clearing indicators:*

$$c_m \in \{0, 1\}, \quad \forall m \in \mathbb{M}$$

- *Market participant power dispatch levels:*

$$p_m(t), \quad \forall m \in \mathbb{M}, \quad t \in \mathbb{T}$$

¹²See [14] for a discussion of the more general case in which offers can take the form of portfolios consisting of multiple SCs.

¹³As discussed in [14], option SCs seem to be a more suitable vehicle than firm SCs for handling contingency reserve requirements.

¹⁴See Table 5 in the Appendix for definitions of all terms appearing in the following equations. Although power levels $p_m(t)$ for all market participants $m \in \mathbb{M}$ nominally appear in the objective function (2), it will be seen below that the constraints for this SC DAM optimization formulation restrict the power amounts for market participants with non-cleared SCs to be zero.

- *Bus voltage angles:*

$$\theta_b(t), \forall b \in \mathbb{B}, t \in \mathbb{T}$$

The system constraints for the minimization of (2) are as follows:

ISO decision variable bounds:

$$c_m \in \{0, 1\}, \quad \forall m \in \mathbb{M} \quad (3)$$

$$-\pi \leq \theta_b(t) \leq \pi, \quad \forall b \in \mathbb{B}, t \in \mathbb{T} \quad (4)$$

Unit commitment constraints:

$$v_m(t) = c_m \cdot A_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (5)$$

Voltage angle specification at angle reference bus 1:

$$\theta_1(t) = 0, \quad \forall t \in \mathbb{T} \quad (6)$$

Line power transmission constraints:

$$w_\ell(t) = S_o B(\ell) [\theta_{O(\ell)}(t) - \theta_{E(\ell)}(t)], \quad \forall \ell \in \mathbb{L}, t \in \mathbb{T} \quad (7)$$

$$-F_\ell^{max} \leq w_\ell(t) \leq F_\ell^{max}, \quad \forall \ell \in \mathbb{L}, t \in \mathbb{T} \quad (8)$$

Power balance constraints at each bus:

$$\sum_{m \in \mathbb{M}_b} p_m(t) + \sum_{\ell \in \mathbb{L}_{E(b)}} w_\ell(t) = N L_b(t) + \sum_{\ell \in \mathbb{L}_{O(b)}} w_\ell(t), \\ \forall b \in \mathbb{B}, t \in \mathbb{T} \quad (9)$$

Market participant capacity constraints:

$$\underline{p}_m(t) \leq p_m(t) \leq \bar{p}_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (10)$$

$$\bar{p}_m(t) \leq P_m^{max} v_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (11)$$

$$\underline{p}_m(t) \geq P_m^{min} v_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (12)$$

Market participant down/up ramp constraints:

$$\bar{p}_m(t) - p_m(t-1) \leq R_m^U \Delta t v_m(t-1) + P_m^{max} [1 - v_m(t-1)], \\ \forall m \in \mathbb{M}, \forall t = 2, \dots, T \quad (13)$$

$$\begin{aligned} p_m(t-1) - \underline{p}_m(t) &\leq R_m^D \Delta t v_m(t) + P_m^{max}[1 - v_m(t)], \\ \forall m \in \mathbb{M}, \forall t = 2, \dots, T \end{aligned} \quad (14)$$

System-wide down/up spinning reserve requirement constraints:

$$\sum_{m \in \mathbb{M}} \bar{p}_m(t) \geq \sum_{b \in \mathbb{B}} NL_b(t) + RR^U(t), \quad \forall t \in \mathbb{T} \quad (15)$$

$$\sum_{m \in \mathbb{M}} \underline{p}_m(t) \leq \sum_{b \in \mathbb{B}} NL_b(t) - RR^D(t), \quad \forall t \in \mathbb{T} \quad (16)$$

5.3 More Detailed Explanations of Key Terms

The absolute value terms $|p_m(t)|$ appear in the objective function (2) because a market participant m with dispatchable resources might be called upon to provide power curtailments $p_m(t) < 0$ as well as power injections $p_m(t) > 0$ in support of period- t net load balancing requirements. The power curtailments provided by m are assumed to be compensated at the same flat rate $\phi_m(t)$ as m 's power injections.¹⁵

The contract clearing indicator $c_m \in \{0, 1\}$ indicates whether SC_m has been cleared (1) or not (0). The offer service indicator $A_m(t) \in \{0, 1\}$ indicates whether time period t is (1) or is not (0) within the contract service times covered by SC_m .

Note that $A_m(t)$ is a derived value, calculated by the ISO from the information provided within SC_m . Consider, for example, the numerical SC example presented in Section 2. In this example, a market participant m submits an SC consisting of an offer to provide service between 8:00am and 10:00am during the following day. Thus,

$$A_m(t) = \begin{cases} 1 & \text{if } t = 8, 9 \\ 0 & \text{if } t = 1, \dots, 7, 10, \dots, 24 \end{cases}$$

As seen in Section 5.2, the unit commitment constraints take the form

$$v_m(t) = c_m \cdot A_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (17)$$

¹⁵The absolute value terms $|p_m(t)|$ in the objective function (2) do not pose any computational difficulty. Because the goal is to minimize this objective function, these absolute value terms can equivalently be represented in terms of linear inequality constraints, as follows. First, introduce new decision variables for the ISO: $p_m^a(t), \forall m \in \mathbb{M}, t \in \mathbb{T}$. Second, in the objective function (2), replace $|p_m(t)|$ by $p_m^a(t), \forall m \in \mathbb{M}, t \in \mathbb{T}$. Third, include the following additional linear inequality constraints in the constraint set: $p_m^a \geq p_m$ and $p_m^a \geq -p_m, \forall m \in \mathbb{M}, t \in \mathbb{T}$. Any solution for the resulting constrained minimization problem will then require $p_m^a(t) = |p_m(t)|, \forall m \in \mathbb{M}, t \in \mathbb{T}$.

The unit commitment $v_m(t) \in \{0, 1\}$ for each market participant $m \in \mathbb{M}$ in each time period t is thus determined by two factors:

- (a) Has SC_m been cleared by the ISO or not?
- (b) Does SC_m include service for time period t or not?

The contract clearing indicator $c_m \in \{0, 1\}$ represents condition (a), and the offer service indicator $A_m(t) \in \{0, 1\}$ represents condition (b). If conditions (a) and (b) are both met, then m is available to provide service in time period t . Otherwise, if at most one of these conditions is met, m is not available to provide service in time period t .

The market participant capacity constraints take the form

$$\underline{p}_m(t) \leq p_m(t) \leq \bar{p}_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (18)$$

$$\bar{p}_m(t) \leq P_m^{\max} v_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (19)$$

$$\underline{p}_m(t) \geq P_m^{\min} v_m(t), \quad \forall m \in \mathbb{M}, t \in \mathbb{T} \quad (20)$$

Also, the market participant down/up ramp constraints take the form

$$\begin{aligned} \bar{p}_m(t) - p_m(t-1) &\leq R_m^U \Delta t v_m(t-1) + P_m^{\max} [1 - v_m(t-1)], \\ \forall m \in \mathbb{M}, \forall t = 2, \dots, T \end{aligned} \quad (21)$$

$$\begin{aligned} p_m(t-1) - \underline{p}_m(t) &\leq R_m^D \Delta t v_m(t) + P_m^{\max} [1 - v_m(t)], \\ \forall m \in \mathbb{M}, \forall t = 2, \dots, T \end{aligned} \quad (22) \quad (23)$$

The terms $\underline{p}_m(t)$ and $\bar{p}_m(t)$ appearing in constraints (18) through (23) are derived values; they give the run-time lower and upper bounds on down/up power availability from market participant $m \in \mathbb{M}$ in each time period $t = 2, \dots, T$ as a function of the ISO's unit commitment decisions $v_m(t-1)$ and $v_m(t)$.

To see this, note from (18)–(20) that $v_m(t) = 0$ implies $p_m(t) = 0$ for each $t \in \mathbb{T}$. Also, the binary unit commitment vector $(v_m(t-1), v_m(t))$ can take on only one of the four possible value combinations for $t = 2, \dots, T$, namely, $(0,0)$, $(1,0)$, $(0,1)$, or $(1,1)$. Given each of these four possible value combinations, it is straightforward to show that constraints (18) through (23) reduce to a distinct set of restrictions on $(\underline{p}_m(t), \bar{p}_m(t))$ for $t = 2, \dots, T$, as indicated in Table 1.

Finally, it is interesting to note that an “inherent reserve range” can be derived for the power system in each time period t , as a function of the solution for the SC DAM optimization. Define

$$RR^{\max}(t) = \sum_{m \in \mathbb{M}} \bar{p}_m(t), \quad \forall t \in \mathbb{T} \quad (24)$$

$$RR^{\min}(t) = \sum_{m \in \mathbb{M}} \underline{p}_m(t), \quad \forall t \in \mathbb{T} \quad (25)$$

Table 1 Min/max available power output from m under different unit commitment combinations

$v_m(t)$	0	0	1	1
$v_m(t-1)$	0	1	0	1
$\bar{p}_m(t)$	0	0	$\bar{p}_m(t) \leq P_m^{max}$	$\bar{p}_m(t) \leq P_m^{max}$ $\bar{p}_m(t) \leq p_m(t-1) + R_m^U \Delta t$
$\underline{p}_m(t)$	0	0	$\underline{p}_m(t) \geq P_m^{min}$	$\underline{p}_m(t) \geq P_m^{min}$ $\underline{p}_m(t) \geq p_m(t-1) - R_m^D \Delta t$

By construction, the MW amounts $RR^{max}(t)$ and $RR^{min}(t)$ are the maximum and minimum amounts of power available for the system in each time period t during implementation of the SC DAM optimization solution. The *inherent reserve range* (*IRR*) for time period t thus takes the form

$$IRR(t) = [RR^{min}(t), RR^{max}(t)]. \quad (26)$$

5.4 Size Comparison with Standard DAM SCUC Formulations

As noted in Section 3, two optimizations are undertaken in current US ISO/RTO-managed DAMs to determine unit commitment, economic dispatch, and pricing solutions, namely, *security-constrained unit commitment (SCUC)* and *security-constrained economic dispatch (SCED)*. SCUC is formulated as a *mixed-integer linear programming (MILP)* problem, and SCED is formulated as a linear programming problem.

Instead of conducting two optimizations, our proposed new SC DAM optimization uses a single optimization process to determine which SCs are cleared; hence which dispatchable market participants are obligated (committed) to ensure service availability for the following day. As seen in Section 5.2, this SC DAM optimization is formulated as a MILP problem.

The sizes of the standard DAM SCUC MILP problem and the SC DAM MILP problem can be approximately measured by the number of integer decision variables and constraints in their problem formulations. To permit direct comparisons, suppose the current day is d and the planning horizon for each problem consists of all 24 hours h of day $d+1$.

Consider, first, the relative number of integer decision variables. For the DAM SCUC MILP problem, the ISO has 24 integer decision variables (unit commitment indicators) for each market participant m with dispatchable resources, one for each hour h of day $d+1$. In contrast, for the SC DAM MILP problem, the ISO has one integer decision variable (contract clearing indicator) for each market participant m with dispatchable resources that covers the entire 24 hours of day $d+1$.

Now consider the relative number of constraints. For the standard DAM SCUC MILP problem, unit commitment restrictions (e.g., start-up/shut-down, minimum

down/up time) must be included among the MILP problem constraints. In contrast, for the SC DAM MILP problem, each market participant m is responsible for ensuring the physical feasibility of SC_m , its offered swing contract, which requires in particular that all services offered in SC_m must satisfy m 's unit commitment restrictions. Thus, unit commitment restrictions are implicitly imposed through the forms of the submitted SCs; they do not appear among the MILP problem constraints.

Consequently, measured in terms of integer decision variables and numbers of constraints, the size of the SC DAM optimization formulation is smaller than the size of the standard DAM SCUC optimization formulation, substantially so if the number of dispatchable market participants is large.

6 Illustrative Example

This section reports illustrative SC DAM optimization findings for a simple power system with three dispatchable GenCos and no transmission congestion. Each GenCo m submits one swing contract SC_m to the ISO-managed SC DAM, as depicted in Table 2.

Time periods t are measured in hours, and the net load $NL(t)$ for each hour t of the following day is as depicted in Figure 5. The system-wide down/up spinning reserve requirements are set at 10MW below/above net load for each hour t , i.e., $RR^D(t) = RR^U(t) = 10\text{MW}$, for each hour t .

The ISO applies a MILP solver to determine an SC DAM optimization solution for the following day, conditional on the three submitted SCs. Simulation results show that the SCs submitted by GenCo 2 and GenCo 3 are cleared: i.e., $c_{m1} = 0$, $c_{m2} = 1$, and $c_{m3} = 1$. The optimal unit commitment $v_m(t)$ and dispatch level $p_m(t)$ for each GenCo m in each hour t are shown in Tables 3 and 4, respectively.

The DAM prices for the cleared SCs are their submitted availability prices, and the payments to be received for any actual services performed under these SCs the following day are based on the energy prices specified by the cleared SC performance payment methods: that is, $\phi_{m2} = \$10/\text{MWh}$ and $\phi_{m3} = \$20/\text{MWh}$.

The results show that GenCo 2 serves as base load due to its relatively low-performance price, similar to a coal or nuclear plant. The reasons why GenCo 3's

Table 2 SCs submitted by the three GenCos in the illustrative example

GenCo	Service period [$t_s, t_e]$	Power range [P^{min}, P^{max}] (MW)	Ramp rate range [- R^D, R^U] (MW/h)	Performance price ϕ (\$/MWh)	Availability price α (\$)
1	[1, 24]	[0, 80]	[-60, 60]	25	1500
2	[1, 24]	[0, 200]	[-30, 30]	10	2000
3	[8, 24]	[0, 120]	[-50, 50]	20	1000

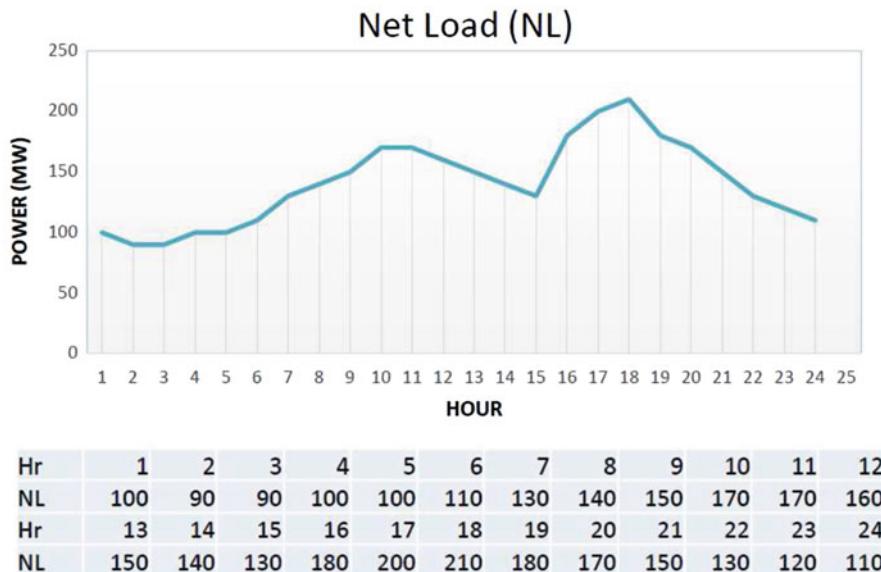


Fig. 5 24-hour net load profile for the illustrative example

Table 3 Optimal SC DAM unit commitments for the illustrative example

		Periods																							
GenCo		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3		0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

submitted SC is also cleared are as follows: First, there is a big ramp-up in net load from hour 15 to hour 16. Due to GenCo 2's limited ramp capability, the maximum available power output for GenCo 2 at hour 16 is 160MW. Thus, GenCo 3 is cleared although it is relatively more expensive. Second, the net load for hour 18 is 210MW, which exceeds GenCo 2's upper output limit 200MW. Thus, GenCo 3 is needed to provide additional power.

Although GenCo 3's available power is not used until hour 16, the unit commitment for GenCo 3 in fact spans from hour 8 to hour 24. The reason for this is that GenCo 3's SC commits this GenCo to be available to provide power from hour 8 through hour 24. Thus, if the ISO clears the contract, GenCo 3 must be synchronized to the grid during each of these hours.¹⁶

Figure 6 depicts the inherent reserve range resulting from the cleared SCs for GenCo 2 and GenCo 3, together with the down/up spinning reserve requirements.

¹⁶As in any market, increased competition among SC providers should reduce the need of an ISO to clear SCs that entail excess resource availability.

Table 4 Optimal SC DAM dispatch schedule (MWs) for the illustrative example

GenCo	Periods											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	0	0	0	0	0	0	0
2	100	90	90	100	100	110	130	140	150	170	160	150
3	0	0	0	0	0	0	1	1	1	1	1	1

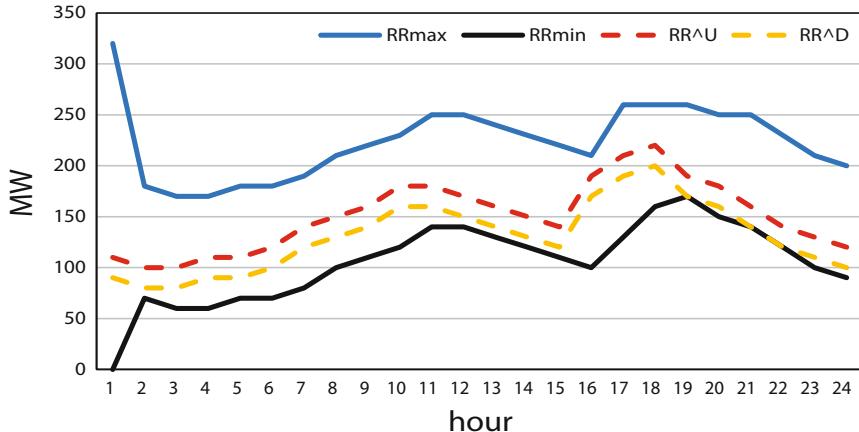


Fig. 6 Comparison of the 24-hour inherent reserve range $IRR = [RR^{min}, RR^{max}]$ depicted by solid lines with the 24-hour down/up spinning reserve requirements RR^D and RR^U depicted by dashed lines

Note that the inherent reserve range satisfies the down/up spinning reserve requirements while at the same time providing valuable additional flexibility to the ISO for use in real-time balancing operations.

7 Conclusion

A new mixed-integer linear programming (MILP) optimization formulation has been developed for an ISO-managed day-ahead market (DAM) based on swing contracting that could facilitate the flexible provision and efficient pricing of power and reserve services. A limitation of the current study is that we have not yet implemented and tested our proposed new SC market design for large-scale systems or for systems involving a DAM and a real-time market (RTM) operating in parallel.

In future work, we will extend our SC market design formulation to encompass combined DAM/RTM operations, and we will undertake systematic feasibility and cost comparisons with existing DAM/RTM operations. We will also explore the potential of swing contracts, offered into wholesale power markets by managers of distributed energy resources, to facilitate the integrated operation of transmission and distribution systems.

Acknowledgements This research has been supported in part by grants from the ISU Electric Power Research Center (EPRC), the Sandia National Laboratories (Contract No. 1163155), and the Department of Energy (DE-AR0000214 and DE-OE0000839). The authors are grateful to the editors and four reviewers for constructive comments and to Zhaoyu Wang and Shanshan Ma for helpful discussions related to the topic of this study.

Appendix

Table 5 Nomenclature table listing symbols and symbol descriptions

Symbol	Description
<i>Sets and intervals:</i>	
\mathbb{B}	Set of bus indices b
$\mathbb{L} \subset \mathbb{B} \times \mathbb{B}$	Set of transmission line indices ℓ
$\mathbb{L}_{O(b)} \subset \mathbb{L}$	Subset of lines ℓ originating at bus b
$\mathbb{L}_{E(b)} \subset \mathbb{L}$	Subset of lines ℓ ending at bus b
\mathbb{M}	Set of indices m for market participants with dispatchable resources
$\mathbb{M}_b \subset \mathbb{M}$	Market participants at bus b with dispatchable resources
\mathcal{P}	Interval of power levels p offered in a swing contract
\mathcal{R}	Interval of ramp rates r offered in a swing contract
\mathbb{T}	Set of time period indices $t = 1, \dots, T$
<i>Parameters and functions:</i>	
$A_m(t)$	1 if m in time period t is within its contract service period; 0 otherwise
$B(\ell)$	Inverse of reactance (pu) for line ℓ
$E(\ell)$	End bus for line ℓ
F_{ℓ}^{\max}	Power limit (MW) for line ℓ
$NL_b(t)$	Net load (MW) at bus b in time period t
$O(\ell)$	Originating bus for line ℓ
P_m^{\min}	Lower power limit (MW) of m
P_m^{\max}	Upper power limit (MW) of m
R_m^D	Ramp-down limit (MW/ Δt) of m
R_m^U	Ramp-up limit (MW/ Δt) of m
$RR^D(t)$	System-wide down spinning reserve requirement (MW) in time period t
$RR^U(t)$	System-wide up spinning reserve requirement (MW) in time period t
S_o	Positive base power (in three-phase MVA)
t_e	Power delivery end time offered in a swing contract
t_s	Power delivery start time offered in a swing contract

(continued)

Table 5 (continued)

Symbol	Description
Δt	Time period length
α_m	Availability price (\$) requested by m for a swing contract that offers service availability
ϕ	Performance payment method for real-time service offered in a swing contract
$\phi_m(t)$	Energy price (\$/MW Δt) used in illustrative SC examples as a simple form of performance payment method for the compensation of real-time down/up power services performed by a market participant m
SC DAM optimization variables:	
c_m	1 if the swing contract offered by m is cleared; 0 otherwise
$v_m(t)$	1 if m is online in time period t ; 0 otherwise
$p_m(t)$	Power output (MW) of m in time period t
$\bar{p}_m(t)$	Maximum available power output (MW) of m in time period t
$\underline{p}_m(t)$	Minimum available power output (MW) of m in time period t
$\theta_b(t)$	Voltage angle (radians) at bus b in time period t
$w_\ell(t)$	Line power (MW) for line ℓ in time period t

References

- Barooah P, Bušić A, Meyn S (2015) Spectral decomposition of demand-side flexibility for reliable ancillary services in a smart grid. In: 48th Hawaii international conference on system sciences (HICSS)
- Bidwell M (2005) Reliability options: a market-oriented approach to long-term adequacy. *Electr J* 18(5):11–25
- Bunn D (2004) Structural and behavioural foundations of competitive electricity prices. In: Bunn D (ed) Modelling prices in competitive electricity markets. Wiley, New York, pp 1–17
- Bushnell JB (2013) JP Morgan and market complexity. Energy economics exchange blog. Online: <http://energyathaas.wordpress.com/2013/08/12/jp-morgan-and-market-complexity>.
- Capitanescu F, Ramos JLM, Panciatici P, Kirschen D, Marcolini AM, Platbrood L, Wehenkel L (2011) State-of-the-art, challenges, and future trends in security constrained optimal power flow. *Electr Power Syst Res* 81(8):1731–1741
- Carrión M, Arroyo J (2006) A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem. *IEEE Trans Power Syst* 21(3):1371–1378
- Chao HP, Wilson R (2002) Multi-dimensional procurement auctions for power reserves: robust-incentive compatible scoring and settlement rules. *J Regul Econ* 22(2):161–183
- EIA (2017) U.S. Energy Information Administration (EIA). <http://www.eia.gov/>
- Ela E, Milligan M, Kirby B (2011) Operating reserves and variable generation. National Renewable Energy Laboratory, Technical report, NREL/TP-5500-51976
- Ellison JF, Tesfatsion LS, Loose VW, Byrne RH (2012) A survey of operating reserve markets in U.S. ISO/RTO-Managed Electric Energy Regions. Sandia National Laboratories report (SAND2012-1000)
- FERC (2011) Frequency regulation compensation in the organized wholesale power markets. Order No. 755: Final Rule
- Hand MM, Baldwin S, DeMeo E, Reilly JM, Mai T, Arent D, Porro G, Meshek M, Sandor D (eds) (2012) Renewable electricity futures study, 4 vols. National Renewable Energy Laboratory (NREL), NREL/TP-6A20-52409

13. Hedman K, Korad AS, Zhang M, Dominguez-Garcia A, Jiang X (2014) The application of robust optimization in power systems. Final report to the power systems engineering research center. PSERC Publication 14-6
14. Heo DY, Tesfatsion LS (2015) Facilitating appropriate compensation of electric energy and reserve through standardized contracts with swing. *J Energy Mark* 8(4):93–121
15. Hinman C (2015) Pay for performance regulation (FERC Order 755) updated with year one design changes California Independent System Operator (CAISO)
16. Kirschen D, Strbac G (2004) Fundamentals of power system economics. Wiley, New York, NY.
17. Li W, Tesfatsion L (2016) Market provision of flexible energy/reserve contracts: optimization formulation. In: Proceedings of the IEEE power and energy society general meeting, Boston, MA (electronic)
18. NAS (2016) Analytic research foundations for the next-generation electric grid. National Academies of Science. The National Academies Press, Washington, DC
19. Oren SS (2005) Generation adequacy via call options obligations: safe passage to the promised land. *Electr J* 18(9):28–42
20. Padhy NP (2004) Unit commitment: a bibliographical survey. *IEEE Trans Power Syst* 19(2):1196–1205
21. Tesfatsion L (2009) Auction basics for wholesale power markets: objectives and pricing rules. In: Proceedings of the IEEE power and energy society general meeting, Calgary, Alberta, CA (electronic)
22. Tesfatsion LS, Silva-Monroy CS, Loose VW, Ellison JF, Elliott RT, Byrne RH, Guttromson RT (2013) New wholesale power market design using linked forward markets. Sandia National Laboratories report (SAND2013-2789)
23. Wood AJ, Wollenberg BF, Sheblé GB (2013) Power generation, operation, and control, 3rd edn. Wiley, New York, NY
24. Zheng QP, Wang J, Liu AL (2015) Stochastic optimization for unit commitment: a review. *IEEE Trans Power Syst* 30(4):1913–1924

A Dynamic Framework for Electricity Markets



Anuradha Annaswamy and Stefanos Baros

Abstract The current transformation towards a cyber-enabled power grid consists of two dominant features, one of which is a high penetration of distributed energy resources (DER) and the other is an increasing participation of demand response (DR), the concept of adjustable power consumption. As much of the DERs are renewable energy resources, their increased penetration introduces intermittencies and uncertainties, which encompass a large range of timescales. DR-compatible devices bring in new degrees of freedom that lead to decision-making over multiple timescales as well. Both of these features necessitate the need for revisiting the electricity market structure, its mechanisms and its overall coupling with the physical power grid. In this paper, we propose a dynamic framework for market mechanisms in the wholesale electricity market at fast timescales. In particular, we propose a dynamic market mechanism and a dynamic regulation market mechanism for the operation of a real-time market and a regulation market, respectively. Taking into account various physical constraints for generation, transmission and consumption, we design these mechanisms so that efficient market equilibrium can be realized. Performance metrics that reflect the social cost as well as physical costs of frequency regulation and area control errors are taken into account. Both market mechanisms are shown to be implementable and exhibit good performance through case studies on a modified IEEE 118-bus system and a three-area system where each of the areas is a modified IEEE 300-bus system.

1 Introduction

Electricity markets represent an important building block of the foundation for a reliable and affordable electricity infrastructure. Electricity markets lie in the intersection of two systems, the financial and the physical, as the products and services

A. Annaswamy (✉) · S. Baros
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: aanna@mit.edu; sbaros@mit.edu

transacted in power system markets have to interact with the physical grid and satisfy its constraints. While economic theory is the underlying tool utilized in order to govern the principles of electricity markets, such a tool alone is not sufficient, especially in the context of the current transformation that the grid is experiencing towards a cyber-enabled architecture with increased penetration of renewable energy resources and increased participation of customers in flexible consumption. This transformation is therefore providing cause for revisiting the electricity market structure, its mechanisms and its overall coupling with the physical power grid.

Following efforts at deregulation, which occurred in Latin America, the UK and the USA around the 1990s, electricity markets began to play an important role in power systems. In the USA, electricity markets appeared in various zones such as California, New York and New England [28]. The market goals herein were to ensure efficient pricing of electricity generation and use as well as incentivizing enhanced grid services and infrastructure maintenance. Efforts to realize these goals led to a hierarchical power system architecture where wholesale electricity market transactions conducted by an independent system operator (ISO) form the top level and automated closed-loop control constitutes the lower levels [10, 17, 29].

An efficient market is one where electricity is traded at a price that minimizes the cost of generation while supplying the entire demand. Current practice in the USA is for the ISO to determine, usually through auction mechanisms, power and energy allocations for three markets that function at different timescales that include day-ahead, real-time and regulation markets. These market clearings include, among other decisions, set points for generating companies, who subsequently produce power that follows these set points. The overall market goals are to ensure efficient pricing of electricity generation and incentivize enhanced grid services and infrastructure maintenance. The outputs of the electricity market can therefore be viewed as set points for the actual units that generate or consume electricity. As electricity cannot be stored in large quantities at the current cost of energy storage, the amount of electricity generated must match the demand at every instant of time to ensure reliability. It is therefore not surprising that these electricity markets range over a broad timescale, from years to seconds, to accommodate planning as well as operations. Examples include markets for forward capacity, energy and ancillary services. With the current transformation that the grid is witnessing, towards modernization, towards a cyber-enabled grid, with new stakeholders that are subject to various dynamic constraints and uncertainties, the question is if the same market structure as above is sufficient. Two major players that are driving this transformation are renewable generation and flexible demand. Motivated by environmental and sustainability concerns, the penetration of renewable energy resources (RER) such as wind and solar is expected to increase significantly world over in the coming decades. The intermittent nature of these resources introduces challenges across all aspects of power system planning and operations. The typical operation of a power grid consists of achieving power balance where load is assumed to be fixed, and generation assets are assembled to equal the load, with voltage and frequency control achieved through inertial and terminal voltage stabilization of a large number of synchronous generators. The very first step in this operation, of

power balance, is directly affected by the introduction of RERs due to the fact that power generation from RERs is subject to uncertainties and intermittencies.

One of the most promising concepts that is being increasingly discussed is demand response (DR), a concept which allows demand to be adjustable, to cope with variations in RERs. A fairly vast literature exists on demand response, its potential and associated challenges and opportunities [4, 5, 8, 15, 20–22]. The concept of introducing flexible consumption in market operations has long been recognized as a highly beneficial one [16, 22]. The idea is then to offset uncertainties in generation with control of consumption. The question that arises is if a coordination of RERs and DRs can be enabled so as to ensure an optimal economic dispatch of generation for power balance.

The introduction of intermittency and uncertainty in a smart grid as well as the increasing potential of adjustable demand via DR necessitates a dynamic framework to address the operation, scheduling and financial settlements of electricity markets. The former brings in issues of strong intermittency and uncertainty and the latter a feedback control structure where demand can be modulated over a range of timescales. Both of these components are dictating a new look at market mechanisms, with a control viewpoint enabling a novel framework for analysis and synthesis. This paper proposes one such alternative to the current structures of real-time market and regulation markets in the form of dynamic market mechanisms.

1.1 Current Practice

Figure 1 shows typical timescales of the most commonly found markets in the USA with respect to other power system planning and operation processes.

Because of the multi-year lead times for building electric power plants and transmission projects, planning markets exist in many places in the USA in order to ensure that the overall supply of electricity will be able to meet projected demand. Markets that govern operation, termed day-ahead (DA) and real-time (RT) markets, ensure that the instantaneous supply of and demand for electric power are balanced

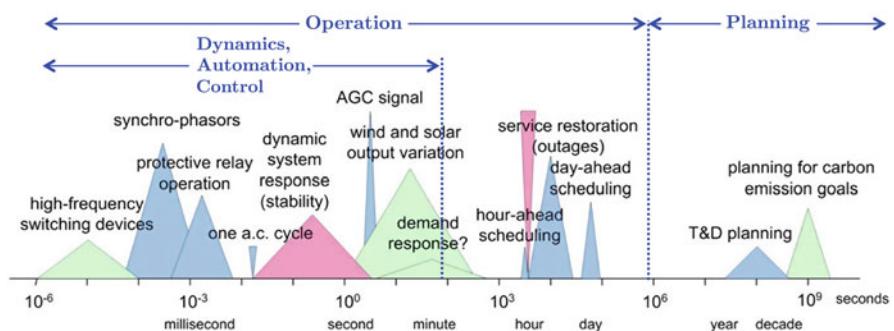


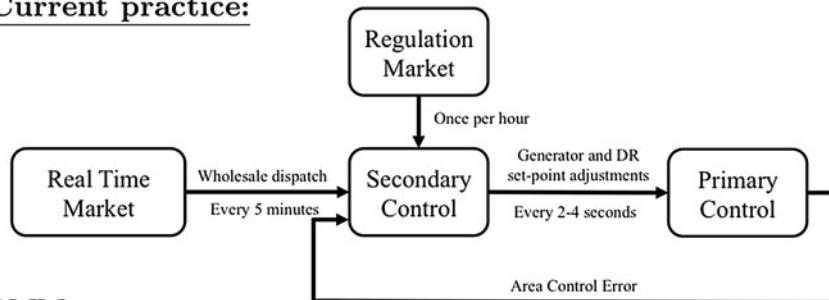
Fig. 1 Illustration of typical planning and operation market timescales

in a least-cost manner. The DA market clears a day prior to operation for 24 hourly intervals, while the RT market clears an hour ahead of operation for 5- to 15-minute intervals [7]. The DA energy market also needs to ensure that certain security constraints are accommodated so as to keep line flows within specified limits. The results of the DA market are a binding agreement to buy, sell or reserve the cleared energy, defined as the specified power set point over a 1-hour interval, at the locational marginal price (LMP).

The focus of the real-time (RT) markets is to augment the DA market actions so as to accommodate the physical operation of the transmission network in real time. This augmentation occurs in the form of set-point commands issued every 5 minutes (e.g. in ISO-NE) to generation units (and in a few regions in the USA, consumption units) that bid into this market. Regulation markets allow units to bid for capacity and service, with the former to enable robust operation around the RT set points and the latter to ensure that the units are dispatched in real time in order to ensure power balance, which is typically carried out through frequency regulation. Frequency regulation is attained through automatic generation control (AGC) which occurs at a secondary control level, at the timescale of 2–4 seconds, and serves as a secondary loop following the primary control loop that ensures stability at the sub-second level.

In this paper, we propose an alternative structure to the RT market and the regulation market that allows decision and control to occur at faster timescales (see Figures 2 and 3). Rather than using the solutions of a relevant constrained

- Current practice:**



- DMM:**

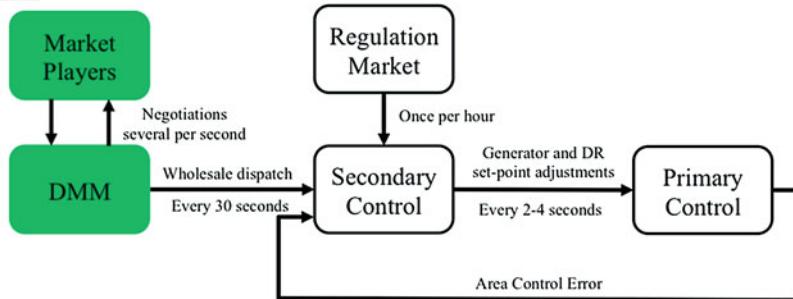
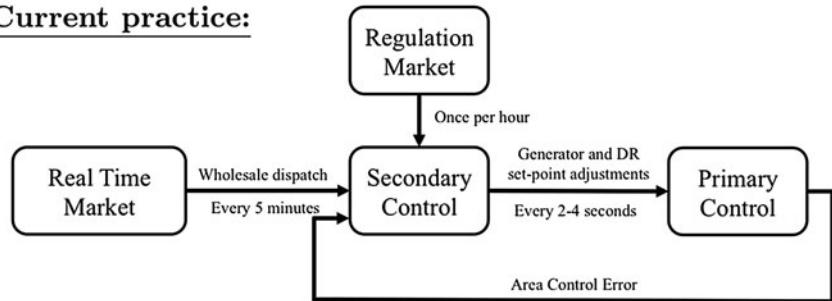


Fig. 2 Real-time market: current practice and the proposed dynamic market mechanism

- Current practice:



- DRMM:

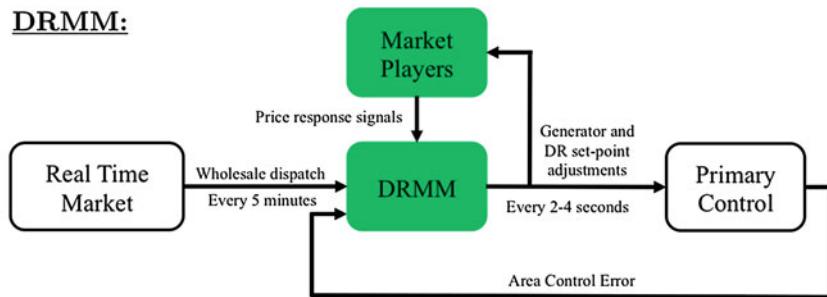


Fig. 3 Regulation market: current practice and the proposed dynamic regulation market mechanism. Reproduced from [26] © 2018 IEEE and used with permission

optimization problem as the market set points, we propose an iterative solution of the optimization problem itself as a dynamic market mechanism. This mechanism is represented as a set of negotiations between the market stakeholders including generation and consumption units that bid into the market and ISO that publishes their schedules and determines their prices. The counterparts to real-time market and regulation market, denoted as dynamic market mechanism (DMM) and dynamic regulation market mechanism (DRMM), are described in detail in the subsequent sections. As is evident from the discussions below and results reported, both DMM and DRMM have the potential to provide an improved response, both from a control-centric and an economic point of view, which is mainly enabled through their ability to incorporate real-time information and mitigate the effect of uncertainties and intermittencies. While the details of the DMM and DRMM can be found in references [25] and [24], respectively, this paper provides a comprehensive discussion of an overall dynamic framework for electricity markets.

The importance of such a dynamic framework is underscored by two reasons. The first is the increasingly renewable-rich environment that is anticipated in the power grid. As forecast errors that are associated with renewables drastically reduce in near real time, the use of a dynamic framework can translate into better load following and reduced needs for frequency regulation. Second, compared with static centralized mechanisms, a dynamic framework including mechanisms such as the

DMM and DRMM has the ability to distribute the computation among all the market participants and significantly reduce the computational effort imposed on the single central entity, the ISO. Finally, the DRMM requires the generators and consumers to implement the intermediate values of the set points at every iteration as actual set points instead of anticipating convergence of the algorithm. This enables the units to provide optimal frequency regulation by dynamically tracking the system optimal point. That means in the intra-dispatch intervals, the set points deployed by the generators and DR units get close to the optimal ones, and the system as a whole operates at a point close the overall optimal.

2 Dynamic Market Mechanisms

In order to describe the DMM, we start with market clearing, the procedure by which economic dispatch is carried out [9]. Participants in this market are generation companies (GenCos) that seek to recover their fuel costs, consumer companies (ConCos) who seek to procure electric energy for their needs and ISO who acts on behalf of ConCos and GenCos, maximizing the utility of ConCos while minimizing the cost of GenCos. The objective function that the ISO uses to solve this optimization problem is commonly termed social welfare, denoted here by S_W and defined as

$$S_W = \sum_{i \in Dr} U_{Dr_i}(P_{Dr_i}) - \sum_{i \in Gc} C_{Gc_i}(P_{Gc_i}) - \sum_{i \in Gr} C_{Gr_i}(P_{Gr_i}),$$

where quadratic utility curves of flexible consumers and quadratic cost curves of conventional and renewable generators are given in (1), (2) and (3), respectively.

$$U_{Dr_i}(P_{Dr_i}) = b_{Dr_i} P_{Dr_i} + \frac{c_{Dr_i}}{2} P_{Dr_i}^2, \quad \forall i \in Dr \quad (1)$$

$$C_{Gc_i}(P_{Gc_i}) = b_{Gc_i} P_{Gc_i} + \frac{c_{Gc_i}}{2} P_{Gc_i}^2, \quad \forall i \in Gc \quad (2)$$

$$C_{Gr_i}(P_{Gr_i}) = b_{Gr_i} P_{Gr_i} + \frac{c_{Gr_i}}{2} P_{Gr_i}^2, \quad \forall i \in Gr \quad (3)$$

Parameters $b_{(.)}$ and $c_{(.)}$ in the above equations reflect base and incremental cost-utility parameters, respectively, while P_{Dr_i} is the power demand of flexible consumers, and P_{Gc_i} and P_{Gr_i} are the power supplied by conventional and renewable generators. Letting the decision variables be collectively denoted by the vector $x = [\delta^T \ P_{Dr}^T \ P_{Gc}^T \ P_{Gr}^T]^T$, the overall market clearing can then be written as the following optimization problem:

$$\underset{x}{\text{minimize}} - S_W \quad (4)$$

subject to

$$\begin{aligned} & \hat{P}_{Dc_n} + \sum_{i \in \phi_n} P_{Dr_i} - \sum_{i \in \theta_n} P_{Gc_i} - \sum_{i \in \vartheta_n} P_{Gr_i} \\ & + \sum_{m \in \Omega_n} B_{nm}(\delta_n - \delta_m) = 0, \quad \forall n \in V \end{aligned} \quad (5)$$

$$\underline{P}_{Dr_i} \leq P_{Dr_i} \leq \bar{P}_{Dr_i} \quad \forall i \in D_r \quad (6)$$

$$\underline{P}_{Gc_i} \leq P_{Gc_i} \leq \bar{P}_{Gc_i} \quad \forall i \in G_c \quad (7)$$

$$\underline{P}_{Gr_i} \leq P_{Gr_i} \leq \hat{\bar{P}}_{Gr_i} \quad \forall i \in G_r \quad (8)$$

$$B_{nm}(\delta_n - \delta_m) \leq \bar{P}_{nm} \quad \forall n \in V, \quad \forall m \in \Omega_n \quad (9)$$

In this problem, prediction of inflexible power consumption at node n is denoted by \hat{P}_{Dc_n} , and δ_n denotes the voltage angle at node n . One of the nodes must be designated as the reference node that is also known as slack bus, at which the voltage angle is defined to be zero. This is a typical assumption for finding a unique solution for power flow problems. Since the true wind generation is not known beforehand, based on wind predictions, an estimate $\hat{\bar{P}}_{Gr_i}$ is used in (8). Here, we consider that the renewable energy generators bid, in the same manner as conventional generators, to the market which is happening close to real time. This is entirely reasonable as close to real time, the wind forecast errors, which can be argued are stochastic, are minimized and the available wind power becomes known. The motivation behind allowing renewable energy generators bid to the market is that, rather than inject all of the power that they produce directly to the grid, it is much more advantageous to have them bid (and deliver) as needed and utilize this extra degree of freedom towards a more efficient grid.

The optimization problem is then solved to compute the solution vector x , where all of the variables comprising vector x are themselves vectors, e.g. $\delta = [\delta_1, \dots, \delta_n]^T$. Throughout this paper we assume vector notation by omitting the corresponding index.

One can accommodate the inequality constraints through the use of barrier functions [3]. For this purpose, we express the constraints in (6)–(8) as $g_{1i}(x) \geq 0, i \in D_r \cup G_c \cup G_r$ and note that these constraints stem from GenCos and ConCos. We express the constraints in (9) as $g_{2j}(x) \geq 0, j \in V$, which stem from the transmission system limitations. We now define barrier functions $\beta_1(x)$ and $\beta_2(x)$ as

$$\beta_1(x) = \sum_{i \in D_r \cup G_c \cup G_r} \frac{v}{g_{1i}(x)}, \quad \beta_2(x) = \sum_{j \in V} \frac{v}{g_{2j}(x)} \quad (10)$$

where ν is a small positive constant used to adjust the slope of the barrier. We observe that the barrier functions get infinitely steep at zero, and so with a sufficiently small step size, they force the algorithm to search for a solution in the feasible set where the constraints are positive. Using these barrier functions, which penalize the solution as it approaches the boundaries of the constraints in (6)–(9), we construct a modified cost function of the form

$$f(x) = -S_W(x) + \beta_1(x) + \beta_2(x). \quad (11)$$

The power balance in (5) can be written as

$$h(x) = \hat{P}_{Dc} + [A^T B_{\text{line}} A_r A_{Dr} - A_{Gc} - A_{Gr}]x = 0 \quad (12)$$

in which system parameter matrices A , A_r and B_{line} , defined in the nomenclature, are a part of a standard DC power flow representation in matrix form [11]. Since our focus is on the wholesale market and primarily on active power, we focus on the DC optimal power flow where the parameter matrices A_{Dr} , A_{Gc} and A_{Gr} are incidence matrices used to map generation and consumption to their respective nodes. For example, $A_{Dr}(i, j) = 1$ if flexible consumer j is at node i and 0 otherwise. The reader is referred to [11] for further details.

The overall underlying optimization problem can be summarized as

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad f(x) \\ & \text{subject to } h(x) = 0. \end{aligned} \quad (13)$$

By choosing ν to be small, the optimal solution of (13) approaches that of problem (4)–(9). The Lagrangian of the optimization problem can be expressed as

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{n \in V} \lambda_n h_n(x) \quad (14)$$

The typical procedure in the OPF method is to find the optimal solution x and the corresponding Lagrangian multipliers λ_n to determine the optimal dispatch and locational marginal prices.

Today it is common practice for the RTO to run the markets on an operating hour schedule, where bids and offers are collected for an entire hour at once. The market will close at least 30 minutes before the operating hour begins, meaning all inputs from market participants must be collected before this time. Throughout the operating hour, the ED is solved, typically every 5 minutes, to determine the dispatch and LMP. Figure 4 shows the timeline for real-time market operation and dispatch of the second operating interval in the operating hour.

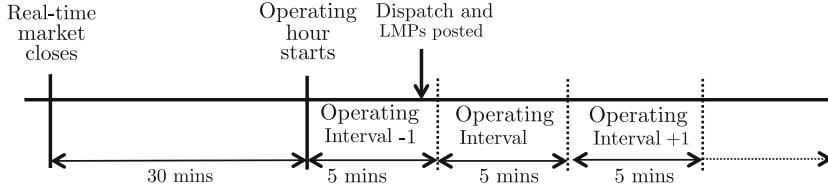


Fig. 4 Simplified RT market operation timeline

2.1 Derivation of DMM

The simplest numerical solution of the optimization problem in (13) can be obtained using a primal-dual gradient algorithm in the form of

$$x^{k+1} = x^k - \alpha_x \cdot \nabla_x \mathcal{L}(x^k, \lambda^k) \quad (15)$$

$$\lambda^{k+1} = \lambda^k + \alpha_\lambda \cdot \nabla_\lambda \mathcal{L}(x^k, \lambda^k) \quad (16)$$

where α_x and α_λ are suitably chosen step sizes. Under convexity of f and the fact that the constraints h are affine, it follows that as $k \rightarrow \infty$, (15) and (16) will converge to the unique global optimum equilibrium point of (13) [3]. Our proposal is to simply use the numerical iteration as the dynamic market mechanism, where the iterations x^k, λ^k at k correspond to the suggested generation, consumption and price in real time, at time t_k prior to the operating interval. That is, at time t_k , we assume that the suggested generation P_{Gc}^k and P_{Gr}^k and consumption, P_{Dr}^k , correspond to a communication from ConCos and GenCos to the ISO and a suggested price λ^k is sent from the ISO to the GenCos and ConCos. Such a real-time communication between the main participants of the real-time market constitutes the proposed DMM. The convergence properties of the DMM have been articulated in several publications including [9, 13, 14, 25] with the primal-dual gradient algorithm suitably replaced by second-order methods in later publications. The reason for using a second-order instead of a first-order method is that first-order gradient-based algorithms require significantly more iterations for convergence. In particular, the algorithm given by (15) and (16) in [9] required on the order of 100,000 iterations for convergence which can be reduced significantly through a Newton's method [25]. To apply the latter, one needs to compute the Hessian and gradient of the Lagrangian, which can be determined as

$$\nabla^2 \mathcal{L}(x^k, \lambda^k) = \begin{bmatrix} H^k & N^k \\ N^{kT} & 0 \end{bmatrix} \quad (17)$$

$$\nabla \mathcal{L}(x^k, \lambda^k) = \begin{bmatrix} \nabla_x \mathcal{L}(x^k, \lambda^k) \\ h(x^k) \end{bmatrix} \quad (18)$$

where $H^k = \nabla_{xx}^2 \mathcal{L}(x^k, \lambda^k)$ and $N^k = \nabla h(x^k)$. It follows by inspection that $N^k = N$, a constant, for all k . To ensure that the Hessian remains non-singular, the Hessian H^k is modified to $\bar{H}^k = H^k + cNN^T$ where c is a positive scalar chosen so that \bar{H}^k is positive definite [2]. This results in a DMM given by

$$x^{k+1} = x^k - \alpha_x \cdot \bar{H}^{k-1} \nabla_x \mathcal{L}(x^k, \hat{\lambda}^k) \quad (19)$$

$$\lambda^{k+1} = \hat{\lambda}^k - \alpha_\lambda \cdot c \cdot h(x^k) \quad (20)$$

where

$$\hat{\lambda}^k = (N^T \bar{H}^{k-1} N)^{-1} (h(x^k) - N^T \bar{H}^{k-1} \nabla f(x^k)) \quad (21)$$

In [25], we have shown that the DMM in (19)–(21) converges to the unique global optimum of problem (13).

2.2 Distributed Implementation of DMM

Since the DMM in (19)–(21) is to be implemented by the ISO, GenCos and ConCos in a distributed manner, we discuss exactly what is involved in each iteration of DMM. We emphasize here that, by distributed, we mean that some of the computations will be performed by the market participants in parallel and in a distributed manner, not all. In fact, the ISO will still be responsible for performing some computations—the ones that involve terms not separable among the participants, e.g. the Hessian. We begin by analysing the gradient of Lagrangian given by

$$\begin{aligned} \nabla_x \mathcal{L}(x^k, \hat{\lambda}^k) &= \nabla_x (-S_W(x^k)) + \nabla_x (\beta_1(x^k)) \\ &\quad + \nabla_x (\beta_2(x^k)) + N\hat{\lambda}^k. \end{aligned} \quad (22)$$

Each of the terms in (22) is now expanded. The term $\nabla_x (-S_W(x^k))$ is composed of cost-utility curves which belong to generators and consumers as

$$\nabla_x (-S_W(x^k)) = \begin{bmatrix} 0 \\ -b_{Dr} - c_{Dr} P_{Dr}^k \\ b_{Gc} + c_{Gc} P_{Gc}^k \\ b_{Gr} + c_{Gr} P_{Gr}^k \end{bmatrix}. \quad (23)$$

Second, $\nabla_x (\beta_1(x^k))$ is composed of barrier functions used to incorporate limits of generation and demand as

$$\nabla_x(\beta_1(x^k)) = \begin{bmatrix} 0 \\ \nabla_{P_{Dr}}(\beta_1(x^k)) \\ \nabla_{P_{Gc}}(\beta_1(x^k)) \\ \nabla_{P_{Gr}}(\beta_1(x^k)) \end{bmatrix}. \quad (24)$$

where

$$\nabla_{P_\ell}(\beta_1) = \frac{\nu}{(P_\ell^k - \bar{P}_\ell)^2} - \frac{\nu}{(P_\ell^k - \underline{P}_\ell)^2} \quad (25)$$

for $\ell \in D_r \cup G_c \cup G_r$. Each row in (23) and (24) is evaluated by a corresponding generator or consumer. The combination $\nabla_x(-S_W(x^k) + \beta_1(x^k))$ represents the marginal cost of production (or marginal value of utility) for each entity at the current negotiation.

Third, $\nabla_x(\beta_2(x^k))$ is composed of barrier functions used to incorporate transmission system constraints as

$$\nabla_x(\beta_2(x^k)) = [\nabla_\delta(\beta_2(x^k))^T \ 0 \ 0 \ 0]^T \quad (26)$$

where

$$\begin{aligned} \nabla_{\delta_n}(\beta_2(x^k)) &= \sum_{m \in \Omega_n} \frac{\nu B_{nm}}{(B_{nm}(\delta_n(k) - \delta_m(k)) - \bar{P}_{nm})^2} \\ &\quad - \frac{\nu B_{nm}}{(B_{nm}(\delta_n(k) - \delta_m(k)) + \bar{P}_{nm})^2}. \end{aligned} \quad (27)$$

The gradient in (26) is entirely evaluated by the ISO.

Additionally, each iteration of DMM requires the updated Hessian:

$$\begin{aligned} H^k &= \nabla_{xx}^2 \mathcal{L}(x^k, \lambda^k) \\ &= -\nabla_{xx}^2 S_W(x^k) + \nabla_{xx}^2 \beta_1(x^k) + \nabla_{xx}^2 \beta_2(x^k). \end{aligned} \quad (28)$$

The last two terms may require private information about the barriers on generation and consumption. To avoid this, the Hessian is approximated as \hat{H} where

$$\hat{H} \approx -\nabla_{xx}^2 S_W(x^k) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -c_{Dr} & 0 & 0 \\ 0 & 0 & c_{Gc} & 0 \\ 0 & 0 & 0 & c_{Gr} \end{bmatrix}. \quad (29)$$

This approximation is of high accuracy when DMM state variables are away from the barriers, since $\nabla_{xx}^2(\beta_1) \approx 0$ and $\nabla_{xx}^2(\beta_2) \approx 0$. We note that even with the approximated Hessian, the DMM is still guaranteed to converge locally. To avoid

sharing incremental cost-utility coefficients $c_{(\cdot)}$, we propose that the ISO uses estimates for the values in \hat{H} , which could be a modification of the ones from the day-ahead market clearing or obtained by inference based on the type of generator or consumer. Regardless of the approach taken to obtain these coefficients, it is important to note that the equilibrium of the DMM will not be affected, but the paths that the state variables take, and hence the convergence time, will be altered. Since the approximation \hat{H} remains constant for all k , a single offline inversion of \hat{H} is sufficient to implement the proposed market clearing mechanism.

We see in (19) that to update x^k to x^{k+1} , we require the product of \hat{H}^{-1} and the vector $\nabla_x \mathcal{L}(x^k, \hat{\lambda}^k)$. As this vector contains information about market players' costs, care should be taken to not make $\nabla_x \mathcal{L}(x^k, \hat{\lambda}^k)$ public. Also, communicating $\nabla_x \mathcal{L}(x^k, \hat{\lambda}^k)$ to all players at every iteration requires extensive communication infrastructure. Thus, to preserve privacy and to simplify the communication structure, we propose the following two-stage implementation for the DMM in (19)–(21): i) the ISO sends out to each market player i participating in DMM a negotiation state x_i^k ; ii) each market player responds with $\nabla_{x_i} f(x_i^k)$ which is the marginal cost, including barrier functions, of market player i for the negotiation x^k . After this two-stage communication, the ISO can evaluate $\nabla_x(\beta_2(x^k))$ for the transmission constraints and can compute the next negotiation state x^{k+1} . This negotiation process continues until convergence is achieved. Such a two-stage implementation, however, implies that every negotiation requires communication from the ISO to the market players and back. This in turn implies that round trip communication is required at each iteration. Ultimately, the DMM implemented through the above process and described by (19)–(21) converges to the global optimum set points $P_{Dr}^*, P_{Gc}^*, P_{Gr}^*$ which solve problem (13).

2.3 Integration with AGC

In addition to allowing the incorporation of real-time information, another advantage of the DMM is its ability to readily integrate other control-centric objectives such as the AGC. The purpose of AGC, as mentioned in the introduction, is to deal with fluctuations in load and generation that occur at a faster timescale than those of the real-time market. Here we associate with the fast AGC timescale the index K and with the RT market timescale the index k where $t_K - t_{K-1} \approx 2$ secs and $t_k - t_{k-1} \approx 30$ msecs. As the fluctuations in load and generation lead directly to deviations in frequencies, the focus of AGC is to restore power balance at fast timescales using frequency deviations as a measure. Typically labelled as a secondary control, the underlying model that captures the effect of the underlying dynamics is of the form

$$\omega^{K+1} = \omega^K + \frac{T_K}{J_{eq}}(P_{G_{total}}^K - P_{D_{total}}^K - ACE^K) \quad (30)$$

where ACE is the area control error given by

$$(ACE)^K = P_{TL} \Delta_{TL} + B_{eq}(\omega^K - \omega_{ref}), \quad (31)$$

where P_{TL} is the scheduled tie-line power flow (negative meaning *into* balancing area) and Δ_{TL} is the error in tie-line flow. Conventional generators automatically adjust to changes in system frequency. This automatic adjustment is known as primary control, the goal of which is to arrest changes in frequency following a power imbalance. The sum of these effects is captured by the equivalent frequency bias of the balancing area, denoted by B_{eq} . Since this quantity is difficult to measure accurately, most balancing authorities use a bias of 1% of peak load [27]. Also,

$$P_{G_{total}}^K = \sum_{i \in Gc} P_{Gc_i, true}^K + \sum_{i \in Gr} P_{Gr_i, true}^K \quad (32)$$

$$P_{D_{total}}^K = \sum_{n \in N} P_{Dc_n, true}^K + \sum_{i \in Dr} P_{Dr_i, true}^K, \quad (33)$$

and J_{eq} is the total inertia of the balancing area, and it is assumed that $\omega^0 = \omega_{ref}$. Suppose that all conventional generators that are dispatched in the real-time market participate in providing regulation services by supplying an incremental quantity of power $P_{reg_i}^K$ in response to ACE at each AGC update period t_K . The total regulation power is

$$P_{reg}^{K+1} = P_{reg}^K - K_R ACE^K \quad (34)$$

where K_R is a control gain chosen so that closed-loop stability is ensured and the settling time T_s and response time T_r are within certain specifications [17]. The true power produced by a conventional generator then includes the market dispatch P_{Gc}^* as well as frequency regulation, that is,

$$P_{Gc, true}^K = P_{Gc}^* + P_{reg}^K \quad (35)$$

Suppose we assume that DR resources adhere to their market dispatch, so $P_{Dr, true} = P_{Dr}^*$.

2.3.1 Feedback from AGC to DMM

As the penetration of RER increases, even with the above secondary control approach, a significant ACE may remain. In order to address this, we suggest an aggregated measure of ACE to be fed back to the DMM. For this purpose, we propose that the frequency error is aggregated over the period of a single DMM market clearing T_m as

$$\omega^m = \frac{1}{|\Psi_m|} \sum_{K \in \Psi_m} \omega^K. \quad (36)$$

We note that all DMM negotiations at $t_k \in [t_{m-1}, t_m]$ use information obtained over the previous period T_{m-1} , and hence the aggregated frequency information available at t_{m-1} corresponds to ω^{m-2} . The aggregated ACE for this period is $\overline{ACE}^{m-2} = P_{TL} \Delta_{TL} + B_{eq}(\omega^{m-2} - \omega_{ref})$. The resulting \overline{ACE}^{m-2} is fed back into the DMM in the form of a modified power balance $h'(x) = 0$ that replaces $h(x) = 0$, defined below:

$$h'(x^k) = h(x^k) - K_L A_{Gc} \bar{B} \overline{ACE}^{m-2} \quad (37)$$

where \bar{B} is a vector of elements $\bar{B}_n = \sum_{i \in \theta_n} B_i, \forall n \in V$. Intuitively, the role of vector \bar{B} is to disaggregate frequency error and to distribute it to the nodes with generators who committed to provide regulation. Such implementation allows the aggregated frequency error to be met optimally by the market. This means that demand response resources can also participate in regulating grid frequency in an economically efficient manner. In this paper, we assume that renewable generators do not provide regulation.

The choice of the feedback gain K_L in (37) is dictated, in general, by conditions of stability and optimality. That is, K_L should be chosen so that stability of the combined DMM+AGC system is ensured while also ensuring that the quantity e_{ACE} is as small as possible, which is defined as

$$e_{ACE} = \sqrt{\frac{1}{|\tau|} \sum_{i \in \tau} (ACE^i)^2} \quad (38)$$

An analytical procedure for determining K_L that guarantees stability can be established along the lines of [14]. However, for the purposes of this paper, we select K_L empirically, such that both of the above requirements are satisfied. Details of this choice are provided Section 4.

With the above feedback from AGC, the integrated DMM consists of solving (13) with the equality constraint replaced by $h'(x) = 0$. The iterates of the integrated DMM at the timescale t_k take the form

$$x^{k+1} = x^k - \hat{H}^{-1} \nabla_x \mathcal{L}(x^k, \hat{\lambda}^k) \quad (39)$$

$$\lambda^{k+1} = \hat{\lambda}^k - \alpha \cdot c \cdot h'(x^k) \quad (40)$$

$$\hat{\lambda}^k = (N^T \hat{H}^{-1} N)^{-1} (h'(x^k) - N^T \hat{H}^{-1} \nabla f(x^k)). \quad (41)$$

Equations (39)–(41) together with the AGC iterates (30) and AGC aggregation defined by (36) constitute the overall DMM. In this DMM formulation, x^k now

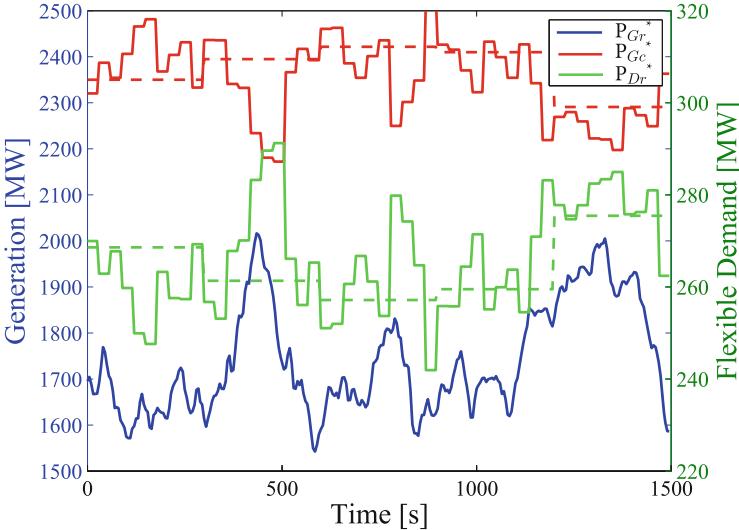


Fig. 5 Power schedules over 25 minutes using the DMM (solid lines) and OPF (dashed lines). The iterations between any two market clearings are not shown. Reproduced from [25] © 2018 IEEE and used with permission

corresponds to the intermediate generation set points at t_k prior to the operating interval whose converged values jointly satisfy the predicted demand for the next time horizon and eliminate the aggregated average ACE that has arisen since the previous market clearing in an optimal manner. In other words, x^k here corresponds to the negotiated state variables that take into account both effects of economic dispatch and average frequency regulation. The latter is realized by incorporating a feedback from the aggregated ACE into the real-time market through the power balance constraints which results in the modified equality constraints $h'(x^k)$ in (40). Ultimately, through this correction, the proposed DMM described by Equations (39)–(41) leads to tighter bounds on the real-time ACE which can significantly reduce the burden on the AGC system.

In [25], we applied the proposed DMM on the modified 118 bus test case which includes wind generation and flexible consumption. Next, we summarize the key results while the reader can find a more comprehensive and detailed analysis in [25]. In Figure 5, we demonstrate that the DMM is able to dynamically and rapidly schedule conventional generators and flexible consumers in response to wind predictions at the market level, through their set points P_{Gc}^* , P_{Gr}^* and P_{Dr}^* over a 25-minute period. Further in Figure 6, we zoom in and illustrate the evolution of the individual DMM iterations for the flexible consumers within one market clearing. Specifically, we show that, at 390 secs, the DMM utilizing new predictions about the wind and demand that become available initiates a new iterative process that is completed at 420 secs (when the market clearing occurs) leading to an increase

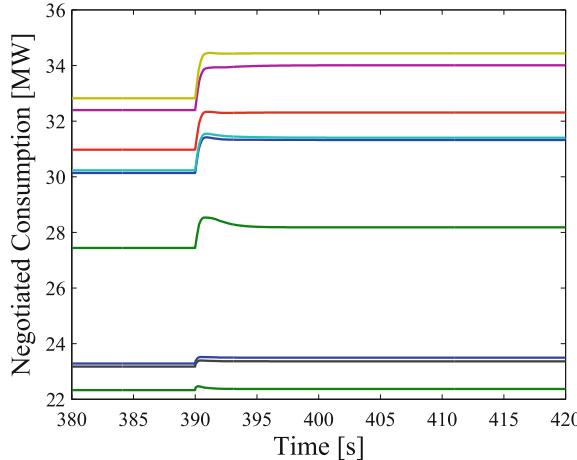


Fig. 6 Sample iterations of flexible consumers for market clearings at $t = 390$ s and $t = 420$ s.
Reproduced from [25] © 2018 IEEE and used with permission

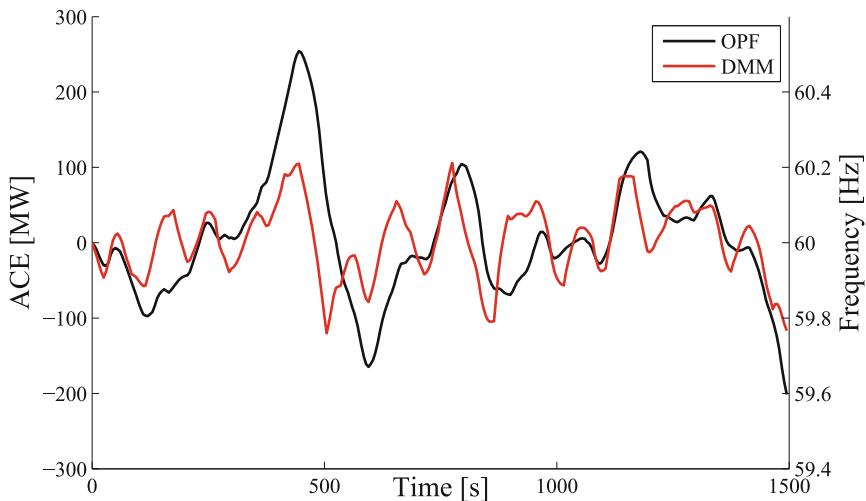


Fig. 7 ACE using OPF and DMM ($K_L = 1$) over a 25-minute window. Reproduced from [25] © 2018 IEEE and used with permission

in the set points of the flexible consumers, $P_{D_r}^k$. It is important to realize that convergence of the DMM is reached well in advance of the 30 secs dispatch window. Finally, in Figure 7, we show that the DMM can alleviate the burden on the AGC system as the peaks in ACE appear to be less severe in the simulations obtained using the DMM than the ones using OPF.

3 Dynamic Regulation Market Mechanisms

The regulation market is the mechanism for selecting and compensating market participants to provide regulation—the capability of specially equipped generators and other energy sources to increase or decrease output or consumption every 4 seconds. Participants allow their automatic generation control (AGC) resources to be controlled by the ISO using automated signals to balance both second-by-second variations in demand and the system frequency, which must be kept constant [6]. Financial settlements include both payments and charges. The real-time market provides an *energy payment* to generators and accepts *energy charges* from DR resources, where generators and DR resources are units that participate in the real-time market. Current practice in the regulation market is to allocate and compensate for capacity and service separately. For instance, in ISO New England, the system operator clears both a capacity price and a service price once every hour, based on estimates of the capacity and service needed to maintain system reliability [12]. These are denoted as *reserve capacity payments* and *service payments* to all its participants [19]. Our thesis is that a dynamic regulation market mechanism that keeps frequency errors in check at a faster timescale and simultaneously ensures that the financial settlements are efficient needs to be designed as the dynamic signature in a power grid becomes more and more dominant (see Figure 3 for a comparison of the existing structure and the proposed DRMM). The structure of the DRMM that we propose is described in more detail below.

3.1 Our Approach

Described in more detail in [23, 24], the DRMM achieves frequency regulation in the presence of various physical disturbances in an optimal manner. The starting point for this model is the description of the linearized physical dynamics of a power system. The most important state variables of this system include the voltage angle deviations $\Delta\delta_i$ and electrical frequency deviations $\Delta\omega_i$ dynamics (swing dynamics) of the synchronous machines buses, the voltage angle deviations $\Delta\delta_i$ dynamics (swing dynamics) of the load buses, the speed-governor valve position deviations ΔY_i dynamics of a synchronous generator and its power output deviations $\Delta P_{M,i}$ dynamics. As our problem statement includes flexible consumers as well, we also include as an additional state, the power consumption $\Delta P_{C,i}$ of the demand response (DR) resources, which is assumed to respond to the command from say a DR aggregator, with a lag. Altogether, they can be compactly stated in discrete-time state-space form as

$$\psi_{K+1} = \Phi\psi_K + \Gamma_B u_K + \Gamma_E \Delta P_{L,K} \quad (42)$$

where $\psi = [\Delta\omega^T \ \Delta\delta^T \ \Delta Y^T \ \Delta P_M^T \ \Delta P_C^T]^T$ is the state vector that denotes deviations of frequencies, phase angles, valve positions, mechanical power generation and flexible consumption from their equilibrium values and $\Phi, \ \Gamma_B, \ \Gamma_E$

are constant matrices. We refer the detailed derivation of these matrices and (42) to [24]. In (42), the control input is given by $u = [\Delta P_{Gc}^T \ \Delta P_{Dr}^T]^T$ where $\Delta P_{Gc} = (P_{Gc} - P_{Gc}^*)$, $\Delta P_{Dr} = (P_{Dr} - P_{Dr}^*)$ are the secondary control set points for the generators and DR units, respectively. In the current practice today in power systems, only regulation of the P_{Gc} 's of the generators is considered at the secondary control level, and its objective is to regulate the frequency of a balancing area back to its nominal value. This is attained by designing the control input ΔP_{Gc} as an integral control feedback of the area control error (ACE) which, for an area β , is defined as a weighted sum of the frequency error $\bar{\omega}_\beta$ and the tie-line error P_T :

$$e_{C_K}^{(\beta)} = B^{(\beta)} \bar{\omega}_\beta + N_T^{(\beta)} \Delta P_{T,K} = N_\psi^{(\beta)} \psi_K \quad (43)$$

where $N_T^{(\beta)}$, $N_\psi^{(\beta)}$ are constant matrices. The specific generators that participate in the secondary control and their corresponding regulation capacities are defined through a regulation market which is cleared once every hour. Further, the slow-varying set points P_{Gc}^* , P_{Dr}^* are provided every 5 minutes by the real-time market, often determined through the solution of an optimal power flow (OPF) problem.

In contrast to this practice, our proposed DRMM will result to *optimal values* for both ΔP_{Gc} , ΔP_{Dr} by iteratively minimizing both the cost of generation and the disutility of deferred consumption described by the function

$$f^{(a)} = -S_W^{(a)} \quad (44)$$

$$S_W^{(a)} := \sum_{i \in \mathcal{D}^{(a)}} U_{Dr_i}(\Delta P_{Dr_i}) - \sum_{i \in \mathcal{G}^{(a)}} C_{Gc_i}(\Delta P_{Gc_i}) \quad (45)$$

subject to constraints in the network in the fast AGC timescale, i.e. solving the following optimization problem (for an area a):

$$\underset{\xi}{\text{minimize}} f^{(a)} \quad (46)$$

subject to

$$\begin{aligned} \Delta \hat{P}_{L_i} + \sum_{j \in \mathcal{D}_i^{(a)}} \Delta P_{Dr_j} - \sum_{j \in \mathcal{G}_i^{(a)}} \Delta P_{Gc_j} \\ + \beta_i \rho^{(a)} + \sum_{(i,j) \in \mathcal{E}^{(a)}} T_{ij}(\Delta \theta_i - \Delta \theta_j) = 0 \quad \forall i \in \mathcal{N}^{(a)} \end{aligned} \quad (47)$$

$$\Delta P_{ij}^{min} \leq T_{ij}(\Delta \theta_i - \Delta \theta_j) \leq \Delta P_{ij}^{max} \quad \forall (i, j) \in \mathcal{E}^{(a)} \quad (48)$$

$$\Delta P_{Gc_i}^{min} \leq \Delta P_{Gc_i} \leq \Delta P_{Gc_i}^{max} \quad \forall i \in \mathcal{G}^{(a)} \quad (49)$$

$$\Delta P_{Dr_i}^{min} \leq \Delta P_{Dr_i} \leq \Delta P_{Dr_i}^{max} \quad \forall i \in \mathcal{D}^{(a)} \quad (50)$$

$$\Delta E_{Dr_i} = 0, \quad \forall i \in \mathcal{D}^{(a)} \quad (51)$$

where the decision variables are given by the vector $\xi = [\Delta\theta^T \ \Delta P_{Gc}^T \ \Delta P_{Dr}^T]^T$ and ΔE_{Dr_i} denotes the energy deviations of the DR units and is defined as the integral of ΔP_{Dr_i} . Nodal power balance is enforced by (47), transmission line flow constraints are enforced by (48), and generator and DR unit loading constraints are enforced by (49)–(50) and energy payback for the DR units by (51). Since the load disturbance ΔP_L may not be known exactly, the market uses an estimate $\hat{\Delta P}_L$. Note that $\Delta\theta_i$ is a “virtual” variable as it is used only in the optimization algorithm. This is distinct from $\Delta\delta_i$ which refers to the actual, physical voltage angle. The term $\beta_i \rho^{(a)}$ represents a feedback from primary control of frequency deviation, described in greater detail next. The above optimization problem can be compactly written as

$$\underset{\xi}{\text{minimize}} \ f^{(a)} \quad (52)$$

subject to

$$h^{(a)} = 0 \quad (53)$$

$$g^{(a)} \leq 0 \quad (54)$$

$$\Delta E_{Dr}^{(a)} = 0 \quad (55)$$

To solve the above problem, we propose a DRMM [23, 24] which is implemented as an ongoing negotiation process between generators, DR units and the ISO that allows both generators and DR resources to bid for regulation services in real time while also ensuring optimal allocation of these services. The negotiations are realized through the iterative market dynamics stated below which, via a Newton-like method, drive the market to the solution of Problem (52). At the same time, by incorporating an ACE signal into the power balance equation (53), they realize real-time optimal secondary control, while additionally, they guarantee energy payback of the DR units through an additional energy equality constraint (55). Using a Newton-like method, the DRMM iterative dynamics can be expressed as [23, 24]:

Set-point dynamics

$$\xi_{K+1} = \xi_K - a \hat{H}_\gamma^{-1} (\pi_K + N_h \hat{\lambda}_K) \quad (56)$$

Multipplier dynamics

$$\mu_{K+1} = \max\{0, \mu_K + K_\mu g_K\} \quad (57)$$

$$v_{K+1} = K_v v_K + K_E \Delta E_{Dr,K} + K_\eta \eta_{D,K} \quad (58)$$

Auxiliary dynamics

$$\Delta E_{Dr,K+1} = \Delta E_{Dr,K} + N_E \xi_K \quad (59)$$

$$\eta_{D,K+1} = \eta_{D,K} + \Delta E_{Dr,K} \quad (60)$$

Regulation signal dynamics

$$\rho_{K+1} = \rho_K - K_f N_\psi \psi_K \quad (61)$$

where

$$\pi_K = \nabla_{\xi} f_K + N_g \mu_K + N_E \nu_K \quad (62)$$

$$\hat{\lambda}_K = (N_h^T \hat{H}_\gamma^{-1} N_h)^{-1} (h_K(\xi_K, \rho_K) - N_h^T \hat{H}_\gamma^{-1} \pi_K) \quad (63)$$

DRMM realizes frequency regulation by incorporating the integral of ACE ρ_K into the market dynamics through the modified power balance equality constraints $h_k(\xi_K, \rho_K)$ in (63) as described by the function:

$$h_k(\xi_K, \rho_K) := N_h^T \xi_K + \beta \rho_K \quad (64)$$

The vector ξ represents the set points and μ , ν the multipliers of the inequality and equality constraints, respectively. Moreover, the vector ΔE_{Dr} represents the energy deviations state variables of the DR units, η_D the integrals of these states, ρ the regulation signal, π_K a price response signal and \hat{H}_γ an estimate of the Hessian of the Lagrangian function. Altogether, Equations (56)–(60) can be compactly stated as

$$\zeta_{K+1} = A_{\zeta_K} \zeta_K + B_\rho \rho_K + B_{P_L} \hat{P}_{L,K} + B_{l_g} l_g \quad (65)$$

where $\zeta = [\xi^T \ \mu^T \ \nu^T \ \Delta E_{Dr}^T \ \eta_D^T]^T$. The DRMM for each balancing area is executed as follows. The independent system operator (ISO) provides the set points $\Delta P_{Gc,K}$ and $\Delta P_{Dr,K}$ obtained by the Equation (56) to the generators and DR units every Δt_K seconds. When they receive these set points, the generators and DR units use their multipliers μ_K , ν_K to compute their price response signal π_K which they communicate to the ISO. The ISO updates the regulation signal ρ_K and h_K by computing the ACE in its area and finally, upon receiving all π_K , computes the next set of set points ξ_{K+1} with the whole process repeating in the same manner as depicted in Figure 3. For a power system where the set of balancing areas is denoted by \mathcal{B} and each area indexed by β , the physical, DRMM and regulation signal dynamics can be stated as [23, 24]:

Physical dynamics

$$\psi_{K+1} = \Phi\psi_K + \Gamma_B \sum_{\beta \in \mathcal{B}} N_\zeta^{(\beta)} \zeta_K^{(\beta)} + \Gamma_E \Delta P_{L,K} \quad (66)$$

DRMM dynamics

$$\zeta_{K+1}^{(\beta)} = A_{\zeta_K}^{(\beta)} \zeta_K^{(\beta)} + B_\rho^{(\beta)} \rho_K^{(\beta)} + B_{P_L}^{(\beta)} \Delta \hat{P}_{L,K}^{(\beta)} + B_{l_g}^{(\beta)} l_g^{(\beta)} \quad (67)$$

Regulation signal dynamics

$$\rho_{K+1}^{(\beta)} = \rho_K^{(\beta)} - K_f^{(\beta)} N_\psi^{(\beta)} \psi_K \quad (68)$$

Collectively, the dynamical model describing the interconnected physical and dynamic regulation market dynamics can be compactly written as

$$\chi_{K+1} = A_\chi \chi_K + B_\chi v_K \quad (69)$$

where $\chi = [\psi^T \ \zeta^{(1)T} \ \dots \ \zeta^{(|\mathcal{B}|)T} \ \rho^{(1)} \ \dots \ \rho^{(|\mathcal{B}|)}]^T$ and $v = [\Delta P_L^T \ \Delta \hat{P}_L^T \ l_g^T]^T$. Observe that the combined physical/market system (69) defines a cyber-physical energy system (CPES) due to the two-way real-time cyber communication of the set points ξ and price signals π among the ISO, generators and DR units.

The DRMM in (69) evolves at each K , which corresponds to a time instant t_K . The cyber components of information and communication determine the sampling interval $T_K = t_{K+1} - t_K$. With technological advances in the underlying ICT infrastructure, it is entirely possible that this period can be made shorter than the current 2 to 4 seconds that is in vogue for AGC to perhaps a second or even smaller.

3.2 Results

The performance of the DRMM control strategy is verified by running a series of tests on a 900 bus power system with three areas. Each area is a modified IEEE 300 bus test system with 411 transmission lines and 56 active generators [24]. Thirty DR units were added at random locations in each area. The results of the aggregate frequency measurements of three connected areas are shown in Figure 8. To assess the cost of regulation, we define a *regulation service cost* function by combining the social welfare function (44) and a quadratic function that includes the ACE deviation. In more detail, we define the service cost function that corresponds to an area a , $S^{(a)}$, as

$$S^{(a)} = f^{(a)} + c_{ACE}(e_C^{(a)})^2 \quad (70)$$

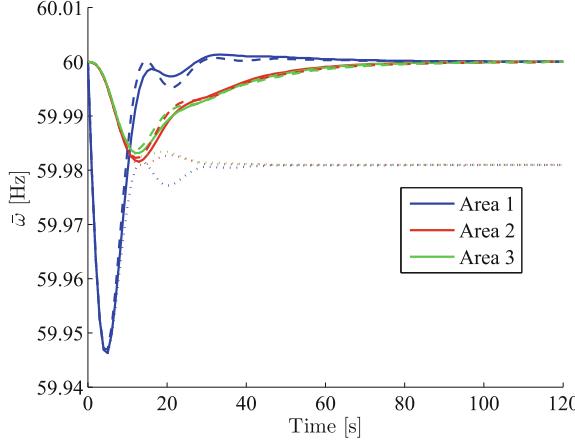


Fig. 8 Aggregate frequency measurements of three connected areas. Solid lines indicate the response with DRMM, dashed lines indicate the response with conventional AGC, and dotted lines indicate the response with primary control only. Reproduced from [26] © 2018 IEEE and used with permission

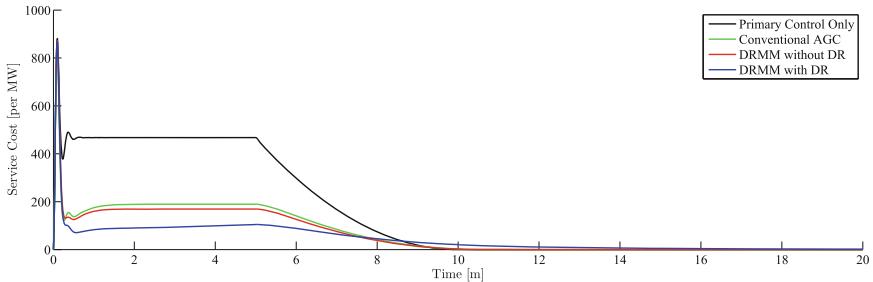


Fig. 9 Impact of DRMM and DR units on regulation service costs. In this plot, the DR payback time is 20 minutes. Reproduced from [26] © 2018 IEEE and used with permission

where c_{ACE} captures the reflected cost to the system due to nonzero ACE which here is taken to be $c_{ACE} = 0.02$. This latter term is required in order to effectively compare the system costs that are resulted by the secondary control and the primary control alone (otherwise the optimal strategy would be to let the frequency fall). In Figure 9, it is shown that the resulted operating cost using the DRMM (with and without DR units) is less than using the conventional AGC.

The above results verify that the proposed DRMM is indeed a feasible and viable approach that is practically realizable. In addition, it results in satisfactory performance both with respect to control-centric metrics such as frequency realization (see Figure 8) and with respect to market-centric metrics such as service cost (see Figure 9) that may become increasingly important with high penetration of RERs. In the following section, we connect this DRMM to an energy and reserve capacity co-optimization framework that is currently carried out in the real-time market and show that the performance of the DRMM results in improved financial metrics [1, 18, 30].

3.3 A *Co-optimization Framework for DRMM*

In order for the DRMM to effectively realize optimal frequency regulation, the generators that participate in frequency regulation need to have adequate reserve capacity levels. Today, the current practice for realizing this goal is by co-optimizing energy and reserve capacities at the real-time market. Such a co-optimization can be incorporated into the DRMM as well, as shown in [7]. This ensures that the generators participating in the DRMM maintain enough reserve capacities to carry out optimal frequency regulation, i.e. that the DRMM optimization problem is always feasible. Our results show that a combination of the standard co-optimization problem formulation and the DRMM can yield both optimal frequency regulations with good performance (using standard control metrics) but also significantly reduced regulation costs as shown in Figures 8 and 9, respectively. In particular, due to the introduction of an optimization framework in secondary control, one can realize lower make-whole payments in accordance with the reduced service costs (see Figure 9), which correspond to payment to generators that provide frequency regulation by the ISO (see [7] for further details).

4 Summary

With increasing DERs and DRs, there is a need to revisit decisions of electricity markets and control of key physical variables in the design of a cyber-enabled grid, especially at near real-time. In this paper, we described a dynamic framework for market mechanisms in the wholesale electricity market at fast timescales including a dynamic market mechanism and a dynamic regulation market mechanism. Taking into account various physical constraints for generation, transmission and consumption, we design these mechanisms so that efficient market equilibrium can be realized. Rather than using the solutions of a relevant constrained optimization problem as the market set points, we propose an iterative solution of the optimization problem itself as a dynamic market mechanism. This mechanism is represented as a set of negotiations between the market stakeholders including generation and consumption units that bid into the market and ISO that publishes their schedules and determines their prices. In the sections above, it was shown that the dynamic market mechanisms have the potential to lead to improved performance metrics that reflect the social cost as well as physical costs of frequency regulation and area control errors. Both the DMM and DRMM were validated through case studies of a modified IEEE-118 bus and a four-area system, each of which is a modified IEEE-300 bus.

Several extensions remain to be carried out of the dynamic framework introduced in this paper. Given that the penetration of DERs is shifting challenges substantially towards the distribution grid, the extension of such dynamic mechanisms to distribution grids and microgrids is essential and poses a highly nontrivial problem. Given

the absence of a decision-maker such as the RTO in retail markets in the USA, how market-based decisions can be integrated into voltage control and reactive power injection needs to be explored. The accommodation of a large number of individual consumers with varying constraints on their flexibility in consumption and the notion of transactive control [14] that incentivizes these consumers while retaining grid-level objectives of volt/var control are all grand challenges and tremendous opportunities for advanced systems and control concepts.

Acknowledgements This work was supported in part by the NSF Award no. EFRI-1441301.

References

1. Beaude O, He Y, Hennebel M (2013) Introducing decentralized ev charging coordination for the voltage regulation. In: Innovative smart grid technologies Europe (ISGT EUROPE), 2013 4th IEEE/PES. IEEE, Piscataway, NJ, pp 1–5
2. Bertsekas DP (1995) Dynamic programming and optimal control, vol 1. Athena Scientific, Belmont, MA
3. Bertsekas DP (1999) Nonlinear programming. Athena Scientific, Belmont, MA
4. Biegel B, Andersen P, Pedersen T, Nielsen K, Stoustrup J, Hansen L (2013) Smart grid dispatch strategy for ON/OFF demand-side devices. In: Proceedings of the European control conference
5. Brooks A, Lu E, Reicher D, Spirakis C, Weihl B (2010) Demand dispatch: using real-time control of demand to help balance generation and load. IEEE Power Energ Mag 8:20–29
6. Federal Energy Regulatory Commission (2011) Order 755: frequency regulation compensation in the organized wholesale power markets
7. Garcia MJ, Nudell TR, Annaswamy AM (2017, to appear) A dynamic regulation market mechanism for improved financial settlements. In: Proceedings of the American control conference
8. Hansen J, Knudsen J, Annaswamy AM (2014) Demand response in smart grids: participants, challenges, and a taxonomy. In: Proceedings of the conference on decision and control
9. Hansen J, Knudsen J, Annaswamy AM (2016) A dynamic market mechanism for integration of renewables and demand response. IEEE Trans Control Syst Technol 24(3):940–955
10. Ilić MD (2007) From hierarchical to open access electric power systems. Proc IEEE 95(5):1060–1084
11. Ilic M, Zaborszky J (2000) Dynamic and control of large electric power systems. Wiley, New York
12. Iso New England Inc. Transmission, markets, and services tariff (iso tariff), section iii: Market rule 1. <http://www.iso-ne.com/participate/rules-procedures/tariff/market-rule-1>. Accessed 21 Mar 2017
13. Kiani A, Annaswamy AM (2014) A dynamic mechanism for wholesale energy market: stability and robustness. IEEE Trans Smart Grid 5(6):2877–2888
14. Kiani A, Annaswamy AM, Samad T (2014) A hierarchical transactive control architecture for renewables integration in smart grids: analytical modeling and stability. IEEE Trans Smart Grid 5(4):2054–2065
15. Kim JJ, Yin R, Kilicotte S (2013) Automated price and demand response demonstration for large customers in New York city using openADR. Technical report 6472E, LBNL
16. Kirschen DS, Strbac G, Cumperayot P, de Paiva Mendes D (2000) Factoring the elasticity of demand in electricity prices. IEEE Trans Power Syst 15(2):612–617
17. Kundur P (1997) Power system stability and control. McGraw-Hill, New York
18. Manski CF (1977) The structure of random utility models. Theor Decis 8(3):229–254

19. New England Independent System Operator (2016) Market rule 1 standard market design. Section III, ISO New England. Inc FERC Electric Tariff
20. Olsen D, Goli S, Faulkner D, McKane A (2012) Opportunities for automated demand response in wastewater treatment facilities in California - southeast water pollution control plant case study. Technical report 6056E, LBNL
21. Petersen MK, Hansen LH, Bendtsen J, Edlund K, Stoustrup J (2013) Market integration of virtual power plants. In: Proceedings of the conference on decision and control
22. Scheppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. Kluwer Academic Publishers, Boston, MA
23. Shiltz DJ (2016) Integrating automatic generation control and demand response via a dynamic regulation market mechanism. Master's thesis, Massachusetts Institute of Technology
24. Shiltz D, Annaswamy AM (2016) A practical integration of automatic generation control and demand response. In: Proceedings of the American control conference
25. Shiltz DJ, Cvetković M, Annaswamy AM (2016) An integrated dynamic market mechanism for real-time markets and frequency regulation. IEEE Trans Sustainable Energy 7(2):875–885
26. Shiltz D, Baros S, Cvetkovic M, Annaswamy AM (2017) Integration of automatic generation control and demand response via a dynamic regulation market mechanism. IEEE Trans Control Syst Technol **PP**(99):1–16. <https://doi.org/10.1109/TCST.2017.2776864>
27. Subcommittee NR (2011) Balancing and frequency control. Technical report, North American Electric Reliability Corporation
28. Wang G, Negrete-Pincetic M, Kowli A, Shafieeoorfard E, Meyn S, Shanbhag UV (2011) Dynamic competitive equilibria in electricity markets. Springer, Berlin, pp 35–62
29. Wollenberg AWB (1996) Power generation operation and control, 2nd edn. Wiley and Sons, Hoboken, NJ
30. Zhong J, He L, Li C, Cao Y, Wang J, Fang B, Zeng L, Xiao G (2014) Coordinated control for large-scale EV charging facilities and energy storage devices participating in frequency regulation. Appl Energy 123:253–262

Fast Market Clearing Algorithms



Arvind U. Raghunathan, Frank E. Curtis, Yusuke Takaguchi,
and Hiroyuki Hashimoto

Abstract Real-time electricity markets are the main transaction platforms for providing necessary balancing services, where the market clearing (nodal or zonal prices depending on markets) is very close to real-time operations of power systems. We present single and multiple time period decentralized market clearing models based on the DC power flow model. The electricity market we study consists of a set of generation companies (GenCos) and a set of Distribution System Operators (DSOs). The Independent System Operator (ISO) determines the market clearing generation and demand levels by coordinating with the market participants (GenCos and DSOs). We exploit the problem structure to obtain a decomposition of the market clearing problem where the GenCos and DSOs are decoupled. We propose a novel semismooth Newton algorithm to compute the competitive equilibrium. Numerical experiments demonstrate that the algorithm can obtain several orders of magnitude speedup over a typical subgradient algorithm with no modification to the existing communication protocol between the ISO and market participants.

1 Introduction

Electricity markets are commodity markets where (i) suppliers (electricity generators) and consumers (electricity customers) are spread across a network and (ii) the flow of the commodity (electricity) is dictated by physical laws [14]. Competition in electricity markets allows to establish a market price that is acceptable to

A. U. Raghunathan (✉)

Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, USA

e-mail: raghunathan@merl.com

F. E. Curtis

Department of Industrial & Systems Engineering, Lehigh University, Bethlehem, PA 18015, USA
e-mail: frank.e.curtis@gmail.com

Y. Takaguchi · H. Hashimoto

Advanced Technology Center, Mitsubishi Electric Corporation, Hyogo 661-8661, Japan

all market participants, whereby the market is said to have reached *equilibrium*. The design of appropriate market or pricing mechanisms is governed by the theory of general equilibria. The nonexistence of such an equilibrium implies the possibility of some market participants that can unilaterally affect the prices to their own advantage [28]. For example, competition [4] and active participation (e.g., demand response) [27] in electricity markets are known to significantly enhance efficiency and reduce prices. Given the importance of an efficient and reliable grid infrastructure, the modeling and subsequent analysis of electricity markets have seen extensive research. Hobbs and Helman [12] study market equilibria via competitive equilibrium models. Oligopolistic price equilibria were investigated by Hobbs, Metzler, and Pang [13] for Direct Current (DC) power flow networks using supply functions. Baldick [2] compares Cournot and supply function equilibrium models of bid-based electricity markets. Day, Hobbs, and Pang [7] investigate conjectured supply function models of competition among power generators on a DC power flow network. The impact of network constraints on the market performance is analyzed in [4] under different game theory models. Weber and Overbye [29] study Nash equilibria for electricity markets by employing a full representation of the transmission system. Motto et al. [19] formulate a multiperiod electricity auction market tool for a DC power flow network accounting for the transmission congestion, losses, and unit commitment constraints as a mixed integer program. A mathematical framework to construct dynamic models for electricity markets and study their competitive equilibria using DC power flow models is provided in Wang et al. [28].

A realistic market model is associated with important nonlinearities, arising from non-convex utility functions and nonlinear network constraints [3]. For instance, the DC power flow model may not be appropriate when voltage constraints or reactive power constraints are considered. Motto et al. [20] proposed a single time period decentralized electricity market clearing model that includes reactive power and demand responsiveness, based on the Alternating Current (AC) power flow network. More recently, Lavaei and Sojoudi [15] also investigate market efficiency for AC power flow networks by leveraging the zero duality gap of certain OPF formulations [16, 31].

While research has focused largely on aspects of electricity market design, there has been little work on algorithmic and computational aspects. This is especially important in the current context of grid infrastructure modernization and increased penetration of distributed generation. For instance, the DOE agency ARPA-E envisions a future grid infrastructure that incorporates diverse, distributed generation and storage sources and that is operated under a distributed control architecture [1]. In that context it is important to develop decentralized or distributed algorithms that scale with network size and have little overhead in communication. This serves as the motivation and focus of this chapter.

1.1 Our Focus

We consider a multiple time period pool-based electricity market consisting of generation companies (GenCo), load entities called the Distribution System Operators (DSO), and an Independent System Operator (ISO). We assume that (a) the DC power flow model is used by the ISO to model the power flow in the transmission system, (b) the DSOs are modeled as a single node neglecting the underlying distribution network, (c) the DSOs have the ability to defer loads, and (d) the GenCos and DSOs are price-taking and unwilling to share their cost function to the ISO. Maintaining privacy of the individual market participants motivates the development of a decentralized framework, whereby the ISO only transmits price signals to the individual participants and obtains price-sensitive optimal actions from them. Using such information, the ISO could employ a subgradient algorithm to converge to an equilibrium. However, the convergence rate for subgradient algorithms is known to be quite slow [11, 26]. Hence, these algorithms require significant numbers of message communications with the individual participants. This is undesirable in the current context of rapidly changing grid infrastructure which aims to incorporate intermittent distributed generation and distributed architectures for control and operation [1]. In such a distributed setting, reduction in communication overhead is important. Hence, fast convergence to an equilibrium is desirable for robust grid operation.

In this chapter, we exploit the problem structure to obtain decentralized optimization problems in the context of multiple time period market clearing. In such a scheme, the ISO transmits a price signal to the individual participants, who in turn solve their individual optimization problems, the solutions of which are communicated back to the ISO, so they may update the price. With this information, we propose that the ISO solves its market clearing problem by solving an implicit complementarity problem (ICP) as introduced in Curtis and Raghunathan [5]. To solve the ICP, Curtis and Raghunathan [5] propose a semismooth Newton algorithm for accelerating convergence when solving structured quadratic programs. We employ the same algorithm for solving the ISO's market clearing problem. The algorithm requires the computation of sensitivity of the market participants' solution to the price. We exploit the underlying communication protocol to additionally compute the sensitivity of their solution to changes in the price. Note that this requires no change in the computations performed by the GenCos and DSOs. We demonstrate through numerical experiments that our approach leads to orders of magnitude fewer function evaluations as compared to a subgradient method. The authors previously investigated the approach for single time period market clearing in [24] in which the GenCos and DSOs were also assumed to provide the sensitivity information. The approach described in this chapter removes this assumption.

We note that a similar approach has been considered when an AC power flow model is used and only equality constraints are present; see Motto et al. [19]. In this work, the authors propose applying dual decomposition to a non-convex nonlinear program for which the guarantees of finding a solution with zero duality gap do

not exist. Further, [19] employs a pure Newton strategy which does not have global convergence guarantees [22]. By contrast, in this chapter we consider the simpler DC power flow model which is convex, and, hence, there exists no duality gap. Further, we allow for inequality constraints and also argue that from a computational standpoint, the problem is better posed than the equality-constrained formulation. We also present numerical results showing that a pure Newton strategy, such as in [19], is not robust in converging to the solution. Our approach can also be extended to consider AC power flow models as in [19]. In fact, [5] also proposed a semismooth Newton algorithm for solving non-convex structured quadratic programs using semismooth Newton algorithms. The framework of [5] can be extended to the case of AC power flow models and will be investigated in a separate study.

1.2 Organization of the Chapter

The rest of the chapter is organized as follows. Models of the market participants and the notions of competitive equilibrium are presented in Section 2. An implicit complementarity formulation of the ISO’s market clearing problem is presented in Section 3. We describe the semismooth formulation and algorithm in Section 4. Numerical results demonstrating the efficacy of the method are presented in Section 5 followed by conclusions in Section 6.

2 Competitive Equilibrium

In this section, we describe the optimization problems related to each of the market participants: generation companies (GenCos), Distribution System Operators (DSOs), and the Independent System Operator (ISO). Based on these, we present the notion of competitive equilibrium and social welfare maximization. In what follows, \mathcal{N} denotes the set of buses in the transmission network of the ISO, while \mathcal{N}^G and \mathcal{N}^D (with $\mathcal{N} = \mathcal{N}^G \cup \mathcal{N}^D$), respectively, denote the nodes connected to GenCos and DSOs. Further, \mathcal{L} denotes the set of lines in the transmission network. We assume there are T time periods, and the set $\mathcal{T} = \{1, \dots, T\}$ represents the set of time periods. The electricity price at a node $i \in \mathcal{N}$ and time period $t \in \mathcal{T}$ is denoted by $\lambda_{i,t}$. We use boldface to denote vector quantities: $\boldsymbol{\lambda}_{i,\cdot} = (\lambda_{i,1}, \dots, \lambda_{i,T}) \in \mathbb{R}^T$ is the vector of prices over all time periods at the node i , $\boldsymbol{\lambda}_{\cdot,t} = (\lambda_{1,t}, \dots, \lambda_{|\mathcal{N}|,t}) \in \mathbb{R}^{|\mathcal{N}|T}$ is the vector of prices over the entire set of nodes in the time period t , and $\boldsymbol{\lambda} = (\lambda_{\cdot,1}, \dots, \lambda_{\cdot,T}) \in \mathbb{R}^{|\mathcal{N}|T}$ is the vector of all nodal prices for all time periods. Note that $(\lambda_{1,\cdot}, \dots, \lambda_{|\mathcal{N}|,\cdot})$ is a different ordering of the vector $\boldsymbol{\lambda}$. The power injection into the network at the node i at time period t is denoted by $P_{i,t}^*(\boldsymbol{\lambda})$. Similarly, $\boldsymbol{P}_{i,\cdot}^* \in \mathbb{R}^T$ is the vector of power injections at a node i over all time periods, and $\boldsymbol{P}_{\cdot,t}^* \in \mathbb{R}^{|\mathcal{N}|}$ is the vector of all nodal power injections in the time period t .

2.1 Generation Company (GenCo)

The generation company located at node $i \in \mathcal{N}^G$ chooses its optimal power generation level $\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot})$ over all time periods given the set of nodal prices over the time periods $\lambda_{i,\cdot}$ from the ISO by solving the optimization problem

$$\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}) = \arg \min_{P_1, \dots, P_T} \sum_{t \in \mathcal{T}} (c_i(P_t) - \lambda_{i,t} P_t) \quad (1a)$$

$$\text{s.t. } \underline{P}_i^G \leq P_t \leq \overline{P}_i^G, \forall t \in \mathcal{T} \quad (1b)$$

$$|P_{t+1} - P_t| \leq \Delta \overline{P}_i^G, \forall t \in \mathcal{T} \setminus \{T\} \quad (1c)$$

where $c_i(\cdot)$ is the cost of generation associated with the GenCo, \underline{P}_i^G and \overline{P}_i^G are, respectively, the minimum and maximum generation levels and $\Delta \overline{P}_i^G$ represents the limit on the change in power generation over successive time periods. We assume the following on the cost function of the GenCo, which implies that (1) has a unique solution.

Assumption 1 *The function $c_i(\cdot)$ is strictly convex.*

The optimization problem in (1) assumes that the cost function is independent of time periods. This is done for the sake of simplicity of exposition and is not a restriction of the approach. When considering multiperiod operations, GenCos schedule to operate additional units of generation which typically incurs a start-up cost. In addition, there are minimum down (up) periods for generation units once they are shut down (started up). Modeling such operations requires the introduction of binary variables which renders the GenCo problem non-convex. However, these non-convexities can be handled by relaxing the binary variables to be continuous and replacing the feasible region by the convex hull. We do not pursue this further, but refer the interested reader to [6, 25].

2.2 Distribution System Operator (DSO)

The DSO located at node $i \in \mathcal{N}^D$ chooses its optimal power consumption level $-\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot})$ over all time periods given the set of nodal prices over the time periods $\lambda_{i,\cdot}$ from the ISO by solving the optimization problem

$$\mathbf{P}_i^*(\lambda_{i,\cdot}) = \arg \min_{P_1, \dots, P_T} \sum_{t \in \mathcal{T}} (-\lambda_{i,t} P_t - u_i(-P_t)) \quad (2a)$$

$$\text{s.t. } \underline{P}_i^D \leq -P_t \leq \overline{P}_i^D, \quad (2b)$$

$$-\sum_{t \in \mathcal{T}} P_t \geq \underline{P}_i^{D,\text{tot}} \quad (2c)$$

where $u_i(\cdot)$ is the utility function of the DSO, \underline{P}_i^D and \overline{P}_i^D are minimum and maximum power consumption levels in a time period and $\underline{P}_i^{D,\text{tot}}$ represents a minimum total power consumption over the multiple time periods. Note that $P_i^*(\lambda_{i,\cdot})$ is negative since it represents power being withdrawn from the electrical network. We assume the following on the utility function of the DSO which ensures that (2) has a unique solution.

Assumption 2 *The function $u_i(\cdot)$ is strictly concave.*

The optimization problem in (2) assumes that the utility function is independent of time periods. This is done for the sake of simplicity of exposition and is not a restriction of the approach. Under Assumption 2 a DSO's optimization problem (2) is strictly convex and, hence, has an unique solution.

2.3 Independent System Operator (ISO)

The ISO is responsible for maintaining balance between the GenCos and DSOs, and ensuring that the power flows in the network are within certain limits. Given a vector of nodal prices λ over all time periods, the ISO chooses the optimal power injection levels by solving the optimization problem

$$\mathbf{P}^{\text{ISO}}(\lambda) = \arg \min_{(\mathbf{P}_{\cdot,1}, \dots, \mathbf{P}_{\cdot,T})} \sum_{t \in \mathcal{T}} \lambda_{\cdot,t}^T \mathbf{P}_{\cdot,t} \quad (3a)$$

$$\text{s.t. } \mathbf{1}^T \mathbf{P}_{\cdot,t} = 0, \forall t \in \mathcal{T} \quad (3b)$$

$$-\overline{\mathbf{P}} \leq \mathbf{A} \mathbf{P}_{\cdot,t} \leq \overline{\mathbf{P}}, \forall t \in \mathcal{T} \quad (3c)$$

where $\overline{\mathbf{P}} \in \mathbb{R}^{|\mathcal{L}|}$ denotes the vector of power limits on the lines in the network, $\mathbf{1} \in \mathbb{R}^{|\mathcal{N}|}$ is a vector of all ones, and \mathbf{A} is the matrix of power distribution factors for the ISO's transmission network. The constraint (3b) imposes power balance between the GenCos and DSOs at each time period. The DC power flow model appears in (3c) through the power distribution factors [30].

2.4 Competitive Equilibrium

A pair $(\widehat{\mathbf{P}}, \widehat{\lambda})$ is said to achieve *competitive (or Walrasian)* equilibrium for an electricity market if:

- (a) $\widehat{\mathbf{P}}_{i,\cdot} = \mathbf{P}_{i,\cdot}^*(\widehat{\lambda}_{i,\cdot}) \forall i \in \mathcal{N}^G$,
- (b) $\widehat{\mathbf{P}}_{i,\cdot} = \mathbf{P}_{i,\cdot}^*(\widehat{\lambda}_{i,\cdot}) \forall i \in \mathcal{N}^D$, and
- (c) $\widehat{\mathbf{P}} = \mathbf{P}^{\text{ISO}}(\widehat{\lambda})$.

By the well-known first and second fundamental theorems of welfare economics [18], we have the following:

- A competitive equilibrium is Pareto optimal.
- Every Pareto optimal allocation can be decentralized as a competitive equilibrium.

By the second fundamental theorem of welfare economics [18, 28], a competitive equilibrium can be characterized by maximizing social welfare given as

$$\min_{\mathbf{P}} \sum_{t \in \mathcal{T}} \left(\sum_{i \in \mathcal{N}^G} c_i(P_{i,t}) - \sum_{i \in \mathcal{N}^D} u_i(-P_{i,t}) \right) \quad (4a)$$

$$\text{s.t. } \mathbf{1}^T \mathbf{P}_{\cdot,t} = 0, \forall t \in \mathcal{T} \quad (4b)$$

$$-\bar{\mathbf{P}} \leq \mathbf{A} \mathbf{P}_{\cdot,t} \leq \bar{\mathbf{P}}, \forall t \in \mathcal{T} \quad (4c)$$

$$\underline{P}_i^G \leq P_{i,t} \leq \bar{P}_i^G, \forall i \in \mathcal{N}^G, t \in \mathcal{T} \quad (4d)$$

$$|P_{i,t+1} - P_{i,t}| \leq \Delta \bar{P}_i^G, \forall i \in \mathcal{N}^G, t \in \mathcal{T} \setminus \{T\} \quad (4e)$$

$$\underline{P}_i^D \leq -P_{i,t} \leq \bar{P}_i^D, \forall i \in \mathcal{N}^D, t \in \mathcal{T} \quad (4f)$$

$$-\sum_{t \in \mathcal{T}} P_{i,t} \geq \underline{P}_i^D, \forall i \in \mathcal{N}^D. \quad (4g)$$

Social welfare maximization achieves Pareto optimal allocation only under certain assumptions. Any electricity dispatch and pricing system is Pareto optimal only if prices are “right” and maximizes welfare only if all the important costs and benefits are priced into the system. For instance, it is well known that electric generation shifts some costs to society such that they are not priced in this market. Furthermore, even when prices are right, welfare is only maximized if willingness to pay is an accurate measure of utility. We do not delve further into these aspects but refer the reader to the texts [14, 18].

The social welfare maximization formulation in (4) is a *centralized formulation*. This does not lend itself to preserving privacy of the market participants. However, the formulation in (4) serves as the starting point for deriving the decentralized formulation which we do next.

3 Decentralized Market Formulation

We develop the decentralized market formulation based on Lagrange dualization of the coupling constraints in (4). For ease of presentation, we represent the power balance constraint in (4b) as two inequalities:

$$-\mathbf{1}^T \mathbf{P}_{\cdot,t} \leq 0, \mathbf{1}^T \mathbf{P}_{\cdot,t} \leq 0 \quad \forall t \in \mathcal{T}. \quad (1.4b')$$

Introducing multipliers $\underline{\xi}_t, \bar{\xi}_t$ for the power balance constraints in (4b) and $\underline{\zeta}_{l,t}, \bar{\zeta}_{l,t} \forall l \in \mathcal{L}$ for the line limit constraints in (4c), the partial Lagrangian for (4) can be written as

$$\begin{aligned} & L(\mathbf{P}, \underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}) \\ &= \sum_{t \in \mathcal{T}} \left(\sum_{i \in \mathcal{N}^G} c_i(P_{i,t}) - \sum_{i \in \mathcal{N}^D} u_i(-P_{i,t}) + (-\underline{\xi}_t + \bar{\xi}_t)(\mathbf{1}^T \mathbf{P}_{\cdot,t}) \right) \\ &\quad + \sum_{t \in \mathcal{T}} \left(\underline{\zeta}_t^T (-\bar{\mathbf{P}} - \mathbf{A} \mathbf{P}_{\cdot,t}) + \bar{\zeta}_t^T (\mathbf{A} \mathbf{P}_{\cdot,t} - \bar{\mathbf{P}}) \right). \end{aligned} \quad (5)$$

The Lagrangian dualization is restricted to the constraints that fall under the purview of the ISO's optimization problem (3). Using the partial Lagrangian in (5) and duality theory of convex optimization [17], we can decentralize the welfare maximization problem in (4) as explained below. Define the Lagrange dual function as

$$\begin{aligned} g(\underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}) &= \min_{\mathbf{P}} L(\mathbf{P}, \underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}) \\ \text{s.t. (4d) -- (4g).} \end{aligned} \quad (6)$$

From the definition of the partial Lagrangian in (5), it is easy to see that the objective function and constraints in (6) are separable by the GenCos and DSOs. Indeed, we can express the dual function as

$$\begin{aligned} & L(\mathbf{P}, \underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}) \\ &= \sum_{t \in \mathcal{T}} \left(\sum_{i \in \mathcal{N}^G} (c_i(P_{i,t}) - \lambda_{i,t} P_{i,t}) + \sum_{i \in \mathcal{N}^D} (-\lambda_{i,t} P_{i,t} - u_i(-P_{i,t})) \right) \\ &\quad - \sum_{t \in \mathcal{T}} (-\underline{\xi}_t + \bar{\xi}_t)^T \bar{\mathbf{P}} \end{aligned} \quad (7)$$

where $\lambda_{\cdot,t}$, the vector of nodal prices at time period t , is defined as

$$\lambda_{\cdot,t} = (\underline{\xi}_t - \bar{\xi}_t) \mathbf{1} + \mathbf{A}^T (\underline{\zeta}_t - \bar{\zeta}_t). \quad (8)$$

With this definition of the vector of nodal prices $\lambda_{\cdot,t}$, the optimization problem in (6) is precisely the set of optimization problems for GenCos (1) and DSOs (2). Thus, the Lagrangian dualization yields a decentralized formulation in which the ISO interacts with GenCos and DSOs through a price signal, thereby allowing the market participants to maintain the privacy of their optimization data.

To obtain the solution to (4), we rely on convex duality [17] which states the equivalence between (4) and its dual optimization problem given as

$$\begin{aligned} \max_{\underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}} & g(\underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}) \\ \text{s.t. } & (\underline{\xi}, \bar{\xi}, \underline{\zeta}, \bar{\zeta}) \geq 0. \end{aligned} \quad (9)$$

The solution to (9) could be obtained through a projected subgradient algorithm [11, 26] which is stated in Algorithm 1. For ease of presentation we introduce

$$\mathbf{v}_t = \begin{pmatrix} \underline{\xi}_t \\ \bar{\xi}_t \\ \underline{\zeta}_t \\ \bar{\zeta}_t \end{pmatrix}, \mathbf{F}_t(\mathbf{v}) = \begin{pmatrix} -\mathbf{1}^T \mathbf{P}_{\cdot,t}^*(\lambda) \\ \mathbf{1}^T \mathbf{P}_{\cdot,t}^*(\lambda) \\ \mathbf{A} \mathbf{P}_{\cdot,t}^*(\lambda) + \bar{\mathbf{P}} \\ -\mathbf{A} \mathbf{P}_{\cdot,t}^*(\lambda) + \bar{\mathbf{P}} \end{pmatrix} \forall t \in \mathcal{T} \quad (10)$$

$$\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_T), \mathbf{F}(\mathbf{v}) = (\mathbf{F}_1(\mathbf{v}), \dots, \mathbf{F}_T(\mathbf{v}))$$

where \mathbf{F} denotes the vector of dualized constraints and \mathbf{v} the corresponding multipliers. Note that \mathbf{F}_t is denoted as a function of \mathbf{v} , multipliers over all time periods, since the optimization problem for GenCos (1) and DSOs (2) are coupled across time periods. The update step for the multipliers in Algorithm 1 only requires access to the optimal solution of the GenCos and DSOs. Thus, they are quite simple to implement and fit the decentralized framework very well. The typical number of iterations required for convergence of the algorithm to a solution that is within ϵ of the optimal solution is $O(\frac{1}{\epsilon^2})$. Thus, a large number of communication rounds are required between the ISO and the market participants (GenCos and DSOs) to achieve convergence. This renders the algorithm quite slow in practice.

Algorithm 1: Subgradient Algorithm

- 1 Let $\epsilon \in (0, 1)$ be a desired convergence tolerance
 - 2 Choose $\mathbf{v}_t^0 = (\underline{\xi}_t^0, \bar{\xi}_t^0, \underline{\zeta}_t^0, \bar{\zeta}_t^0)$ for all $t \in \mathcal{T}$ as the initial guess.
 - 3 Set $k = 0$.
 - 4 **repeat**
 - 5 Set λ^k according to (8)
 - 6 For $i \in \mathcal{N}^G$, solve (1) with $\lambda_{i,\cdot} = \lambda_{i,\cdot}^k$.
 - 7 For $i \in \mathcal{N}^D$, solve (2) with $\lambda_{i,\cdot} = \lambda_{i,\cdot}^k$.
 - 8 Choose γ^k (typically $\gamma^k = \frac{a}{k+1}$ for some some $a > 0$)
 - 9 Set $\mathbf{v}^{k+1} = \max(0, \mathbf{v}^k - \gamma^k \mathbf{F}(\mathbf{v}^k))$
 - 10 Set $e^{k+1} = \|\min(\mathbf{v}^k, \mathbf{F}(\mathbf{v}^k))\|_\infty$
 - 11 Set $k = k + 1$
 - 12 **until** $e^k \leq \epsilon$
-

4 Semismooth Equation Approach

We describe the semismooth equation approach of Curtis and Raghunathan [5] for computing the competitive equilibrium. The optimality conditions [17] for the ISO's problem for all $t \in \mathcal{T}$ are

$$\lambda_{\cdot,t} = (\underline{\xi}_t - \bar{\xi}_t)\mathbf{1} + A^T(\underline{\xi}_t - \bar{\xi}_t) \quad (11a)$$

$$0 \leq \underline{\xi}_t \perp (\mathbf{1}^T \mathbf{P}_{\cdot,t}) \geq 0 \quad (11b)$$

$$0 \leq \bar{\xi}_t \perp (-\mathbf{1}^T \mathbf{P}_{\cdot,t}) \geq 0 \quad (11c)$$

$$0 \leq \underline{\xi}_t \perp (A \mathbf{P}_{\cdot,t} + \bar{\mathbf{P}}) \geq 0 \quad (11d)$$

$$0 \leq \bar{\xi}_t \perp (-A \mathbf{P}_{\cdot,t} + \bar{\mathbf{P}}) \geq 0 \quad (11e)$$

where for a pair of vectors $\{a, b\}$, the expression $0 \leq a \perp b \geq 0$ represents the conditions $a_i \geq 0$, $b_i \geq 0$, and $a_i b_i = 0$ for all i . The constraints in (11b)–(11e) are the so-called *complementarity constraints* [17]. Following the definition in Section 2.4, competitive equilibrium is attained when the conditions in (11) hold for $\mathbf{P} = \mathbf{P}^*(\lambda)$. Following [5], we pose the ISO's market clearing problem as the following *implicit complementarity problem* (ICP):

$$0 \leq \mathbf{v} \perp \mathbf{F}(\mathbf{v}) \geq 0 \quad (12)$$

where $(\mathbf{v}, \mathbf{F}) \in \mathbb{R}^m \times \mathbb{R}^m$ are as defined in (10) with $m = (2 + 2|\mathcal{L}|)T$. We call the complementarity problem in (12) as *implicit* since $\mathbf{P}^*(\lambda)$, which appears in computation of $\mathbf{F}(\mathbf{v})$, is obtained by solving a set of optimization problems. Observe that the evaluation of $\mathbf{P}^*(\lambda)$ only requires communication with the GenCos and DSOs through transmission of the price vector λ . Thus, the ICP (12) has the desired property of decoupling by participants and allows the participants to maintain *privacy* of their optimization problem.

The following theorem formalizes the equivalence between the ICP (12) and the competitive equilibrium.

Theorem 1 *The following are equivalent:*

- (a) $(\widehat{\mathbf{P}}, \widehat{\lambda})$ is a competitive equilibrium;
- (b) $\widehat{\mathbf{v}}$ solves the ICP (12) with $\widehat{\lambda}_{\cdot,t} = (\widehat{\underline{\xi}}_t - \widehat{\bar{\xi}}_t)\mathbf{1} + A^T(\widehat{\underline{\xi}}_t - \widehat{\bar{\xi}}_t)$.

Proof First, we show that (a) implies (b). Suppose (a) holds. From the definition of competitive equilibrium in Section 2.4, $\widehat{\mathbf{P}} = \mathbf{P}^*(\widehat{\lambda})$. Since $\widehat{\mathbf{P}}$ solves the ISO's problem (3), we have that there exists multipliers $(\widehat{\underline{\xi}}, \widehat{\bar{\xi}}, \widehat{\underline{\xi}}, \widehat{\bar{\xi}})$ satisfying the optimality conditions in (11) with $\mathbf{P} = \mathbf{P}^*(\widehat{\lambda})$. Thus, (a) holds. Now, suppose (b) holds. By the preceding arguments, we have that first-order stationarity conditions

of the ISO's problem (3) hold. Since (3) is convex, a first-order stationary point is also a minimizer [17]. This completes the proof.

To solve the ICP, we rewrite the complementarity system using the Fischer-Burmeister operator as

$$\Phi^{\text{FB}}(\mathbf{v}) = \begin{pmatrix} \phi(v_1, F_1(\mathbf{v})) \\ \vdots \\ \phi(v_m, F_m(\mathbf{v})) \end{pmatrix}, \quad (13)$$

where, given scalars a and b , the Fischer-Burmeister function [10] has the form

$$\phi(a, b) = \sqrt{a^2 + b^2} - a - b. \quad (14)$$

Clearly, this latter function satisfies

$$\phi(a, b) = 0 \iff \{a \geq 0, b \geq 0, \text{ and } ab = 0\}. \quad (15)$$

The articles [9, 21] discuss regularity properties and sophisticated implementations of semismooth Newton algorithms for complementarity problems using the Fischer-Burmeister function. However, our formulation here is different in the sense that, in our context, the complementarity components \mathbf{v} and $\mathbf{F}(\mathbf{v})$ are both functions of \mathbf{v} ; hence, our formulation is somewhat more straightforward.

At each iteration k of the semismooth Newton algorithm [23], the step $d\mathbf{v}^k$ is obtained as the solution of

$$\Phi^{\text{FB}}(\mathbf{v}^k) + H^k d\mathbf{v}^k = 0, \quad (16)$$

where H^k represents the first-order variation of the function Φ^{FB} at the point \mathbf{v}^k . We postpone the discussion on the computation of the matrix H^k to Section 4.3 and instead focus on the local convergence properties and algorithmic details. The step $d\mathbf{v}^k$ obtained by solving (16) is called the *semismooth Newton* step.

4.1 Fast Local Convergence

Semismooth functions such as Φ^{FB} are almost everywhere differentiable [23]. Further, at points of non-differentiability, Φ^{FB} is directionally differentiable and can be approached through a sequence of differentiable points. Consequently, for any sequence of directions $d\mathbf{v} \rightarrow 0$ with associated Jacobians $H \in \partial\Phi(\mathbf{v} + d\mathbf{v})$ and directional derivatives $(\Phi^{\text{FB}})'(\mathbf{v}; d\mathbf{v})$, we have that

$$\|Hd\mathbf{v} - (\Phi^{\text{FB}})'(\mathbf{v}; d\mathbf{v})\| = o(\|d\mathbf{v}\|). \quad (17)$$

This Taylor-series-like property is sufficient to show that iterations defined by (16) can converge locally superlinearly.

Theorem 2 ([23]) Suppose that \mathbf{F} is continuously differentiable and \mathbf{v}^* satisfies $\Phi^{\text{FB}}(\mathbf{v}^*) = 0$ such that all $H \in \partial\Phi^{\text{FB}}(\mathbf{v}^*)$ are non-singular. Then, for any \mathbf{v}^k in a sufficiently small neighborhood of \mathbf{v}^* , it follows that $\|\mathbf{v}^{k+1} - \mathbf{v}^*\| \leq C\|\mathbf{v}^k - \mathbf{v}^*\|^{1+\gamma}$ for some $C > 0$ and $\gamma > 0$.

In the present setting, \mathbf{F} is not continuously differentiable, only piecewise differentiable (PC^1) since $\mathbf{P}_{i,\cdot}^*(\cdot)$ are PC^1 [5]. The main result in [5] proves local superlinear convergence for $\mathbf{F} \in PC^1$. Hence, the semismooth Newton iteration [5] converges fast locally, unlike a conventional subgradient approach. We provide numerical evidence for this in Section 5.

4.2 Algorithm

To promote global convergence, we employ a line-search based on the merit function $\Psi^{\text{FB}}(\mathbf{v}) := \|\Phi^{\text{FB}}(\mathbf{v})\|_2^2$, the 2-norm of the vector $\Phi^{\text{FB}}(\mathbf{v})$. Observe that the minimum of $\Psi^{\text{FB}}(\mathbf{v})$ is 0 corresponding to a solution of the ICP (12). Thus, reduction of the merit function $\Psi^{\text{FB}}(\mathbf{v})$ can be used to certify that the steps of the algorithm ultimately decrease the distance to a solution of the ICP. Given a direction $d\mathbf{v}^k$, the step length α^k is determined as the largest $\alpha^k \in (0, 1]$ such that the sufficient decrease condition

$$\Psi^{\text{FB}}(\mathbf{v}^k + \alpha^k d\mathbf{v}^k) \leq \Psi^{\text{FB}}(\mathbf{v}^k) + \eta\alpha^k \nabla\Psi^{\text{FB}}(\mathbf{v}^k)^T d\mathbf{v}^k \quad (18)$$

holds where $\eta \in (0, 1)$; e.g., one typically chooses $\eta = 10^{-4}$. The step length α^k may be obtained using a backtracking line-search starting from 1 and multiplying by a constant factor $\rho \in (0, 1)$ until the sufficient decrease condition (18) holds. The complete steps of the algorithm are provided in Algorithm 2. At each iteration of the algorithm, the ISO computes the price vector λ (Step 5) and communicates the nodal price vector $\lambda_{i,\cdot}^k$ to the GenCos and DSOs to obtain their optimal power generation and consumption levels (Steps 6 and 7). The sensitivity of these power levels to the nodal prices is computed by finite difference in Step 8. To compute the sensitivity of a particular participant, $i \in \mathcal{N}$ requires $2 \cdot T$ calls to the participant to solve the respective optimization problems (1) or (2) for different perturbations of the price vector. This operation can be performed in parallel for each participant $i \in \mathcal{N}$. We emphasize again that the computation of the sensitivity does not require any modification in the optimization problems of the market participants. The computation of the matrix H^k in Step 9 is described in Section 4.3.

Algorithm 2: Semismooth Newton Algorithm

-
- 1 Choose a convergence tolerance $\epsilon \in (0, 1)$.
 2 Choose an initial guess $\mathbf{v}_t^0 = (\underline{\xi}_t^0, \bar{\xi}_t^0, \underline{\zeta}_t^0, \bar{\zeta}_t^0)$ for all $t \in \mathcal{T}$. Choose $\{\eta, \rho\} \subset (0, 1)$.
 3 Set $k = 0$.
- 4 repeat**
- 5 Set λ^k according to (8).
 6 For $i \in \mathcal{N}^G$, compute $\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}^k)$ from (1)
 7 For $i \in \mathcal{N}^D$, compute $\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}^k)$ from (2).
 8 For $i \in \mathcal{N}$, $t \in \mathcal{T}$ compute $\frac{\partial \mathbf{P}_{i,\cdot}^*}{\partial \lambda_{i,t}}$ as

$$\text{Set } \lambda_{i',t'}^\pm = \begin{cases} \lambda_{i',t'} & \text{for } i' \neq i, t' \neq t \\ \lambda_{i,t} \pm \delta & \text{for some } \delta > 0. \end{cases}$$
 Compute $\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}^+), \mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}^-)$ from (1) for $i \in \mathcal{N}^G$ or (2) for $i \in \mathcal{N}^D$

$$\text{Set } \frac{\partial \mathbf{P}_{i,\cdot}^*}{\partial \lambda_{i,t}} = \frac{\mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}^+) - \mathbf{P}_{i,\cdot}^*(\lambda_{i,\cdot}^-)}{2\delta}.$$
- 9 Compute H^k using (19) and $d\mathbf{v}^k$ using (16).
 10 Find the smallest integer $n \geq 0$ such that (18) holds for $\alpha^k = \rho^n$.
 11 Set $\mathbf{v}^{k+1} = \mathbf{v}^k + \alpha^k d\mathbf{v}^k$ and $k = k + 1$
12 until $\|\Phi^{\text{FB}}(\mathbf{v}^k)\|_\infty \leq \epsilon$
-

4.3 Computing H^k

The matrix H^k is defined as

$$H^k = D_{\mathbf{v}}^k + D_{\mathbf{F}}^k \nabla_{\mathbf{v}} \mathbf{F}(\mathbf{v}^k)^T \quad (19)$$

where

$$\nabla_{\mathbf{v}} \mathbf{F}(\mathbf{v}^k) = [\nabla_{\mathbf{v}} F_1(\mathbf{v}^k) \cdots \nabla_{\mathbf{v}} F_m(\mathbf{v}^k)] \quad (20)$$

with $F_j(\cdot)$ in (20) denoting the j th component of \mathbf{F} and $\nabla_{\mathbf{v}} F_j(\cdot)$ is the gradient of $F_j(\cdot)$ with respect to \mathbf{v} . Note that $F_j(\cdot)$ is different from the boldface notation $\mathbf{F}_j(\cdot)$ used in (10) and is being used to simplify the presentation of the matrices $D_{\mathbf{v}}^k$ and $D_{\mathbf{F}}^k$. Likewise, v_j represents the j th component of the m -dimensional vector \mathbf{v} and is different from the boldface notation \mathbf{v}_j in (10). The matrices $D_{\mathbf{v}}^k$ and $D_{\mathbf{F}}^k$ are diagonal and are defined as described next. Introducing the set $\beta^k = \{j \mid v_j^k = 0 = F_j(\mathbf{v}^k)\}$, the diagonal matrices can be obtained as

$$[D_{\mathbf{v}}^k]_{jj} = \begin{cases} \left(\frac{v_j^k}{\|(v_j^k, F_j(\mathbf{v}^k))\|} - 1 \right) & \forall j \notin \beta^k \\ \left(\frac{z_j}{\|(z_j, z^T \nabla F_j(\mathbf{v}^k))\|} - 1 \right) & \forall j \in \beta^k \end{cases}$$

$$[D_{\mathbf{F}}^k]_{jj} = \begin{cases} \left(\frac{F_j(\mathbf{v}^k)}{\|(v_j^k, F_j(\mathbf{v}^k))\|} - 1 \right) & \forall j \notin \beta^k \\ \left(\frac{z^T \nabla F_j(\mathbf{v}^k)}{\|(z_j, z^T \nabla F_j(\mathbf{v}^k))\|} - 1 \right) & \forall j \in \beta^k \end{cases}$$

where z is chosen such that $z_j = 1$ for $j \in \beta^k$ and 0 otherwise [8].

To present the expression for the matrix $\nabla_{\mathbf{v}} \mathbf{F}(\mathbf{v}^k)^T$, we recall from (8) and (10) that the vectors \mathbf{v} and \mathbf{F} have the following structure:

$$\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_T), \quad \lambda_{\cdot,t} = \mathbf{B} \mathbf{v}_t,$$

$$\text{and } \mathbf{F}(\mathbf{v}) = (F_1(\mathbf{v}), \dots, F_T(\mathbf{v})), \quad F_t(\mathbf{v}) = \mathbf{B}^T \mathbf{P}_{\cdot,t}^*(\lambda) + \mathbf{b}, \quad (21)$$

$$\text{where } \mathbf{B} = [-\mathbf{1} \ \mathbf{1} \ -\mathbf{A}^T \ \mathbf{A}^T], \quad \mathbf{b}^T = [0 \ 0 \ \overline{\mathbf{P}}^T \ \overline{\mathbf{P}}^T].$$

Then the change in the function $F_t(\mathbf{v}^k)$ due to a perturbation $\Delta \mathbf{v}$ in the variables \mathbf{v}^k can be approximated to the first order as

$$\begin{aligned} F_t(\mathbf{v}^k + \Delta \mathbf{v}) - F_t(\mathbf{v}^k) &= \mathbf{B}^T (\mathbf{P}_{\cdot,t}^*(\lambda^k + \Delta \lambda) - \mathbf{P}_{\cdot,t}^*(\lambda^k)) \\ &\approx \mathbf{B}^T \left(\sum_{s=1}^T \nabla_{\lambda_{\cdot,s}} \mathbf{P}_{\cdot,t}^*(\lambda^k)^T \Delta \lambda_{\cdot,s} \right) = \mathbf{B}^T \left(\sum_{s=1}^T \frac{\partial \mathbf{P}_{\cdot,t}^*(\lambda^k)}{\partial \lambda_{\cdot,s}} \mathbf{B} \Delta \mathbf{v}_{\cdot,s} \right) \end{aligned}$$

where in the last equality, we have used $\Delta \lambda_{\cdot,s} = \mathbf{B} \Delta \mathbf{v}_s$ by (21) and $\frac{\partial \mathbf{P}_{\cdot,t}^*(\lambda^k)}{\partial \lambda_{\cdot,s}}$ is a diagonal matrix with $\left[\frac{\partial \mathbf{P}_{\cdot,t}^*(\lambda^k)}{\partial \lambda_{\cdot,s}} \right]_{jj} = \frac{\partial P_{j,t}^*(\lambda^k)}{\partial \lambda_{j,s}}$ for $j = 1, \dots, |\mathcal{N}|$ and is obtained using the computed sensitivities (step 8 in Algorithm 2). Thus, the Jacobian $\nabla_{\mathbf{v}} \mathbf{F}(\mathbf{v}^k)^T$ can be expressed as

$$\nabla_{\mathbf{v}} \mathbf{F}(\mathbf{v}^k)^T = \begin{bmatrix} \mathbf{B}^T \frac{\partial \mathbf{P}_{\cdot,1}^*(\lambda^k)}{\partial \lambda_{\cdot,1}} \mathbf{B} & \cdots & \mathbf{B}^T \frac{\partial \mathbf{P}_{\cdot,T}^*(\lambda^k)}{\partial \lambda_{\cdot,T}} \mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{B}^T \frac{\partial \mathbf{P}_{\cdot,1}^*(\lambda^k)}{\partial \lambda_{\cdot,1}} \mathbf{B} & \cdots & \mathbf{B}^T \frac{\partial \mathbf{P}_{\cdot,T}^*(\lambda^k)}{\partial \lambda_{\cdot,T}} \mathbf{B} \end{bmatrix}.$$

The matrix H^k is dense, and there is no discernible structure that can be exploited in the solution of the linear system (16).

5 Numerical Results

We consider IEEE networks for testing the performance of the Algorithms 1 and 2. The algorithms were implemented in MATLAB and executed on a machine with a 3.2 GHz Intel Core i7-3930K CPU with 32 GB RAM. Table 1 presents information on the number of GenCos, DSOs, and lines in the different test cases. The rest of the section is organized as follows: Section 5.1 presents the results for single time period market clearing, while Section 5.2 presents multiple time period market clearing.

5.1 Single-Period Market Clearing

We choose the cost function for the GenCos as a strictly convex quadratic function, $c_i(P) = c_{1i}P + c_{2i}P^2$ where $c_{2i} > 0$. The values for the coefficients c_{1i} and c_{2i} are generated randomly. The utility function for the DSOs is chosen as a strictly concave quadratic function, $u_i(-P) = u_{i1}(-P) + u_{2i}P^2$ where $u_{2i} < 0$. The coefficient values u_{i1} and u_{2i} are generated randomly. The demands at the buses are allowed to vary between 80% and 120% of the nominal demand specified in the test cases available in MATPOWER [32]. Table 2 summarizes the performance statistics of Algorithm 2 versus Algorithm 1 (a subgradient algorithm) for a single time period ($T = 1$) market clearing problem. The reported numbers are averaged over ten different runs in which the DSO's utility functions and demands are varied. The convergence tolerance for both algorithms was set to $\epsilon = 10^{-6}$. Note that the error measures used for Algorithms 1 and 2 are distinct but equivalent measures of the error in satisfying the ICP (12). The subgradient algorithm hits the iteration limit of 100000 on most instances, whereas Algorithm 2 solves the problems in very few iterations with modest function evaluation counts. Further, Algorithm 2 is two–three orders of magnitude faster than the subgradient algorithm in terms of CPU time. The number of function evaluations in Table 2 also includes those required for the sensitivity matrices $\partial \mathbf{P}_{\cdot,i}^*/\partial \lambda_{\cdot,t}$ in Step 8 of Algorithm 2.

Figure 1 plots the typical progress of the error ($\|\Phi^{\text{FB}}(\mathbf{v}^k)\|_2$) in satisfying ICP (12) against the iteration index. The semismooth Newton algorithm dominates the subgradient method for all tolerance levels. Further, the convergence rate is indeed superlinear as predicted by Theorem 2 and is key to explaining the observed acceleration in convergence over the subgradient method.

Table 1 Problem size information for the test instances

Name	$ \mathcal{N}^G $	$ \mathcal{N}^D $	$ \mathcal{L} $
case9	3	3	9
case14	5	11	20
case30	6	20	41
case39	10	21	46
case57	7	42	80
case118	54	99	186
case300	69	191	411

Table 2 Results for the single time period ($T = 1$) market clearing problem using algorithms. m , size of the vector v ; Avg. #Iters., average number of iterations; Avg. #Fcn., average number of function evaluations; Avg. CPU (s), average CPU time in seconds

Name	m	Algorithm 2—Semismooth			Algorithm 1—Subgradient	
		Avg. #Iters.	Avg. #Fcn.	Avg. CPU (s)	Avg. #Iters.	Avg. CPU (s)
case9	20	5.4	28.7	0.03	100000	1.8
case14	42	5.7	59.0	0.06	100000	2.1
case30	84	5.2	26.5	0.05	100000	3.1
case39	94	10.0	109.7	0.13	43262	1.6
case57	162	6.8	33.1	0.12	100000	2.7
case118	374	6.2	42.0	0.86	100000	4.5
case300	824	7.2	28.7	4.03	100000	20.1

Fig. 1 Plot of error against iteration index for the algorithms

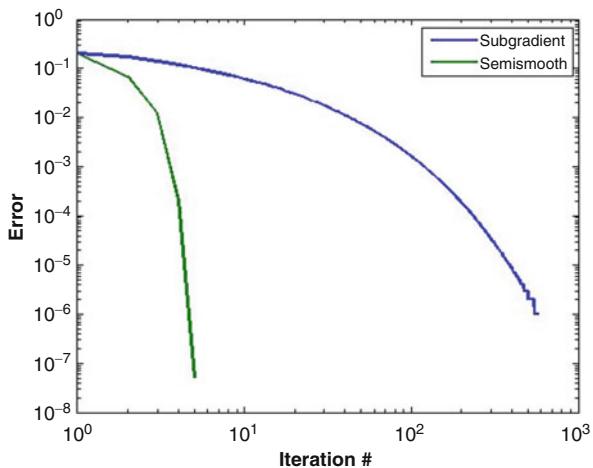


Table 3 Results for the single time period ($T = 1$) market clearing problem. Avg. #Iters., average number of iterations; Avg. #Fcn., average number of function evaluations; Avg. CPU (s), average CPU time in seconds

Name	Algorithm in Motto et al. [19]		
	Avg. #Iters.	Avg. #Fcn.	Avg. CPU (s)
case9	5.0	16.0	0.02
case14	1000.0	3000.0	4.71
case30	901.4	2705.2	6.58
case39	8.0	25.0	0.07
case57	1000.0	3000.0	13.91
case118	1000.0	3000.0	83.31
case300	24.1	73.3	13.64

As mentioned in the introduction, Motto et al. [19] had also proposed an approach that is quite similar to the implicit equation approach. The authors employed a pure Newton strategy without any line-search. Table 3 presents the results using the pure Newton algorithm of [19]. The algorithm was set a limit of 1000 iterations. Table 3 clearly shows that employing the pure Newton strategy is not robust. The

algorithm of [19] stops on attaining the iteration limit on all instances of case14, case57, and case118 and on nine of the ten instances of case30. On the test cases where all the instances were solved – case9, case39, and case300 – the iteration count is comparable to that of the semismooth Newton algorithm proposed in this paper. Thus, it is quite evident that in single-period market clearing, the semismooth Newton algorithm (Algorithm 2) based on the implicit complementarity (ICP) formulation (12) is computationally efficient and robust in its performance.

5.2 Multiperiod Market Clearing

We now explore the computational performance of the semismooth Newton algorithm when solving multiperiod market clearing problems as the number of time periods is varied. We consider five different time periods $T \in \{2, 4, 8, 16, 32\}$. In the multiperiod setting, we impose that $\Delta \bar{P}_i^G = 0.25(\bar{P}_i^G - \underline{P}_i^G)$ and $\underline{P}_i^{D,\text{tot}}$ to be the nominal demand specified in the input file multiplied by the number of time periods. Table 4 lists the size of the vector of unknowns $v \in \mathbb{R}^m$ in the implicit complementarity formulation (12) for the different problem instances and time periods. The size of the problem m dictates the number of floating point operations required to solve the linear system in (16) in order to compute the Newton step $d v^k$ at each iteration of Algorithm 2. Since the matrix H^k is expected to be dense, the number of floating point operations required scales as m^3 and will be reflected in the computational time of the algorithm. We will highlight this aspect in our discussion on CPU times.

Tables 5 and 6 list the number of iterations and function evaluations taken by Algorithm 2 on the different instances and time periods. From the tables it is clear that the number of iterations of Algorithm 2 is independent of the increase in the number of time periods. This is a very desirable feature for practical algorithms. However, the number of function evaluations scales linearly with the number of time periods.

Table 7 lists the CPU time in seconds taken by the algorithm on the different problem instances and time periods. The reported times include the time performing the step computation in (16) and also for the function evaluations. The number reported in the parenthesis is the percentage of time that is spent in computing the sensitivity matrices $\partial P_i^*/\partial \lambda_i$. In our implementation the sensitivity computations

Table 4 Summary of the number of constraints in the implicit complementarity problem (ICP) formulation (12) for the different instances and time periods

Name	$T = 2$	$T = 4$	$T = 8$	$T = 16$	$T = 32$
case9	40	80	160	320	640
case14	84	168	336	672	1344
case30	168	336	672	1344	2688
case39	188	376	752	1504	3008
case57	324	648	1296	2592	5184
case118	748	1496	2992	5984	11968
case300	1648	3296	6592	13184	26368

Table 5 Summary of the iterations taken by the semismooth Newton algorithm (Algorithm 2) to solve the multiperiod market clearing problem

Name	$T = 2$	$T = 4$	$T = 8$	$T = 16$	$T = 32$
case9	5	5	5	5	5
case14	6	6	6	5	5
case30	5	5	5	4	4
case39	10	10	10	10	10
case57	4	4	4	4	4
case118	5	5	5	5	5
case300	7	6	6	6	6

Table 6 Summary of the function evaluations taken by the semismooth Newton algorithm (Algorithm 2) to solve the multiperiod market clearing problem

Name	$T = 2$	$T = 4$	$T = 8$	$T = 16$	$T = 32$
case9	36	76	156	316	636
case14	94	142	238	367	687
case30	41	81	161	258	514
case39	154	234	394	714	1354
case57	29	61	125	253	509
case118	49	89	169	329	649
case300	74	115	211	403	787

Table 7 Summary of the CPU time in seconds taken by the semismooth Newton algorithm (Algorithm 2) to solve the multiperiod market clearing problem. The number in the parenthesis is the percentage of time spent in evaluating the sensitivities

Name	$T = 2$	$T = 4$	$T = 8$	$T = 16$	$T = 32$
case9	0.2 (93.6%)	0.5 (96.9%)	1.0 (98.0%)	2.2 (97.5%)	5.0 (96.9%)
case14	0.7 (86.7%)	1.6 (92.0%)	3.4 (93.8%)	6.3 (93.6%)	14.6 (90.7%)
case30	0.9 (96.3%)	2.1 (96.4%)	4.9 (95.7%)	9.3 (93.8%)	22.9 (86.6%)
case39	2.1 (90.1%)	4.9 (93.2%)	11.3 (92.7%)	26.5 (88.6%)	67.9 (79.2%)
case57	1.4 (97.6%)	3.8 (95.6%)	8.9 (92.8%)	22.0 (85.6%)	70.9 (74.7%)
case118	7.3 (96.0%)	20.8 (93.3%)	49.5 (88.4%)	169.9 (77.4%)	736.4 (64.1%)
case300	36.3 (93.7%)	75.0 (88.2%)	278.2 (78.4%)	1253.0 (64.9%)	9182.8 (37.4%)

for all the participants are performed serially. If these computations are performed in parallel as will be the case in a practical implementation, then the expected speedups are reported in Table 8. The speed up is computed as

$$\text{speedup} = \frac{\tau_{cpu}}{\tau_{cpu} - \tau_{sen} + \tau_{sen}/|\mathcal{N}|}$$

where τ_{cpu} is the total CPU time taken by Algorithm 2 as reported in Table 7 and τ_{sen} is the CPU time spent in sensitivity evaluation. Note that this computation does include the communication overheads that are typically involved in a parallel computing framework. From Table 8 it is evident that we can attain almost an order of magnitude speedup up to time periods $T \leq 4$ on the larger instances. However, as

Table 8 Summary of the potential speedup in computations when parallel computations are taken into consideration

Name	$T = 2$	$T = 4$	$T = 8$	$T = 16$	$T = 32$
case9	4.54	5.19	5.46	5.33	5.19
case14	5.34	7.27	8.29	8.17	6.70
case30	13.48	13.73	12.49	10.21	5.96
case39	7.78	10.19	9.69	7.00	4.29
case57	22.83	15.72	11.02	6.19	3.73
case118	21.73	13.64	8.22	4.32	2.75
case300	15.12	8.24	4.56	2.83	1.59

the number of time periods increases, the time involved in the step computation (16) dominates the overall CPU time, and as a consequence the speedups are not significant.

6 Conclusions

In this paper, we have presented a novel semismooth Newton algorithm for multiperiod electricity markets. The approach is decentralized in that it only requires the GenCos and DSOs to communicate their optimal response to the price signal from the ISO. The proposed approach is shown to be robust in converging to a tight tolerance of 10^{-6} . For single-period market clearing, the proposed algorithm requires about four orders of magnitude fewer function evaluations than a subgradient algorithm. Our numerical experiments demonstrate that the algorithm scales very well with the number of time periods. The communication requirement for the semismooth Newton algorithm (Algorithm 2) is identical to that of the subgradient algorithm (Algorithm 1). Hence, the proposed approach can be readily implemented in practice.

There are a number of extensions for this work. We outline some of them below.

- In the current paper, the GenCo problem (1) does not include startup or shutdown costs and minimum up or down time for generators. Modeling such operations requires the introduction of binary variables which renders the GenCo problem non-convex. Our algorithm can be easily extended to the GenCo problem resulting from relaxing the binary variables to be continuous and replacing the feasible region by the convex hull [6, 25].
- The current chapter assumes a lumped model for DSOs and no distributed generation. It is also possible to extend our approach to situations in which the electrical network of each DSO is modeled explicitly. We believe this is a straightforward extension.
- We will also investigate the applicability of the approach when the DSO's power flow is modeled using AC power flow equations. In this context, we will also explore the convex SDP relaxation [15] which has shown to have zero duality gap in a number of instances.

Acknowledgements We are grateful to the referees for a careful reading of the manuscript and bringing to our attention the subtleties of social welfare maximization.

References

1. Advanced Research Project Agency - Energy (ARPA-E) (2015) Network optimized distributed energy systems (NODES). Technical report DE-FOA-0001289, ARPA-E
2. Baldick R (2002) Electricity market equilibrium models: the effect of parametrization. *IEEE Trans Power Syst* 17(4):1170–1176
3. Bautista G, Anjos MF, Vannelli A (2007) Formulation of oligopolistic competition in AC power networks: an NLP approach. *IEEE Trans Power Syst* 22(1):105–115
4. Bompard E, Ma YC, Napoli R, Gross G, Guler T (2010) Comparative analysis of game theory models for assessing the performances of network constrained electricity markets. *IET Gener Transm Distrib* 4(3):386–399
5. Curtis FE, Raghunathan AU (2017) Solving nearly-separable quadratic optimization problems as nonsmooth equations. *Comput Optim Appl* 67(2):317–360
6. Damci-Kurt P, Küçükyavuz S, Rajan D, Atamtürk A (2016) A polyhedral study of production ramping. *Math Program* 158(1–2):175–205
7. Day CJ, Hobbs BF, Pang JS (2002) Oligopolistic competition in power networks: a conjectured supply function approach. *IEEE Trans Power Syst* 17(3):597–607
8. De Luca T, Facchinei F, Kanzow C (1996) A semismooth equation approach to the solution of nonlinear complementarity problems. *Math Program* 75:407–439
9. Facchinei F, Fischer A, Kanzow C (1998) Regularity properties of a semismooth reformulation of variational inequalities. *SIAM J Optim* 8:850–869
10. Fischer A (1992) A special Newton-type optimization method. *Optimization* 24:269–284
11. Goffin J-L (1977) On the convergence rate of subgradient optimization methods. *Math Program* 13:329–347
12. Hobbs BF, Helman U (2004) Complementarity-based equilibrium modeling for electric power markets. In: Modeling prices in competitive electricity markets. Wiley series in financial economics. Wiley, West Sussex, pp 69–98
13. Hobbs BF, Metzler CB, Pang JS (2000) Strategic gaming analysis for electric power systems: an MPEC approach. *IEEE Trans Power Syst* 15(2):638–645
14. Kirschen DS, Strbac G (2004) Fundamentals of power system economics. John Wiley & Sons, Chichester
15. Lavaei J, Low SH (2012) Zero duality gap in optimal power flow problem. *IEEE Trans Power Syst* 27(1):92–107
16. Lavaei J, Sojoudi S (2012) Competitive equilibria in electricity markets with nonlinearities. In: American control conference, pp 3081–3088
17. Mangasarian OL (1969) Nonlinear programming. Classics in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, MA
18. Mas-Colell A, Whinston MD, Green JR (1991) Microeconomic theory. Oxford University Press, Oxford
19. Motto, AL, Galiana FD, Conejo AJ, Arroyo JM (2002) Network constrained multiperiod auction for a pool-based electricity market. *IEEE Trans Power Syst* 17(3):646–653
20. Motto AL, Galiana FD, Conejo AJ, Huneault M (2002) On Walrasian equilibrium for pool-based electricity markets. *IEEE Trans Power Syst* 17(3):774–781
21. Munson TS, Facchinei F, Ferris M, Fischer A, Kanzow C (2001) The Semismooth algorithm for large scale complementarity problems. *J Comput* 13:294–311
22. Nocedal J, Wright SJ (1999) Numerical optimization. Springer series in operations research and financial engineering. Springer, New York
23. Qi L, Sun J (1993) A nonsmooth version of Newton's method. *Math Program* 58:353–368

24. Raghunathan AU, Curtis FE, Takaguchi Y, Hashimoto H (2016) Accelerating convergence to competitive equilibrium in electricity markets. In: IEEE power & energy society general meeting PESGM2016–000221
25. Rajan D, Takriti S (2005) Minimum up/down polytopes of the unit commitment problem with start-up costs. Technical report IBM research report RC23628, IBM, Yorktown Heights, NY
26. Shor NZ (1985) Minimization methods for non-differentiable functions. Springer, Berlin
27. Su CL, Kirschen D (2009) Quantifying the effect of demand response on electricity markets. *IEEE Trans Power Syst* 24(3):1199–1207
28. Wang G, Negrete-Pincetic M, Kowli A, Shafieeoorfard E, Meyn S, Shanbhag UV (2012) Dynamic competitive equilibria in electricity markets. In: Chakrabortty A, Ilić MD (eds) Control and optimization methods for electric smart grids. Power electronics and power systems, vol 3. Springer, New York, pp 35–62
29. Weber JD, Overbye TJ (2002) An individual welfare maximization algorithm for electricity markets. *IEEE Trans Power Syst* 17(3):590–596
30. Wood AJ, Wollenberg BF (1996) Power generation operation and control, 2nd edn. Wiley-Interscience, Hoboken, NJ
31. Zhang B, Tse D (2013) Geometry of the injection region of power networks. *IEEE Trans Power Syst* 28(2):788–797
32. Zimmerman RD, Murillo-Sánchez CE, Thomas RJ (2011) MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Trans Power Syst* 26(1):12–19

Small Resource Integration Challenges for Large-Scale SCUC



Cuong Nguyen, Lei Wu, Muhammad Marwali, and Rana Mukerji

Abstract Recent regulatory initiatives, technological advancements, and public policies such as New York's Reforming the Energy Vision (REV) and California's Energy Storage Mandate have incentivized the development of smaller, cleaner, and more distributed energy resources. As part of the day-ahead market clearing process, the ISOs/RTOs today have to solve more computationally intensive mixed integer programming (MIP)-based security-constrained unit commitment (SCUC) models to accommodate these small-scale resources within a short time window. This chapter will discuss the MIP solution performance challenges in dealing with the increasing penetration of small resources in the ISO/RTO day-ahead market in terms of both practicality and theory aspects.

1 Introduction

It has been 20 years since the Federal Energy Regulatory Commission (FERC) Order No. 888 mandated the establishment of unbundled electricity markets to promote competition in the wholesale bulk power marketplace and to bring more efficient, lower-cost power to electricity consumers. The order led to the formation of six independent system operators/regional transmission organizations (ISOs/RTOs) across the USA – California ISO (CAISO), Southwest Power Pool (SPP), Mid-continent ISO (MISO), PJM Interconnection (PJM), New York ISO (NYISO), and ISO New England (ISO-NE). Although not subject to FERC jurisdiction, Electric Reliability Council of Texas (ERCOT) was formed in 1996, before any other ISO in the USA. Some key statistics of the US wholesale electricity markets from [7] are

C. Nguyen (✉) · M. Marwali · R. Mukerji

New York Independent System Operator, Rensselaer, NY, USA

e-mail: cnguyen@nyiso.com; mmarwali@nyiso.com; rmukerji@nyiso.com

L. Wu

Clarkson University, Potsdam, NY, USA

e-mail: lwu@clarkson.edu

Table 1 US wholesale electricity market

Market	CAISO	SPP	ERCOT	MISO	PJM	NYISO	ISO-NE
Peak load (MW)	47,000	45,000	69,000	12,700	165,000	34,000	28,000
Gen capacity (MW)	60,000	80,000	77,000	175,000	184,000	39,000	31,000
Generating units	760	600	550	1,400	1,300	400	350
Annual energy (TWh)	230	230	340	525	780	160	135
Transmission (Miles)	26,000	56,000	43,000	65,000	72,000	11,000	8,000
Market volume (\$B)	10	12	34	37	35–50	7.5	7

shown in Table 1. The numbers are given in the order of magnitude for the sake of comparison. Peak load represents the historic maximum value to date. Market volume shows the total revenue flowing through the ISO/RTO markets from consumers to suppliers for all market products. Finally transmission miles show the estimated magnitude of the bulk electrical network monitored and secured by the ISOs/RTOs.

While there are differences in standard market design, most ISOs/RTOs have the same core suite of market products – day-ahead and real-time energy, operating reserve, and regulation reserve. All utilize locational marginal price (LMP) of energy and some co-optimize energy and reserve procurement across those markets.

As the electric power system strives to reduce its environmental impact, foster cleaner, renewable resources, and promote energy efficiency, attention increasingly turns to the potential of distributed energy resources (DERs).

In the traditional model of the centralized power system, electricity is said to flow “downhill” from large power plants to a widespread set of residential, commercial, and industrial customers. The emergence and growth of distributed resources are changing the landscape of the electric power system. DERs include an array of power generation and storage resources that are typically located on or near an end user’s property and supply all or a portion of the end user’s electricity. Such resources may also deliver power to the grid. Distributed energy technologies include solar photovoltaic (PV), combined heat and power (CHP) systems, microgrids, wind turbines, microturbines, backup generators, and energy storage devices.

A growing number of customer-sited PV installations are connected to the grid and take advantage of available net metering opportunities. Net metering enables customers to provide power generated by their distributed energy system to their host utility in return for credits on their electric bill. In New York, it is available on a first-come, first-served basis to customers of New York State’s major electric utilities, subject to technology, system size, and aggregate capacity limitations.

Examining grid-connected distributed resources, the Electric Power Research Institute (EPRI) clarified an important distinction between DERs that are connected to the grid and resources that are truly integrated into grid operations. The study stated, “...rapidly expanding deployments of DERs are connected to the grid but not integrated into grid operations, which is a pattern that is unlikely to be sustainable,” according to [6].

A 2014 report conducted for the NYISO by [5] assessed the state of distributed technologies and their prospects for growth in New York State. According to that report, New York's DER base was led by small-scale CHP with 57 percent of the state's distributed generation capacity. In other states, solar PV is the dominant DER technology. Solar PV ranked second in New York at 41 percent. Energy storage accounted for the remaining two percent.

The New York State Public Service Commission's "Reforming the Energy Vision" (REV) proceeding was first initiated in 2014 [3]; see Figure 1. The REV initiative is expected to lead to a wider deployment of small "distributed" energy resources, such as microgrids, rooftop solar and other on-site power supplies, and storage. A forecast of the potential REV impacts to the NYISO by 2025 is summarized in Table 2.

From the electricity marketplace standpoint, DERs offer the potential to make loads more dynamic and responsive to wholesale market price signals, potentially improving overall system efficiencies. Given the growing number of DERs with smaller MW capacities and operation costs relative to traditional resources, it is technically and operationally challenging to integrate them into the electricity

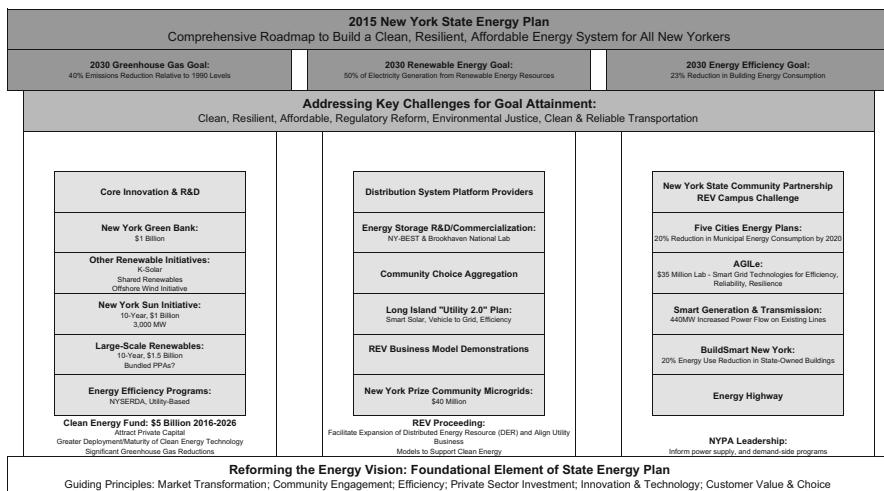


Fig. 1 New York State Policy influence on New York electricity market

Table 2 Potential REV impacts to NYISO by 2025

Market	Low	Expected	High
Solar (MW)	1,000	5,000	9,000
Wind (MW)	1,000	2,000	4,000
Efficiency (MW)	+1% Demand Growth	Flat Load Growth	-1% Demand Growth
Active DERs (MW)	1,000	2,000	4,000

markets. This stresses the need for developing more advanced metering, complex tools, and mathematical techniques. This chapter will address some mathematical challenges.

2 Large-Scale Security-Constrained Unit Commitment

The electricity market provides a mechanism for market participants to buy and sell energy at prices established through a competitive auction process designed to meet energy demands or “loads” and system reliability requirements with the least-cost resources available or, through contractual, bilateral transactions where quantities and prices are arranged directly between wholesale suppliers and “load-serving entities” (LSEs) such as utilities. For energy purchases arranged through the ISO/RTO markets, the ISO/RTO administers day-ahead and real-time auctions, resulting in a two-settlement process that sets the price of energy based on market and grid conditions at specific times. Further, the ISO/RTO auctions reflect geographic conditions, establishing LMP for energy that reflect local demand and supply conditions as well as any constraints that may bind when moving energy across the grid to meet demand. The first settlement is based upon the day-ahead bids and the corresponding schedule and prices or day-ahead commitment. The second settlement is based upon the real-time bids and the corresponding real-time commitment and dispatch. Market participants may participate in the day-ahead market (DAM) and/or the real-time market. For the NYISO markets, roughly 94% of energy is scheduled in the day-ahead market, while the remaining 6% is accounted for in the real-time market. The DAM allows for more certainty in prices due to its reduced volatility, ensuring that the resources are online in time for when they are needed and provide financial entities with greater liquidity within the market. Figure 2 for instance shows the high-level SCUC market flow in the NYISO market.

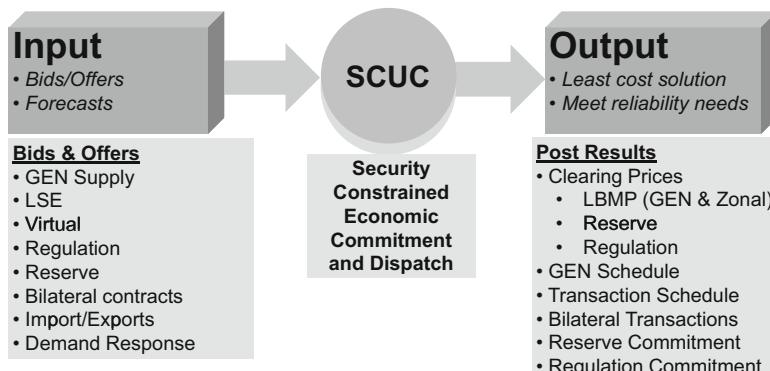


Fig. 2 NYISO’s security-constrained unit commitment process, where GEN is generation and LSE is load-serving entity

As a result of the day-ahead commitment process, a set of generators is scheduled to be available for dispatch in each hour of the next day, and a set of LSEs is scheduled to buy a certain amount of load at day-ahead prices. The generators designated to be available for the next day are scheduled against the LSE bid-in load and transmission losses. From the schedule, LMPs are computed, and forward contracts are established for generation and load accordingly. Subsequently, during real-time operations, changes in operating conditions, the influence of additional real-time supply offers, and variations in actual loads will cause the real-time schedules and prices to be different from the day-ahead schedules and prices. Differences in generation levels and in load consumption, as compared to the first settlement values, are settled at the second settlement or the real-time prices.

The discussion in this chapter is dedicated to the DAM where the vast majority of the energy is settled.

2.1 Traditional SCUC Formulation

Mathematically, SCUC is a nonconvex, nonlinear, large-scale, mixed integer programming problem with thousands to hundreds of thousand variables (binary decision variables, continuous and discrete control variables) and thousands to tens of thousands nonelectrical (such as emission allowance, wheeling contracts, water flow, fuel consumption) and electrical constraints (such as voltage, line flow, stability) as discussed in [8]. According to [10], the traditional MIP-based SCUC formulation with prevailing objective function terms and constraints can be presented as follows:

Objective function is to minimize total production cost, including startup cost SU_{it} (\$) and shutdown cost SD_{it} (\$), no-load cost N_i (\$/h), and incremental cost $c_{ik} \cdot P_{ikt}$ (\$/h) for unit i , segment k at hour t :

$$\text{Min}_{I_{it}, P_{it}} \sum_t \sum_i [C_{it} + SU_{it} + SD_{it}] \quad (1)$$

where C_{it} (\$/h) is the operation cost of unit i at hour t , I_{it} is the commitment decision of unit i at hour t , and P_{it} (MW) is the generation dispatch of unit i at hour t .

Subject to the following constraints:

Energy balance constraint

$$\sum_i P_{it} = D_t \quad (2)$$

where D_t (MW) is the system load at hour t .

Generation constraints

$$\begin{aligned}
 C_{it} &= N_i \cdot I_{it} + \sum_{k=1}^K c_{ik} \cdot P_{ikt} \\
 P_{it} &= P_i^{min} \cdot I_{it} + \sum_{k=1}^K P_{ikt} \\
 0 \leq P_{ikt} &\leq P_{ik}^{max} \\
 P_{it} &\leq P_i^{max} \cdot I_{it} \\
 I_{it} &\in \{0, 1\}
 \end{aligned} \tag{3}$$

where k is the index of generation segment, K is the number of piecewise linear generation segments, c_{ik} (\$/MWh) is the incremental cost of unit i at segment k , P_{ikt} (MW) is the dispatch of unit i at hour t at segment k , P_{ik}^{max} (MW) is the maximum capacity for segment k of unit i , and P_i^{min} (MW)/ P_i^{max} (MW) is the minimum/maximum capacity of unit i .

Generation ramping up/down constraints

$$P_{it} - P_{i(t-1)} \leq UR_i \cdot I_{i(t-1)} + UP_i \cdot I_{i(t-1)} + P_i^{max} \cdot (1 - I_{it}) \tag{4}$$

$$P_{i(t-1)} - P_{it} \leq DR_i \cdot I_{it} + DP_i \cdot (I_{i(t-1)} - I_{it}) + P_i^{max} \cdot (1 - I_{i(t-1)}) \tag{5}$$

where UR_i (MW/h) / DR_i (MW/h) is the ramping up/down limit of unit i and UP_i (MW/h)/ DP_i (MW/h) is the startup/shutdown ramp limit of unit i .

Transmission constraint

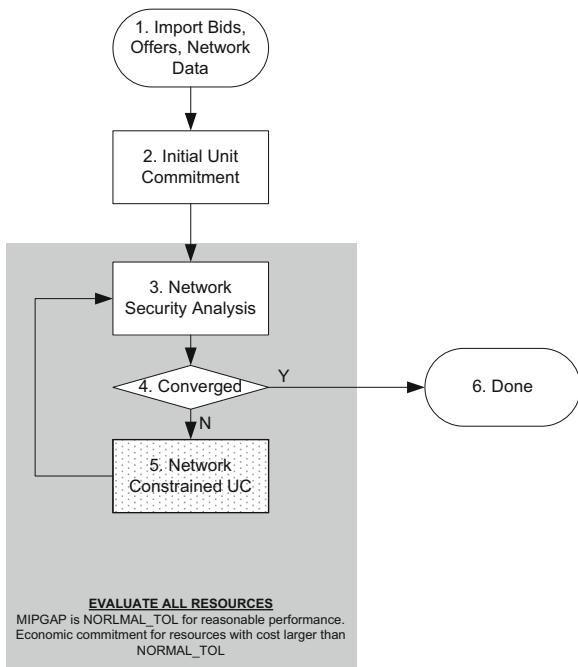
$$-PL_l^{max} \leq \sum_m LSF_{lt}^m \cdot \left(\sum_{i \in U(m)} P_{it} - D_{mt} \right) \leq PL_l^{max} \tag{6}$$

where PL_l^{max} (MW) is the maximum capacity of transmission line l , LSF_{lt}^m is the linear sensitivity factor of DC-based power flow on line l to the power injection at bus m at hour t , $U(m)$ is the set of generators located at bus m , and D_{mt} is the load at bus m at hour t .

Other constraints including minimum on/off time, spinning/non-spinning reserves, and regulation can be found in [11].

Due to the scale in size and the complexity in scope, the practical solution to the original SCUC problem today in some ISOs/RTOs is usually carried out by an iterative process. At every desired number of iterations, system nonlinearities such as transmission line losses and bus voltages can be linearized. Steps to solve the traditional SCUC problem are illustrated in Figure 3 and outlined below:

Fig. 3 Traditional approach for solving unit commitment with small resources



- Step 1** Import hourly load bids, unit offers, and network data. Set MIP solution tolerance to a desired NORMAL_TOL. As a practical matter, the tolerance is typically set at the lowest cost of committing a unit and running that unit at its minimum generation level. For example, NORMAL_TOL can be set to the cheapest unit cost of \$10,000 including \$6,000 of start-up and \$4,000 for running the unit at minimum output level in 1 hour.
- Step 2** Starting with the initial unit commitment (IUC) where all network constraints are relaxed, a MIP-based commitment and dispatch solution can be obtained. This initial solution is called free-flow solution.
- Step 3** Network security analysis (NSA) performs security analysis using the latest unit schedule determined from **Step 2** for initial iteration or from **Step 5** for subsequent iterations.
- Step 4** The convergence flag is set based on violation check (constraint flow is over the limit, voltage exceeds operating range, or any other special operating nomogram violations). If any violation is found, then calculate newly violated constraints, and go to **Step 5**; else go to **Step 6**.
- Step 5** Network-constrained UC (NCUC) solves MIP optimization problem as formulated in Equations 1–6 with the MIP gap set at NORMAL_TOL, respecting existing constraints plus the newly generated constraints from **Step 3**. Calculate unit schedule, then go to **Step 3**.
- Step 6** Done and final commitment and dispatch schedule is posted.

2.2 Two-Stage Approach for Solving Small Resources

As discussed earlier in Section 1, in the traditional model of the centralized power system, electricity is said to flow downhill from large power plants through an interconnected high-voltage transmission network (typically 110 kV and above) to a widespread set of residential, commercial, and industrial customers via the distribution network. The transmission operator manages transmission network, while the distribution company manages lower voltage lines from the point of transmission interconnection to the end-use customer.

Since customer electricity usage patterns are considered fairly static from day to day, ISO/RTO energy management system (EMS) and market optimization software typically model the system from the generation point up to the transmission-distribution system boundary. The increased development of smaller resources such as DERs will pose more challenges to today's market software.

With size ranging from a fractional MW to a few MWs, small resources typically incur a fairly small operation cost compared to traditional resources of tens to hundreds of MWs. Meanwhile a practical SCUC process uses branch-and-bound technique with a reasonable "MIP gap" to achieve a solution within the limited time window. Since initial unit commitment may not be physically feasible, SCUC must iterate to achieve a least-cost unit commitment while respecting all system security constraints. As the SCUC solution is achieved at a certain MIP gap tolerance, the impact of small resources on the solution may be within this tolerance, hence randomly committed. In other words, a branch-and-bound MIP solution may not accurately reflect whether such resources' commitment will enhance efficiency. As discussed in [2], even a MIP solution with a small optimality gap may result in market participant dispute. Commitment of small resources therefore will require more computer processing power, better optimization techniques, and/or a more active role from distribution system operators. This chapter will focus on new optimization techniques.

The two-stage approach flowchart is illustrated in Figure 4. Stage 1 (**Step 1 to Step 5**) and stage 2 (**Step 6 to Step 9**) are iterative and outlined as follows:

- Step 1** Import hourly load bids, unit offers, and network data. Set MIP solution tolerance of stage 1 to desired NORMAL_TOL, stage 2 to desired SMALL_TOL.
- Step 2** Starting with the initial unit commitment (IUC) where all network constraints are relaxed, an MIP-based commitment and dispatch solution can be obtained. This initial solution is called free-flow solution.
- Step 3** Network security analysis (NSA) performs security analysis using the latest unit schedule determined from **Step 2** for initial iteration or from **Step 5** for subsequent iterations.
- Step 4** The convergence flag is set based on violation check (constraint flow is over the limit, voltage exceeds operating range, or any other special

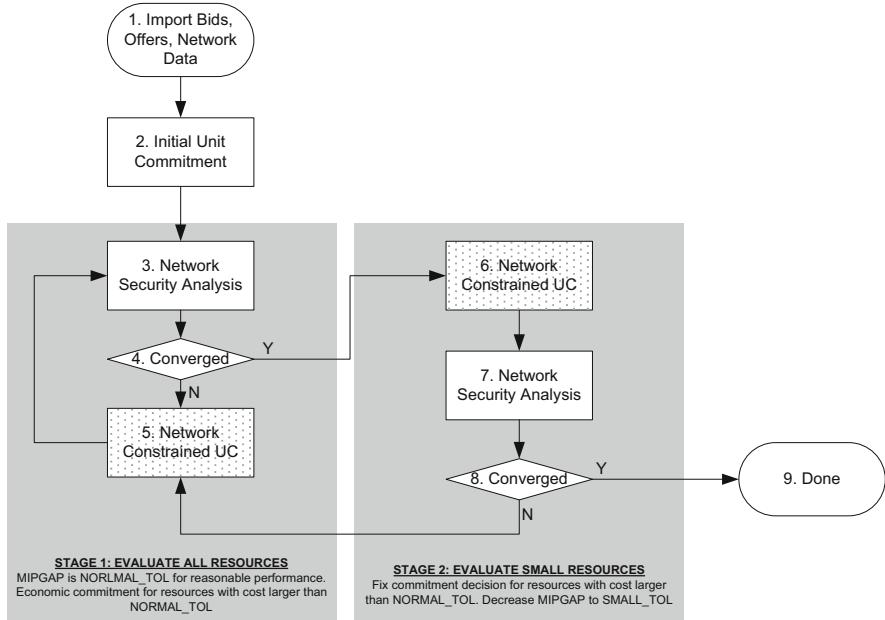


Fig. 4 Two-stage approach for solving unit commitment with small resources

operating nomogram violations). If any violation is found, then calculate newly violated constraints, and go to **Step 5**; else go to **Step 6**.

Step 5 Network-constrained UC (NCUC) solves MIP optimization problem as formulated in Equations 1–6 with the MIP gap set at NORMAL_TOL, respecting existing constraints plus the newly generated constraints from **Step 3**. Calculate unit schedule then go to **Step 3**.

Step 6 Network-constrained UC (NCUC) solves MIP optimization problem as formulated in Equations 1–6 with the MIP gap decreased to SMALL_TOL, respecting existing constraints plus the newly generated constraints from **Step 3**. Calculate unit schedule, and then go to **Step 7**.

Step 7 Network security analysis (NSA) performs security analysis using latest unit schedule determined from **Step 6**.

Step 8 The convergence flag is set based on violation check (constraint flow is over the limit, voltage exceeds operating range, or any other special operating nomogram violations). If any violation is found, then calculate newly violated constraints, and go to **Step 5**; else go to **Step 9**.

Step 9 Done and final commitment and dispatch schedule is posted.

2.3 Locally Ideal Formulation for Solving Small Resources

The traditional SCUC formulation in subsection 2.1 is widely used by ISOs/RTOs. The solution is usually attained by a generic branch-and-bound MIP solver such as GUROBI, see [9], and CPLEX, see [4]. Since these solvers are plug-and-play and built for general purposes, they usually do not recognize the unique characteristics of the SCUC problem thus unavoidably lead to longer solution time.

Recent advances in SCUC algorithms have proved to transform the traditional SCUC formulation to a tighter MIP model. Discussion in this subsection is focused on recent optimization theory advances that reformulate the generation constraints in Equation 3.

The traditional generation energy cost in the ISOs/RTOs today is modeled as monotonically nondecreasing piecewise linear function which results in a large branch-and-bound MIP solution search space. Moreover, the increasing trend in the number of small resources such as DERs, virtual resources, and transactions will further put pressure on the ISOs/RTOs to get the solution posted in the limited time windows. A recent SCUC market day with a large number of sub-MW resource bids in [2] highlighted the challenge that the MISO was facing in reaching an MIP solution at a reasonable MIP gap tolerance before its post time.

An alternative reformulation of Equation 3 is shown in Equation 7; see [10]. The reformulation recognizes the fact that the only binary variable I_{it} in the SCUC problem Equations 1–6 is closely related to the dispatch variable P_{it} , and the binary solution I_{it} is largely dependent on the operation cost C_{it} . Therefore, a tighter reformulation can be achieved to dramatically reduce computational burden. A special ordered set of type 2 (SOS2), discussed in detail in [1], can be used to transform a nonlinear function into a piecewise linear approximation function in a linear program. In Equation 7, δ_{ikt} is an ordered set of type 2 of nonnegative variables of which at most two can be nonzero, and if two are nonzero, these must be consecutive in the set.

$$\begin{aligned}
 C_{it} &= \sum_{k=1}^{K+1} F_{ik} \cdot \delta_{ikt} \\
 P_{it} &= \sum_{k=1}^{K+1} P_{ik} \cdot \delta_{ikt} \\
 \sum_{k=1}^{K+1} \delta_{ikt} &= I_{it} \\
 0 \leq \delta_{ikt} & \\
 0 \leq I_{it} \leq 1 & \\
 \delta_{ikt} \text{ are SOS2} &
 \end{aligned} \tag{7}$$

where F_{ik} (\$/h) and P_{ik} (MW) are the operation cost and generation dispatch of unit i at segment k 's starting point and δ_{ikt} is a nonnegative continuous variable to indicate the dispatch portion of unit i at hour t at segment k .

3 Case Study

A three-bus system, as in Figure 5, is studied to demonstrate the effectiveness of the new solution approaches. The system includes eight units, one load, and three transmission lines. Generator data are given in Table 3 where operation costs are approximated via two-segment piecewise linear curves. For example, the incremental cost of G1 is 10 \$/MWh in the dispatch range of [150 MW, 225 MW] and is 12 \$/MWh in the dispatch range of [225 MW, 300 MW]. Transmission line capacity is shown in Figure 5. All transmission lines have the same reactance of 0.1 p.u. In addition, generation unit G1–G5 are considered traditional large units, while G6–G8 are DERs with small capacity and cost.

A 1-hour SCUC problem is studied to determine the least-cost commitment and dispatch solution that satisfies the system load of 450 MW at bus 3. Other constraints, such as reserve and regulation requirements, generator minimum on and off time constraints, and ramping up and down limits, are not considered.

The following three cases are studied. CPLEX solver, see [4], is used as the solution engine.

Case 1 The original SCUC model in subsection 2.1

Case 2 Two-stage SCUC model in subsection 2.2

Case 3 SCUC model with the locally ideal formulation in subsection 2.3

It is important to note that all cases describe the same SCUC optimization problem and the major difference among them is how the problem constraints are formulated and how the corresponding MIP model is solved. Thus, all cases studies

Fig. 5 Three-bus system

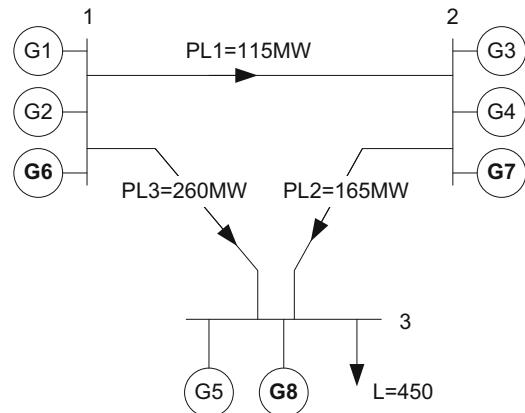
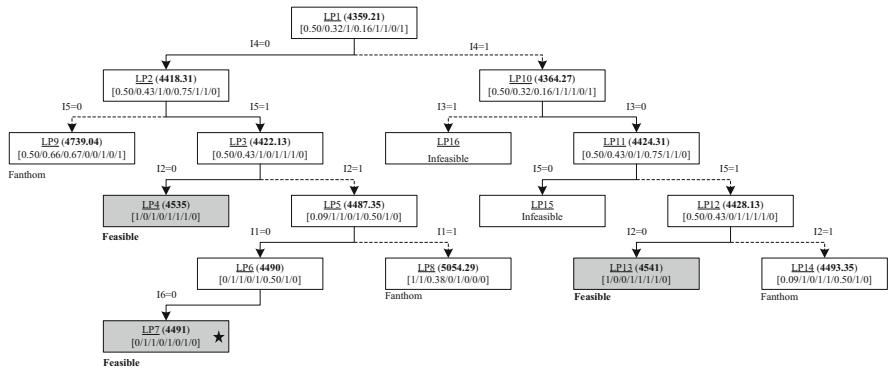


Table 3 Generator parameters of the three-bus system

Unit	Startup Cost (\$)	Mingen Cost (\$/h)	Pmin (MW)	Pmax (MW)	Incremental Cost (\$/MWh)	
G1	500	1270	150	300	10	12
G2	700	1660	205	305	10	11
G3	200	310	105	135	12	21
G4	200	316	105	135	12	19
G5	300	430	55	95	13	23
G6	3	8	1	3	10	11
G7	2	3	1	2	12	20
G8	6	8	1	2	13	20

**Fig. 6** Branch-and-bound solution for the original SCUC MIP

if solved at \$0 MIP gap tolerance would attain the same optimal solutions for unit commitment, generation dispatch, and operation costs, while their computational performance could be different.

The branch-and-bound (BAB) method is used to illustrate the solution procedure of the three cases, as shown in Figures 6, 7, and 8, respectively. While there are many ways to implement the BAB, in this chapter the depth-first search strategy is chosen for the node selection operation, and the least fractional strategy is adopted to choose binary variables for the branching operation. That is, at each node, the binary variable that satisfies $\arg\min_i \min_{I_i} [I_i, 1 - I_i]$ will be branched first. Priorities of the two branches $I_i \leq 0$ and $I_i \geq 1$ are determined by whether the current non-integer solution is closer to 0 or 1. In Figures 6, 7, and 8, solid and dash lines represent the first and the second branch directions of a binary variable. Values in the square brackets $[I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8]$ represent optimal solutions of $G_1, G_2, G_3, G_4, G_5, G_6, G_7$, and G_8 , respectively, at each node. Gray highlights indicate integer feasible solutions.

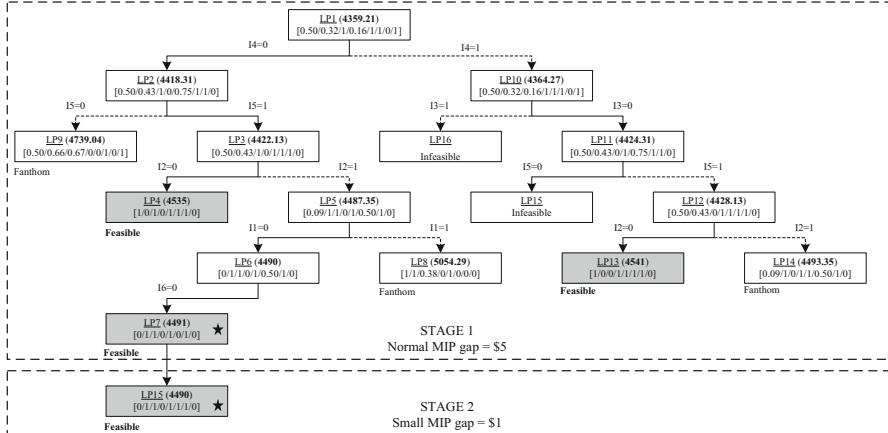


Fig. 7 Branch-and-bound solution for the two-stage SCUC MIP

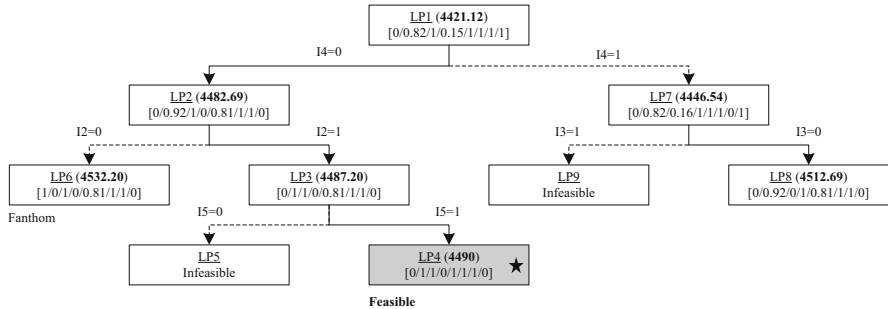


Fig. 8 Branch-and-bound solution for the original SCUC MIP

3.1 Case 1: The Original SCUC Model

The branch-and-bound solution for the original SCUC MIP problem is shown in Figure 6. While setting the solution tolerance MIP gap to \$0 is desirable, it is usually impossible to achieve a solution to the real-world MIP problem given a tight time window such as SCUC in the electricity markets. Table 4 shows the time between when US day-ahead electricity markets close until the time the solution is posted and the solution time is calculated as the difference between these two time stamps. As an example, the solution time window between market closing and market posting in PJM is only 3 hours. For practical matters, the absolute MIP gap can be set to the smallest sum of [*StartupCost*(\$)+*MingenCost*(\$/h)] among all units, which is \$5 for generator G7 in this example.

Figure 6 indicates that the solver traverses 16 relaxed linear programming (LP) nodes to obtain the optimal solution satisfying the \$5 MIP gap. The solution flowchart is illustrated in Figure 3 and detailed in subsection 2.1. The SCUC solution

Table 4 US wholesale electricity market time line

Market	CAISO	SPP	ERCOT	MISO	PJM	NYISO	ISO-NE
Name	IFM*	DAM**	DAM	DAM	DAM	DAM	DAM
DAM closing	10:00 PT	11:00 CT	10:00 CT	10:30 EPT [^]	10:30 ET	05:00 ET	10:00 ET
DAM posting	13:00 PT	16:00 CT	13:30 CT	13:30 EPT [^]	13:30 ET	11:00 ET	13:30 ET

* Integrated forward market (IFM)

** Day-ahead market (DAM)

[^] Eastern “Prevailing” Time (EPT) based on MISO terminology

process starts with the root node LP1 where all integrality constraints are relaxed. $[0.50/0.32/1/0.16/1/1/0/1]$ is the optimal solution. The optimal objective cost of the root \$4359.21 is a lower bound of the integer program. Since the LP1 solution is not integral, the BAB solution method is applied. Since $I_4 = 0.16$ is the least fractional value, it is branched into two sub-nodes LP2 and LP10. The non-integer solution is closer to its lower bound 0, so the branching direction is first performed as shown in the solid line $I_4 = 0$. That is, LP2 is solved after LP1. The BAB solution process subsequently continues to sub-node LP4 where a first integer solution $[1/0/1/0/1/1/1/0]$ is found. The LP4 solution is commonly termed the incumbent solution and serves as an upper bound of the integer program. The incumbent solution is the new upper bound to the integer program.

The BAB solution process repeatedly traverses through 16 nodes by applying variable branching and direction priority strategies. The incumbent solution is updated at the LP7 node. Nodes LP14 and LP15 are fathomed because of infeasibility (a solution does not exist given integrality constraints). Nodes LP4, LP7, and LP13 are not further explored because integer feasible solutions have been obtained. Nodes LP8, LP9, and LP14 are pruned because their optimal objective values are worse than the best lower bound of \$4490 obtained at node LP7.

In this case, solution does not explore the $I_6 = 1$ branch from LP6 sub-node since the MIP gap after solving LP7 is $\$4491(LP7) - \$4490(LP6) = \$1 < \5 tolerance. The final optimal solution at \$5 MIP gap is found with the objective value of \$4491 and $P_1 = 0$ MW, $P_2 = 283$ MW, $P_3 = 105$ MW, $P_4 = 0$ MW, $P_5 = 61$ MW, $P_6 = 0$ MW, $P_7 = 1$ MW, and $P_8 = 0$ MW.

3.2 Case 2: Two-Stage SCUC Model

The solution flowchart is illustrated in Figure 4 and detailed in subsection 2.2. Stage 1 is performed exactly to what is described in subsection 3.1. Once the stage 1 solution of $[0/1/1/0/1/0/1/0]$ is obtained at LP7, commitment status of committed units G2, G3, G5, and G7 are frozen. In addition, commitment status of large and uncommitted units G1 and G4 are also frozen. Therefore the units remaining for stage 2 evaluation are G6 and G8.

The stage 2 solution is carried out with a smaller MIP gap, for instance \$1. For this small test case, the integer solution is obtained right at the root node of stage 2. However this may not be generally true in the real-world cases.

Compared to the stage 1 solution, the only observed change in the commitment is unit G6. The final optimal solution at a \$1 MIP gap is found with the objective of \$4490 and $P_1 = 0$ MW, $P_2 = 280$ MW, $P_3 = 105$ MW, $P_4 = 0$ MW, $P_5 = 61$ MW, $P_6 = 3$ MW, $P_7 = 1$ MW, and $P_8 = 0$ MW.

It is worthwhile mentioning that the small resource G6 is now committed and dispatched at full output, displacing 3 MW from the large resource G2. This results in \$1 decrease in total objective function cost to \$4490. The total production cost improvement can be much more pronounced for the real-world test cases. It is observed that while the two-stage solution approach may still not guarantee a global optimal solution, it can be effective in fine-tuning the original MIP solution. The two-stage solution approach can improve objective function cost and give a fairer treatment of small resource commitment in the large-scale security-constrained unit commitment problem.

3.3 Case 3: SCUC Model with the Locally Ideal Formulation

As discussed in subsection 2.3, locally ideal formulation starts with transforming the original energy cost curve from Equation 3 to Equation 7.

Taking generator G3 in Table 3 as an example. The no-load cost is $200 + 310 = 510$ \$/h. The cost curve is made up of two incremental segments of 12 \$/MWh for the dispatch range of [105, 120] MW and 21 \$/MWh for the dispatch range of [120, 135] MW. The minimum and maximum capacity of G3 are 105 MW and 135 MW, respectively.

In the traditional formulation, G3's energy cost can be modeled by a two-segment incremental cost curve as shown in Equation 8. The monotonically increasing generation costs will guarantee convexity of the linearized SCUC formulation; thus P_{i1t} will be dispatched before P_{i2t} .

The locally ideal formulation for G3 is shown in Equation 9. There are three break points in G3's energy cost curve. Firstly, at $P_{i1} = 105$ MW operation cost is $F_{i1} = 510$ \$/h. Secondly, at $P_{i2} = 120$ MW operation cost is $F_{i2} = 510 + 12 \cdot 15 = 690$ \$/h. Thirdly, at $P_{i3} = 135$ MW operation cost is $F_{i3} = 690 + 21 \cdot 15 = 1005$ \$/h.

$$\begin{aligned}
 C_{it} &= 510 \cdot I_{it} + 12 \cdot P_{i1t} + 21 \cdot P_{i2t} \\
 P_{it} &= 105 \cdot I_{it} + P_{i1t} + P_{i2t} \\
 0 \leq P_{i1t} &\leq 15, 0 \leq P_{i2t} \leq 15 \\
 P_{it} &\leq 135 \cdot I_{it} \\
 I_{it} &\in \{0, 1\}
 \end{aligned} \tag{8}$$

Table 5 Comparison of the three test cases

Case	Case 1	Case 2	Case 3
Number of binary variables	8	8	8
Number of continuous variables	33	33	52
Number of equality constraints	18	18	26
Number of inequality constraints	14	14	6
Number of explored nodes	14	15	9
Absolute MIP gap at optimal solution (\$)	1	0	0
First feasible node	4	4	4

$$\begin{aligned}
 C_{it} &= 510 \cdot \delta_{i1t} + 690 \cdot \delta_{i2t} + 1005 \cdot \delta_{i3t} \\
 P_{it} &= 105 \cdot \delta_{i1t} + 120 \cdot \delta_{i2t} + 135 \cdot \delta_{i3t} \\
 \delta_{i1t} + \delta_{i2t} + \delta_{i3t} &= I_{it} \\
 0 \leq \delta_{i1t}, \delta_{i2t}, \delta_{i3t} &\leq 1 \\
 \{ \delta_{i1t}, \delta_{i2t}, \delta_{i3t} \} &\text{ are SOS2}
 \end{aligned} \tag{9}$$

Figure 8 shows the BAB procedures for this case, which solves nine relaxed LP nodes to obtain the same optimal solution as shown in subsection 3.2. In addition, the root node objective value is \$ 4421.12, which is higher than \$ 4359.21 of **Case 1** and **Case 2**. That means, the locally ideal reformulation in Equation 7 can tighten the SCUC formulation and, thus, achieve better computational performance.

For comparison, Table 5 summarizes formulation characteristics and computational performances of **Case 1–Case 3**.

4 Conclusion

The emergence and growth of distributed resources have put more pressure on the day-ahead market clearing process to obtain a unit commitment solution and dispatch schedule that is more efficient and fair to all participating resources within a tight time frame.

This chapter presents two formulation and solution strategies for enhancing the computational performance of SCUC problems, namely, the two-stage approach and the locally ideal formulation. Case studies on a three-bus system show that the two-stage solution approach can efficiently fine-tune the original solution (from stage 1 with normal MIP gap) to a more optimal solution (stage 2 with smaller MIP gap). By tightening the search space, the locally ideal formulation can reach the solution faster than the traditional formulation.

Recognizing the special characteristics of the SCUC problem and limitations of standard off-the-shelf solvers such as GUROBI and CPLEX, researchers are actively looking at ways to reformulate a tighter and more compact SCUC problem. The efficient and robust SCUC formulation is the fundamental technical building block for integrating smaller, cleaner, and more distributed energy resources into the ISO/RTO markets.

References

1. Beale E, Forrest JJ (1976) Global optimization using special ordered sets. *Math Program* 10(1):52–69
2. Chen Y, Casto A, Wang F, Wang Q, Wang X, Wan J (2016) Improving large scale day-ahead security constrained unit commitment performance. *IEEE Trans Power Syst* 31(6):4732–4743. <https://doi.org/10.1109/TPWRS.2016.2530811>
3. Commission NYSPS (2014) Reforming the energy vision (REV) proceeding. <http://documents.dps.ny.gov/public/Common/ViewDoc.aspx?DocRefId=%7B5A9BDBBD-1EB7-43BE-B751-0C1DAB53F2AA%7D>
4. Cplex II (2016) 12.6 User's manual. https://www.ibm.com/support/knowledgecenter/en/SSSA5P_12.6.2/ilog.odms.studio.help/pdf/usrcplex.pdf
5. DNV-GL (2014) A review of distributed energy resources, DNV-GL for the New York independent system operator
6. EPRI (2014) The integrated grid: realizing the full value of central and distributed energy resources, Palo Alto, CA
7. EPRI (2016) Wholesale electricity market design initiatives in the United States: survey and research needs
8. Fu Y, Li Z, Wu L (2013) Modeling and solution of the large-scale security-constrained unit commitment. *IEEE Trans Power Syst* 28(4):3524–3533
9. Optimization G (2016) Gurobi reference manual. <http://www.gurobi.com/documentation/7.0/refman/index.html>
10. Wu L (2016) Accelerating NCUC via binary variable-based locally ideal formulation and dynamic global cuts. *IEEE Trans Power Syst* 31(5):4097–4107. <https://doi.org/10.1109/TPWRS.2015.2502594>
11. Wu L, Shahidehpour M (2010) Accelerating the benders decomposition for network-constrained unit commitment problems. *Energy Syst* 1(3):339–376

Multi-Grid Schemes for Multi-Scale Coordination of Energy Systems



Sungho Shin and Victor M. Zavala

Abstract We discuss how multi-grid computing schemes can be used to design hierarchical coordination architectures for energy systems. These hierarchical architectures can be used to manage multiple temporal and spatial scales and mitigate fundamental limitations of centralized and decentralized architectures. We present the basic elements of a multi-grid scheme, which includes a smoothing operator (a high-resolution decentralized coordination layer that targets phenomena at high frequencies) and a coarsening operator (a low-resolution centralized coordination layer that targets phenomena at low frequencies). For smoothing, we extend existing convergence results for Gauss-Seidel schemes by applying them to systems that cover unstructured domains. This allows us to target problems with multiple timescales and arbitrary networks. The proposed coordination schemes can be used to guide transactions in decentralized electricity markets. We present a storage control example and a power flow diffusion example to illustrate the developments.

1 Motivation and Setting

We consider the following optimization problem:

$$\min_z \quad \frac{1}{2} z^T Q z - c^T z \tag{1a}$$

$$\text{s.t.} \quad Az + Bd = 0, \quad (\nu) \tag{1b}$$

$$\Pi z = 0. \quad (\lambda) \tag{1c}$$

S. Shin · V. M. Zavala (✉)

Department of Chemical and Biological Engineering, University of Wisconsin-Madison,
1415 Engineering Dr, Madison, WI 53706, USA
e-mail: shin79@wisc.edu; zavalatejeda@wisc.edu

Here, $z \in \mathbb{R}^{N \cdot n_z}$ are decision or primal variables (including states and controls), and $d \in \mathbb{R}^{N \cdot n_d}$ is the data (including disturbances and system parameters). These variable vectors contain elements that are distributed over a mesh with $N \in \mathbb{Z}$ points that covers a certain temporal or spatiotemporal domain of interest Ω . We define the set of points in the mesh as \mathcal{N} with $|\mathcal{N}| = N$. The matrix $Q \in \mathbb{R}^{N \cdot n_z \times N \cdot n_z}$ is positive definite, and $c \in \mathbb{R}^{N \cdot n_z}$ is a cost vector. The constraint (1b) (with associated dual variables $v \in \mathbb{R}^m$) is defined by the matrices $A \in \mathbb{R}^{m \times N \cdot n_z}$ and $B \in \mathbb{R}^{m \times N \cdot n_d}$, and the matrix A is assumed to have full row rank. The constraints may include discretized dynamic equations (in space and time) and other physical constraints. The constraints (1c) (with associated dual variables $\lambda \in \mathbb{R}^p$) are defined by the matrix $\Pi \in \mathbb{R}^{p \times N \cdot n_z}$. This constraint models coupling (connectivity) between the primal variables at different mesh points and can also be used to model boundary conditions. Problem (1) captures formulations used in optimization-based control strategies such as model predictive control (MPC) [26].

We assume that the dimension of the mesh N describing problem (1) is so large that the problem cannot be solved in a *centralized* manner. This is often the case in systems that cover large temporal and spatial domains and/or multiple scales. In an electrical network, for instance, a large number of nodes and harmonics might need to be captured, rendering centralized control impractical. An alternative to address this complexity is to partition the problem into subdomains to create *decentralized* control architectures. We begin by defining a partitioned version of problem (1):

$$\min_{z_k} \quad \sum_{k \in \mathcal{K}} \frac{1}{2} z_k^T Q_k z_k - c_k^T z_k \quad (2a)$$

$$\text{s.t.} \quad A_k z_k + B_k d_k = 0, \quad k \in \mathcal{K} \quad (v_k) \quad (2b)$$

$$\sum_{k' \in \mathcal{K}} \Pi_{kk'} z_{k'} = 0, \quad k \in \mathcal{K} \quad (\lambda_k) \quad (2c)$$

We denote this problem as \mathcal{P} . Here, \mathcal{K} is a set for partitions (subdomains) of the set \mathcal{N} , and we define the number of partitions as $K := |\mathcal{K}|$. Each partition $k \in \mathcal{K}$ contains mesh elements $\mathcal{N}_k \subseteq \mathcal{N}$ satisfying $\cup_{k \in \mathcal{K}} \mathcal{N}_k = \mathcal{N}$ and $\mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset$ for all $k, k' \in \mathcal{K}$ and $k \neq k'$. The number of elements in a partition is denoted as $N_k := |\mathcal{N}_k|$. The variables and data (z_k, d_k) are defined over the partition $k \in \mathcal{K}$. We represent the cost function as a sum of the partition cost functions with associated positive definite matrices Q_k and cost vectors c_k . The constraints are also split into individual partition constraints with associated matrices A_k , B_k , and we link the partition variables by using the coupling constraints (2c) and associated matrices $\Pi_{kk'}$, $k, k' \in \mathcal{K}$. As we will discuss later, we can always obtain such a representation by introducing duplicate decision variables in each partition and by adding suitable coupling constraints. This procedure is known as *lifting* [1].

To *avoid centralized coordination*, a wide range of *decomposition* schemes (we also refer to them as *decentralized* coordination schemes) can be used. A popular approach used in the solution of partial differential equations (PDEs)

and decomposition methods such as the alternating direction method of multipliers (ADMM) is the *Gauss-Seidel* (GS) coordination scheme [7, 8]. Here, the problem in each partition k (often called a control *agent*) is solved independently from the rest and exchanges information with its neighbors to coordinate. For a lifted problem of the form (2), we will derive a decentralized GS scheme that solves problems over individual partitions $k \in \mathcal{K}$ of the form:

$$z_k^{\ell+1} = \underset{z_k}{\operatorname{argmin}} \quad \frac{1}{2} z_k^T Q_k z_k - z_k^T \left(c_k - \sum_{k'=1}^{k-1} \Pi_{k'k}^T \lambda_{k'}^{\ell+1} - \sum_{k'=k+1}^N \Pi_{k'k}^T \lambda_{k'}^\ell \right) \quad (3a)$$

$$\text{s.t. } A_k z_k + B_k d_k = 0 \quad (3b)$$

$$\Pi_{kk} z_k + \sum_{k'=1}^{k-1} \Pi_{kk'} z_{k'}^{\ell+1} + \sum_{k'=k+1}^K \Pi_{kk'} z_{k'}^\ell = 0 \quad (\lambda_k). \quad (3c)$$

We denote this partition subproblem as \mathcal{P}_k^ℓ that is solved at the update step $\ell \in \mathbb{Z}_+$ (that we call here the *coordination step*). From the solution of this problem, we obtain the updated primal variables $z_k^{\ell+1}$ and dual variables $\lambda_k^{\ell+1}$ (corresponding to the coupling constraints (3c)). Here, $z_{k'}^\ell$ and $\lambda_{k'}^\ell$ are primal and dual variables for neighboring partitions connected to partition k and that have not been updated, while $z_{k'}^{\ell+1}$ and $\lambda_{k'}^{\ell+1}$ are primal and dual variables for neighboring partitions that have already been updated. We refer to the variables communicated between partitions as the *coordination variables*. We note that partition k cannot update its primal and dual variables until the variables of a subset of the partitions connected to it have been updated. Consequently, the GS scheme is sequential and synchronous in nature. We also note that the connectivity topology (induced by the coupling matrices $\Pi_{kk'}$) determines the communication structure. We highlight, however, that the order of the updates presented in (3) is lexicographic (in the order of the partition number), but this choice of update order is arbitrary and can be modified. We will see that the update order can be designed to derive parallel schemes (i.e., in which certain partitions can proceed independently of others) but that the order can affect performance. In Figure 1 we illustrate the configuration of a GS scheme over a 1-D mesh, while in Figure 2 we present a configuration for a 2-D mesh. For the 2-D mesh, we note that the nodes spanning the domain are grouped into sets of the form $\mathcal{N}_{m,n}$, and we note that the information is exchanged using the state and dual variables in the boundary of the partitions. In Section 3 we discuss this approach in more detail.

In the next sections, we derive and analyze a decentralized GS scheme to solve problem \mathcal{P} . The analysis seeks to illustrate how the structure of the partition subproblem \mathcal{P}_k^ℓ arises and to highlight how information of the coordination variables propagates throughout the partitions. We then discuss how to create *coarse representations* of the full-resolution problem \mathcal{P} to obtain approximations for the

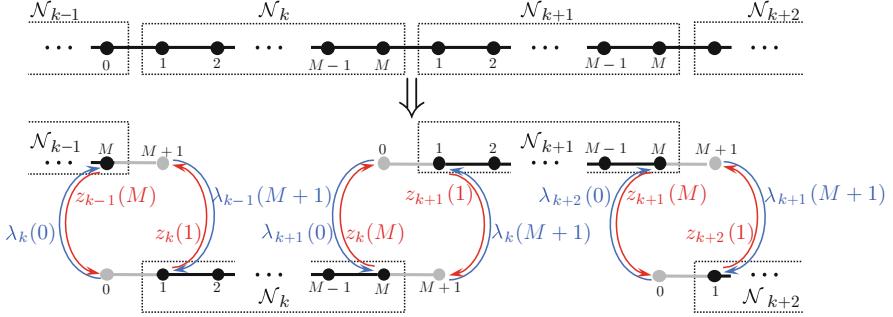


Fig. 1 Configuration of a GS scheme over a 1-D mesh.

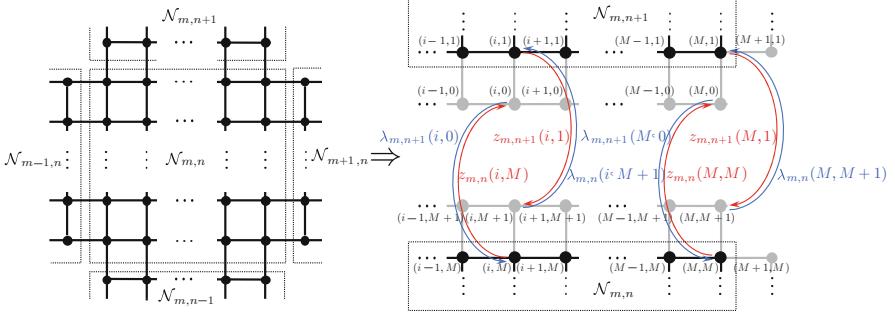


Fig. 2 Configuration of a GS scheme over a 2-D mesh.

coordination variables and with this accelerate the decentralized GS scheme. This gives rise to the concept of *multi-grid schemes* that can be used to design hierarchical coordination architectures. Our analysis is performed on convex quadratic programs (QPs), which will reveal important features of multi-grid schemes.

The concepts discussed in this paper seek to extend existing literature on decentralized and hierarchical MPC. Many strategies have been proposed to address the complexity of centralized MPC such as spatial and temporal decomposition methods [10, 15, 21, 28, 31], fast inexact schemes [12, 35, 36], and reduced-order modeling techniques [4]. Decentralized control manages the entire system by coordinating multiple controllers, each operating in a different node or subnetwork. Decentralization also enables *resiliency and asynchronicity*, which are key practical advantages over centralized MPC. Different schemes for coordinating MPC controllers have been devised [28]. Lagrangian dual decomposition is a technique where Lagrange multipliers are used for coordination. This technique is popular in electricity markets because the dual variables can be interpreted as prices that are used as a coordination mechanism [3, 18, 24]. Techniques based on coordinate minimization schemes and distributed gradient methods have also been proposed to coordinate MPC controllers in general settings [23, 32, 33]. An important limitation

of decentralized schemes is that coordination of subsystems tends to be slow (e.g., convergence rates of existing schemes are at best linear) [8, 14, 16]. This slow convergence has been reported in the context of energy networks in [3]. Moreover, spatial decentralization by itself does not address the complexity induced by multiple timescales. In particular, timescales and prediction horizons of different decentralized controllers might be different. To the best of our knowledge, no coordination schemes currently exist to handle such settings.

Hierarchical control seeks to overcome fundamental limitations of decentralized and centralized control schemes. Fundamental concepts of hierarchical control date as far back as the origins of automatic control itself [17]. Complex industrial control systems such as the power grid are structured hierarchically in one way or another to deal with multiple time and spatial scales. Existing hierarchies, however, are often constructed in ad hoc manners by using objectives, physical models, and control formulations at different levels that are often *incompatible*. For instance, an independent system operator (ISO) solves a hierarchy of optimization problems (unit commitment, economic dispatch, optimal power flow) that use different physical representations of the system. This can lead to lost economic performance, unreachable/infeasible command signals, and instabilities [5, 27]. Hierarchical MPC provides a general framework to tackle dynamics and disturbances occurring at multiple timescales [17, 28, 29, 34] and spatial scales [13]. In a traditional hierarchical MPC scheme, one uses a high-level controller to compute coarse control actions that are used as targets (commands) by low-level controllers. This approach has been used recently in microgrids and multi-energy systems [22, 37]. More sophisticated MPC controllers use robustness margins of the high-level controller that are used by the lower level controller to maintain stability [29]. Significant advances in the analysis of multi-scale dynamical systems have also been made, most notably by the use of singular perturbation theory to derive reduced-order representations of complex networks [11, 19, 25, 30]. The application of such concepts in hierarchical MPC, however, has been rather limited. In particular, the recent review on hierarchical MPC by Scattolini notices that systematic design methods for hierarchical MPC are still lacking [28]. More specifically, no hierarchical MPC schemes have been proposed that aggregate and *refine* trajectories at multiple scales. In addition, existing schemes have been tailored to achieve feasibility and stability but *do not have optimality guarantees*. This is important in systems where both economic efficiency and stability must be taken into account. To the best of our knowledge, no attempt has been made to combine hierarchical and decentralized MPC schemes to manage spatial and temporal scales simultaneously. The multi-grid computing concepts presented in this work seek to take a first step toward creating more general hierarchical control architectures. The proposed multi-grid schemes provide a framework to *coordinate decentralized electricity markets*. This is done by exchanging state and price (dual information) at the interfaces of the agents domain. The ability to do this hierarchically enables coordination over multiple spatial schemes, in particular, provides a framework to cover large geographical regions that might involve many market players.

2 Analysis of Gauss-Seidel Schemes

This section presents basic concepts and convergence results for a GS scheme under a general convex QP setting. The results seek to highlight how the structure of the coupling between partition variables as well as the coordination sequence affect the performance of GS schemes.

2.1 Illustrative Setting

To introduce notation, we begin by considering a convex QP with two variable partitions (i.e., $\mathcal{K} = \{1, 2\}$). To simplify the presentation, we do not include internal partition constraints and focus on coupling constraints across partitions. Under these assumptions, we have the following lifted optimization problem:

$$\min_{z_1, z_2} \frac{1}{2} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^T \begin{bmatrix} Q_1 & \\ & Q_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} - \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}^T \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad (4a)$$

$$\text{s.t. } \underbrace{\begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}}_{\Pi} \underbrace{\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}}_z = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{array}{l} (\lambda_1) \\ (\lambda_2) \end{array} \quad (4b)$$

Here, the matrices Π_{11} , Π_{12} , Π_{21} , and Π_{22} capture coupling between the variables z_1 and z_2 . We highlight that, in general, $\Pi_{12} \neq \Pi_{21}$ (the coupling between partitions is not symmetric). The matrices Q_1 and Q_2 are positive definite, and we assume that Π_{11} and Π_{22} have full row rank and that the entire coupling matrix Π has full row rank. By positive definiteness of Q_1 and Q_2 and the full rank assumption of Π , we have that the feasible set is non-empty, and the primal and dual solution of (4) is unique. The coupling between partition variables is two-directional (given by the structure of the coupling matrix Π). This structure is found in 1-D linear networks. When the coupling is only one-directional, as is the case of temporal coupling, we have that $\Pi_{12} = 0$, and thus Π is block lower triangular, indicating that the primal variables only propagate forward in time. We will see, however, that backward propagation of information also exists but in the space of the dual variables.

The first-order KKT conditions of (4) are given by the linear system:

$$\begin{bmatrix} Q_1 & \Pi_{11}^T & \Pi_{21}^T \\ \Pi_{11} & \Pi_{12} & \\ \Pi_{12}^T & Q_2 & \Pi_{22}^T \\ \Pi_{21} & \Pi_{22} & \end{bmatrix} \begin{bmatrix} z_1 \\ \lambda_1 \\ z_2 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \\ c_2 \\ 0 \end{bmatrix} \quad (5)$$

To avoid solving this system in a centralized manner, we use a decentralized GS scheme with *coordination update* index $\ell \in \mathbb{Z}_+$. The scheme has the form:

$$\begin{bmatrix} Q_1 & \Pi_{11}^T \\ \Pi_{11} & \end{bmatrix} \begin{bmatrix} z_1^{\ell+1} \\ \lambda_1^{\ell+1} \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \end{bmatrix} - \begin{bmatrix} & \Pi_{21}^T \\ \Pi_{12} & \end{bmatrix} \begin{bmatrix} z_2^\ell \\ \lambda_2^\ell \end{bmatrix} \quad (6a)$$

$$\begin{bmatrix} Q_2 & \Pi_{22}^T \\ \Pi_{22} & \end{bmatrix} \begin{bmatrix} z_2^{\ell+1} \\ \lambda_2^{\ell+1} \end{bmatrix} = \begin{bmatrix} c_2 \\ 0 \end{bmatrix} - \begin{bmatrix} & \Pi_{12}^T \\ \Pi_{21} & \end{bmatrix} \begin{bmatrix} z_1^{\ell+1} \\ \lambda_1^{\ell+1} \end{bmatrix} \quad (6b)$$

This scheme requires an initial guess for the variables of the second partition (z_2^0, λ_2^0) that is used to update (z^1, λ^1) , and we then proceed to update the variables of the second partition. Note, however, that we have picked this *coordination order* arbitrarily (lexicographic order). In particular, one can start with an initial guess of the first partition to update the second partition and then update the first partition (reverse lexicographic order). This scheme has the form:

$$\begin{bmatrix} Q_2 & \Pi_{22}^T \\ \Pi_{22} & \end{bmatrix} \begin{bmatrix} z_2^{\ell+1} \\ \lambda_2^{\ell+1} \end{bmatrix} = \begin{bmatrix} c_2 \\ 0 \end{bmatrix} - \begin{bmatrix} & \Pi_{12}^T \\ \Pi_{21} & \end{bmatrix} \begin{bmatrix} z_1^\ell \\ \lambda_1^\ell \end{bmatrix} \quad (7a)$$

$$\begin{bmatrix} Q_1 & \Pi_{11}^T \\ \Pi_{11} & \end{bmatrix} \begin{bmatrix} z_1^{\ell+1} \\ \lambda_1^{\ell+1} \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \end{bmatrix} - \begin{bmatrix} & \Pi_{21}^T \\ \Pi_{12} & \end{bmatrix} \begin{bmatrix} z_2^{\ell+1} \\ \lambda_2^{\ell+1} \end{bmatrix} \quad (7b)$$

We will see that the coordination order affects the convergence properties of the GS scheme. In problems with many partitions, we will see that a large number of coordination orders are possible. Moreover, we will see that coordination orders can be designed to enable sophisticated parallel and asynchronous implementations. A key observation is that the linear systems (6) in the GS scheme are the first-order KKT conditions of the following partition problems \mathcal{P}_1 and \mathcal{P}_2 , respectively:

$$z_1^{\ell+1} = \underset{z_1}{\operatorname{argmin}} \quad \frac{1}{2} z_1^T Q_1 z_1 - z_1^T (c_1 - \Pi_{21}^T \lambda_2^\ell) \quad (8a)$$

$$\text{s.t.} \quad \Pi_{11} z_1 + \Pi_{12} z_2^\ell = 0 \quad (\lambda_1) \quad (8b)$$

$$z_2^{\ell+1} = \underset{z_2}{\operatorname{argmin}} \quad \frac{1}{2} z_2^T Q_2 z_2 - z_2^T (c_2 - \Pi_{12}^T \lambda_1^{\ell+1}) \quad (8c)$$

$$\text{s.t.} \quad \Pi_{22} z_2 + \Pi_{21} z_1^{\ell+1} = 0 \quad (\lambda_2) \quad (8d)$$

The relevance of this observation is that one can implement the GS scheme by *directly solving optimization problems*, as opposed to performing intrusive linear algebra calculations [34]. This has practical benefits, as one can use algebraic modeling languages and handle sophisticated problem formulations. This also reveals that *both primal and dual variables* are communicated between partitions.

Primal information enters in the coupling constraints. The dual variables enter as cost terms in the objective and highlights the fact that dual variables can be interpreted as *prices of the primal variable information exchanged* between partitions.

We now seek to establish conditions guaranteeing convergence of this simplified GS scheme. To do so, we define the following matrices and vectors:

$$A_1 := \begin{bmatrix} Q_1 & \Pi_{11}^T \\ \Pi_{11} & \end{bmatrix}, \quad x_1^\ell := \begin{bmatrix} z_1^\ell \\ \lambda_1^\ell \end{bmatrix}, \quad b_1 := \begin{bmatrix} c_1 \\ 0 \end{bmatrix}, \quad B_{12} := \begin{bmatrix} & -\Pi_{21}^T \\ -\Pi_{12} & \end{bmatrix} \quad (9a)$$

$$A_2 := \begin{bmatrix} Q_2 & \Pi_{22}^T \\ \Pi_{22} & \end{bmatrix}, \quad x_2^\ell := \begin{bmatrix} z_2^\ell \\ \lambda_2^\ell \end{bmatrix}, \quad b_2 := \begin{bmatrix} c_2 \\ 0 \end{bmatrix}, \quad B_{21} := \begin{bmatrix} & -\Pi_{12}^T \\ -\Pi_{21} & \end{bmatrix}. \quad (9b)$$

The partition matrices A_1 and A_2 are nonsingular because the matrices Q_1 and Q_2 are positive definite, and Π_{11} and Π_{22} have full row rank. Nonsingularity of A_1 and A_2 implies that the partition optimization subproblems \mathcal{P}_1 and \mathcal{P}_2 have a unique solution for any values of the primal and dual variables of the neighboring partition. We can now express $(x_1^{\ell+1}, x_2^{\ell+1})$ in terms of (x_1^ℓ, x_2^ℓ) to obtain a recursion of the form:

$$\begin{bmatrix} A_1 & \\ -B_{21} & A_2 \end{bmatrix} \begin{bmatrix} x_1^{\ell+1} \\ x_2^{\ell+1} \end{bmatrix} = \begin{bmatrix} B_{12} \\ B_{21} \end{bmatrix} \begin{bmatrix} x_1^\ell \\ x_2^\ell \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \quad (10)$$

We can write this system in compact form by defining:

$$w^\ell := \begin{bmatrix} x_1^\ell \\ x_2^\ell \end{bmatrix}, \quad S := \begin{bmatrix} A_1 & \\ -B_{21} & A_2 \end{bmatrix}^{-1} \begin{bmatrix} B_{12} \\ B_{21} \end{bmatrix}, \quad r := \begin{bmatrix} A_1 & \\ -B_{21} & A_2 \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \quad (11)$$

The solution of the ℓ -th update step of (6) can be represented as:

$$w^{\ell+1} = Sw^\ell + r. \quad (12a)$$

$$= S^\ell w^0 + (I + S + \cdots + S^{\ell-1})r. \quad (12b)$$

The solution of the QP (4) (i.e., the solution of the KKT system (5)) solves the implicit system $w = Sw + r$, which can also be expressed as $(I - S)w = r$ or $w = (I - S)^{-1}r$. Consequently, we note that the eigenvalues of matrix S play a key role in the convergence of the GS scheme (8). We discuss this in more detail in the following section.

2.2 General Setting

We now extend the previous analysis to a more general QP setting with an arbitrary number of partitions. Here, we seek to illustrate how to perform *lifting* in a general case where coupling is implicit in the model and how to derive a GS scheme to solve such a problem. Our discussion is based on the following convex QP:

$$\min_z \quad \frac{1}{2} z^T Q z - c^T z. \quad (13)$$

Here $z = (z(1), z(2), \dots, z(N)) \in \mathbb{R}^N$ are the optimization variables, $Q \in \mathbb{R}^{N \times N}$ is a positive definite matrix, and $c \in \mathbb{R}^N$ is the cost vector. For simplicity (and without loss of generality), we assume that $n_z = 1$ (there is only one variable per node). We let Q_{ij} represent the (i, j) -th component of matrix Q , and we let $\mathcal{N} = \{1, 2, \dots, N\}$ be the variable indices (in this case also the node indices). Problem (13) has a unique solution because the matrix Q is positive definite.

We focus our analysis on the QP (13) because we note that equality constraints $\bar{A}z + \bar{B}d = 0$ in (1) (with $\bar{A}^T = [A^T \Pi^T]$ and $\bar{B}^T = [B^T 0]$) can be eliminated by using a null-space projection procedure. To see how this can be achieved we note that, if A has full row rank, we can always construct a matrix $Z \in \mathbb{R}^{N \times \tilde{N}}$ whose columns span the null-space of \bar{A} (i.e., $\bar{A}Z\tilde{z} = 0$ for any $\tilde{z} \in \mathbb{R}^{\tilde{N}}$) and with $\tilde{N} < N$. Similarly, we can construct a matrix $Y \in \mathbb{R}^{N \times (N-\tilde{N})}$ whose columns span the range space of \bar{A}^T (i.e., $Y\tilde{y} \in \text{Range}(\bar{A}^T)$ for any $\tilde{y} \in \mathbb{R}^{N-\tilde{N}}$). We can express any $z \in \mathbb{R}^N$ as $z = Z\tilde{z} + Y\tilde{y}$. We thus have that $\bar{A}z = \bar{A}Y\tilde{y}$ for all \tilde{z} and thus $\tilde{y} = -(\bar{A}Y)^{-1}\bar{B}d$ and $z = -Y(\bar{A}Y)^{-1}\bar{B}d + Z\tilde{z}$ satisfies $\bar{A}z = -\bar{B}d$ for any \tilde{z} . With this, we can express the quadratic objective as $\frac{1}{2}z^T Q z - c^T z$ as $\frac{1}{2}\tilde{z}^T Z^T Q Z \tilde{z} - c^T Z \tilde{z} - (Y(\bar{A}Y)^{-1}\bar{B}d)^T Q Z \tilde{z} + \kappa$, where κ is a constant. We thus obtain a QP of the same form as (13) but with matrix $Q \leftarrow Z^T Q Z$, reduced cost $c \leftarrow Z^T c + Z^T Q Y (\bar{A}Y)^{-1} \bar{B}d$, and variable vector $z \leftarrow \tilde{z}$. We highlight that this reduction procedure does not need to be applied in a practical implementation, but we only use it to justify that the formulation (13) is general.

We thus have that the variables z in (13) are coupled implicitly via the matrix Q , and we seek to express this problem in the lifted form. We proceed to partition the set \mathcal{N} into a set of partitions $\mathcal{K} = \{1, \dots, K\}$ to give the partition sets $\mathcal{N}_1, \mathcal{N}_2, \dots, \mathcal{N}_K \subseteq \mathcal{N}$ satisfying $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2 \cup \dots \cup \mathcal{N}_K$ and $\mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset$ for all $k, k' \in \mathcal{K}$ and $k \neq k'$. Coupling between variables arises when $Q_{ij} \neq 0$ for $i \in \mathcal{N}_k, j \in \mathcal{N}'_k$, and $k \neq k'$. We perform lifting by defining index sets:

$$\underline{\mathcal{N}}_k := \{j \in \mathcal{N} \setminus \mathcal{N}_k \mid \exists i \in \mathcal{N}_k \text{ s.t. } Q_{ij} \neq 0\}, \quad \overline{\mathcal{N}}_k := \mathcal{N}_k \cup \underline{\mathcal{N}}_k \quad (14)$$

The set $\underline{\mathcal{N}}_k$ includes all the coupled variables in the partition set \mathcal{N}_k that are not in partition k . The set $\overline{\mathcal{N}}_k$ includes all variables in partition \mathcal{N}_k and its coupled variables. These definitions allow us to express problem (13) as:

$$\min_z \quad \frac{1}{2} \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}_k} \sum_{j \in \overline{\mathcal{N}}_k} Q_{ij} z(i)z(j) - \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}_k} c_i z(i). \quad (15)$$

To induce lifting, we introduce a new set of variables $\{z_1, z_2, \dots, z_K\}$ defined as:

$$z_k := \begin{bmatrix} z(k_1) \\ \vdots \\ z(k_{N_k}) \end{bmatrix}, \quad \underline{z}_k = \begin{bmatrix} z(k_{N_k+1}) \\ \vdots \\ z(k_{N_k+\underline{N}_k}) \end{bmatrix}, \quad \bar{z}_k := \begin{bmatrix} z_k \\ \underline{z}_k \end{bmatrix} \quad (16)$$

where $\mathcal{N}_k = \{k_1, k_2, \dots, k_{N_k}\}$, $\underline{\mathcal{N}}_k = \{k_{N_k+1}, k_{N_k+2}, \dots, k_{N_k+\underline{N}_k}\}$, N_k is the number of variables in partition \mathcal{N}_k , and \underline{N}_k is the number of variables coupled to partition k . With this, we can express problem (15) in the following *lifted* form:

$$\min_z \quad \sum_{k \in \mathcal{K}} \frac{1}{2} \bar{z}_k^T \bar{Q}_k \bar{z}_k - \bar{c}_k^T \bar{z}_k. \quad (17a)$$

$$\text{s.t.} \quad \Pi_{kk} \bar{z}_k + \sum_{k' \in \mathcal{K} \setminus \{k\}} \Pi_{kk'} \bar{z}_{k'} = 0, \quad k \in \mathcal{K}. \quad (17b)$$

Here, $\bar{Q}_k \in \mathbb{R}^{(N_k+\underline{N}_k) \times (N_k+\underline{N}_k)}$ and $\bar{c}_k \in \mathbb{R}^{(N_k+\underline{N}_k)}$ are given by:

$$(\bar{Q}_k)_{ij} = \begin{cases} Q_{k_i k_j}, & \text{for } k_i, k_j \in \mathcal{N}_k \\ \frac{1}{2} Q_{k_i k_j}, & \text{for } k_i \in \mathcal{N}_k \text{ and } k_j \notin \mathcal{N}_k \\ \frac{1}{2} Q_{k_i k_j}, & \text{for } k_i \notin \mathcal{N}_k \text{ and } k_j \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases}, \quad (\bar{c}_k)_i = \begin{cases} c_{k_i}, & \text{for } k_i \in \mathcal{N}_k \\ 0, & \text{for } k_i \notin \mathcal{N}_k \end{cases} \quad (18a)$$

The coefficient matrices $\Pi_{kk} \in \mathbb{R}^{N_k \times (N_k+\underline{N}_k)}$ are given by:

$$\Pi_{kk} = \begin{bmatrix} e_{N_k+1}^T \\ \vdots \\ e_{N_k+\underline{N}_k}^T \end{bmatrix}, \quad (19)$$

where $e_i \in \mathbb{R}^{N_k+\underline{N}_k}$ are elementary column vectors. Note that for $Q_{ij} \neq 0$, $i \in \mathcal{N}_k$, $j \in \mathcal{N}_k$, and $k \neq k'$, $Q_{ij} z(i)z(j)$ can be included in the objective function of either partition k , partition k' , or both. In the lifting scheme shown in (18a), we assume that these terms are equally divided and included in each partition. However, this approach is arbitrary, and other lifting schemes are possible. In other words, we can manipulate the lifting scheme to set the partition sets to satisfy either $j \in \mathcal{N}_k$ or $j \notin \mathcal{N}_k$. Interestingly, one can show that the solution of the lifted problem is unique and is the same as that of problem (13). The proof of this assertion is intricate and will not be discussed here due to the lack of space. To simplify the notation, we

express the lifted variables \bar{z}_k , matrices \bar{Q}_k , and cost vectors \bar{c}_k in (17) simply as z_k, Q_k, c_k .

The primal-dual solution of (17) can be obtained by solving the KKT conditions:

$$\begin{bmatrix} Q_1 & \Pi_{11}^T & \Pi_{21}^T & \cdots & \Pi_{K1}^T \\ \Pi_{11} & \Pi_{12} & \cdots & \Pi_{1K} & \\ & \Pi_{12}^T & Q_2 & \Pi_{22}^T & \vdots \\ & \Pi_{21} & \Pi_{22} & \vdots & \\ & \vdots & \ddots & \vdots & \\ & \vdots & & \vdots & \\ & \Pi_{1K}^T & \cdots & \cdots & \Pi_{KK}^T \\ \Pi_{K1} & \cdots & \cdots & \Pi_{KK} & \end{bmatrix} \begin{bmatrix} z_1 \\ \lambda_1 \\ z_2 \\ \lambda_2 \\ \vdots \\ z_K \\ \lambda_K \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \\ c_2 \\ 0 \\ \vdots \\ c_K \\ 0 \end{bmatrix} \quad (20)$$

By exploiting the structure of this system, we can derive a GS scheme of the form:

$$\begin{bmatrix} Q_k & \Pi_{kk}^T \\ \Pi_{kk} & \end{bmatrix} \begin{bmatrix} z_k^{\ell+1} \\ \lambda_k^{\ell+1} \end{bmatrix} = \begin{bmatrix} c_k \\ 0 \end{bmatrix} - \sum_{k'=1}^{k-1} \begin{bmatrix} 0 & \Pi_{k'k}^T \\ \Pi_{kk'} & 0 \end{bmatrix} \begin{bmatrix} z_{k'}^{\ell+1} \\ \lambda_{k'}^{\ell+1} \end{bmatrix} - \sum_{k'=k+1}^K \begin{bmatrix} 0 & \Pi_{k'k}^T \\ \Pi_{kk'} & 0 \end{bmatrix} \begin{bmatrix} z_{k'}^\ell \\ \lambda_{k'}^\ell \end{bmatrix}. \quad (21)$$

Here, we have used a lexicographic coordination order. We note that the solution of the linear system (21) solves the optimization problem:

$$z_k^{\ell+1} = \underset{z_k}{\operatorname{argmin}} \quad \frac{1}{2} z_k^T Q_k z_k - z_k^T \left(c_k - \sum_{k'=1}^{k-1} \Pi_{k'k}^T \lambda_{k'}^{\ell+1} - \sum_{k'=k+1}^K \Pi_{k'k}^T \lambda_{k'}^\ell \right) \quad (22a)$$

$$\text{s.t.} \quad \Pi_{kk} z_k + \sum_{k'=1}^{k-1} \Pi_{kk'} z_{k'}^{\ell+1} + \sum_{k'=k+1}^K \Pi_{kk'} z_{k'}^\ell = 0 \quad (\lambda_k). \quad (22b)$$

From this structure, we can see how the primal and dual variables propagate forward and backward relative to the partition k , due to the inherent block triangular nature of the GS scheme. We now seek to establish a condition that guarantees convergence of the GS scheme in this more general setting. To do so, we define:

$$A_k := \begin{bmatrix} Q_k & \Pi_{kk}^T \\ \Pi_{kk} & \end{bmatrix}, \quad x_k^\ell := \begin{bmatrix} z_k^\ell \\ \lambda_k^\ell \end{bmatrix}, \quad b_k := \begin{bmatrix} c_k \\ 0 \end{bmatrix}, \quad B_{kk'} := \begin{bmatrix} -\Pi_{k'k}^T \\ -\Pi_{kk'} \end{bmatrix}. \quad (23)$$

By using (23), we can express (21) as:

$$A_k x_k^{\ell+1} = b_k + \sum_{k'=1}^{k-1} B_{kk'} x_{k'}^{\ell+1} + \sum_{k'=k+1}^K B_{kk'} x_{k'}^\ell. \quad (24)$$

This can be expressed in matrix form as:

$$\begin{bmatrix} A_1 & & \\ -B_{21} & A_2 & \\ \vdots & \ddots & \ddots \\ -B_{K1} & \cdots & -B_{KK-1} & A_K \end{bmatrix} \begin{bmatrix} x_1^{\ell+1} \\ x_2^{\ell+1} \\ \vdots \\ x_K^{\ell+1} \end{bmatrix} = \begin{bmatrix} B_{12} & \cdots & B_{1K} \\ \ddots & \ddots & \vdots \\ & & B_{K-1K} \end{bmatrix} \begin{bmatrix} x_1^\ell \\ x_2^\ell \\ \vdots \\ x_K^\ell \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}. \quad (25)$$

We can see that the partition matrices A_k are nonsingular by inspecting the block structure of Q_k . In particular, by the definition of Q_k in (18a) (we denoted Q_k as \underline{Q}_k here) and the definition of Π_{kk} in (19), we can express A_k as:

$$A_k = \begin{bmatrix} \hat{Q}_k & \underline{Q}_k^T \\ \underline{Q}_k & I \end{bmatrix} \quad (26)$$

where each components of $\hat{Q}_k \in \mathbb{R}^{N_k \times N_k}$ and $\underline{Q}_k \in \mathbb{R}^{\frac{N_k}{2} \times N_k}$ are defined in (18a). Since Q is positive definite, \hat{Q}_k is also positive definite. Noting that \hat{Q}_k is nonsingular, we can see that the columns of A_k are linearly independent, and thus A_k is nonsingular as well. This implies that the block lower triangular matrix on the left-hand side of (25) is also nonsingular. To simplify notation we define:

$$w^\ell := \begin{bmatrix} x_1^\ell \\ x_2^\ell \\ \vdots \\ x_K^\ell \end{bmatrix}, \quad r := \begin{bmatrix} A_1 & & & \\ -B_{21} & A_2 & & \\ \vdots & \ddots & \ddots & \\ -B_{K1} & \cdots & -B_{KK-1} & A_K \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} \quad (27a)$$

$$S := \begin{bmatrix} A_1 & & & \\ -B_{21} & A_2 & & \\ \vdots & \ddots & \ddots & \\ -B_{K1} & \cdots & -B_{KK-1} & A_K \end{bmatrix}^{-1} \begin{bmatrix} B_{12} & \cdots & B_{1K} \\ \ddots & \ddots & \vdots \\ & & B_{K-1K} \end{bmatrix}. \quad (27b)$$

We express (25) by using the compact form $w^{\ell+1} = Sw^\ell + r$ or $w^{\ell+1} = S^\ell w^0 + (I + S + \cdots + S^{\ell-1})r$. The solution of (20) satisfies $w = Sw + r$. This implies that the solution of the lifted problem also solves the original problem (13). We now formally establish the following convergence result for the GS scheme.

Proposition 1 *The GS scheme (22) converges to the solution of (13) if all the eigenvalues of the matrix:*

$$\Sigma := \begin{bmatrix} Q_1 & & \Pi_{11}^T & \cdots & \Pi_{K1}^T \\ & Q_2 & \Pi_{12}^T & \Pi_{22}^T & \\ & & \ddots & \ddots & \ddots \\ & & \Pi_K & \Pi_{1K}^T & \cdots & \Pi_{K-1K}^T & \Pi_{KK}^T \\ \Pi_{11} & & & & & & -\Pi_{21}^T & \cdots & -\Pi_{K1}^T \\ \Pi_{21} & \Pi_{22} & & & & & \vdots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & & & -\Pi_{12} & \cdots & -\Pi_{1K} \\ \Pi_{K1} & \cdots & \Pi_{K-1K} & \Pi_{KK} & & & \vdots & \ddots & \vdots \\ & & & & & & -\Pi_{K-1K} & & \end{bmatrix}^{-1}$$

have magnitude less than one.

Proof Matrix Σ is a permutation of matrix S . The permutation is a similarity transformation, and thus S has the same set of eigenvalues as Σ . Consequently, all the nonzero eigenvalues of S have a magnitude of less than one. This implies that all the eigenvalues of S^ℓ decay exponentially as $\ell \rightarrow \infty$. Furthermore, this indicates that the series $I + S + S^2 + \cdots$ converges to $(I - S)^{-1}$. Note that $I - S$ is invertible since all the eigenvalue of S have magnitude less than one. Accordingly, the solution of the scheme (22) satisfies $\lim_{\ell \rightarrow \infty} w^\ell = (I - S)^{-1}r$. This convergence value $(I - S)^{-1}r$ satisfies equation $w = Sw + r$. Note that this is a unique solution to equation $w = Sw + r$ and thus solves (17). Problem (17) has same solution of problem (13), and both problems have unique solutions. Thus, the solution of the GS scheme (22) converges to the solution of (13). \square

Convergence is achieved without any assumptions on the initial guess w^0 . The error between the solution of problem (13) and the solution of the ℓ -th coordination updates of the GS scheme (21) (i.e., $(I - S)^{-1}r - w^\ell$) is given by $\epsilon^\ell = S^\ell((I - S)^{-1}r - w^0)$. We will call $\|\epsilon^\ell\|$ the *error* of the ℓ -th coordination step. If we express the initial error term as $\|\epsilon^0\|$, we can write $\|\epsilon^\ell\| = \|S^\ell\epsilon^0\|$. Moreover, if we define $\rho(S) = \lambda_{max}(S)$ then, for any $\delta > 0$, there exists $\kappa > 0$ such that the bound $\|S^\ell\| \leq \kappa|\rho(S) + \delta|^\ell$ holds for all ℓ and where $\|S\|$ is a matrix norm of S . Using this inequality we can establish that $\|\epsilon^\ell\| \leq \kappa|\rho(S) + \delta|^\ell\|\epsilon^0\|$. Since we can choose δ to be arbitrarily small, we can see that the decaying rate of the error is $O(\rho(S)^\ell)$. Although the error decays exponentially we note that, if $\rho(S)$ is close to one, convergence can be slow and thus having a good initial guess w^0 is essential. We also note that $\rho(S)$ is *tightly related to the topology of the coupling* between partitions, indicating that the partition structure contributes to the convergence rate.

We highlight that the GS concepts discussed here focus on problems with no inequality constraints. In practice, however, GS schemes can also be applied to such problems by using projected GS schemes [6, 34, 35].

2.3 Coordination Orders

In the lexicographic coordination order proposed in (21), we use the update sequence $k = 1, 2, \dots, K$. We note that the order affects the structure of the matrix Σ and thus convergence can be affected as well. To see this, consider a new

order given by the sequence $\sigma(1), \sigma(2), \dots, \sigma(K)$ where $\sigma : \{1, 2, \dots, K\} \rightarrow \{1, 2, \dots, K\}$ is a bijective mapping. We can use this mapping to rearrange the partition variables as:

$$[z_1, \lambda_1, z_2, \lambda_2, \dots, z_K, \lambda_K] \rightarrow [z_{\sigma(1)}, \lambda_{\sigma(1)}, z_{\sigma(2)}, \lambda_{\sigma(2)}, \dots, z_{\sigma(K)}, \lambda_{\sigma(K)}]. \quad (28)$$

This gives the reordered matrix:

$$S = \begin{bmatrix} A_{\sigma(1)} & & & & \\ -B_{\sigma(2)\sigma(1)} & A_{\sigma(2)} & & & \\ \vdots & \ddots & \ddots & & \\ -B_{\sigma(K)\sigma(1)} & \cdots & -B_{\sigma(K)\sigma(K-1)} & A_{\sigma(K)} & \end{bmatrix}^{-1} \begin{bmatrix} B_{\sigma(1)\sigma(2)} & \cdots & B_{\sigma(1)\sigma(K)} \\ \ddots & \ddots & \vdots \\ & B_{\sigma(K-1)\sigma(K)} & \end{bmatrix} \quad (29)$$

Importantly, the *change in update order is not necessarily a similarity transformation* of matrix S , and thus the eigenvalues will be altered. The GS scheme converges as long as eigenvalues of the reordered matrix S have magnitude less than one, but the convergence rate will be affected. Interestingly, GS schemes are highly flexible and allow for a large number of update orders. For instance, in some cases one can derive ordering sequences that enable parallelization. As an example, the 1-D spatial problem has a special structure with $B_{kk'}=0$ for any (k, k') such that $|k - k'| \geq 2$. The GS scheme becomes:

$$A_k x_k^{\ell+1} = b_k + B_{kk-1} x_{k-1}^{\ell+1} + B_{kk+1} x_{k+1}^{\ell} \quad (30)$$

Instead, we consider the following ordering $\sigma(i) = 2i - 1$ for $1 \leq i \leq \frac{K}{2}$ and $\sigma(i) = 2i - K$ for $\frac{K}{2} + 1 \leq i \leq K$. Here we assume that the number of partitions is even. This is called a *red-black ordering* and is widely popular in the solution of PDEs. By changing the index using $\sigma(\cdot)$, we can express (30) as:

$$A_{\sigma(i)} x_{\sigma(i)}^{\ell+1} = b_{\sigma(i)} + B_{\sigma(i)\sigma(i+\frac{K}{2})} x_{\sigma(i+\frac{K}{2})}^{\ell}, \quad i = 1 \quad (31a)$$

$$A_{\sigma(i)} x_{\sigma(i)}^{\ell+1} = b_{\sigma(i)} + B_{\sigma(i+\frac{K}{2}-1)\sigma(i+\frac{K}{2}-1)} x_{\sigma(i+\frac{K}{2}-1)}^{\ell} + B_{\sigma(i)\sigma(i+\frac{K}{2})} x_{\sigma(i+\frac{K}{2})}^{\ell},$$

$$2 \leq i \leq \frac{K}{2} \quad (31b)$$

and

$$\begin{aligned} A_{\sigma(i)} x_{\sigma(i)}^{\ell+1} &= b_{\sigma(i)} + B_{\sigma(i)\sigma(i-\frac{K}{2})} x_{\sigma(i-\frac{K}{2})}^{\ell+1} + B_{\sigma(i)\sigma(i-\frac{K}{2}+1)} x_{\sigma(i-\frac{K}{2}+1)}^{\ell+1}, \\ \frac{K}{2} + 1 \leq i &\leq K - 1. \end{aligned} \quad (32a)$$

$$A_{\sigma(i)} x_{\sigma(i)}^{\ell+1} = b_{\sigma(i)} + B_{\sigma(i)\sigma(i-\frac{K}{2})} x_{\sigma(i-\frac{K}{2})}^{\ell+1}, \quad i = K. \quad (32b)$$

We can thus see that the solution of (31) can proceed *independently* for any $1 \leq i \leq \frac{K}{2}$. This is because these partitions only depend on the solutions of (32) but not

Fig. 3 Sketch of 1-D ordering methods.

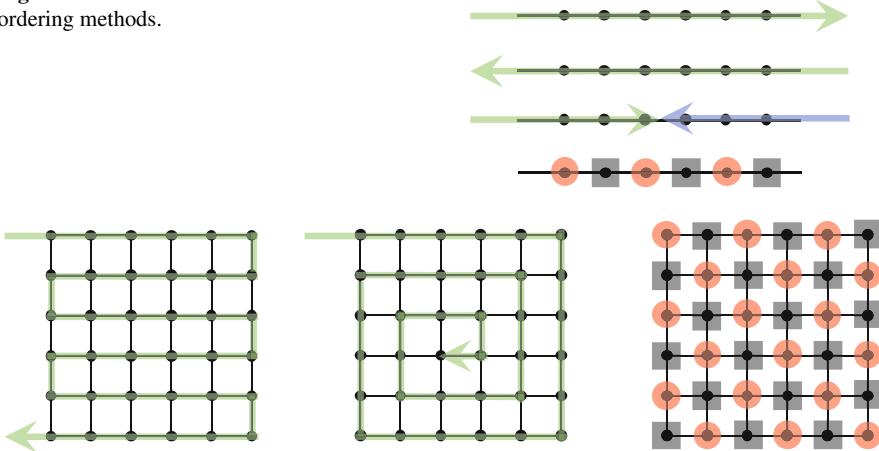


Fig. 4 Sketch of 2-D ordering methods.

on the solutions of (31). Likewise, solving (32) can be done independently for any $\frac{K}{2} + 1 \leq i \leq K$. Red-black ordering thus enables parallelism. Different coordination orders for 1-D and 2-D meshes are presented in Figures 3 and 4.

2.4 Coarsening

Low-complexity coarse versions of the full-resolution problem (1) can be solved to obtain an initial guess for the GS scheme and with this accelerate coordination. To illustrate how this can be done, we use the following representation of (1):

$$\min_z \quad \frac{1}{2} z^T Q z - c^T z \quad (33a)$$

$$\text{s.t.} \quad \underbrace{\begin{bmatrix} A \\ \Pi \end{bmatrix}}_{\tilde{A}} z + \underbrace{\begin{bmatrix} B \\ 0 \end{bmatrix}}_{\tilde{B}} d = 0 \quad (\nu, \lambda) \quad (33b)$$

Our goal is to obtain a substantial reduction in the dimension of this problem by introducing a *mapping* from a coarse variable space to the original space. This is represented by the linear mapping $z = T\tilde{z}$ where $\tilde{z} \in \mathbb{R}^{N_c \cdot n_c}$ is the coarse variable, and we assume that the mapping $T \in \mathbb{R}^{N \cdot n_c \times N_c \cdot n_c}$ (called a restriction operator) has full column rank and $N_c < N$. We can thus pose the low-dimensional coarse problem:

$$\min_{\tilde{z}} \quad \frac{1}{2} \tilde{z}^T T^T Q T \tilde{z} - c^T T \tilde{z} \quad (34a)$$

$$\text{s.t.} \quad U^T \bar{A} T \tilde{z} + U^T \bar{B} d = 0 \quad (\tilde{v}, \tilde{\lambda}) \quad (34b)$$

A key issue that arises in coarsening is that the columns of matrix $\bar{A} T$ do not necessarily span the entire range space of \bar{B} (i.e., we might not be able to find a coarse variable \tilde{z} that satisfies $\bar{A} T \tilde{z} + \bar{B} d = 0$). Consequently, we introduce a *constraint aggregation* matrix U that has full column rank to ensure that $U^T \bar{A} T$ spans the range space of $U^T \bar{B}$. With this, we can ensure that the feasible set of (34) is non-empty.

After solving the coarse problem (34), we can project the primal-dual solution from the coarse space to the original space. We note that the dimension of the dual space is also reduced because we performed constraint aggregation. The projection for the primal solution can be done by using $z = T \tilde{z}$, while the projection for the dual solution can be obtained as $(v, \lambda) = U(\tilde{v}, \tilde{\lambda})$. The derivation of coarsened representations is *application-dependent* and often requires domain-specific knowledge. In particular, coarsening can also be performed by using reduced-order modeling techniques such as proper orthogonal decompositions [2] or coherency-based network aggregation schemes [20]. In the following sections, we demonstrate how to derive coarse representations in certain settings.

2.5 Multi-Grid Schemes and Hierarchical Coordination

Multi-grid serves as a bridge between a fully centralized and a fully decentralized coordination schemes. In particular, a fully centralized scheme would aim to find a solution of the full-resolution problem (2) by gathering all the information in a single processing unit. A fully decentralized scheme such as GS, on the other hand, would proceed by finding solutions to subproblems (3) over each partition, and information would only be shared between the connected partitions through the coordination variables. A drawback of a decentralized scheme is that a potentially large number of coordination steps might be needed to reach a solution for the full-resolution problem, particularly when many partitions are present.

In a multi-grid scheme, we seek to aid the decentralized scheme by using information from a low-resolution central problem that oversees the entire domain. In our context, the information is in the form of states and dual variables defined over the partition interfaces (i.e., the coupling variables and constraints). The key idea of this hierarchical arrangement is that the coarse central scheme can capture effects occurring at low global frequencies, while the agents in the decentralized schemes can handle effects occurring at high local frequencies. As can be seen, multi-grid provides a framework to design hierarchical control architectures by leveraging existing and powerful reduced-order modeling techniques such as coherency-based aggregation and decentralized control schemes.

Multi-grid is a widely studied computational paradigm. The framework proposed here presents basic elements of this paradigm, but diverse extensions are possible [7, 9]. For instance, the scheme proposed here involves only a coarse and a fine resolution level, but one can create multilevel schemes that transfer information between multiple scales recursively by using meshes of diverse resolution. This can allow us to cover a wider range of frequencies present in the system. In the following section, we illustrate how sequential coarsening can be beneficial.

From an *electricity markets* perspective, we highlight that multi-grid schemes provide a framework to coordinate transactions at multiple spatial and temporal scales. To see this, consider the case of spatial coordination of electricity markets. Under such setting, we can interpret each partition of the spatial domain as a market player (e.g., a microgrid). The market players have internal resources (e.g., distributed energy resources) that they manage to satisfy their internal load. The players, however, can also transact energy with other players in order to improve their economic performance. The proposed GS scheme provides a mechanism to handle intra-partition decision-making (by solving the partition subproblems) and inter-partition transactions by exchanging state (voltages) and dual information (nodal prices). If transaction information is exchanged multiple times (corresponding to multiple GS iterates), the GS scheme will converge to an equilibrium point corresponding to the solution of the centralized economic maximization problem (e.g., the social welfare problem). This is a useful property of decentralized coordination because centralization of information and decision-making is often impractical. If the players only exchange information once (or a handful of times), they might not reach an optimal equilibrium, and an inefficiency will be introduced. Moreover, when a disturbance affects the system, many GS iterations might be needed to reach the new optimal equilibrium. This is where hierarchical optimization becomes beneficial, because one can solve and aggregated spatial representation of the system (in which each partition is treated as a node) to compute approximate dual variables and states at the interfaces of the partitions. This approximation can be used to aid the convergence of the decentralized GS scheme (by conveying *global* spatial information to *local* market players). One can think of the coarse high-level problem as a system operator (supervisor) problem (e.g., at the distribution level). The operator might, at the same time, need to coordinate with other system operators (each of which oversees its own set of market players). These system operators can at the same time be aggregated into a higher level which would represent, for instance, a transmission or regional operator. We can thus see that hierarchical multi-grid schemes enable scalable coordination of a potentially large number of market players over large geographical regions. The hierarchical multi-grid scheme can also be applied to handle multiple timescales of a single market player that might need to manage, for instance, assets with different dynamic characteristics.

3 Case Studies

We now present numerical case studies to demonstrate the concepts in the context of temporal and spatial management of energy systems. We use a multi-scale (in time) optimization problem with features of a storage management problem and a multi-scale (in space) optimization problem that considers power flow dispatch over a network.

3.1 Multi-Scale Temporal Control

We use a multi-grid scheme to solve the following temporal planning problem \mathcal{P} :

$$\min_{x,u} \quad \sum_{i \in \mathcal{N}} (x(i)^2 + u(i)^2) \quad (35a)$$

$$\text{s.t. } x(i+1) = x(i) + \delta(u(i+1) + d(i+1)), \quad i \in \mathcal{N} \quad (35b)$$

$$x(0) = 0. \quad (35c)$$

This problem has a state, a control, and a disturbance defined over $N = M \cdot K$ time points contained in the set \mathcal{N} . The state and control are grouped into the decision variable $z(i) = (x(i), u(i))$. The distance between mesh points is given by δ . The structure of this problem resembles that of an inventory (storage) problem in which the disturbance $d(i)$ is a load and $u(i)$ is a charge/discharge flow from the storage. In these types of problems, the load might have frequencies covering multiple timescales (e.g., seasonal, daily, and down to seconds). Consequently, the time mesh δ has to be rather fine to capture all the frequencies. Moreover, the planning horizon (the time domain $N \cdot \delta$) might need to be long so as to capture the low-frequency components in the load. We partition the problem into K partitions, each containing M points. The set of inner points in the partition is defined as \mathcal{M} . The optimization problem over a partition k solved in the GS scheme is given by:

$$\min_{x_k, u_k} \quad \sum_{i \in \mathcal{M}} (x_k(i)^2 + u_k(i)^2) + x_k(M)\lambda_{k+1}^\ell \quad (36a)$$

$$\text{s.t. } x_k(i+1) = x_k(i) + \delta(u_k(i+1) + d_k(i+1)), \quad i \in \mathcal{M} \quad (36b)$$

$$x_k(0) = x_{k-1}^{\ell+1}(M) \quad (\lambda_k). \quad (36c)$$

In our numerical experiments, we set $K = 10$ and $M = 100$ to give $N = 1,000$ points. The time mesh points were set $t(i) = i \cdot \delta$ with $\delta = 0.1$. We use a disturbance signal composed of a low and a high frequency $d(i) = 4 \sin\left(\frac{4\pi i}{N}\right) + \sin\left(\frac{24\pi i}{N}\right)$, $i \in \mathcal{N}$. The disturbance signal and its frequency components are shown in Figure 5.

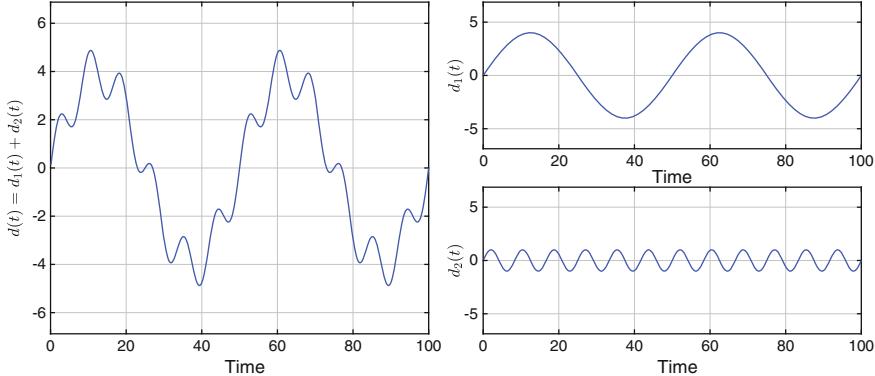


Fig. 5 Disturbance profile (left) and frequency components (right) for temporal problem.

We solve a coarse problem to create a hierarchical structure that aids the GS scheme (which operates on the partitions at high resolution). To perform coarsening, we aggregate $M_c = 4$ internal grid points (i.e., collapse 25 points into one coarse point). The set of inner coarse points is \mathcal{M}_c . The projection is $x_k = T_x \tilde{x}_k$ and $u_k = T_u \tilde{u}_k$ with:

$$T_x = T_u = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \ddots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad (37a)$$

The projection matrices T_x and T_u have full column rank. The fine to coarse restriction can also be expressed as:

$$\tilde{x}_k(\tilde{i}) = x_k(i), \quad \tilde{u}_k(\tilde{i}) = u_k(i), \quad \text{if } \tilde{i} = \lfloor \frac{i-1}{M/M_c} \rfloor + 1 \quad (38)$$

where $\lfloor \cdot \rfloor$ is a round-down operator. We denote $\varphi(i) = \lfloor \frac{i-1}{M/M_c} \rfloor + 1$. The coarse problem can then be stated in terms of the coarse variables as follows:

$$\min_{\tilde{x}_k, \tilde{u}_k} \sum_{j \in \mathcal{M}_c} (\tilde{x}_k(j)^2 + \tilde{u}_k(j)^2) + \tilde{x}_k(M_c) \lambda_{k+1}^\ell \quad (39a)$$

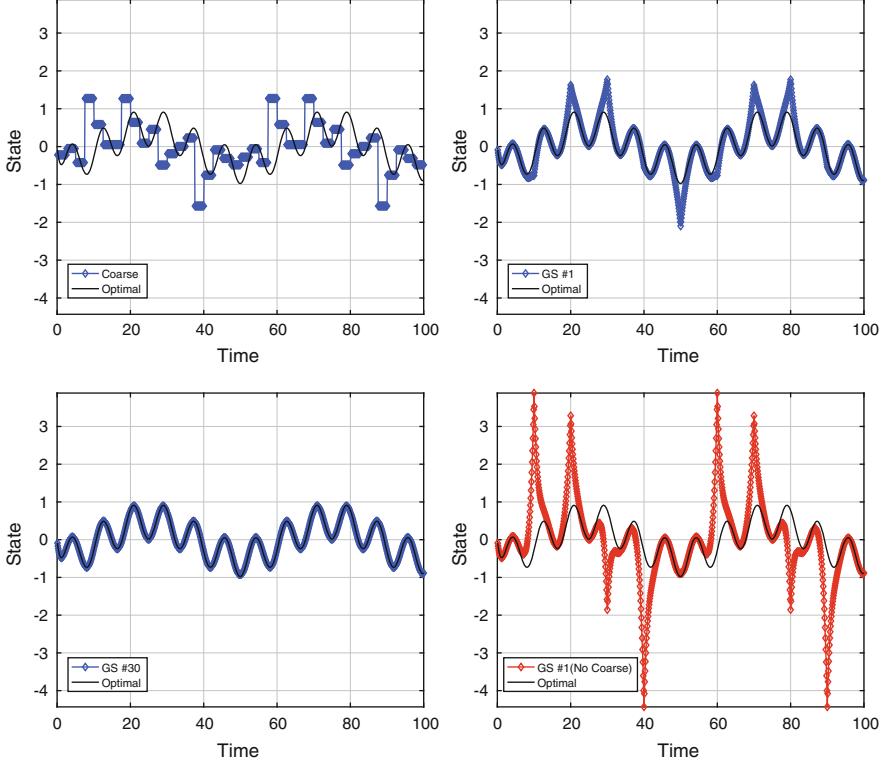


Fig. 6 (Top-left) Solution of coarse problem, (top-right) solution of first GS step with coarsening, (bottom-left) solution of 30th GS step with coarsening, (bottom-right) solution of first GS step without coarsening.

$$\text{s.t. } \tilde{x}_k(j+1) = \tilde{x}_k(j) + \frac{M}{M_c} \delta(\tilde{u}_k(j+1) + \tilde{d}_k(j+1)), \quad j \in \mathcal{M}_c \quad (39b)$$

$$\tilde{x}_k(0) = \tilde{x}_{k-1}^{\ell+1}(M_c) \quad (\lambda_k), \quad (39c)$$

where $\tilde{d}_k(j) = \frac{M_c}{M} \sum_{i \in \varphi^{-1}(j)} d_k(i)$ is the coarsened disturbance signal. The dynamic equations are defined over a smaller dimensional space (defined over \mathcal{M}_c) which results from aggregating the dynamic equations in the full-resolution problem (defined over \mathcal{M}). We solve the full-resolution problem (35) and compare its solution against that of the pure GS scheme, the one of the coarse low-resolution problem, and the one of the hierarchical scheme that solves the coarse problem to coordinate the GS scheme. We also solve the coarse problem by using a GS scheme. Figure 6 shows the results. We note that the solution to the coarse problem (36) captures the general long-term trend of the solution but misses the high frequencies. The GS scheme refines this solution and converges in around 30 coordination steps.

By comparing Figure 6 (top-right) and (bottom-right), we can see that initializing GS scheme with the coarse solution significantly reduces the initial error and demonstrates the benefit of the hierarchical scheme.

3.2 Multi-Scale Spatial Control

Useful insights on how to use multi-grid schemes to create hierarchical network control structures result from *interpreting network flows as diffusive processes*. To illustrate this, we consider a network with nodes defined on a rectangular mesh. A node (i, j) in the network exchanges flows with its four neighboring nodes $(i, j + 1)$, $(i, j - 1)$, $(i + 1, j)$, and $(i - 1, j)$ (this is called a stencil). The flows $f(i, j)$ are a function of the node potentials $p(i, j)$ and given by:

$$f(i, j; i, j + 1) = D(p(i, j) - p(i, j + 1)) \quad (40a)$$

$$f(i, j; i, j - 1) = D(p(i, j) - p(i, j - 1)) \quad (40b)$$

$$f(i, j; i + 1, j) = D(p(i, j) - p(i + 1, j)) \quad (40c)$$

$$f(i, j; i - 1, j) = D(p(i, j) - p(i - 1, j)). \quad (40d)$$

Here, $D \in \mathbb{R}$ is the diffusion constant (i.e., the flow resistance) of the link connecting the nodes. At each node (i, j) , we have a load $d(i, j)$ and a source $u(i, j)$ that is used to counteract (balance) the load. This gives the flow balance conservation equation:

$$\begin{aligned} & f(i, j; i, j + 1) + f(i, j; i, j - 1) + f(i, j; i + 1, j) + f(i, j; i - 1, j) \\ &= u(i, j) + d(i, j). \end{aligned}$$

This can also be written in terms of the potentials as:

$$\begin{aligned} & D(4 \cdot p(i, j) - p(i - 1, j) - p(i + 1, j) - p(i, j - 1) - p(i, j + 1)) \\ &= u(i, j) + d(i, j). \end{aligned}$$

We assume that we have fixed 2-D spatial domain $\Omega := [0, X] \times [0, Y]$ that is discretized using $M \cdot P$ nodes in each direction. The sets $\mathcal{N}^x = \mathcal{N}^y := \{1, 2, \dots, P \cdot M\}$ are the sets of points in each direction. The set of total mesh points is $\mathcal{N} = \mathcal{N}^x \times \mathcal{N}^y$ and thus $N = (P \cdot M) \cdot (P \cdot M)$. As the number of nodes increases, the node potentials form a continuum described by the 2-D diffusion equation:

$$D \left(\frac{\partial^2 p(x, y)}{\partial x^2} + \frac{\partial^2 p(x, y)}{\partial y^2} \right) = u(x, y) + d(x, y), \quad (x, y) \in \Omega. \quad (41)$$

Using this analogy, we consider the following full-space problem:

$$\min_{p,u} \sum_{(i,j) \in \mathcal{N}} (p(i,j)^2 + u(i,j)^2) \quad (42a)$$

$$\text{s.t. } D(4 \cdot p(i,j) - p(i-1,j) - p(i+1,j) - p(i,j-1) - p(i,j+1)) \\ = u(i,j) + d(i,j), \quad (i,j) \in \mathcal{N} \quad (42b)$$

$$p(0,j) = 0, \quad j \in \mathcal{N}^y \quad (42c)$$

$$p(M \cdot P + 1, j) = 0, \quad j \in \mathcal{N}^y \quad (42d)$$

$$p(i,0) = 0, \quad i \in \mathcal{N}^x \quad (42e)$$

$$p(i, M \cdot P + 1) = 0, \quad i \in \mathcal{N}^x \quad (42f)$$

The goal of the optimization problem is, given the loads $d(i,j)$, to control the potentials in the network nodes $p(i,j)$ by using the sources $u(i,j)$. The decision variables at every node are $z(i,j) = (p(i,j), u(i,j))$. The presence of multiple frequencies in the 2-D *disturbance load field* $d(i,j)$ might require us to consider fine meshes, making the optimization problem intractable. One can think of the *disturbance field* as spatial variations of electrical loads observed over a geographical region. If the loads have high-frequency spatial variations, it would imply that we need high control resolution (i.e., we need sources at every node in the network to achieve tight control). This can be achieved, for instance, by installing distributed energy resources (DERs). Moreover, if the load has low-frequency variations, it would imply that the DERs would have to cooperate to counteract global variations.

In our experiments, the size of mesh was set $P = 10$ and $M = 10$, which results in $N = 10,000$ mesh points. To address this complexity, we partition the 2-D domain into $K = P \cdot P$ partitions each with $M \cdot M$ points, and we label each element in the partition as $k = (n,m) \in \mathcal{K}$. We can think of each partition $k \in \mathcal{K}$ as a region of the network. We define inner index sets by: $\mathcal{M}^x = \mathcal{M}^y := \{1, 2, \dots, M\}$ and $\mathcal{M} := \mathcal{M}^x \times \mathcal{M}^y$. The GS scheme for partition (n,m) is given by:

$$\min_{p_{n,m}, u_{n,m}} \sum_{(i,j) \in \mathcal{M}} (p_{n,m}(i,j)^2 + u_{n,m}(i,j)^2) \quad (43a)$$

$$+ \sum_{j \in \mathcal{M}^y} p_{n,m}(1,j) \lambda_{n-1,m}^{\ell+1}(M+1,j) + \sum_{j \in \mathcal{M}^y} p_{n,m}(M,j) \lambda_{n+1,m}^{\ell}(0,j)$$

$$+ \sum_{i \in \mathcal{M}^x} p_{n,m}(i,1) \lambda_{n,m-1}^{\ell+1}(i,M+1) + \sum_{i \in \mathcal{M}^x} p_{n,m}(i,M) \lambda_{n,m+1}^{\ell}(i,0)$$

$$\text{s.t. } D(4 \cdot p_{n,m}(i,j) - p_{n,m}(i-1,j) - p_{n,m}(i+1,j) - p_{n,m}(i,j-1) \\ - p_{n,m}(i,j+1)) = u_{n,m}(i,j) + d_{n,m}(i,j), \quad (i,j) \in \mathcal{M} \quad (43b)$$

$$p_{n,m}(0,j) = p_{n-1,m}^{\ell+1}(M,j), \quad (\lambda_{n,m}(0,j)) \quad (43c)$$

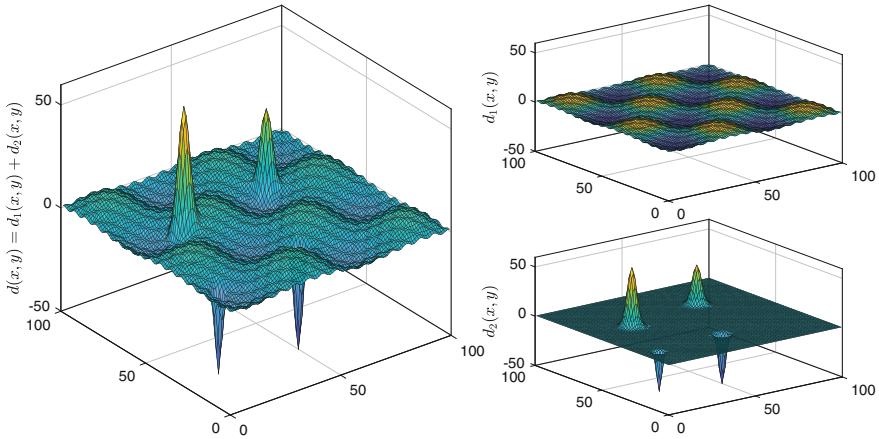


Fig. 7 Disturbance field (left) and its components (right) for spatial optimization problem.

$$p_{n,m}(M+1, j) = p_{n+1,m}^\ell(1, j), \quad (\lambda_{n,m}(M+1, j)) \quad (43d)$$

$$p_{n,m}(i, 0) = p_{n,m-1}^{\ell+1}(i, M), \quad (\lambda_{n,m}(i, 0)) \quad (43e)$$

$$p_{n,m}(i, M+1) = p_{n,m+1}^\ell(i, 1), \quad (\lambda_{n,m}(i, M+1)) \quad (43f)$$

The constraint indices for the constraints (43c)–(43f) run over $j \in \mathcal{M}^y$ and $i \in \mathcal{M}^x$.

To perform coarsening, a mesh of $(M/M_c) \cdot (M/M_c)$ points is collapsed into a single coarse point, and the mapping from the coarse space to the original space is:

$$\tilde{p}_{n,m}(\tilde{i}, \tilde{j}) = p_{n,m}(i, j) \quad \text{if } \tilde{i} = \lfloor \frac{i-1}{M/M_c} \rfloor + 1, \quad \tilde{j} = \lfloor \frac{j-1}{M/M_c} \rfloor + 1. \quad (44)$$

As with the temporal case, we also perform aggregation of the constraints in the partition to obtain a coarse representations. In our experiments, we used $M_c = 2$ as default. The disturbance field is given by a linear combination of a 2-D sinusoidal and of a Gaussian function. The shape of the load field illustrated in Figure 7. Figure 8 shows the optimal potential field obtained with the coarse problem and that obtained with the GS scheme at the first and tenth steps (initialized with the coarse field). Note that the coarse field error captures the global structure of the load field but misses the high frequencies, while the GS scheme corrects the high-frequency load imbalances. In Figure 9 we again illustrate that the hierarchical scheme outperforms the decentralized GS scheme.

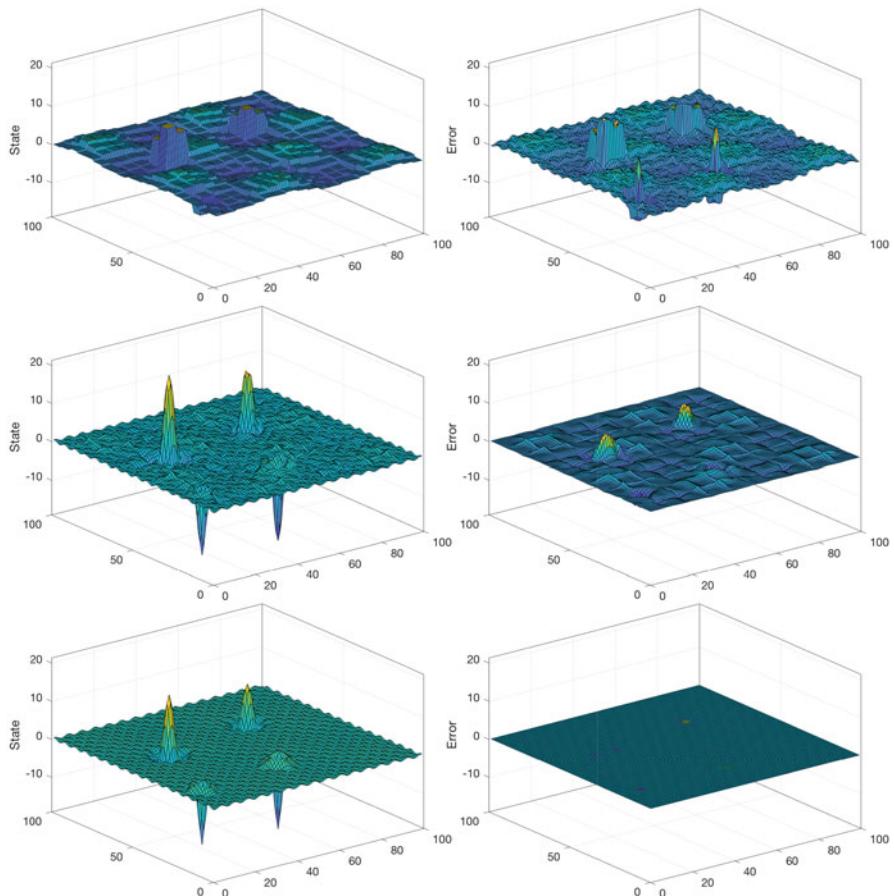


Fig. 8 (Top) Potential field solution and error of coarse problem, (middle) solution and error of first GS update, and (bottom) solution and error of tenth GS update.

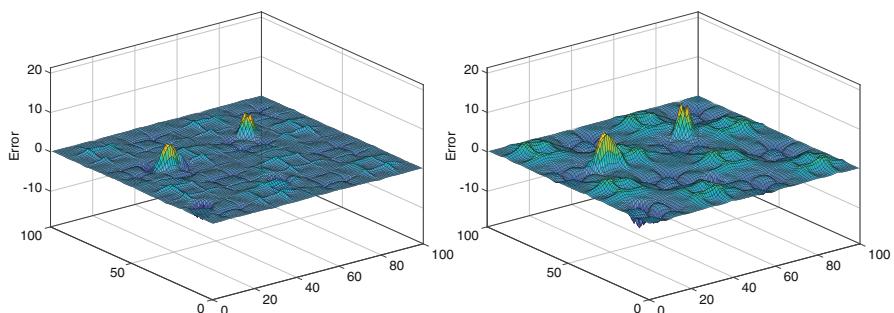


Fig. 9 (Left) Error for first GS update with coarsening and (right) error without coarsening.

3.3 Effect of Coarsening Strategy

We now compare the efficiency of coarsening with different resolutions on the performance of GS. The results are given in Figure 10 (top left and right) and reveal that using a higher resolution for the coarse problem does not necessarily result in better performance of the GS scheme. This is particularly evident in the temporal case, while for the spatial case increasing the mesh resolution does help. We attribute this difference to the asymmetric nature of the temporal problem compared to the symmetric mesh of the spatial case. For the temporal problem, we have found that the most effective coarsening strategy is to solve a sequence of coarse problems with increasing resolution. At each coarsening level, however, we only perform a single GS coordination step, and the resulting coordination variables are used to initialize the GS coordination at the next level. The error evolution of this sequential coarsening scheme is shown in Figure 10 (top left). We sequentially solved the coarse problems by using $M_c = 1, M_c = 2, M_c = 4, M_c = 5, M_c = 10, M_c =$

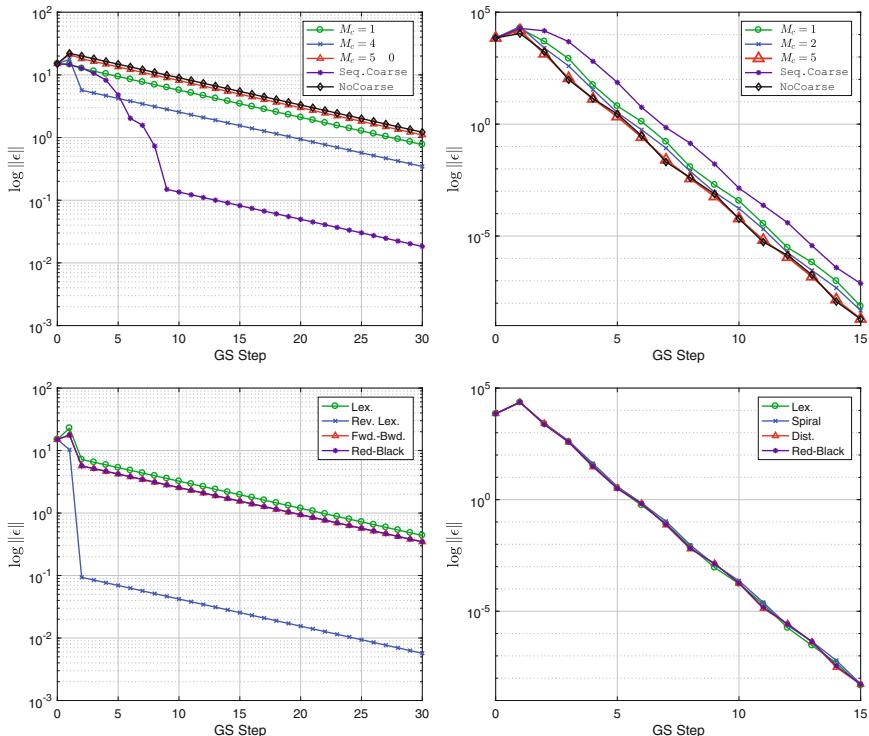


Fig. 10 (Top-left) Error of temporal control with different coarsening schemes, (top-right) error of spatial control with different coarsening schemes, (bottom-left) error of temporal control with different ordering schemes, and (bottom-right) error of spatial control with different ordering schemes.

20 , $M_c = 25$, and $M_c = 50$ (this gives a total of eight GS steps). We note that, at the tenth GS step, the solution from this sequential coarsening scheme is about seventy times smaller than that obtained with no coarsening. This can be attributed to the ability of the single step GS schemes to cover a wider range of frequencies.

3.4 Effect of Coordination Order

We solved the multi-scale temporal control problem (36) and the spatial control problem (43) with four different ordering methods for each problem. For temporal control problem, ordering method 1 was a lexicographic ordering, ordering method 2 was a reverse lexicographic ordering, ordering method 3 was a forward-backward ordering, and ordering method 4 was the red-black scheme. For the spatial control problem, ordering method 1 was a lexicographic ordering, ordering method 2 was a spiral-like ordering, ordering method 3 was the red-black ordering, and ordering method 4 was set by ordering the partitions based on the magnitude of the disturbance. The results are presented in Figure 10 (top left and right). As can be seen, in the temporal problem, the performance of reverse lexicographic ordering is significantly better than that achieved by other methods. This can be attributed to the asymmetry of the coupling topology. In particular, in the temporal problem, the primal variable information is propagated in forward direction, while the dual information is propagated in reverse direction. It can be seen that dual information plays an important role in the convergence of the temporal problem. In the spatial problem, the performance of the different orderings is virtually the same. The red-black ordering (which enables parallelism) achieves the same performance as the rest. We attribute this to the symmetry of the spatial domain.

4 Conclusions and Directions of Future Work

We have presented basic elements of multi-grid computing schemes and illustrated how to use these to create hierarchical coordination architectures for complex systems. In particular, we discuss how Gauss-Seidel schemes can be seen as decentralized coordination schemes that handle high-frequency effects, while coarse solution operators can be seen as low-resolution centralized coordination schemes that handle low-frequency effects. We believe that multi-grid provides a powerful framework to systematically construct hierarchical coordination architectures, but diverse challenges need to be addressed. In particular, it is important to understand convergence properties of GS schemes in more complex settings with nonlinear effects and inequality constraints. Moreover, it is necessary to develop effective coarsening (aggregation) schemes that can retain useful information while reducing complexity. Moreover, it is desirable to combine hierarchical coordination schemes and existing control theory to analyze stability and robustness properties.

Acknowledgements We acknowledge funding from the National Science Foundation under award NSF-EECS-1609183.

References

1. Albersmeyer J, Diehl M (2010) The lifted newton method and its application in optimization. *SIAM J Optim* 20(3):1655–1684
2. Antoulas AC, Sorensen DC, Gugercin S (2001) A survey of model reduction methods for large-scale systems. *Contemp Math* 280:193–220
3. Arnold M, Negenborn R, Andersson G, De Schutter B (2010) Distributed predictive control for energy hub coordination in coupled electricity and gas networks. In: Intelligent Infrastructures. Springer, Berlin, pp 235–273
4. Baldea M, Daoutidis P (2007) Control of integrated process networks multi-time scale perspective. *Comput Chem Eng* 31(5):426–444
5. Biegler LT, Zavala VM (2009) Large-scale nonlinear programming using IPOPT: an integrating framework for enterprise-wide dynamic optimization. *Comput Chem Eng* 33(3):575–582
6. Borzì A, Kunisch K (2005) A multigrid scheme for elliptic constrained optimal control problems. *Comput Optim Appl* 31(3):309–333
7. Borzì A, Schulz V (2009) Multigrid methods for PDE optimization. *SIAM Rev* 51(2):361–395 (2009)
8. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* 3(1):1–122
9. Brandt A (1986) Algebraic multigrid theory: the symmetric case. *Appl Math Comput* 19(1):23–56
10. Camponogara E, Jia D, Krogh BH, Talukdar S (2002) Distributed model predictive control. *IEEE Control. Syst.* 22(1):44–52
11. Chow JH, Kokotovic PV (1985) Time scale modeling of sparse dynamic networks. *IEEE Trans Autom Control* 30(8):714–722
12. Diehl M, Bock HG, Schlöder JP, Findeisen R, Nagy Z, Allgöwer F (2002) Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. *J Process Control* 12(4):577–585
13. Farina M, Zhang X, Scattolini R (Preprint, 2017) A hierarchical MPC scheme for interconnected systems. arXiv:1703.02739
14. Fisher ML (2004) The lagrangian relaxation method for solving integer programming problems. *Manag Sci* 50(12 suppl.):1861–1871
15. Giselsson P, Doan MD, Keviczky T, De Schutter B, Rantzer A (2013) Accelerated gradient methods and dual decomposition in distributed model predictive control. *Automatica* 49(3):829–833
16. Hong M, Luo ZQ (Preprint, 2012) On the linear convergence of the alternating direction method of multipliers. arXiv:1208.3922
17. Jamshidi M, Tarokh M, Shafai B (1992) Computer-aided analysis and design of linear control systems. Prentice-Hall, Inc., Upper Saddle River
18. Joo JY, Ilic MD (2013) Multi-layered optimization of demand resources using lagrange dual decomposition. *IEEE Trans Smart Grid* 4(4):2081–2088
19. Kokotovic P (1981) Subsystems, time scales and multimodeling. *Automatica* 17(6):789–795
20. Kokotovic P, Avramovic B, Chow J, Winkelman J (1982) Coherency based decomposition and aggregation. *Automatica* 18(1):47–56
21. Kozma A, Frasch JV, Diehl M (2013) A distributed method for convex quadratic programming problems arising in optimal control of distributed systems. In: 2013 IEEE 52nd annual conference on decision and control (CDC). IEEE, New York, pp 1526–1531

22. Lefort A, Bourdais R, Ansaney-Alex G, Guéguen H (2013) Hierarchical control method applied to energy management of a residential house. *Energ Buildings* 64:53–61
23. Liu C, Shahidehpour M, Wang J (2010) Application of augmented lagrangian relaxation to coordinated scheduling of interdependent hydrothermal power and natural gas systems. *IET Gener Transm Distrib* 4(12):1314–1325
24. Negenborn RR, De Schutter B, Hellendoorn J (2007) Efficient implementation of serial multi-agent model predictive control by parallelization. In: 2007 IEEE international conference on networking, sensing and control. IEEE, New York, pp 175–180
25. Peponides GM, Kokotovic PV (1983) Weak connections, time scales, and aggregation of nonlinear systems. *IEEE Trans Autom Control* 28(6):729–735
26. Rawlings JB, Mayne D (2008) Model predictive control. Noble Hill Publishing, Madison
27. Rawlings JB, Bonné D, Jørgensen JB, Venkat AN, Jørgensen SB (2008) Unreachable setpoints in model predictive control. *IEEE Trans Autom Control* 53(9):2209–2215
28. Scattolini R (2009) Architectures for distributed and hierarchical model predictive control—a review. *J Process Control* 19(5):723–731
29. Scattolini R, Colaneri P (2007) Hierarchical model predictive control. In: 2007 46th IEEE conference on decision and control. IEEE, New York, pp 4803–4808
30. Simon HA, Ando A (1961) Aggregation of variables in dynamic systems. *Econometrica* 29(2):111–138
31. Stewart BT, Venkat AN, Rawlings JB, Wright SJ, Pannocchia G (2010) Cooperative distributed model predictive control. *Syst Control Lett* 59(8):460–469
32. Stewart BT, Wright SJ, Rawlings JB (2011) Cooperative distributed model predictive control for nonlinear systems. *J Process Control* 21(5):698–704
33. Summers TH, Lygeros J (2012) Distributed model predictive consensus via the alternating direction method of multipliers. In: 2012 50th annual Allerton conference on communication, control, and computing (Allerton). IEEE, New York, pp 79–84
34. Zavala VM (2016) New architectures for hierarchical predictive control. *IFAC-PapersOnLine* 49(7):43–48
35. Zavala VM, Anitescu M (2010) Real-time nonlinear optimization as a generalized equation. *SIAM J Control Optim* 48(8):5444–5467
36. Zavala VM, Biegler LT (2009) The advanced-step NMPC controller: optimality, stability and robustness. *Automatica* 45(1):86–93
37. Zhu D, Yang R, Hug-Glazmann, G (2010) Managing microgrids with intermittent resources: a two-layer multi-step optimal control approach. In: North American power symposium (NAPS). IEEE, New York, pp 1–8

Graphical Models and Belief Propagation Hierarchy for Physics-Constrained Network Flows



Michael Chertkov, Sidhant Misra, Marc Vuffray, Dvijotham Krishnamurthy, and Pascal Van Hentenryck

Abstract We review new ideas and the first results from the application of the graphical models approach, which originated from statistical physics, information theory, computer science, and machine learning, to optimization problems of network flow type with additional constraints related to the physics of the flow. We illustrate the general concepts on a number of enabling examples from power system and natural gas transmission (continental scale) and distribution (district scale) systems.

M. Chertkov

Theoretical Division, T-4, CNLS, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Energy System Center, Skoltech, Moscow 143026, Russia

e-mail: chertkov@lanl.gov

S. Misra

Theoretical Division, T-5, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

e-mail: sidhant@lanl.gov

M. Vuffray

Theoretical Division, T-4, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

e-mail: vuffray@lanl.gov

D. Krishnamurthy (✉)

Pacific Northwest National Laboratory, PO Box 999, Richland, WA 99352, USA

e-mail: krishnamurthy.dvijotham@pnnl.gov; dvij@cs.washington.edu

P. Van Hentenryck

University of Michigan, Department of Industrial and Operations Engineering, Ann Arbor, MI 48109, USA

e-mail: pvanhent@umich.edu

1 Introductory Remarks

In this chapter we discuss optimization problems describing flows over networks constrained by the physical nature of the flows that occur in the context of electric power systems (see, e.g., [25, 43] and natural gas applications (see, e.g., [10] and references therein). Other examples of physical flows where similar optimization problems arise include pipe flow systems, such as district heating [14, 76] and water [55], as well as traffic systems [39]. We aim to show that the network flow optimization problems can be stated naturally in terms of the so-called graphical models (GMs). In general, GMs for optimization and inference are widespread in statistical disciplines such as applied probability, machine learning and artificial intelligence [9, 13, 28, 31, 50, 54], information theory [56], and statistical physics [46].

The main benefit of adapting GM methodology to the physics-constrained network flows is the modularity and flexibility of the approach—any new constraints, any set of new variables, and any modification of the optimization objective can be incorporated in the GM formulation with ease. Furthermore, if all (or at a least majority of) constraints and modifications are factorized, i.e., can be stated in terms of a small subset of variables, underlying GM optimization or GM statistical inference problems can be solved exactly or approximately with the help of an emerging set of techniques, algorithms, and computational approaches collectively coined belief propagation (BP) (see, e.g., an important original paper [75] and recent reviews [46, 56, 67]). It is also important to emphasize that an additional benefit of the GM formulation is its principal readiness for generalizations. Even though we limit our discussion to application of the GM and BP framework to deterministic optimizations, many probabilistic and/or mixed generalizations (largely not discussed in this paper) fit very naturally in this universal framework as well.

We will focus on optimization problems associated with physics-constrained network flow (PCNF) problems. Structure of the networks will obviously be inherited in the GM formulation; however, this takes place indirectly, through graph and variable transformations and modifications. Specifically, Section 2 is devoted solely to stating a number of example energy system formulations in GM terms. Thus, in Sections 2.1 and 2.2, we consider dissipation optimal and, respectively, general PCNF problems. In particular, Section 2.2 includes discussion of the power flow problems in both power-voltage (Section 2.2.1) and current-voltage (Section 2.2.2) formats, as well as discussion of the gas flow formulation (Section 2.2.3) and the general k-component PCNF problem (Section 2.2.4). Section 2.3 describes problems of the next level of complexity, including those involving optimization over resources. In particular, the general optimal PCNF problem is discussed in Section 2.3.1, and more specific cases of optimal flows involving the optimal power flow problem (in both power flow and current-voltage formulations) and the optimal gas flow problem are discussed in Sections 2.3.2, 2.3.3, and 2.2.3, respectively. Section 2.4 introduces a number of feasibility problems, all stated as special kinds of optimizations. Here we discuss the so-called instanton (Section 2.4.1), containment (Section 2.4.2), and state estimation (Section 2.4.3) formulations. The section concludes with a discussion in Section 2.5 of an example even more complex optimization involving split of resources between participants/aggregators.

In Section 3 we describe how any of the aforementioned PCNF and optimal PCNF problems can be restated in the universal GM format.

Then, in Section 4, we take advantage of the factorized form of the PCNF GM and illustrate how BP methodology can be used to solve the optimization problems exactly and/or approximately. Specifically, in Section 4.1 we restate the optimization (maximum likelihood) GM problem as linear programming (LP) in the space of beliefs (proxies for probabilities). The resulting LP is generally difficult because it involves working with all variables in a combination. We take advantage of the GM factorization and introduce in Section 4.2 the so-called LP-BP relaxation, providing a provable lower bound for the optimal. Finally, in Section 4.3 we construct a tractable relaxation of LP-BP based on an interval partitioning of the underlying space.

Section 5 discusses hierarchies that allow us to generalize, and thus improve, LP-BP. The so-called LP-BP hierarchies, related to earlier papers on the subject [29, 61, 66] are discussed in Section 5.1. Then, the relation between the LP-BP hierarchies and classic LP-based Sherali-Adams [60] and semidefinite programming-based Lasserre hierarchies [35, 36, 40, 53] is discussed in Section 5.2.

Section 6 discusses the special case of a GM defined over a tree (graph without loops). In this case LP-BP is exact, equivalent to the so-called dynamic programming (DP) approach, and as such it provides a distributed alternative to the global optimization through a sequence of graph-element-local optimizations. However, even in the tree case, the exact LP-BP and/or DP are not tractable for GM stated in terms of physical variables, such as flows, voltages, and/or pressures, drawn from a continuous set. Next, [20] we discuss how the problem can be resolved with a proper interval-partitioning (discretization).

We conclude the manuscript by presenting a summary and discussing a path forward in Section 7.

2 Problems of Interest: Formulations

In this section we formulate a number of PCNF problems that we will attempt to analyze and solve with the help of GM/BP approaches/techniques in the following sections.

2.1 Dissipation-Optimal Network Flow

We start by introducing/discussing network flows constrained by a minimum dissipation principle, i.e., one that can be expressed as an unconstrained optimization/minimization of an energy function (potential).

Consider a static flow of a commodity over an undirected graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, described through the following network flow equation:

$$i \in \mathcal{V} : q_i = \sum_{j:(i,j) \in \mathcal{E}} \phi_{ij}, \quad (1)$$

where q_i stands for injection ($q_i > 0$) or consumption ($q_i < 0$) of the flow at the node i and $\phi_{ij} = -\phi_{ji}$ stands for the value of the flow through the directed edge (i, j) in the direction from i to j .¹ We consider a balanced network, $\sum_{i \in \mathcal{V}} q_i = 0$.

We constrain the flow, requiring that the minimum dissipation principle is obeyed:

$$\min_{\phi} \sum_{\{i, j\} \in \mathcal{E}} E_{ij}(\phi_{ij}) \quad , \quad \text{Equation (1)} \quad (2)$$

where $\phi \doteq (\phi_{ij} = -\phi_{ji} | \{i, j\} \in \mathcal{E})$ and $E_{ij}(x)$ are local (energy) functions of their arguments for all $\{i, j\} \in \mathcal{E}$. The local edge terms in the energy function $E_{ij}(x)$, which we call local energy functions, are required to be convex at least over a restricted domain. We call the sum of local energy functions, $E(\phi) = \sum_{\{i, j\} \in \mathcal{E}} E_{ij}(\phi_{ij})$, the global energy function or simply the energy function. Versions of this problem appear in the context of the feasibility analysis of the dissipative network flows, which are flows whose redistribution over the network is constrained by potentials, e.g., voltages or pressures in the context of resistive electric networks and gas flow networks, respectively [19, 47, 65]. Note that the formulation (2) can also be supplemented by additional flow or potential constraints.

Imposing Karush-Kuhn-Tucker (KKT) stationary point conditions on the optimization problem stated in Equation (2) leads to the following set of equations:

$$\forall \{i, j\} \in \mathcal{E} : \quad E'_{ij}(\phi_{ij}) = \lambda_i - \lambda_j, \quad (3)$$

where λ_i is a Lagrangian multiplier corresponding to the i 's equation (1). The problem becomes fully defined by the pair of Equations (1, 3), which can also be restated solely in terms of the λ -variables:

$$i \in \mathcal{V} : \quad q_i = \sum_{j: \{i, j\} \in \mathcal{E}} \left(E'_{ij} \right)^{-1} (\lambda_i - \lambda_j). \quad (4)$$

2.2 General Physics-Constrained Network Flows

We call a network flow “unconstrained” if only the flow conservations, described by Equation (1), are enforced. Contrariwise, a network flow is “physics-constrained” if constrains relating line flows to physics potentials, such as voltages and pressures, are enforced in addition to the flow conservation constraints.

A particular example of the PCNF was discussed above in Section 2.1. However, this example is special, representing a line flow as a gradient of a scalar energy

¹In the following we will use the notation $\{i, j\}$ for the undirected edge and (i, j) for the respective directed edge. When the meaning is clear, we slightly abuse notations, denoting both the set of undirected and directed edges by \mathcal{E} .

function. Aiming to discuss the general case, where a PCNF problem cannot be stated as one following from minimization of a scalar energy function, we find it useful to start below with an example of the AC electric power flow and then proceed to discussing an abstract general case.

2.2.1 AC Network Flow: Power-Voltage Formulation

An AC power flow (PF) system of equations provides the main example of the PCNF. We thus start by discussing the AC flow, stated in terms of the set of algebraic equations over the graph \mathcal{G} :

$$\forall i \in \mathcal{V} : P_i = \sum_{j:(i,j) \in \mathcal{E}} \phi_{ij}, \quad (5)$$

$$\forall (i, j) \in \mathcal{E} : \phi_{ij} = V_i \left(\frac{V_i - V_j}{z_{ij}} \right)^* \quad (6)$$

where all the characteristics take values over complex numbers, V_i is the complex voltage potential, ϕ_{ij} is the complex power leaving node i in the direction to node j , P_i is the complex injection/consumption at the node i , $z_{ij} = z_{ji}$ is the complex impedance of the line $\{i, j\}$ (assumed known), and y^* stands for the complex conjugate of y . One formulation/problem of interest is as follows: given P_i at all nodes but one (e.g., $i = 0$, called slack bus) and fixing the voltage at the slack bus, e.g., $V_0 = 1$, find V_i at $\forall i \in \mathcal{V} \setminus 0$.

In general, Equations (5, 6) cannot be stated as KKT conditions imposed on a scalar function of voltages. However, such representation is possible in a number of special cases, e.g., when one ignores the resistance of lines (in comparison with the inductance) or when all the lines of the system are characterized by a constant inductance-to-resistance ratio.

Power systems may include transformers of different types, e.g., standard voltage transformers or phase transformers. These devices can be described as nodes of degree two. For instance, consider multiplicative transformation at a node:

$$\forall i \in \mathcal{V}_T \subset \mathcal{V} : V_{i,in} = \alpha_i V_{i,out} \quad (7)$$

characterized by a complex transformation coefficient, α_i . For phase transformers, $|\alpha_i| = 1$ if losses of actual and reactive power at the transformer are ignored. Other types of transformations, e.g., additive or generally nonlinear, can be easily incorporated in the model. Even though already installed transformers are typically not used for the real-time control in practical transmission systems, the newly installed solid-state transformers are capable of fast and efficient response and thus can actually be used for real-time (even second-scale) controls. High-voltage direct current (HVDC) links are new installations that can also be incorporated into the PF description. The HVDC can be modeled as a pair of points, or multiple points for

multiterminal HVDC, with a zero-net injected/withdrawn active and reactive power if we assume that the devices are lossless.

Finally, let us also mention that lines can be modeled in a more accurate way via the so-called π -model. We will not describe it here in detail but mention that this modeling fits naturally the general graph description of the systems because it requires introducing only two auxiliary nodes at the two ends of a line connected through capacitors to the ground.

2.2.2 AC Network Flow: Current-Voltage Formulation

The PF Equations (5, 6) can also be restated in terms of the linear Kirchoff law relations between currents and voltages:

$$\forall i \in \mathcal{V} : I_i = \sum_{j:(i,j) \in \mathcal{E}} J_{ij}, \quad (8)$$

$$\forall (i, j) \in \mathcal{E} : J_{ij} = \frac{V_i - V_j}{z_{ij}}, \quad (9)$$

where $P_i = V_i I_i^*$ and $\phi_{ij} = V_i J_{ij}^*$. When the focus is on resolving the PF problem—given nodal consumptions and productions of power P , one aims to find voltages V and PFs ϕ over lines—the nonlinear PF formulation due to Equations (5, 6) is primal. However, as argued below, the Kirchoff original formulation in terms of currents (8, 9) may offer some additional computational advantages for posing and solving optimal problems where the power production and consumption is an optimization variable that is not fixed to a predefined value.

2.2.3 Gas Flows

Balanced gas flows (GFs) satisfy the following set of algebraic equations:

$$\forall i \in \mathcal{V} : q_i = \sum_{j:(i,j) \in \mathcal{E}} \phi_{ij}, \quad (10)$$

$$\forall (i, j) \in \mathcal{E} : \phi_{ij} = \gamma_{ij} \frac{|\pi_i - \pi_j + b_{ij}|^{3/2}}{\pi_i - \pi_j + b_{ij}}, \quad (11)$$

$$\forall i \in \mathcal{V}_c \subset \mathcal{V} : \pi_{i;out} = \alpha_i \pi_{i;in}, \quad (12)$$

where $\pi_i \geq 0$ is the squared pressure at node i ; γ_{ij} is a constant characterizing the line or pipe $\{i, j\}$, which depends on the diameter of the pipe, the friction coefficient, the type of gas used, etc.; b_{ij} is a coefficient of an additive compression at the pipe $\{i, j\}$; α_i is a coefficient of a multiplicative compression at the compressor node $i \in \mathcal{V}_c$, which is normally a node of degree two; and $\pi_{i;in}$ and $\pi_{i;out}$ stand for

squared pressures at both sides of the node. Both types of compressors can be placed at a line but not simultaneously and depending on possible operational strategies. Like in the PF case, it is also convenient to assume the existence of a slack bus, which also reflects a practical situation. The slack bus is a special node, $i = 0$, where the pressure is maintained constant, providing a source for the global balance of the GF.

2.2.4 General Physics-Constrained Network Flows

A general k -component PCNF problem becomes

$$\forall k = 1, \dots, K, i \in \mathcal{V} : q_i^{(k)} = \sum_{j:(i,j) \in \mathcal{E}} \phi_{ij}^{(k)}, \quad (13)$$

$$\forall k = 1, \dots, K, (i, j) \in \mathcal{E} : \phi_{ij}^{(k)} = f_{ij}^{(k)}(\pi_i, \pi_j), \quad (14)$$

where $\pi_i \doteq (\pi_i^{(k)} | k = 1, \dots, K)$. Nodal transformers/compressors can be readily included into the model

$$i \in \mathcal{V}_t \subset \mathcal{V} : \pi_{i,out} = T_i(\pi_{i,in}), \quad (15)$$

where $T_i(\cdot)$ can be a general nonlinear transformation and $\pi_i = (\pi_i^{(k)} | k = 1, \dots, K)$.

2.3 Optimal Physics-Constrained Network Flow Problems

The optimal PCNF problem aims to find an optimum over a set of control/optimization parameters that enter the PCNF description.

2.3.1 General Case

In the most general case, one poses the following optimization problem:

$$\min_{q, \pi, \phi, \{T\}} \left(\sum_{i \in \mathcal{V}} C_i(q_i) + \sum_{i \in \mathcal{V}_t} C_i^{(t)}\{T_i\} \right) \Bigg| \begin{array}{ll} \text{Equations (13,14,15)} & \\ \pi_i \in \Pi_i & \forall i \in \mathcal{V} \\ \phi_{ij} \in \Psi_{ij} & \forall (i, j) \in \mathcal{E} \end{array} \quad (16)$$

where Π_i and Ψ_{ij} describe the domains of allowed values for node potentials and edge flows, respectively.

2.3.2 Optimal Power Flow: Power-Voltage Formulation

Standard optimal power flow (OPF) transmission system formulation is stated as follows (see, e.g., [5, 6] and references therein):

$$\min_{P, \Phi, V} \sum_{i \in \mathcal{V}} C_i(P_i) \quad \left| \begin{array}{l} \text{Equations (5,6)} \\ V_i \in U_i \quad \forall i \in \mathcal{V} \setminus 0 \\ \phi_{ij} \in \Psi_{ij} \quad \forall (i, j) \in \mathcal{E} \end{array} \right. \quad (17)$$

where $V_0 = 1$, $C_i(P_i)$ is the cost function that is potentially nonlinear and site dependent and U_i , Ψ_{ij} are domains of allowed values for site voltage and line flows, respectively. There are multiple other extensions (e.g., generalizations accounting for investment and planning of new devices, such as FACTS, HVDC, and transformer devices; see, e.g., [21–23]).

The OPF problem (17) gets simpler in the case of the distribution grid where the grid graph is a tree. Then, voltage is fixed at the head of the tree, $i = 0$, considered as a slack bus, while all other nodes of the system are modeled in the static setting as (p, q) nodes, where p_i is an accumulated consumption and photovoltaic (PV) generation at the node i and q_i is the reactive power consumed/produced at the node. PV power is injected to the grid through inverters, which have a capability to adjust reactive power. This degree of freedom can be used to achieve various objectives, e.g., to minimize (active) power losses in lines subject to voltages in order to stay within predefined safety limits. An example distribution grid for OPF is

$$\min_{q, V} \sum_{\{i, j\} \in \mathcal{E}} \frac{|V_i - V_j|^2}{r_{ij}^2 + x_{ij}^2} r_{ij}, \quad (18)$$

$$\text{s.t.} \quad p_i + iq_i = V_i \sum_{j: \{i, j\} \in \mathcal{E}} \left(\frac{V_i - V_j}{z_{ij}} \right)^*, \quad \forall i \in \mathcal{V} \setminus 0 \\ V_i \in U_i \quad \& \quad q_i \in \mathcal{Q}_i \quad \forall i \in \mathcal{V} \quad (19)$$

where $q \doteq (q_i | i \in \mathcal{V} \setminus 0)$, $V = (V_i \in \mathbb{C} | i \in \mathcal{V} \setminus 0)$ are variable vectors of reactive injections/consumptions and voltages and the vector of active injection/consumption \mathcal{Q}_i describes the allowed range of the nodal reactive power adjustment; $p = (p_i \in \mathbb{R} | i \in \mathcal{V} \setminus 0)$ is assumed fixed; and $U_0 = \{1\}$, i.e., voltage at the head of the line is constrained. Notice that given that the underlying graph is a tree, the PF equations can be rewritten in the so-called Baran-Wu representation [4], stated in terms of both active and reactive power flows flowing through the line segments and voltages at the nodes. Note that the Baran-Wu representation also applies to loopy networks; however, in the loopy case the related system of equations is incomplete, i.e., underdefined.

2.3.3 Optimal Power Flow: Current-Voltage Formulation

Assume that all nodes of the network have some kind of flexibility in terms of the injection/consumption, i.e., I_i is not fixed but is allowed to be drawn from a range Ξ_i that may be node-specific. Then one poses the following current-voltage version of the OPF formulation:

$$\left. \min_{I, J, V} \sum_{i \in \mathcal{V}} C_i(V_i I_i^*) \right| \begin{array}{l} \text{Equations (8,9)} \\ V_i \in U_i \text{ & } I_i \in \Xi_i \quad \forall i \in \mathcal{V} \setminus 0 \\ (V_i - V_j) J_{ij}^* \in \Psi_{ij} \quad \forall (i, j) \in \mathcal{E} \end{array} \quad (20)$$

2.3.4 Optimal Gas Flow

A rather general version of the optimum GF problem is

$$\left. \min_{q, p, \alpha} \left(\sum_{i \in \mathcal{V}} C_i(q_i) + \sum_{i \in \mathcal{V}_\alpha} C_i(\alpha_i) \right) \right| \begin{array}{l} \text{Equations (10,11,12)} \\ \pi_i \in \Pi_i \quad \forall i \in \mathcal{V} \\ \phi_{ij} \in \Psi_{ij} \quad \forall (i, j) \in \mathcal{E} \end{array} \quad (21)$$

where \mathcal{V}_α is the set of the multiplicative compressor nodes, $\alpha = (\alpha_i | i \in \mathcal{V}_\alpha)$ is the vector of compression, and the two contributions to the objective balance deviation of the consumption/injection of gas from the nominal value across the system with the cost of compression. See [2, 51, 52, 65, 72–74] for additional details.

2.4 Feasibility as an Optimal Physics-Constrained Network Flow Problem

The problems discussed below can all be understood as network feasibility problems focusing on describing or characterizing domains of the network operation feasibility. Constructing good algorithms for efficient and accurate solutions of these problems would allow us to monitor the state of the system not as one particular configuration but as a succinct characterization of the domains with good or bad properties. Thus the problem can also be described as guiding, building, or focusing on “extended state evaluations or characterizations.”

2.4.1 Instanton as an Optimal PCNF Problem

An instanton is a special network flow state, $(\phi, \pi)_{inst}$, that is defined as the most probable failure state. Consider, for example, stochastic injections/consumptions, q , drawn from an exogenously known probability distribution, $\mathcal{P}(q)$. The probability is viewed as a distance measure, $D(q; q_0) = \log(\mathcal{P}(q_0)/\mathcal{P}(q))$, from the most probable configuration of the injection/consumption, q_0 . In many practical cases, $D(q; q_0)$ shows nice properties; for example, $D(q; q_0)$ may be a convex function of q . A state, (ϕ, π) , is considered faulty if it is on the boundary, $(\phi, \pi) \in B_{safe}$, of the domain of the safe operation. Therefore the instanton problem, in the case of a general PCNF flow, is a solution of the following optimization problem:

$$\min_{(\phi, \pi)} D(q; q_0) \Big| \begin{array}{l} \text{Equations (13,14)} \\ (\phi, \pi) \in B_{safe} \end{array} \quad (22)$$

Description of the boundary domain, B_{safe} , will depend on what is considered “safe.” Two examples of interest are boundaries of (a) union of the box constraints on line flows, and (b) the PCNF feasibility, i.e., the domain where the determinant of the respective Jacobian is zero. Considering the boundary of the intersection of the two example domains is also of interest. See [11, 12, 30] for additional details.

2.4.2 Containment as an Optimal PCNF Problem

Suppose we identify “desirable properties” in a space of operational parameters, (ϕ, π) , such as voltages, pressures, power flows, etc. The special features of the “desirable” domain, \mathcal{D}_{des} , may allow simpler characterization of the domain. The examples are convexity of an underlying energy function, monotonicity of an underlying operator, piece-wise monotonicity in the response of the system, or simply existence of a solution. Description of \mathcal{D}_{des} may be algebraically nontrivial, e.g., stated as a non-negativity of a matrix, positivity of the largest eigenvalue of a matrix, or positivity of all components of a matrix. On the other hand, we may have an alternative description of a “safety” domain, \mathcal{D}_{safe} , in a space of operational parameters. For example, we may want flows over lines not to exceed respective thresholds, voltages, or pressures to be within bounds, etc. Description of both the “desirable property” and “safety” domains may allow some additional degrees of freedom that will change the domain shape, e.g., make the domain larger or smaller, fit a certain shape within the domain, etc. For example, one may consider a “safety” domain dependent on a rescaling volume factor, V : $\mathcal{D}_{safe}(V)$. The containment problem becomes optimizing the additional degrees of freedom in the description of both the “desirable” domain and/or the “safety” domain, e.g., V , so that the latter is contained within the former. Formally, the containment problem is stated as the following optimization problem:

$$\min_{(\phi, \pi); V} V \left| \begin{array}{l} \text{Equations (13,14)} \\ \mathcal{D}_{\text{safe}}(V) \subseteq \mathcal{D}_{\text{des}} \end{array} \right. \quad (23)$$

See [15–18] for additional details.

2.4.3 State Estimation as an Optimal PCNF Problem

Here we discuss a data-driven state estimation (SE) problem: given deterministic or probabilistic measurements, describe a state or domain of states that is most consistent with the data. For example, consider the observational data, e.g., measured by PMU in the case of power systems, to be a subset of line flows, $\phi_d = (\phi_{ij;d}^{(k)} | (i, j) \in \mathcal{E}_d \subseteq \mathcal{E}; \forall k = 1, \dots, K)$, and potentials, $\pi_d(\pi_{i;d}^{(k)} | i \in \mathcal{V}_d \subseteq \mathcal{V}; \forall k = 1, \dots, K)$ measured at \mathcal{V}_d and \mathcal{E}_d , respectively. Then an example data-most-consistent SE can be found by solving the following optimization problem:

$$\min_{q; (\phi, \pi)} \sum_{k=1, \dots, K; i \in \mathcal{V}} \| q_i^{(k)} - \sum_{j:(i,j) \in \mathcal{E}} \phi_{ij}^{(k)} \| \quad (24)$$

$$\begin{aligned} & \text{s.t. } \forall i \in \mathcal{V}_d, \forall k = 1, \dots, K : \pi_i^{(k)} = \pi_{i;d}^{(k)} \\ & \quad \forall (i, j) \in \mathcal{E}_d, \forall k = 1, \dots, K : \phi_{ij}^{(k)} = \phi_{ij;d}^{(k)} \end{aligned} \quad (25)$$

Equations (14)

2.5 Optimal Physics-Constrained Network Flows with Resources Split Between Aggregators

In some cases, energy resources and energy consumers are flexible and can be redistributed between a group of nodes. For example, an electric vehicle (EV) aggregator may split its EV fleet into two or more groups to be charged at distinct locations. Similarly, mobile battery resources can be redistributed by a battery aggregator between two or more nodes. These types of dependencies can be modeled by introducing additional pair-wise or high-order constraints on the nodal injection/consumptions.

For example, consider the following generalization of the distribution system OPF (18, 19), allowing for resources to be split between a number of aggregators:

$$\min_{q, V, p_c} \sum_{\{i, j\} \in \mathcal{E}} \frac{|V_i - V_j|^2}{r_{ij}^2 + x_{ij}^2} r_{ij}, \quad (26)$$

$$\begin{aligned}
p_i + iq_i &= V_i \sum_{j:\{i,j\} \in \mathcal{E}} \left(\frac{V_i - V_j}{z_{ij}} \right)^*, \quad \forall i \in \mathcal{V} \setminus 0 \\
\text{s.t.} \quad V_i &\in U_i \quad \& q_i \in \mathcal{Q}_i \quad \forall i \in \mathcal{V} \\
\underline{p}_\alpha &\leq |p_i + p_j| \leq \bar{p}_\alpha \quad \forall i, j \sim \alpha \in \mathcal{A}
\end{aligned} \tag{27}$$

where \mathcal{A} stands for the list of the pair-wise aggregators and $i, j \sim \alpha$ indicates that the two distinct nodes i and j are under control of the same aggregator. Generalization to aggregators controlling more than two nodes is straightforward.

3 Graphical Model for a Physics-Constrained Optimal Network Flow

In the case of general optimal PCNF (16), the state/optimization vector, $s \doteq (\pi, \phi, q)$, or simply state, is represented by the vector of potentials, $\pi \doteq (\pi_{ij}|(i, j) \in \mathcal{E})$; the vector of line flows, $\phi \doteq (\phi_{ij}|(i, j) \in \mathcal{E})$, where components are associated with the directed edges and are thus assumed to be computed at the starting node of the edge; and the injection-consumption vector, $q = (q_i|i \in \mathcal{V})$.

Consider the following probabilistic version of the optimization problem (16) where the state s is realized with a probability factorized according to the following distribution function:

$$\mathcal{P}(s) \sim \exp \left(-\beta \sum_{i \in \mathcal{V} \setminus 0} C_i(q_i) \right) \prod_{i \in \mathcal{V}} F_i(q_i; \pi_{\sim i}; \phi_{\sim i}) F_{ij}(\pi_{ij}, \phi_{ij}; \pi_{ji}, \phi_{ji}), \tag{28}$$

$$\forall i \in \mathcal{V} \setminus 0 : \quad F_i(q_i; \pi_{\sim i}; \phi_{\sim i}) \doteq \begin{cases} 1, & (q_i; \pi_{\sim i}; \phi_{\sim i}) \in \Upsilon_i \\ 0, & (q_i; \pi_{\sim i}; \phi_{\sim i}) \notin \Upsilon_i \end{cases} \tag{29}$$

$$\Upsilon_i \doteq (\pi_{ik} = \pi_{ij} = \pi_{il}, \& q_i = \phi_{ij} + \phi_{ik} + \phi_{il}), \tag{30}$$

$$\forall (i, j) \in \mathcal{E} : \quad F_{ij}(\pi_{ij}; \phi_{ij}; \pi_{ji}) \doteq \begin{cases} 1, & (\pi_{ij}; \phi_{ij}; \pi_{ji}) \in \Upsilon_{ij} \\ 0, & (\pi_{ij}; \phi_{ij}; \pi_{ji}) \notin \Upsilon_{ij} \end{cases} \tag{31}$$

$$\Upsilon_{ij} \doteq (\phi_{ij} = f_{ij}(\pi_{ij}, \pi_{ji})), \tag{32}$$

where $\beta > 0$ is an auxiliary parameter sometimes called inverse effective temperature; $\forall i \in \mathcal{V} : \pi_{\sim i} \doteq (\pi_{ij}|(i, j) \in \mathcal{E})$ and $\phi_{\sim i} \doteq (\phi_{ij}|(i, j) \in \mathcal{E})$ are vectors of potentials and flows associated with a vertex; $C_i(q_i)$ is the cost dependent on the consumption/injection, q_i , at the node i ; and $\delta(x)$ is the characteristic function of the logical expression x : $\delta(x)$ is unity if x is true, and it is zero otherwise. Let us assume that all the flow variables, i.e., all components of ϕ and q vectors, are drawn

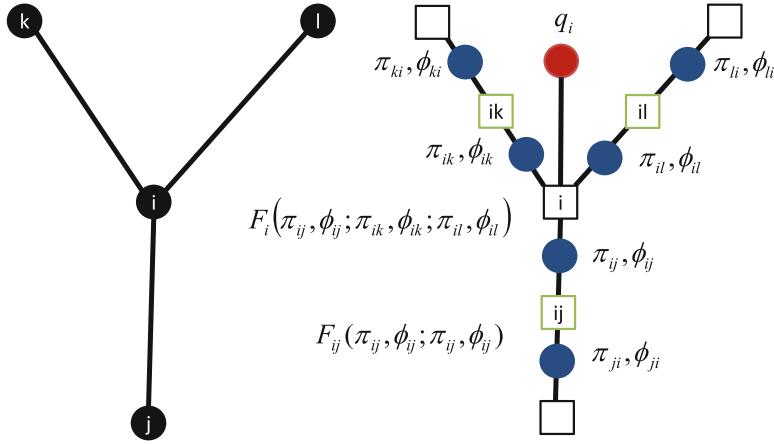


Fig. 1 Illustration of an element of the GM (28), shown on the right, and construction from the respective element of the base physical network graph, shown on the left. Variable nodes of the GM are shown as the circles/nodes. Check/function nodes are shown as squares. Duplicated potentials, e.g., π_{ij} , and flows, e.g., ϕ_{ij} , are associated with the blue circles/nodes, and injections/consumptions, e.g., q_i , are associated with the red circles/nodes. Functions associated with the black and green checks/squares implement duplication and flow conservation, e.g., $F_i(q_i; \pi_{\sim i}; \phi_{\sim i})$, and dissipative relation for the flow drop over line as a function of potentials at the two ends of the line, e.g., $F_{ij}(\pi_{ij}; \phi_{ij}; \pi_{ji})$, defined in Equations (29, 31), respectively.

from a finite alphabet, $\forall i \in \mathcal{V} : q_i \in \Theta$ and $\forall (i, j) \in \mathcal{E} : \phi_{ij} \in \Theta$. Let us also assume that the components of π take values in a finite set, $\forall (i, j) \in \mathcal{E} : \pi_{ij} \in \tilde{\Pi}$, and denote the resulting finite set for s by Σ . The probability of the state $s \in \Sigma$ given by Equation (28) can be understood as representing a GM constructed based on the physical network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the construction is illustrated in Figure 1.

Obviously the maximum likelihood configuration associated with $\mathcal{P}(s)$ from Equation (28) corresponds to the solution of the optimal flow problem (16), which can also be restated in GM terms as follows:

$$E \doteq \min_{s \in \Sigma} \sum_{i \in \mathcal{V} \setminus 0} C_i(q_i) \quad (33)$$

$$\begin{aligned} \text{s.t.} \quad & (q_i; \pi_{\sim i}; \phi_{\sim i}) \in \Upsilon_i \quad \forall i \in \mathcal{V} \setminus 0 \\ & (\pi_{ij}; \phi_{ij}; \pi_{ji}) \in \Upsilon_{ij} \quad \forall (i, j) \in \mathcal{E}. \end{aligned} \quad (34)$$

This PCNF optimization problem is a special case of a general GM optimization problem:

$$\text{OPT: } \min_{x \in \Sigma} \sum_{\alpha \in \bar{\mathcal{V}}_f} f_\alpha(x_\alpha) \quad (35)$$

$$\Sigma \doteq \left(\prod_{i \in \bar{\mathcal{V}}_v} \Sigma_i \right) \cap \left(\prod_{\beta \in \bar{\mathcal{V}}_c} \Sigma_\beta \right) \quad (36)$$

defined over the bipartite graph, $\bar{\mathcal{G}} \doteq (\bar{\mathcal{V}}_v, (\bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c), \bar{\mathcal{E}})$, where $\bar{\mathcal{V}}_v$, $\bar{\mathcal{V}}_f$, $\bar{\mathcal{V}}_c$, and $\bar{\mathcal{E}}$ are the sets of variable nodes, factor-function nodes, constrain-expressing nodes, and edges connecting variable nodes, factor-function nodes, and constrain-expressing nodes to each other. Here in Equation (36), the variable $x \doteq (x_i | i \in \bar{\mathcal{V}}_v)$ is a vector with components, x_i , labeled by i —a variable node from $\bar{\mathcal{V}}_v$ —taking values from the set Σ_i , which can be discrete or continuous, e.g., taken values over reals. The function $f_\alpha(x_\alpha)$ in Equation (36), associated with a factor $\alpha \in \bar{\mathcal{V}}_f$, is a function of $x_\alpha \doteq (x_i | i \sim \alpha)$ —a vector constructed from variable nodes connected to the factor α through an edge; thus, $i \in \alpha$ is a shortcut for $\forall i \in \bar{\mathcal{V}}_v$ s.t. $(i, \alpha) \in \bar{\mathcal{E}}$. We assume that $\forall \alpha \in \bar{\mathcal{V}}_f$ factor function $f_\alpha : \Sigma_\alpha \rightarrow R^+$ maps from $\Sigma_\alpha \doteq \cup_{i \sim \alpha} \Sigma_i$ to the set of nonnegative finite reals. Σ_β in Equation (36), associated with a factor $\beta \in \bar{\mathcal{V}}_c$, is a set of $x_\beta \doteq (x_i | i \sim \beta)$.

In the next section we will describe an LP-BP approach to solving Equation (36), which will obviously apply to the PCNF problem as well, provided (a) transformation from the network graph \mathcal{G} to the auxiliary graph $\bar{\mathcal{G}}$ is done according to Figure 1 and explanations above; (b) x variable in the general formulation (36) is built by combining π, ϕ , and q variables; and (c) the constraints (30, 32) are embedded in the description of the Σ_β constraints.

4 From Nonlinear Programming to Linear Programming: Belief Propagation (LP-BP)

In this section we utilize the GM reformulation of the PCNF problems and discuss transformation from Equation (36) to the so-called LP-BP. The transformation is done in three steps. First, in Section 4.1, we restate Equation (36) as an LP in the space of beliefs (proxies for probabilities). Second, in Section 4.2, we introduce the LP-BP relaxation. Finally, to get a tractable relaxation of LP-BP and thus of the original NP (36), we introduce in Section 4.3 a part-LP-BP scheme based on an interval partitioning of the underlying space.

4.1 Exact Reformulation of the Nonlinear Programming as a Linear Programming in the Space of Beliefs

The optimization problem (36) also allows reformulation as the exact linear programming (ELP):

$$\text{ELP : } \min_{b(x) \in \mathcal{B}} \int_{\Sigma} dx b(x) \sum_{\alpha \in \bar{\mathcal{V}}_f} f_{\alpha}(x_{\alpha}) \quad (37)$$

where $\int_{\Sigma} dx \dots$ stands for integration (or summation when Σ is discrete) in x over Σ and \mathcal{B} is the following exact set (ES):

$$\mathcal{B} \doteq \left(\{b(x)\} \left| \begin{array}{l} 0 \leq b(x) \leq 1, \forall x \in \Sigma \\ \int_{\Sigma} dx b(x) = 1. \end{array} \right. \right) \quad (38)$$

In general, the belief set \mathcal{B} is not tractable because the number of variables and the number of the set-defining constraints are both infinite when Σ contains a continuous subset and the numbers (of variables and constraints) are exponential in the dimension even when all the Σ_i are discrete. This suggests construction of various relaxations of the ELP through constraint generation methods such as the cutting plane procedure, consisting simply of keeping only a finite subset of constraints from \mathcal{B} thus expressed through a subset of beliefs or moments. By construction any of the relaxations shows the following two key features:

- Optimum value of the relaxed optimization provides a low bound on the exact value of Equation (36) (or, equivalently, of Equation (37)). (We can also construct an upper bound presenting a feasible solution.)
- If the optimum solution (argument) of the relaxed optimization Equation (37) is integer (all beliefs are 0 or 1), then this is also an optimal solution of OPT=ELP. In this (lucky) case we say that the gap is closed.

In the following section we will discuss a particular relaxation strategy, called LP-BP, and then (very briefly) comment on the possibility of adaptively and inhomogeneously constructing over the graph a hierarchy of the Sherali-Adams type starting from LP-BP and proceeding with some extra (and more complex) beliefs added. We intend to make construction of the hierarchy adaptive so that the choice of the higher-order constraints to add to the set of active constraints (included in the optimization) depends on the result/output of the preceding step.

4.2 LP-BP Relaxation

ES (38) can be restated as

$$\mathcal{B} \doteq \left(\begin{array}{l} \{b(x), \\ b_\alpha(x_\alpha), \\ b_i(x_i)\} \end{array} \middle| \begin{array}{ll} 0 \leq b(x) \leq 1, & \forall x \in \Sigma \\ b_\alpha(x_\alpha) = \int_{\Sigma \setminus \Sigma_\alpha} d(x \setminus x_\alpha) b(x), & \forall \alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c, \forall x_\alpha \in \Sigma_\alpha \\ b_i(x_i) = \int_{\Sigma_\alpha \setminus \Sigma_i} d(x_\alpha \setminus x_i) b_\alpha(x), \forall i \in \bar{\mathcal{V}}_v, \forall \alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c \text{ s.t. } \alpha \sim i, \forall x_i \in \Sigma_i \\ \int_{\Sigma_i} dx_i b(x_i) = 1 & \forall i \in \bar{\mathcal{V}}_v \end{array} \right) \quad (39)$$

where we simply added/defined beliefs associated with node and factor variables. LP-BP relaxation of the ES, also called (graph) local consistency relaxation in [68], consists of replacing the first two lines in Equation (39) with the range inequalities for beliefs associated with the factor variables:

$$\mathcal{B}_{LP-BP} \doteq \left(\begin{array}{l} \{b_\alpha(x_\alpha), \\ b_i(x_i)\} \end{array} \middle| \begin{array}{ll} 0 \leq b_\alpha(x_\alpha) \leq 1, & \forall \alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c, \forall x_\alpha \in \Sigma_\alpha \\ b_i(x_i) = \int_{\Sigma_\alpha \setminus \Sigma_i} d(x_\alpha \setminus x_i) b_\alpha(x), \forall i \in \bar{\mathcal{V}}_v, \forall \alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c \text{ s.t. } \alpha \sim i, \forall x_i \in \Sigma_i \\ \int_{\Sigma_i} dx_i b(x_i) = 1, & \forall i \in \bar{\mathcal{V}}_v \end{array} \right) \quad (40)$$

Then the relaxed version of the ELP is

$$\text{LP-BP : } \min_{\{b_i, b_\alpha\} \in \mathcal{B}_{LP-BP}} \sum_{\alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c} \int_{\Sigma_\alpha} dx_\alpha b_\alpha(x_\alpha) f_\alpha(x_\alpha). \quad (41)$$

Because LP-BP is relaxation of the ELP, one generally observes a gap between the two:

$$\text{LP-BP} \leq \text{ELP}. \quad (42)$$

Three remarks are in order.

- We call the aforementioned LP relaxation (in the space of probabilities/beliefs) of the optimization problem (36) LP-BP, following the terminology and tradition of the GM and BP community (see, e.g., [67] and references therein). However, exactly the same object was discussed even earlier in the combinatorial optimization community (see [77] and references therein). According to the complementary terminology, Equation (36) describes the valued constrained satisfaction problem, and Equation (41) is called “basic LP relaxation.”
- Even though the set (39) is convex, the LP-BP optimization (41) is still not tractable (in the case of continuous alphabet) because description of the \mathcal{B}_{LP-BP} set includes infinitely many constraints.

- Suboptimality of LP-BP is related to the fact that it ignores global constraints between beliefs by accounting only for explicit relations between factor/constraint beliefs and nodal beliefs. In other words, LP-BP allows us to optimize over only local beliefs.

4.3 Tractable, Interval-Partitioned Relaxation of LP-BP

The semi-infinite nature, and thus intractability, of LP-BP in the case of interest when components of x are continuous (or mixed) calls for developing tractable approximations of LP-BP. Specifically, given that LP-BP is a relaxation, i.e., an outer approximation (lower bound) of the original NP itself, we are interested in finding a tractable lower bound to LP-BP, so that it will also be a lower bound to the NP.

We suggest an approach that consists of partitioning each Σ_i , corresponding to an elementary continuous variable, into a finite number of intervals. Assume that such a partitioning $\Sigma_i = \cup_{a_i \in \mathcal{A}_i} \Sigma_{i;a_i}$, where \mathcal{A}_i is a set of labels for non-overlapping intervals, is given. (Thus leaving discussion of an optimal partitioning for Section 4.3.) Then, one naturally defines a finite set of marginal beliefs associated with each interval of each elementary variable:

$$\forall i \in \bar{\mathcal{V}}_v, \forall a_i \in \mathcal{A}_i : b_{i;a_i} \doteq \int_{\Sigma_{i;a_i}} dx_i b_i(x_i). \quad (43)$$

By construction, $b_{i;a}$ are all properly normalized:

$$\forall i \in \bar{\mathcal{V}}_v : \sum_{a \in \mathcal{A}_i} b_{i;a} = 1. \quad (44)$$

Respective, and also properly normalized, finite-dimensional factor and constraint beliefs are defined according to

$$\forall \alpha \in \bar{\mathcal{V}}_f, \forall a_\alpha = (a_i | i \sim \alpha) : b_{\alpha;a_\alpha} \doteq \int_{\prod_{i \sim \alpha} \Sigma_{i;a_i}} dx_\alpha b_\alpha(x_\alpha), \quad (45)$$

$$\forall \beta \in \bar{\mathcal{V}}_c, \forall a_\beta = (a_i | i \sim \beta) : b_{\beta;a_\beta} \doteq \int_{(\prod_{i \sim \beta} \Sigma_{i;a_i}) \cap \Sigma_\beta} dx_\beta b_\beta(x_\beta). \quad (46)$$

$$\forall \alpha \in (\bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c) : \sum_{a_\alpha} b_{\alpha;a_\alpha} = 1. \quad (47)$$

Marginalization relation between the interval-partitioned node and factor or constraint beliefs are also straightforward:

$$\forall i \in \mathcal{V}_v, \forall a_i \in \mathcal{A}_i, \forall \alpha \in (\bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c), \text{ s.t. } \alpha \sim i : b_{i;a_i} = \sum_{a_\alpha \setminus a_i} b_{\alpha;a_\alpha}. \quad (48)$$

Then we form the following interval-partitioned finite (thus tractable) belief polytope:

$$\mathcal{B}_{Int-Part-LP-BP} \doteq \left(\{b_{i;a_i}, b_{\alpha;a_\alpha}\} \mid \text{Equations (48, 44)} \right), \quad (49)$$

which is, by construction, a relaxation (outer approximation) of the LP-BP polytope (40).

Next we introduce piecewise constant lower bound approximations for the factor functions, $f_\alpha(x_\alpha)$:

$$\forall \alpha \in \bar{\mathcal{V}}_f, \forall a_\alpha, \forall x_\alpha \in \Sigma_\alpha : f_{\alpha;a_\alpha} \leq f_\alpha(x_\alpha). \quad (50)$$

Combining Equation (49) with Equation (50), one constructs the following tractable (finite-dimensional) LP:

$$\text{Int-Part-LP-BP : } \min_{\{b_{i;a_i}, b_{\alpha;a_\alpha}\} \in \mathcal{B}_{Int-Part-LP-BP}} \sum_{\alpha \in \bar{\mathcal{V}}_f, a_\alpha} b_{\alpha;a_\alpha} f_{\alpha;a_\alpha}. \quad (51)$$

which is probably an interval-partitioned relaxation of LP-BP and thus of ELP and OPT, i.e.:

$$\text{Int-Part-LP-BP} \leq \text{LP-BP} \leq \text{ELP} = \text{OPT}. \quad (52)$$

5 Generalization of the LP-BP Relaxation and Associated Hierarchies

We saw in Section 4.2 that the ELP in Equation (38) can be relaxed into a simpler LP using the LP-BP relaxation from Equation (42). The LP-BP relaxation can be generalized and performed in a systematic way, leading asymptotically to the exact result. This generalization results in a relaxation hierarchy of increasing tightness but also of increasing computational complexity.

5.1 LP-BP Hierarchy

The key idea behind the LP-BP hierarchy is to relax Equation (38) with a set of consistent beliefs involving a group of variables of increasing size around more than one factor or constraint node. The LP-BP hierarchy is not unique because there are multiple ways of grouping variable nodes into “super-nodes.” A set of super-nodes $\bar{\mathcal{V}}_S$ is a collection of subsets of variable nodes:

$$\bar{\mathcal{V}}_S \subset \{\gamma \in \mathcal{P}(\bar{\mathcal{V}}_i)\}, \quad (53)$$

where $\mathcal{P}(\cdot)$ denotes the power set of an ensemble. To be an admissible set of super-nodes, $\bar{\mathcal{V}}_S$ should satisfy two conditions. First, any subset of a super-node should also be a considered as a super-node:

$$\forall \gamma \in \bar{\mathcal{V}}_S, \quad \beta \subset \gamma \Rightarrow \beta \in \bar{\mathcal{V}}_S. \quad (54)$$

Second, sets of variable nodes neighboring a factor node or a constrain node are super-nodes:

$$\forall \alpha \in (\bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c), \quad \{i \in \mathcal{V}_i \mid i \sim \alpha\} \in \bar{\mathcal{V}}_S. \quad (55)$$

The generalized LP-BP relaxation of the constraints in Equation (39) based on the set of “super-nodes” $\bar{\mathcal{V}}_S$ reads as follows:

$$\mathcal{B}_{LP-BP}(\bar{\mathcal{V}}_S) \doteq \left(b_\gamma(x_\gamma) \geq 0 \begin{cases} b_\beta(x_\beta) = \int_{\Sigma_\gamma \setminus \Sigma_\beta} d(x_\gamma \setminus x_\beta) b_\gamma(x_\gamma), & \forall \gamma, \beta \in \bar{\mathcal{V}}_S \text{ s.t. } \beta \subset \gamma, \\ \int_{\Sigma_\gamma} dx_\gamma b_\gamma(x_\gamma) = 1, & \forall \gamma \in \bar{\mathcal{V}}_S \end{cases} \right). \quad (56)$$

The union of power sets of variable nodes around factor or constrain nodes is the minimal set of super-nodes:

$$\bar{\mathcal{V}}_{S_{\min}} = \bigcup_{\alpha \in (\bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c)} \mathcal{P}(\{i \in \mathcal{V}_i \mid i \sim \alpha\}), \quad (57)$$

and all possible combinations of variable nodes is the maximal set of super-nodes:

$$\bar{\mathcal{V}}_{S_{\max}} = \mathcal{P}(\mathcal{V}_i). \quad (58)$$

An LP-BP relaxation hierarchy consists of applying the generalized LP-BP relaxation (56) to an increasing collection of super-nodes:

$$\bar{\mathcal{V}}_{S_{\min}} \subset \bar{\mathcal{V}}_{S_1} \subset \bar{\mathcal{V}}_{S_2} \subset \cdots \subset \bar{\mathcal{V}}_{S_{\max}}, \quad (59)$$

which result in LP-BP relaxations of increasing tightness. Note that the number of variables and constraints associated with an LP-BP relaxation is exponential in the

size of the biggest super-node. The challenge in constructing an LP-BP hierarchy is to build small super-node sets that still provide an effective tightening.

The lowest level of LP-BP hierarchies is in general not equal to the LP-BP relaxation introduced in Equation (40) and is always tighter:

$$\mathcal{B}_{LP-BP}(\bar{\mathcal{V}}_{S_{min}}) \subset \mathcal{B}_{LP-BP}. \quad (60)$$

However, if every pair of factor or constrain nodes has at most one variable node as a common neighbor, then the two relaxations are equal. The technical reason behind this discrepancy comes from condition (54), which is needed for generalizing LP-BP to an arbitrary set of variables. The superset of variable nodes that is used to derive the plain LP-BP in Equation (40) contains only sets of variable nodes neighboring factor or constrain nodes and a singleton of one variable node.

Note that the highest level of LP-BP hierarchies is simply the exact set from Equation (38) because it considers beliefs over all variables:

$$\mathcal{B}_{LP-BP}(\bar{\mathcal{V}}_{S_{max}}) = \mathcal{B}. \quad (61)$$

5.2 Relationship to Other Relaxation Hierarchies

The LP-BP relaxation in Equation (56) can be formulated for any set of super-nodes $\bar{\mathcal{V}}_{S_t}$. In particular, the set of super-nodes can be oblivious to any GM structure contained in the problem. Although this is in general not a desirable property, it makes it possible to establish a relationship between the LP-BP relaxation hierarchy and other known hierarchies.

Consider the set of super-nodes consisting of all subsets of variable nodes of size at most $t > 0$:

$$\bar{\mathcal{V}}_{S_t} = \{\gamma \in \mathcal{P}(\bar{\mathcal{V}}_i) \mid t \geq |\gamma|\}. \quad (62)$$

The sets (62) do not take advantage of the graph structure but remain valid as super-node sets. The corresponding LP-BP relaxations $\mathcal{B}_{LP-BP}(\bar{\mathcal{V}}_{S_t})$ form a relaxation hierarchy for increasing t . This hierarchy is exact for levels $t \geq 1 + \omega(G)$ where $\omega(G)$ is the tree-width of the factor graph, potentially leading to a much smaller relaxation than (61). (See related recent discussion of the interval partitioning and tree-width-based solution of the OPF problem in [7].) When variables are binary, this particular LP-BP hierarchy becomes equivalent to the Sherali-Adams hierarchy [60]. However, for variables with discrete alphabet, binary included, this LP-BP hierarchy is not comparable to the Lasserre moments hierarchy [35] based on semidefinite matrices. Note that it can be shown that for any given level of the LP-BP hierarchy $\mathcal{B}_{LP-BP}(\bar{\mathcal{V}}_{S_t})$, there exists a level for which the Lasserre moments hierarchy is tighter. For more information on the relationship between LP-BP and other hierarchies, we refer the reader to [67].

6 Exactness in Trees and Distributed Message Passing

In the special case when the graph $\bar{\mathcal{G}}$ is a tree, it is well known that the LP-BP relaxation to ELP is tight (see [67] and references therein). However, when Σ contains a continuous subset, we still need to discretize the continuous domains as in Section 4.3 to obtain a tractable lower bound given by the Int-Part-LP-BP. Now, the only inexactness, and hence the lower bound, arises from the error due to discretization:

$$\text{Int-Part-LP-BP} \leq \text{LP-BP} = \text{ELP} = \text{OPT}. \quad (63)$$

The tree structure can also be exploited to design a DP-based algorithm to solve the Int-Part-LP-BP. The resulting algorithm has a complexity of $O(n)$. Following [20], we present here an implementation of the DP that involves a single forward and backward sweep over the tree and can be written in the form of the following message-passing algorithm.

Let $p(j)$ denote the parent of a node j and $\mathcal{C}(i)$ denote the set of children of a node i . Let \mathcal{L} denote the set of leaves.

Forward Pass:

Initialization

$$\forall i \in \bar{\mathcal{V}}_v \cap \mathcal{L}, \alpha = p(i), \forall a_i \in \mathcal{A}_i, \kappa_{i \rightarrow \alpha}(a_i) \leftarrow 0, \quad (64)$$

$$\forall \alpha \in \{\bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c\} \cap \mathcal{L}, i = p(\alpha), \forall a_i \in \mathcal{A}_i, \gamma_{\alpha \rightarrow i}(a_i) \leftarrow f_{\alpha; a_{\alpha}}, \quad (65)$$

$$S_{processed} \leftarrow \mathcal{L}. \quad (66)$$

Forward Traverse

$$\text{Repeat until } S_{processed} = \bar{\mathcal{V}}, \quad (67)$$

$$\text{choose } v \notin S_{processed} \text{ s.t. } \mathcal{C}(v) \subseteq S_{processed}, \quad (68)$$

$$\text{if } v = i \in \bar{\mathcal{V}}_v, \alpha = p(i) : \forall a_i \in \mathcal{A}_i, \quad (69)$$

$$\kappa_{i \rightarrow \alpha}(a_i) \leftarrow \sum_{\tilde{\alpha} \in \mathcal{C}(i)} \gamma_{\tilde{\alpha}}(a_i), \quad (70)$$

$$\text{else if } v = \alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c, i = p(\alpha) : \forall a_i \in \mathcal{A}_i, \quad (71)$$

$$\begin{aligned} \gamma_{\alpha \rightarrow i}(a_i) &\leftarrow \min_{a_{\alpha} \in \mathcal{A}_{\alpha} \setminus a_i} \sum_{j \in \mathcal{C}(i)} \kappa_j(a_{\alpha}(j)) \\ &\quad + f_{\alpha}(a_{\alpha}) \end{aligned} \quad (72)$$

Backward Pass:

Initialization

$$r \in \bar{\mathcal{V}}_v = Root, \quad a_r^* = \operatorname{argmin}_{a_r \in \mathcal{A}_r} \sum_{\bar{\alpha} \in \mathcal{C}_r} \gamma_{\bar{\alpha}}(a_r), \quad (73)$$

$$S_{assigned} \leftarrow \{r\}. \quad (74)$$

Backward Traverse

$$\text{Repeat until } S_{assigned} = \bar{\mathcal{V}}, \quad (75)$$

$$\text{choose } v \notin S_{assigned} \text{ s.t. } p(v) \subseteq S_{assigned}, \quad (76)$$

$$\text{if } v = i \in \bar{\mathcal{V}}_v : \quad \text{continue,} \quad (77)$$

$$\text{else if } v = \alpha \in \bar{\mathcal{V}}_f \cup \bar{\mathcal{V}}_c, \quad i = p(\alpha) : \quad a_{\alpha}^* \leftarrow \operatorname{argmin}_{a_{\alpha} \in \mathcal{A}_{\alpha}: a_{\alpha}(i) = a_i^*} f_{\alpha}(a_{\alpha}) \quad (78)$$

$$+ \sum_{j \in \mathcal{C}(i)} \kappa_j(a_{\alpha}(j)). \quad (79)$$

By using a finer partitioning, i.e., increasing the number of partitions in \mathcal{A}_i , it is possible to obtain very accurate lower bounds for the ELP. However, the computational complexity of the Int-Part-LP-BP as well as the corresponding DP increases rapidly as the number of partitions increases. If $|\mathcal{A}_i| \sim t$, then $|\mathcal{A}_{\alpha}| \sim t^{deg(\alpha)}$, where $deg(\alpha)$ is the nodal degree of the factor α . Observing that step (73) is essentially an exhaustive search over $t^{deg(\alpha)}$ elements, the computational time can grow quite quickly for a given accuracy requirement on the lower bound.

Significant computational benefits can be obtained by reducing the size of the Σ_i via pre-processing. This can be accomplished using the so-called *bound tightening* technique, a well-known technique in the field of constraint programming. We will describe the bound tightening scheme in the special case when the domains Σ_i are intervals given by $\Sigma_i = [l_i, u_i]$. Then the bound tightening pre-processing aims at shrinking Σ_i by solving the following optimization problems:

$$l_i^{(t)} = \min_{x \in \Sigma} x_i, \quad u_i^{(t)} = \max_{x \in \Sigma} x_i. \quad (80)$$

The above program infers a tightened bound on each variable by propagating the bounds on the other variables via the constraints. However, the program in (80) can be as difficult as the original ELP. Instead we suggest the local parallelizable sequential bound tightening scheme below.

Let $N(v)$ denote the set of neighbors of vertex v :

$$\text{for } t = 1, 2, \dots, T : \quad (81)$$

$$(l_i^{t+1}, u_i^{t+1}) \leftarrow \min / \max \quad x_i, \quad (82)$$

$$\text{subject to} \quad \forall \alpha \in N(i) \cap \bar{\mathcal{V}}_c, \quad x_\alpha \in \Sigma_\alpha, \quad (83)$$

$$x_j \in [l_j^t, u_j^t], \quad \forall j \in N(\alpha). \quad (84)$$

The bound tightening procedure described above produces a sequence of increasingly tighter bounds in each iteration. One can either continue the procedure until an approximate fixed point is reached or terminated at any earlier stage when desirable tightening has been obtained. There are also various strategies one can use to solve the optimization problem (81). For example, the constraints in (83) can be replaced by a convex relaxation, and the resulting problem can be solved using a convex nonlinear solver, such as IPOPT [26]. This is still a valid bound because the convex relaxation will produce an interval that is a superset of the interval produced by solving (81) exactly. Alternatively, if the number of constraint nodes in $N(i) \cap \bar{\mathcal{V}}_c$ is small (even though $N(i) \cap \{\bar{\mathcal{V}}_c \cup \bar{\mathcal{V}}_f\}$ may be large), then (81) can be solved by discretization similar to Int-Part-LP-BP followed by exhaustive enumeration.

The combination of bound tightening and DP was shown to be very successful in solving the OPF problem in power distribution networks that are naturally tree-structured [20]. Although the DP algorithm does not directly generalize to loopy graphs, the bound tightening scheme in (81) can still be utilized.

7 Conclusions and Path Forward

In this paper we have described ways to represent optimization and inference problems in physical flow networks as GMs. Then, focusing on the optimization (maximum likelihood) problems, we have discussed the LP-BP relaxation of the resulting GM and related hierarchies. We have also discussed the case when the underlying graph of relations is a tree, when LP-BP becomes exact and resolvable via a distributed message-passing algorithm of the DP type.

Even though we believe that the GM approach will help in the future to build efficient and accurate algorithmic solutions of various physical flow problems, the results reviewed and presented in this manuscript are clearly preliminary.

We conclude with an incomplete list of future directions extending the material presented in the manuscript.

- LP-BP provides a provable low bound. However, the resulting gap may be significant. A valuable input may be received by describing classes of physical flow problems solvable exactly by LP-BP. It is known from early works of Schlesinger [59] (see also [3, 69–71]) that LP-BP is exact when factors are

submodular. The class of problems solvable exactly by LP-BP extends to the so-called symmetric fractional polymorphism class [32]. On the other hand, many simple (not constrained by physical potentials) network flow problems are known to be (or conjectured to be) LP-BP gapless as well. (See, e.g., [24] for related discussions of the message-passing approach to solving min-cost network flow problems.) It will be important to extend this line of work (a) to characterize physical flow GM problems that are gapless and (b) to develop an approach that allows us to quantify the gap associated with LP-BP of the difficult physical flow GM formulations.

- The fact that LP-BP provides a provable low bound is powerful. However, the bound does not extend to the more challenging case of statistical inference when LP-BP optimization is substituted by generally nonconvex (due to an added entropy term) minimization of the so-called Bethe free energy functional [75]. The Bethe free energy approach is exact for GM stated over trees (then BP is equivalent to DP), but generally it provides neither lower nor upper bounds on marginal probabilities (or equivalently on the corresponding normalization factors, called partition functions). It would be important to extend bounding techniques based on GM to the physical flow GM inference problem. Approximating the entropy terms via a chain rule stated solely in terms of the marginal beliefs [57] may be an interesting step toward resolving the problem.
- One significant advantage of LP-BP over LP of a general position is related to an expectation that it can be solved efficiently via a distributed message-passing algorithm. However, designing such a provably convergent and sufficiently fast algorithm is not an easy task and has been completed for only a handful of loopy GMs, notably for Gaussian GMs under conditions of walk-summability [45] and matching GMs [1]. Such distributed, efficient, and provably convergent message-passing algorithms are yet to be developed for the physical flow GMs.
- If LP-BP is not optimal, it is natural to consider correcting it by taking into account the noninteger part of the solution, which is known [29] to have a support within a loop of the graph. Once the loopy structure is identified, one may want to modify the GM or equivalently introduce some additional constraint between beliefs associated with the loops and not linked before in the bare LP-BP. This scheme was developed in [33, 34] based on the notion of frustrated cycles and an associated constrained satisfaction problem stated in terms of beliefs optimal for the original LP-BP. Similar but different heuristic approaches were also discussed in [62–64] for a GM of a general position. Such an approach, which can be viewed as an adaptive and graph-related next step (after LP-BP) in the Sherali-Adams hierarchy, has not yet been discussed/tested on examples of the physical flow GMs.
- As discussed above in Section 4.3, interval partitioning is an important step in making LP-BP for GM with continuous valued variables tractable. Taking advantage of the constrained programming approach to condition variables and then partitioning the intervals adaptively constitutes a promising method already tested in [20] on mixed physical flow GM problems over tree graphs. Extending this method to physical flow GM problems over loopy graphs will be our

next natural step/challenge along this line of research. Notice also that finite dimensional parametrization, e.g., via mixture models [48], constitutes another promising alternative (to interval partitioning) for solving the continuous valued physical flow GM problems.

- The GM-based approach (which we have just started to develop) needs to be compared to more approaches. In the context of the OPF optimization (which is by far the most well-studied PCNF optimization problem), we plan a detailed future comparison of the “GM-based LP-BP and beyond” approach, with many new results derived most recently via a diverse set of SDP relaxations and related approaches [5, 27, 37, 38, 41, 42, 44, 49].
- It will be important to extend the GM approach to more complex PCNF problems. Of a particular interest are extensions allowing us to solve PCNF problems of stochastic and optimization type, e.g., stated in the so-called chance-constrained format [8, 58], and problems involving the interaction of different energy systems stated in terms of two (or more) coupled PCNF problems, such as coordinated scheduling for interdependent electric power and natural gas infrastructures discussed in [78].

Acknowledgements The authors are grateful to M. Lubin, N. Ruozzi, and J. B. Lasserre for fruitful discussions and valuable comments. The work at LANL was carried out under the auspices of the National Nuclear Security Administration of the US Department of Energy under Contract No. DE-AC52-06NA25396.

References

1. Ahn S, Park S, Chertkov M, Shin J (2015) Minimum weight perfect matching via blossom belief propagation. In: Neural Information Processing Systems (NIPS) – spotlight presentation
2. Babonneau F, Nesterov Y, Vial JP (2012) Design and operations of gas transmission networks. Oper Res 60(1):34–47
3. Bach F (2015) Submodular functions: from discrete to continuous domains. arXiv preprint arXiv:1511.00394
4. Baran M, Wu F (1989) Network reconfiguration in distribution systems for loss reduction and load balancing. IEEE Trans Power Delivery 4(2):1401–1407. <https://doi.org/10.1109/61.25627>
5. Bienstock D (2013) Progress on solving power flow problems. Math Optim Soc Newsl (Optima) 93:1–7. <http://www.mathopt.org/Optima-Issues/optima93.pdf>
6. Bienstock D (2015) Electrical transmission system cascades and vulnerability: an operations research viewpoint. Society for Industrial and Applied Mathematics, Philadelphia, PA
7. Bienstock D, Munoz G (2015) LP approximations to mixed-integer polynomial optimization problems. arXiv preprint arXiv:1501.00288
8. Bienstock D, Chertkov M, Harnett S (2014) Chance-constrained optimal power flow: risk-aware network control under uncertainty. SIAM Rev 56(3):461–495. <https://doi.org/10.1137/130910312>
9. Bishop CM (2006) Pattern recognition and machine learning, chapter graphical models. Springer, New York, p 359422
10. Borraz-Sánchez C (2010) Optimization methods for pipeline transportation of natural gas. Ph.D. thesis, Bergen University (Norway)

11. Chertkov M, Pan F, Stepanov MG (2011) Predicting failures in power grids: the case of static overloads. *IEEE Trans Smart Grid* 2(1):162–172. <https://doi.org/10.1109/TSG.2010.2090912>
12. Chertkov M, Stepanov M, Pan F, Baldick R (2011) Exact and efficient algorithm to discover extreme stochastic events in wind generation over transmission power grids. In: 2011 50th IEEE conference on decision and control and European control conference, pp 2174–2180. <https://doi.org/10.1109/CDC.2011.6160669>
13. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic networks and expert systems. Springer, Berlin
14. District heating: Wikipedia. https://en.wikipedia.org/wiki/District_heating. Accessed: 2017-01-07
15. Dvijotham K, Turitsyn K (2015) Construction of power flow feasibility sets. arXiv preprint arXiv:1506.07191
16. Dvijotham K, Chertkov M, Low S (2015) A differential analysis of the power flow equations. 2015 54th IEEE Conference on Decision and Control (CDC), Osaka, pp. 23–30. <https://doi.org/10.1109/CDC.2015.7402082>
17. Dvijotham K, Low S, Chertkov M (2015) Convexity of energy-like functions: theoretical results and applications to power system operations. arXiv preprint arXiv:1501.04052
18. Dvijotham K, Low S, Chertkov M (2015) Solving the power flow equations: a monotone operator approach. arXiv preprint arXiv:1506.08472
19. Dvijotham K, Vuffray M, Misra S, Chertkov M (2015) Natural gas flow solutions with guarantees: a monotone operator theory approach. arXiv preprint arXiv:1506.06075
20. Dvijotham K, Chertkov M, Van Hentenryck P, Vuffray M, Misra S (2016) Graphical models for optimal power flow. Constraints 1–26. <https://doi.org/10.1007/s10601-016-9253-y>
21. Frolov V, Backhaus S, Chertkov M (2014) Efficient algorithm for locating and sizing series compensation devices in large power transmission grids: I. model implementation. *New J Phys* 16(10), 105015. <http://stacks.iop.org/1367-2630/16/i=10/a=105015>
22. Frolov V, Backhaus S, Chertkov M (2014) Efficient algorithm for locating and sizing series compensation devices in large power transmission grids: II. solutions and applications. *New J Phys* 16(10):105016. <http://stacks.iop.org/1367-2630/16/i=10/a=105016>
23. Frolov V, Thakurta PG, Backhaus S, et al. (2016) Optimal placement and sizing of FACTS devices to delay transmission expansion. arXiv preprint arXiv:1608.04467
24. Gamarnik D, Shah D, Wei Y (2010) Belief propagation for min-cost network flow: convergence and correctness. In: Proceedings of the twenty-first annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp 279–292. <http://dl.acm.org/citation.cfm?id=1873601.1873625>
25. Glover JDD, Sarma MS (2001) Power system analysis and design, 3rd edn. Brooks/Cole Publishing, Pacific Grove, CA
26. Interior point optimizer. <https://projects.coin-or.org/Ipopt>. Accessed: 2017-07-23
27. Jabr R (2012) Exploiting sparsity in SDP relaxations of the OPF problem. *IEEE Trans Power Syst* 27(2):1138–1139. <https://doi.org/10.1109/TPWRS.2011.2170772>
28. Jensen F (1996) An introduction to Bayesian networks. Springer, Berlin
29. Johnson J (2008) Convex relaxation methods for graphical models: lagrangian and maximum entropy approaches. Ph.D. thesis, MIT. http://ssg.mit.edu/~jasonj/johnson_phd.pdf
30. Kersulis J, Hiskens I, Chertkov M, Backhaus S, Bienstock D (2015) Temperature-based instanton analysis: identifying vulnerability in transmission networks. In: 2015 IEEE Eindhoven PowerTech, pp 1–6. <https://doi.org/10.1109/PTC.2015.7232816>
31. Koller D, Friedman N (2009) Probabilistic graphical models. MIT Press, Massachusetts
32. Kolmogorov V, Thapper J, Zivny S (2015) The power of linear programming for general-valued CSPs. *SIAM J Comput* 44(1):1–36
33. Kudekar S, Johnson J, Chertkov M (2011) Linear programming based detectors for two-dimensional intersymbol interference channels. In: 2011 IEEE international symposium on information theory proceedings (ISIT), pp 2999–3003. <https://doi.org/10.1109/ISIT.2011.6034129>

34. Kudekar S, Johnson J, Chertkov M (2013) Improved linear programming decoding using frustrated cycles. In: 2013 IEEE international symposium on information theory proceedings (ISIT), pp 1496–1500. <https://doi.org/10.1109/ISIT.2013.6620476>
35. Lasserre J (2001) Global optimization with polynomials and the problem of moments. SIAM J Optim 11(3):796–817
36. Lasserre J (2010) Moments, positive polynomials and their applications. Imperial College Press, London
37. Lavaei J, Low S (2012) Zero duality gap in optimal power flow problem. IEEE Trans Power Syst 27(1):92–107. <https://doi.org/10.1109/TPWRS.2011.2160974>
38. Lesieutre B, Molzahn D, Borden A, DeMarco C (2011) Examining the limits of the application of semidefinite programming to power flow problems. In: 2011 49th annual Allerton conference on communication, control, and computing (Allerton), pp 1492–1499. <https://doi.org/10.1109/Allerton.2011.6120344>
39. Lieu H, Gartner N, Messer CJ, et al. (1999) Traffic flow theory. Public Roads 62:45–47
40. Lovasz L, Schrijver A (1991) Cones of matrices and set-functions and 0–1 optimization. SIAM J Optim 1(12):166–190
41. Low S (2014) Convex relaxation of optimal power flow; part i: formulations and equivalence. IEEE Trans Control Netw Syst 1(1):15–27. <https://doi.org/10.1109/TCNS.2014.2309732>
42. Low S (2014) Convex relaxation of optimal power flow; part ii: exactness. IEEE Trans Control Netw Syst 1(2), 177–189. <https://doi.org/10.1109/TCNS.2014.2323634>
43. Machowski J, Bialek JW, Bumby JR (2008) Power system dynamics: stability and control. Wiley, Chichester. <http://opac.inria.fr/record=b1135564>. Rev. ed. of: Power system dynamics and stability/Jan Machowski, Janusz W. Bialek, James R. Bumby (1997)
44. Madani R, Sojoudi S, Lavaei J (2015) Convex relaxation for optimal power flow problem: mesh networks. IEEE Trans Power Syst 30(1):199–211. <https://doi.org/10.1109/TPWRS.2014.2322051>
45. Malioutov D, Johnson J, Willsky A (2006) Walk-sums and belief propagation in gaussian graphical models. J Mach Learn Res 7:2031–2064
46. Mezard M, Montanari A (2009) Information, physics, and computation. Oxford graduate texts. Oxford University Press, Oxford
47. Misra S, Vuffray M, Chertkov M (2015) Maximum throughput problem in dissipative flow networks with application to natural gas systems. arXiv preprint arXiv:1504.02370
48. Mixture model: Wikipedia. https://en.wikipedia.org/wiki/Mixture_model#Gaussian_mixture_model
49. Molzahn D, Hiskens I (2014) Moment-based relaxation of the optimal power flow problem. In: Power systems computation conference (PSCC), 2014, pp 1–7. <https://doi.org/10.1109/PSCC.2014.7038397>
50. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge
51. Misra S, Fisher MW, Backhaus S, Bent R, Chertkov M, Pan F (2015) Optimal compression in natural gas networks: a geometric programming approach. IEEE Trans Control Netw Syst 2(1):47–56. <https://doi.org/10.1109/TCNS.2014.2367360>
52. Osiadacz A (1987) Simulation and analysis of gas networks. Gulf Publishing Company, London. <http://books.google.com/books?id=cMxTAAAMAAJ>
53. Parrilo P (2003) Semidefinite programming relaxations for semialgebraic problems. Math Program Ser B 96(2):293–320
54. Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo, CA
55. Rauschenbach T (2016) Modeling, control and optimization of water systems: systems engineering methods for control and decision making tasks. Springer, Berlin
56. Richardson TJ, Urbanke RL (2008) Modern coding theory. Cambridge University Press, Cambridge
57. Risteski A (2016) How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods. CoRR abs/1607.03183. <http://arxiv.org/abs/1607.03183>

58. Roald L, Misra S, Chertkov M, Andersson G (2016) Optimal power flow with weighted chance constraints and general policies for generation control. In: 2015 54th IEEE conference on decision and control (CDC), pp 6927–6933. IEEE, Piscataway, NJ
59. Schlesinger MI (1976) Syntatic analysis of two-dimensional visual signals in noisy conditions. Kibernetika [in Russian] 4:113–130
60. Sherali HD, Adams WP (1990) A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. SIAM J Discret Math 3(3):411–430. <https://doi.org/10.1137/0403036>
61. Sontag DA (2010) Approximate inference in graphical models using LP relaxations. Ph.D. thesis, MIT. http://www.cs.nyu.edu/~dsontag/papers/sontag_phd_thesis.pdf
62. Sontag D, Jaakkola T (2008) New outer bounds on the marginal polytope. In: Platt J, Koller D, Singer Y, Roweis S (eds) Advances in neural information processing systems 20, pp 1393–1400. MIT Press, Cambridge, MA
63. Sontag D, Meltzer T, Globerson A, Weiss Y, Jaakkola T (2008) Tightening LP relaxations for MAP using message-passing. In: 24th Conference in Uncertainty in Artificial Intelligence, pp 503–510. AUAI Press, Corvallis
64. Sontag D, Choe DK, Li Y (2012) Efficiently searching for frustrated cycles in MAP inference. In: Proceedings of the twenty-eighth conference on uncertainty in artificial intelligence (UAI-12), pp 795–804. AUAI Press, Corvallis
65. Vuffray M, Misra S, Chertkov M (2015) Monotonicity of dissipative flow networks renders robust maximum profit problem tractable: general analysis and application to natural gas flows. In: 2015 54th IEEE conference on decision and control (CDC), pp 4571–4578. <https://doi.org/10.1109/CDC.2015.7402933>
66. Wainwright MJ (2002) Stochastic processes on graphs: geometric and variational approaches. Ph.D. thesis, MIT. http://www.eecs.berkeley.edu/~wainwrig/Papers/Final2_PhD_May30.pdf
67. Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Found Trends Mach Learn 1(1):1–305
68. Wald Y, Globerson A (2014) Tightness results for local consistency relaxations in continuous MRFs. In: 30th conference in uncertainty in artificial intelligence. AUAI Press, Corvallis
69. Werner T (2007) A linear programming approach to max-sum problem: a review. IEEE Trans Pattern Anal Mach Intell 29(7):1165–1179. <https://doi.org/10.1109/TPAMI.2007.1036>
70. Werner T (2008) High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (map-mrf). In: IEEE conference on computer vision and pattern recognition, 2008 (CVPR 2008), pp 1–8. <https://doi.org/10.1109/CVPR.2008.4587355>
71. Werner T (2010) Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction. IEEE Trans Pattern Anal Mach Intell 32(8):1474–1488. <https://doi.org/10.1109/TPAMI.2009.134>
72. Wolf DD, Smeers Y (2000) The gas transmission problem solved by an extension of the simplex algorithm. Manag Sci 46(11):1454–1465. <http://www.jstor.org/stable/2661661>
73. Wong P, Larson R (1968) Optimization of natural-gas pipeline systems via dynamic programming. IEEE Trans Autom Control 13(5):475–481. <https://doi.org/10.1109/TAC.1968.1098990>
74. Wu S, Ros-Mercado R, Boyd E, Scott L (2000) Model relaxations for the fuel cost minimization of steady-state gas pipeline networks. Math Comput Model 31(23):197–220. [http://dx.doi.org/10.1016/S0895-7177\(99\)00232-0](http://dx.doi.org/10.1016/S0895-7177(99)00232-0). <http://www.sciencedirect.com/science/article/pii/S0895717799002320>
75. Yedidia J, Freeman W, Weiss Y (2005) Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Trans Info Theory 51(7):2282–2312. <https://doi.org/10.1109/TIT.2005.850085>
76. Zinger H (1986) Hydraulic and heating regimes of district heating networks [in Russian]. Energoatomizdat, Moscow
77. Živný S, Werner T, Průša DA (2014) The power of LP relaxation for MAP inference, pp 19–42. MIT Press, Cambridge
78. Zlotnik A, Roald L, Backhaus S, Chertkov M, Andersson G (2017) Coordinated scheduling for interdependent electric power and natural gas infrastructures. IEEE Trans Power Syst 32(1):600–610. <https://doi.org/10.1109/TPWRS.2016.2545522>

Profit Maximizing Storage Integration in AC Power Networks



Anya Castillo and Dennice F. Gayme

Abstract This work demonstrates that there is an analytical relationship between nodal price signals and the optimal allocation and operation of distributed energy storage systems (ESSs) in alternating current (AC) power networks. The results are based on a semidefinite relaxation of a multi-period optimal power flow (OPF) with storage problem in which the ESSs provide both real and reactive power to the grid. Strong duality is exploited to define a storage operator subproblem that is used to elucidate the natural duality between minimizing system costs and maximizing the profits of the storage operator in purely competitive markets. We illustrate these theoretical relationships, which require strong duality to hold, and discuss their impact on siting decisions using case studies based on an augmented IEEE benchmark system. We focus on how the provision of reactive power in combination with traditional grid services changes both the ESS allocation strategy and the overall performance of the simulated power network. Our results highlight the tight connections between market design and the financial viability of large-scale storage integration in AC power systems.

1 Introduction

Large-scale integration of intermittent renewables (e.g., wind and solar) requires complementary technologies to provide grid stability, instantaneous power balance capability, and fast-ramping reserves. Grid-scale energy storage systems (ESSs) can provide these services along with other potential benefits such as the opportunity to defer generation and transmission investments, fast-response ancillary services,

A. Castillo
Sandia National Laboratories, Albuquerque, NM, USA
e-mail: arcasti@sandia.gov

D. F. Gayme (✉)
Johns Hopkins University, Baltimore, MD, USA
e-mail: dennice@jhu.edu

load following, load shifting, as well as improvements in power quality and service reliability [43].

The benefits of grid-scale storage are generally discussed in the contexts of energy arbitrage, reserve power, frequency regulation, and mitigating the effect of non-dispatchable renewable energy sources; e.g., see [3, 11, 17]. The provision of reactive power (VAr) support through the power electronics connecting storage technologies such as batteries to the grid or from the generator-motor excitation of a pumped hydro storage system connected to the grid [48] is far less studied. The lack of research on the provision of VAr support by ESSs is due in part to the complexity of the full AC optimal power flow with storage (OPF+S) problem. Many studies instead use a linear approximation, the DC OPF+S formulation [13, 15, 21, 47, 51], which is more computationally tractable but cannot be used to study the dispatch of reactive power. However, even in studies of the AC OPF+S problem, the provision of VAr support by the storage is often either not included [4, 7, 19, 22, 24, 50] or neglected in the case studies and analysis [10, 20].

Neglecting the provision of VAr support represents an important missed factor in evaluating grid-scale storage because reactive power is critical to the efficient and reliable operation of the electric power grid. More specifically, reactive power is used to control voltage levels and ensure grid stability and power quality [11]. ESSs that are grid connected through an inverter are particularly well suited to provide VAr support because they can supply dynamic reactive power, which gives them distinct advantages over transmission equipment such as capacitors and inductors that supply and consume static reactive power. This capability enables ESSs to quickly change the amount of VAr support independent of the voltage level and enables faster response to sudden large voltage surges or drops [9, 27, 42, 44]. Furthermore, renewables are displacing synchronous generators that have historically provided reactive power compensation, engendering concern that these system changes may lead to a reactive power deficiency. Therefore system operators are currently assessing whether these asynchronous resources should be subjected to requirements to provide reactive power compensation, similar to those currently placed on synchronous resources.

This work addresses this gap in the literature by extending the results in [10] to investigate how the provision of VAr support affects ESS siting and the overall quality of the power flow solution. In particular, we investigate how providing VAr support in addition to traditional grid services affects siting and sizing decisions as well as overall grid performance. We overcome the difficulty of the nonconvex OPF+S problem through the use of the semidefinite relaxation (SDR) approach originally presented in [5], further refined in [30] and extended to the OPF+S problem in [20]. We refer to this convex problem as the SDR-OPF+S problem and note that the SDR-OPF+S is equivalent to the original OPF+S problem when the relaxation is exact, which has been shown to be true for many practical power system examples [45] including all of the IEEE test cases [30]. Developing a full characterization for the classes of OPF problem instances for which a convex relaxation is guaranteed is an active area of research; see, e.g., [8, 31, 32, 46].

Our theoretical results leverage the fact that the Karush-Kuhn-Tucker (KKT) point to the SDR-OPF+S problem is the global optimum of the OPF+S problem when the relaxation is exact. The Lagrangian dual of the SDR is used to form a storage operator subproblem, which isolates the primal and dual variables related to storage. This problem defines the value of storage as a function of its cost and revenue streams from charging, discharging, and VAr support. Our main result exploits the properties of strong duality to prove that maximizing the profits to the storage operator is equivalent to minimizing system costs in a purely competitive market.

We illustrate the effects of providing VAr support from ESSs through case studies based on an augmented IEEE-14 transmission system with wind power integration. In order to ensure that the numerical studies are well posed, we also provide a sufficient condition for preventing simultaneous charging and discharging of individual ESSs governed by linear storage dynamics without having to incorporate mixed integer or nonlinear constraints as in [1, 39, 53]. We verify that the SDR is exact in all cases reported to ensure that we have obtained the global optimum of the OPF+S problem. Our results show that including VAr support as one of the services provided by ESSs greatly changes the optimal allocation of resources in the network versus the same system where ESSs provide only traditional grid services. We also show that having ESSs provide reactive power leads to more efficient power system operations, resulting in market settlements that clear at a lower operating cost. We then briefly discuss how a nodal payment mechanism for VAr support compares to current market mechanisms that pay for reactive power capability. The combination of theoretical results and case studies provide insight that can be used to develop better models of ESS integration and dispatch, improve market design, and determine optimal scheduling policies for ESS units.

The remainder paper is organized as follows. The OPF+S formulation is presented in Section 2. Our main theoretical results are presented in Section 3 followed by supporting numerical results in Section 4. We summarize with final remarks in Section 5.

2 Problem Formulation

Consider a power network with a set of buses $\mathcal{N} := \{1, \dots, N\}$ with a subset of generator buses $\mathcal{G} \subseteq \mathcal{N} := \{1, \dots, G\}$ and a subset of wind farm buses $\mathcal{W} \subseteq \mathcal{N} := \{1, \dots, W\}$. We denote the set of branches as $\mathcal{K} := \{1, \dots, K\}$ and indicate bidirectional flows between interconnected nodes $i, k \in \mathcal{N}$ by $k(n, i), k(i, n) \subseteq k(\cdot)$, where the order of the indices signifies the direction of flow. The associated nodal and branch admittance matrices for the network are defined based on equivalent π models [36] and respectively denoted $Y \in \mathbb{C}^{N \times N}$ and $Y_{k(\cdot)} \in \mathbb{C}^{K \times N}$. The network is assumed to be operating under balanced steady-state conditions over discrete time intervals $t \in \mathcal{T} := \{1, \dots, T\}$.

The net flow of real and reactive power at each bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$ are, respectively, denoted $P_n(t)$ and $Q_n(t)$. Here $P_n(t) + \mathbf{j}Q_n(t) = V_n(t)[I_n(t)]^*$, where $[\cdot]^*$ denotes the complex conjugate and $\mathbf{j} := \sqrt{-1}$. $V_n(t) = V_n^r(t) + \mathbf{j}V_n^i(t)$ and $I_n(t) = I_n^r(t) + \mathbf{j}I_n^i(t)$, respectively, denote the voltage and current phasors. The corresponding vector consisting of the nodal voltages and currents at all buses are, respectively, defined as $\mathbf{I}(t) := [I_1(t) \cdots I_N(t)]^T$ and $\mathbf{V}(t) := [V_1(t) \cdots V_N(t)]^T$, where $[\cdot]^T$ indicates the transpose and $\mathbf{I}(t) = Y\mathbf{V}(t)$.

We now define the optimal power flow with storage problem for the power network described above; we begin by defining the inflows and outflows contributing to the power balance at each node $n \in \mathcal{N}$. The real and reactive power demands at each bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$ are assumed to be known and are, respectively, denoted $P_n^d(t)$ and $Q_n^d(t)$. The real and reactive power injection from conventional generators at each bus $n \in \mathcal{G}$ over time interval $t \in \mathcal{T}$ are, respectively, bounded as

$$P_n^{min} \leq P_n^g(t) \leq P_n^{max} \quad (1)$$

$$Q_n^{min} \leq Q_n^g(t) \leq Q_n^{max}. \quad (2)$$

The rate of real power injection from the generators at each bus $n \in \mathcal{G}$ over time interval $t \in \mathcal{T}$ is constrained as

$$-RR_n \leq P_n^g(t) - P_n^g(t-1) \leq RR_n, \quad (3)$$

where $P_n^g(0)$ is a known real power injection from conventional generation at bus $n \in \mathcal{G}$. The real power injected to the grid from the wind plants at bus $n \in \mathcal{W}$ over time interval $t \in \mathcal{T}$ is limited by

$$0 \leq P_n^w(t) \leq C_n^w(t), \quad (4)$$

where $C_n^w(t)$ is the wind power availability at each bus $n \in \mathcal{W}$ over time interval $t \in \mathcal{T}$.

Each bus $n \in \mathcal{N}$ may also have energy storage, which we model as a single ESS unit with total capacity C_n . The storage level at each bus $n \in \mathcal{N}$ during time interval $t \in \mathcal{T}$ is thus bounded as

$$C_n^{min} \leq s_n(t) \leq C_n, \quad (5)$$

where C_n^{min} represents a technology-dependent depth of discharge in the storage dispatch problem and $C_n^{min} := 0$ in the storage allocation problem, which determines the siting and sizing of ESSs along with an associated dispatch strategy.

The total overall energy storage allocated throughout the network is bounded as

$$\sum_{n \in \mathcal{N}} C_n \leq h, \quad (6)$$

where h is the total storage budget to be allocated.

The per bus storage level over time interval $t \in \mathcal{T}$ is governed by

$$s_n(t) = s_n(t-1) + \eta_n^c r_n^c(t) - r_n^d(t) / \eta_n^d, \quad (7)$$

where $s_n(0)$ is a known initial charge level at bus $n \in \mathcal{N}$. Here $r_n^c(t)$ and $r_n^d(t)$ are the respective real power charge and discharge rates at each bus $n \in \mathcal{N}$ for time interval $t \in \mathcal{T}$, and the coefficients η_n^c and η_n^d denote the corresponding charge and discharge efficiencies. The real power charge and discharge rates are, respectively, bounded as

$$0 \leq r_n^c(t) \leq R_n^c \quad (8)$$

$$0 \leq r_n^d(t) \leq R_n^d. \quad (9)$$

We assume a single operating cycle in which the terminal storage level in the current cycle equals the terminal storage level in the prior operating cycle, that is,

$$\tilde{s}_n(T) - s_n(T) = 0 \quad (10)$$

for all buses $n \in \mathcal{N}$. Therefore, the initial storage level $s_n(0)$ in (7) is parameterized by the terminal storage level in the prior operating cycle. This constraint, which essentially assumes a single finite operating cycle is not a limitation of the method as this operating cycle can be arbitrarily defined. In practice this time interval is generally chosen to approximate naturally occurring cyclic variations in demand and generation profiles, e.g., diurnal or seasonal cycles. We note however that the approach can easily be extended to include several cycles in a single optimization problem or be run over several cycles, e.g., in order to investigate the effect of seasonal variations that are likely to influence siting and sizing decisions, which may be based on worst case or averaged results.

We assume that the ESS at each bus $n \in \mathcal{N}$ has a power converter that enables it to deliver and absorb both real and reactive power. The VAr injection and absorption rates from the ESS at each bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$ are bounded as

$$Z_n^{min} \leq z_n(t) \leq Z_n^{max}, \quad (11)$$

where $z_n(t) > 0$ indicates VAr injection and $z_n(t) < 0$ indicates VAr absorption. Reactive power bounds are specified as a linear function of the energy capacity C_n , i.e., $Z_n^{min} = \alpha C_n$ and $Z_n^{max} = \beta C_n$ for $\alpha, \beta \in \mathbb{R}$.

All of the power injections and withdrawals from the system can be combined to define the respective real and reactive power injections for each bus $n \in \mathcal{N}$ and $t \in \mathcal{T}$ as

$$P_n(t) = \text{Re}\{V_n I_n^*\} = P_n^w(t) + P_n^g(t) - P_n^d(t) - (r_n^c(t) - r_n^d(t)) \quad (12)$$

$$Q_n(t) = \text{Im}\{V_n I_n^*\} = Q_n^g(t) - Q_n^d(t) + z_n(t). \quad (13)$$

Here $\text{Re}\{V_n I_n^*\}$ and $\text{Im}\{V_n I_n^*\}$, respectively, denote the real and imaginary components of $V_n I_n^*$. The corresponding magnitudes of these voltages at each bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$ are bounded as

$$(V_n^{\min}) \leq |V_n(t)| \leq (V_n^{\max}). \quad (14)$$

The apparent power flows on each line $k \in \mathcal{K}$ in time interval $t \in \mathcal{T}$ are limited as

$$(P_{k(\cdot)}^\ell(t))^2 + (Q_{k(\cdot)}^\ell(t))^2 \leq (S_k^{\max})^2, \quad (15)$$

where the flow from bus n to bus i is given by $P_{k(n,i)}^\ell(t) + \mathbf{j}Q_{k(n,i)}^\ell(t) = V_n(t)[I_{k(n,i)}(t)]^*$ for $I_{k(n,i)}(t) = Y_{k(n,i)}\mathbf{V}(t)$.

Remark The generic storage model described by (5)–(11) can be adapted to specific storage technologies through parameterization of the model variables, e.g., R_n^d , C_n^{\min} , C_n , R_n^c , η_n^d , η_n^c , Z_n^{\max} , Z_n^{\min} , α , β , and h .

A number of extensions to the storage model are also possible. For example, the apparent power rating of an ESS can be incorporated through the following constraint:

$$(r_n(t))^2 + (z_n(t))^2 \leq (S_n^{\max})^2,$$

where $r_{n,t} := r_n(t)^c - r_n(t)^d$ and S_n^{\max} denotes the apparent power rating of the technology. In this work we omit this constraint and instead assume a simple box constraint; however, an extension of the results to include this constraint is straightforward as it is of the same mathematical form as (15). In practice the exact bounds will be largely determined by the choice of the power converter, and we leave questions such as the optimization of inverter size and studies related to the influence of the power converter characteristics as directions for future work.

□

2.1 Optimal Power Flow with Storage

We now formulate the OPF+S problem which determines the optimal allocation (i.e., siting and sizing) and operations of ESSs that minimize system costs. In Section 3 we demonstrate that this problem is dual to maximizing storage operator profits in a perfectly competitive market.

Let \mathbf{V} , \mathbf{r}^c , \mathbf{r}^d , \mathbf{z} , \mathbf{s} , \mathbf{P}^g , \mathbf{Q}^g , \mathbf{P}^w , \mathbf{P}^ℓ , \mathbf{Q}^ℓ , and \mathbf{C} , respectively, denote the following arrays of decision variables, where the subscripts indicate the size of the array: voltage phasors, $\{V_n(t)\}_{N \times T}$, storage real power charge rates, $\{r_n^c(t)\}_{N \times T}$, storage real power discharge rates, $\{r_n^d(t)\}_{N \times T}$, storage reactive power injection/withdrawal rates, $\{z_n(t)\}_{N \times T}$, storage energy levels, $\{s_n(t)\}_{N \times T}$, conventional generator real power injections, $\{P_n^g(t)\}_{N \times T}$, conventional generator reactive power injections, $\{Q_n^g(t)\}_{N \times T}$, wind plant real power injections, $\{P_n^w(t)\}_{N \times T}$, real power branch flows, $\{P_{k(\cdot)}^\ell(t)\}_{K \times T}$, reactive power branch flows, $\{Q_{k(\cdot)}^\ell(t)\}_{K \times T}$, and ESS energy capacities, $\{C_n\}_N$. The multi-period OPF+S problem can then be stated as

$$\mathbf{p}^* := \min_{\substack{\mathbf{V}, \mathbf{r}^c, \mathbf{r}^d, \mathbf{z}, \mathbf{s}, \mathbf{P}^g, \\ \mathbf{Q}^g, \mathbf{P}^w, \mathbf{P}^\ell, \mathbf{Q}^\ell, \mathbf{C}}} \sum_{n \in \mathcal{G}} f_n^g(\cdot) + \sum_{n \in \mathcal{N}} f_n^s(\cdot) \quad (16)$$

subject to

$$(1)-(15), \quad (17)$$

where

$$f_n^g(\cdot) := \sum_{t \in \mathcal{T}} c_{n,2}^g (P_n^g(t))^2 + c_{n,1}^g P_n^g(t) \quad (18)$$

is a strictly convex function of real power generation, and

$$f_n^s(\cdot) := \sum_{t \in \mathcal{T}} c_{n,1}^s r_n^d(t) \quad (19)$$

is a linear function of the storage discharge rate. The cost function in (16) assumes negligible marginal cost for wind energy resources in order to reflect the current market paradigm in which wind energy is free.

The optimal power flow with storage problem presented above can be solved as either a storage dispatch problem that finds optimal values for \mathbf{V} , \mathbf{r}^c , \mathbf{r}^d , \mathbf{z} , \mathbf{s} , \mathbf{P}^g , \mathbf{Q}^g , \mathbf{P}^w , \mathbf{P}^ℓ , and \mathbf{Q}^ℓ or a storage allocation problem, which determines both the siting and sizing of ESSs, C_n for each bus $n \in \mathcal{N}$, given a total energy capacity of h . In the storage allocation problem, $\mathbf{C} = \{C_n\}_N$ becomes a decision variable.

3 OPF-Based ESS Integration and the Storage Operator Subproblem

In this section we provide an analytical relationship between a globally optimal solution of the OPF+S problem (16) and (17) and maximizing storage operator profits in particular storage siting and dispatch problems. These relationships are derived by first employing a SDR [5, 30] to convexify the OPF+S problem. We follow the approach that is detailed in [20] and reformulate the problem in a higher dimensional space by lifting the bilinear voltage phasor terms to form the matrix:

$$W(t) := \begin{bmatrix} \mathbf{V}^r(t)^T & \mathbf{V}^j(t)^T \end{bmatrix}^T \begin{bmatrix} \mathbf{V}^r(t)^T & \mathbf{V}^j(t)^T \end{bmatrix},$$

where $\mathbf{V}^r(t)$ and $\mathbf{V}^j(t)$, respectively, indicate column vectors $Re\{V_n(t)\}$ and $Im\{V_n(t)\}$ with each row corresponding to a bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$.

The power balance, voltage magnitude, and power flow constraints in (12)–(15) are then rewritten in terms of $W(t)$ for all $t \in \mathcal{T}$. This matrix is constrained to be positive semidefinite, i.e.,

$$W(t) \succeq 0 \quad \forall t \in \mathcal{T}. \quad (20)$$

We refer to the resulting relaxed primal problem as the SDR-OPF+S problem; its details are provided in Appendix 6.1. The SDR-OPF+S problem is equivalent to the OPF+S problem in (16) and (17) when the rank of $W(t)$ equals one for all $t \in \mathcal{T}$, i.e., the solution of the SDR-OPF+S problem with a rank one $W(t)$ for all $t \in \mathcal{T}$ is equivalent to the global optimum of the OPF+S problem. In what follows we refer to this solution as the *rank one solution* of the SDR-OPF+S problem.

The SDR-OPF+S problem can be solved through its Lagrangian dual. To obtain a solution to the original OPF+S problem, we project the solution of this relaxed problem onto the original problem space; see, e.g., [20, 30]. We refer to the Lagrangian dual of the SDR-OPF+S problem as the LD-OPF+S problem and provide its details in Appendix 6.2. Here we focus on the role of the dual variables for the per bus real and reactive power balance equations (12) and (13), which are, respectively, denoted as $\lambda_n(t)$ and $\varphi_n(t)$ for each bus $n \in \mathcal{N}$ and time interval $t \in \mathcal{T}$ (for details regarding their precise definition; see Appendix 6.2 or [20]). The first of these, $\lambda_n(t)$, is known as the locational marginal price (LMP) or the nodal price at each bus $n \in \mathcal{N}$ and time interval $t \in \mathcal{T}$. Here we analogously refer to $\varphi_n(t)$ for each bus $n \in \mathcal{N}$ and time interval $t \in \mathcal{T}$ as the Q-LMP because it mathematically accounts for the nodal price of VAr support at bus $n \in \mathcal{N}$ and time interval $t \in \mathcal{T}$. These nodal prices play a key role in the storage problem because the energy storage level of the ESS at each bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$ depends on charging and discharging operations in previous time intervals, which are scheduled to minimize system costs. The marginal cost for each ESS over each time interval $t \in \mathcal{T}$ includes the discharge cost plus the cost of the power that is

used to charge it, which is based on the nodal price signal, i.e., the LMP, for that location and time period. The marginal cost for reactive power can be computed in a similar manner.

The dependence of the ESS units on the nodal prices can be exploited to form a subproblem that isolates the interactions among the storage variables and the nodal prices of real and reactive power. We refer to the corresponding optimization problem as the *storage operator subproblem*, which is given by

$$g^s(\lambda, \varphi) := \min_{\mathbf{r}^c, \mathbf{r}^d, \mathbf{z}, \mathbf{s}, \mathbf{C}} \Lambda^s(\cdot) \quad (21)$$

subject to

$$(5)–(11), \quad (22)$$

where

$$\Lambda^s(\cdot) := \sum_{n \in \mathcal{N}} \left\{ f_n^s(\cdot) + \sum_{t \in \mathcal{T}} \left[\lambda_n(t) (r_n^c(t) - r_n^d(t)) - \varphi_n(t) z_n(t) \right] \right\}.$$

This storage operator subproblem is optimal for a rank one solution of the SDR-OPF+S problem for either a single storage operator or multiple storage operators (e.g., a storage operator per technology or per bus, which would be handled through multiple subproblems) due to strong duality. The KKT conditions and Slater's condition for the optimal storage problem [20] ensure that the optimal solution of the subproblem is also globally optimal for the SDR-OPF+S problem. This solution of the SDR-OPF+S problem is also the global optimum of the OPF+S problem when the rank of $W(t)$ is one for all $t \in \mathcal{T}$.

Clearly the use of the Lagrangian dual in forming the storage operator subproblem limits its applicability to cases when the SDR is exact (i.e., there is a rank one solution of the SDR-OPF+S problem). However, this condition is not overly restrictive as the SDR has been shown to be exact for many practical power system examples [45] including all of the IEEE test cases [30]. Classifying the types of OPF problem instances for which a convex relaxation is guaranteed is an active area of research; see, e.g., [8, 31, 32, 46].

3.1 Profit Maximizing Storage Allocation

This subsection describes the main theoretical results connecting profit maximization in the storage operator subproblem in (21) and (22) to optimal storage siting and dispatch decisions for the OPF+S problem in (16) and (17). All of the theory developed here and in the following section can easily be extended to include cost functions with other convex combinations of the system variables; see, e.g., [8]. For

example, the storage discharge function (19) can be replaced by any monotonically nondecreasing convex function in order to add penalties that incorporate ESS life cycle costs and degradation effects.

The marginal profits of the storage operator can be computed as the difference between the total marginal revenues and the total marginal costs from the provision of real and reactive power for the ESS unit at each bus $n \in \mathcal{N}$ over each time interval $t \in \mathcal{T}$. We therefore define these marginal profits as

$$\pi^s(\cdot) := -\Lambda^s = \sum_{n \in \mathcal{N}} \left\{ -f_n^s(\cdot) + \sum_{t \in \mathcal{T}} \left[\lambda_n(t) (r_n^d(t) - r_n^c(t)) + \varphi_n(t) z_n(t) \right] \right\}. \quad (23)$$

Here the costs associated with real power supplied by the ESS units at bus $n \in \mathcal{N}$ over each time interval $t \in \mathcal{T}$ are given by $c_{n,1}^s r_n^d(t) + \lambda_n(t) r_n^c(t)$, where the first term represents discharge costs incurred (these costs can be used to account for, e.g., wear or cycling costs) and the second term represents the LMP-based market settlement, in which revenues from the real power supplied (i.e., discharging), costs from the real power consumed (i.e., charging), and VAr support are accounted for.

The total VAr support revenues and costs at each bus $n \in \mathcal{N}$ over time interval $t \in \mathcal{T}$ are, respectively, given by

$$\left\{ \varphi_n(t) z_n(t) \mid (z_n(t) \geq 0 \wedge \varphi_n(t) \geq 0) \vee (z_n(t) \leq 0 \wedge \varphi_n(t) \leq 0) \right\} \quad (24)$$

and

$$\left\{ \varphi_n(t) z_n(t) \mid (z_n(t) \leq 0 \wedge \varphi_n(t) \geq 0) \vee (z_n(t) \geq 0 \wedge \varphi_n(t) \leq 0) \right\}, \quad (25)$$

where as before $z_n(t) > 0$ indicates that the ESS unit is injecting power into node $n \in \mathcal{N}$ during time interval $t \in \mathcal{T}$ and the value of the Q-LMP is sign indefinite.

The connection between the profits as defined in (23) and the OPF+S-based storage allocation problem defined in (16) and (17) is formalized in the following theorem.

Theorem 1 *For an arbitrary operating cycle $\mathcal{T} := \{1, \dots, T\}$, the energy storage capacity C_n receives the most incremental value at the bus n where*

$$\max_{n \in \mathcal{N}} \left\{ \max_{\lambda, \varphi} \pi_n^s(\cdot)^* \right\}. \quad (26)$$

Here we denote the globally optimal solution of the storage operator subproblem in (21) and (22) with the superscript $$ and the total marginal profits to the storage operator as*

$$\pi_n^s(\cdot)^* := \pi_n^{sP}(\cdot)^* + \pi_n^{sQ}(\cdot)^*, \quad (27)$$

where

$$\pi_n^{sP}(\cdot)^* = -f_n^s(\cdot)^* + \sum_{t \in \mathcal{T}} \lambda_n(t)^* \left[r_n^d(t)^* - r_n^c(t)^* \right], \quad (28)$$

is the storage operator profit from the supply of real power/energy services and

$$\pi_n^{sQ}(\cdot)^* = \sum_{t \in \mathcal{T}} \varphi_n(t)^* [z_n(t)^*] \quad (29)$$

denotes the corresponding profit from VAr support.

Proof Given the KKT point of the LD-OPF+S problem in (37) and (38), the corresponding optimal solution to the storage subproblem in (21) and (22) is

$$\max_{\lambda, \varphi} g^s(\lambda, \varphi), \quad (30)$$

where $g^s(\lambda, \varphi) = g^s(\lambda, \varphi)^*$ and

$$g^s(\lambda, \varphi)^* := \max_{\mathbf{r}^e, \mathbf{r}^d, \mathbf{z}, \mathbf{s}, \mathbf{C}} \sum_{n \in \mathcal{N}} \pi_n^s(\cdot)^*. \quad (31)$$

Therefore the greatest profits are attained at the bus where $\pi_n^s(\cdot)^*$ is maximized. This theorem leads to an important observation about the relationship between maximizing storage operator profits and minimal operational costs (as defined by the global optimal solution of the OPF+S problem); these problems have a natural duality in a purely competitive market as summarized in the following corollary.

Corollary 1 *Given the KKT point of the LD-OPF+S, the profit to the storage at bus n must be nonnegative, i.e., $\pi_n^s(\cdot) \geq 0$, whenever $C_n > 0$. Therefore*

$$g^s(\lambda, \varphi)^* \geq 0. \quad (32)$$

The condition in (32) also holds true for the SDR-OPF+S problem, which in turn minimizes costs in the OPF+S problem for a purely competitive market.

The results in Theorem 1 also point to the role of VAr support in maximizing storage operator profits, a relationship that we further explore in the numerical examples of Section 4.

3.2 Unimodal Storage Dynamics

This subsection provides some properties of the storage dynamics that will be exploited in the numerical examples of Section 4. In particular, we discuss how to prevent the ESS units from simultaneously charging and discharging (i.e., how to

prevent $r_n^c(t)$ and $r_n^d(t)$ from both being positive for the same n and time interval t) in order to ensure that these case studies are well posed. In other words, we want to ensure that profits are not artificially increased by burning energy, which may have economic value, but is unlikely to be beneficial to the overall system.

The linear storage dynamics in (5)–(11) do not explicitly prevent simultaneous charging and discharging, i.e., there is no constraint ensuring that $r_n^c(t)r_n^d(t) = 0$ for all $n \in \mathcal{N}$ and $t \in \mathcal{T}$. Enforcing this condition can be accomplished through the use of mixed integer or nonlinear constraints; e.g., see [1, 39, 53]; however, such constraints are not compatible with the SDR approach that is used to derive our main results. The following theorem provides a condition that ensures $r_n^c(t)r_n^d(t) = 0$ for the storage model in (5)–(11) while maintaining the problem structure assumed in the main results of the work.

Lemma 1 *Consider an ESS unit at bus $n \in \mathcal{N}$ with capacity $C_n > 0$ at the KKT point of the SDR-OPF+S problem. If the Lagrangian multiplier $\lambda_n(t)$ associated with the real power balance (12) at bus $n \in \mathcal{N}$ is nonnegative, i.e., $\lambda_n(t) \geq 0$ for all time intervals $t \in \mathcal{T}$, then $r_n^c(t)r_n^d(t) = 0$ for all $t \in \mathcal{T}$.*

Proof See Appendix 7.1 for the proof, which follows the arguments for a related model in our previous work [12].

The following Lemma 2 states that the condition in Lemma 1 can be enforced by relaxing the real power balance in (12) so that the power inflows can be greater than the power outflows. This new inequality constraint corresponds to allowing the oversatisfaction of loads, which is a common relaxation employed to enforce desirable solution properties in OPF problems; see, e.g., [25, 26, 30]. In practice, this relaxation corresponds to allowing energy spillage.

Lemma 2 *The LMP at bus $n \in \mathcal{N}$ in time interval $t \in \mathcal{T}$ is nonnegative if and only if the load is over-satisfied, i.e., $\lambda_n(t) \geq 0$ if and only if $P_n(t) + P_n^d(t) + r_n^c(t) \leq P_n^g(t) + P_n^w(t) + r_n^d(t)$ for $P_n(t) := \text{tr}\{\Phi_n W(t)\}$ as defined in Appendix 6.2.*

Proof See Appendix 7.2 for the proof, which follows the arguments in our previous work [12].

In the numerical results of the next section, we verify that the condition of Lemma 1 is always satisfied, i.e., $\lambda_n(t) \geq 0$ for all time intervals $t \in \mathcal{T}$, and therefore we do not augment the problem as proposed in Lemma 2.

4 Case Studies

We now present some case studies to illustrate the theoretical results of the previous section. These results also highlight the trade-offs associated with using ESS units to provide both traditional grid services (the supply of real power/energy) and VAr support. In particular, we compare the solution of the problem instance with $z_n(t) = 0$ for all $n \in \mathcal{N}, t \in \mathcal{T}$ (i.e., where the storage is not providing VAr and therefore

$Z_n^{min} = Z_n^{max} = 0$) to the solution of the same problem instance with $z_n(t) \neq 0$ for some n and t . We perform these studies on a 14-bus test system network because its small size makes it an ideal setting to clearly illustrate the theoretical results and investigate the role of VAr support. The validity of Theorem 1 and Lemma 1 was also verified on other datasets and test systems, including the IEEE 118-bus test system [49]; see further details in [12]. In all cases, the results obtained were consistent with the theory presented in Section 3.

All of the results in this section are obtained by solving the SDR-OPF+S problem in (34) and (35) to obtain the global optimum to the OPF+S problem in (16) and (17). We implement the SDR-OPF+S problem in Matlab [33] and solve the resulting semidefinite program with Mosek 7.1 [37] on a 2.2 GHz Intel Core i7 machine with 16 GB 1600 MHz DDR3. In all cases, we verify that our solution corresponds to a rank one $W(t)$ for all $t \in \mathcal{T}$, i.e., that we have also found a globally optimal solution of the OPF+S problem in (16) and (17). We also verify that the condition of Lemma 1 is always satisfied, i.e., $\lambda_n(t) \geq 0$ for all buses $n \in \mathcal{N}$ and time intervals $t \in \mathcal{T}$.

4.1 Test Case Data

The case studies are performed on a test grid that has the topology of the IEEE 14-bus test system [49] but also includes ESS units and wind power plants as shown in Figure 1. We replace the static data from the IEEE 14-bus benchmark system with 24 hour real power demand profiles $P_n^d(t)$ at each bus n based on summer data from 14 Southern California Edison feeders [7]. The demand data is sampled at 30-minute intervals and peak-scaled to match the demand in the test system. The corresponding total aggregate system demand is plotted in Figure 2. We compute the reactive power demand $Q_n^d(t)$ at each bus n and time interval t assuming a power factor of 0.98, i.e., for each bus n and time interval t , i.e.,

$$Q_n^d(t) = \tan(\cos^{-1}(0.98)) P_n^d(t). \quad (33)$$

The real and reactive power bounds along with the corresponding cost function coefficients for the 5 generator buses are provided in Table 1. The per bus generator ramp rates RR_n are set to 15% of the generating unit capacity. We impose line limits S_k^{max} of 80 MVA on the bidirectional flows between buses (1, 2), 40 MVA between buses (1, 5) and (2, 5), and 30 MVA between buses (2, 3) and (2, 4); all of the lines that subject to these capacity constraints are indicated using red markers in Figure 1.

The wind power plants are colocated with the conventional generators at buses 1 and 2. The profiles for the wind power availability $C_n^w(t)$ at each of these buses, shown in Figure 2, are based on data from the NREL western wind resources dataset [38]. This data is peak-scaled so that the total wind power represents 15% of the overall system capacity. More specifically, the total wind power capacity (MW) is

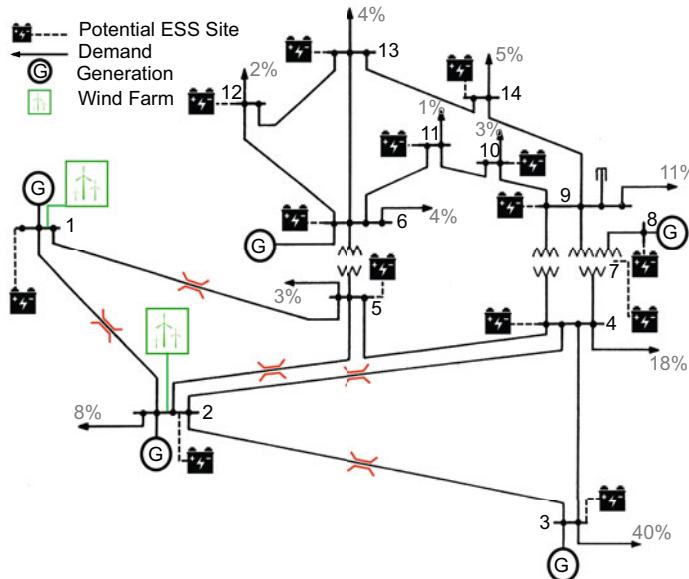
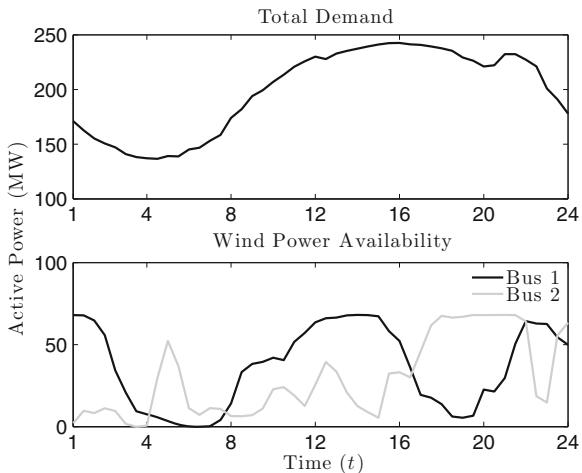


Fig. 1 The IEEE 14-bus benchmark system topology [49] with wind farms, potential ESS sites indicated. The red markings around lines indicate the branches where flow limits are imposed in the numerical examples. The percentage of overall demand is indicated at each bus.

Fig. 2 The top panel shows the total aggregate real power demand (in MW) for the network in Figure 1. The bottom panel shows the wind power availability for the wind farms located at buses 1 and 2. The wind farms each have maximum power availability of 68.15 MW, which represents a total of 15% wind penetration for the test system.



determined as

$$\omega / (1 - \omega) \sum_{n \in \mathcal{G}} P_n^{\max},$$

where $\omega \in (0, 1]$ denotes the wind penetration as a portion of the overall generation capacity. This wind power capacity corresponds to 136.3 MW and is equally divided

Table 1 Conventional generator specifications for the IEEE 14-bus test system obtained from [49, 54].

Bus (n)	P_n^{min} (MW)	P_n^{max} (MW)	Q_n^{min} (MVAr)	Q_n^{max} (MVAr)	$c_{n,2}^g$ (\$/MW ² h)	$c_{n,1}^g$ (\$/MW)
1	0	332.4	0	10	0.043	20
2	0	140	-40	50	0.25	20
3	0	100	0	40	0.01	40
6	0	100	-6	24	0.01	40
8	0	100	-6	24	0.01	40

over the two buses. The reactive power capacity per wind farm is set to 30% of the real power capacity.

We allow ESS units to be placed at any bus and assume a single storage technology with a network aggregate storage capacity of $h = 100 \text{ MW}\Delta t$ where $\Delta t = 30 \text{ minutes}$, i.e., $h = 50 \text{ MWh}$. The storage technology has charging and discharging power rates of 8 MW per time interval and full discharge capability (i.e., $C_n^{min} = 0$). We assume a round trip efficiency of 81% (i.e., $\eta_n^c \times \eta_n^d = 0.81$), which is consistent with a battery technology [11]. For the cases where VAr support is employed, the reactive power range is set to half of the energy storage capacity assuming a full depth of discharge, i.e., $Z_n^{min} = -0.5C_n$ and $Z_n^{max} = 0.5C_n$. Finally, the operational cost of discharging is set to a negligible quantity $c_{n,1}^s = \$0.1/\text{MWh}$, which helps with the conditioning of the optimization problem.

4.2 Numerical Results and Discussion

Table 2 provides the optimal storage allocation and corresponding marginal profits π_n^S for the ESS units at each bus obtained by solving the SDR-OPF+S problem in (34) and (35). We show results for two cases. Case 1, which is labeled *ESS*, corresponds to a simulation where the storage only provides real power to the system, i.e., $z_n(t) = 0$ for all $n \in \mathcal{N}$ and $t \in \mathcal{T}$, and the total profits are computed as (28) since trivially (29) is equal to zero. In Case 2, denoted *ESS+VAr support*, the storage is allowed to provide VAr support, i.e., we allow $z_n(t) \neq 0$ with the bounds given in equation (11), and the total storage operator profit is computed as the sum of (28) and (29). These results demonstrate that the siting decisions change when the storage provides both traditional grid services and reactive power to the system. The results also provide numerical support of Theorem 1 as the storage allocation is directly tied to the marginal profits (either $\pi^{SP}(\cdot)^*$ in Case 1 or $\pi^{SP}(\cdot)^* + \pi^{SQ}(\cdot)^*$ in Case 2). Corollary 1 is also satisfied as all profits are nonnegative.

Table 3 provides the corresponding optimal objective function value (16), real power losses, and total storage profits for the two operating cases. The net system savings of $\$206,164.00 - \$204,697.28 = \$1,466.72$ in total system costs when

Table 2 Nodal storage capacity and profits.

Bus	ESS		ESS+VAr Support	
	c_n (MW/30-min)	$\pi_n^{sP}(\cdot)$ (\$)	c_n (MW/30-min)	$\pi_n^{sP}(\cdot) + \pi_n^{sQ}(\cdot)$ (\$)
1	1.7	20.79	0	0
2	11.4	145.85	6.0	110.87
3	0	0	0	0
4	75.8	553.84	54.1	609.24
5	0	0	36.6	445.84
6–11	0	0	0	0
12	1.0	6.74	0	0
13	3.4	24.09	0	0
14	6.7	46.91	3.3	36.31
Total	100	798.37	100	1,202.39

Table 3 The total system cost, real power losses, and total storage profit.

	Computation	ESS	ESS + VAr Support
Total System Cost (\$)	$\hat{\mathbf{p}}^*$ in (34)	206,164.00	204,697.28
I^2R Losses (MW)	$\sum_{k \in \mathcal{K}, t \in \mathcal{T}} P_{k(\cdot)}^\ell(t)$	151.7	145.1
Total Marginal Profit to ESS (\$)	$\pi^s(\cdot)^*$	798.37	1,202.39

compensating ESS units for VAr support is greater than the $\$1,202.39 - \$798.37 = \$404.02$ increase in total marginal profits for the storage operator(s). These results indicate that when ESS units provide VAr compensation, the total costs and real power losses in the system are lower, i.e., the power system operates more efficiently and the market settlements clear at a lower operating cost. The lower overall power losses arise because the additional reactive power dispatch from ESS units leads to higher overall system voltages, and this operating state requires less current, as illustrated by Figure 3a and b.

Table 4 shows the total costs to consumers and total marginal profits to the generators for cases 1 and 2. Here the total cost to consumers is given by

$$\sum_{t \in \mathcal{T}} \lambda_n(t) P_n^d(t) + \varphi_n(t) Q_n^d(t),$$

and the marginal profit to generators is computed as

$$\sum_{i \in \mathcal{I}(n), t \in \mathcal{T}} \lambda_n(t) P_n^g(t) + \lambda_n(t) P_n^w(t) + \varphi_n(t) Q_n^g(t) - f_n^g(\cdot).$$

The results in Table 4 indicate that the optimal integration of ESSs for both types of services results in a better outcome for all market participants in aggregate, although each individual participant may not have lower costs. For example, the customers at

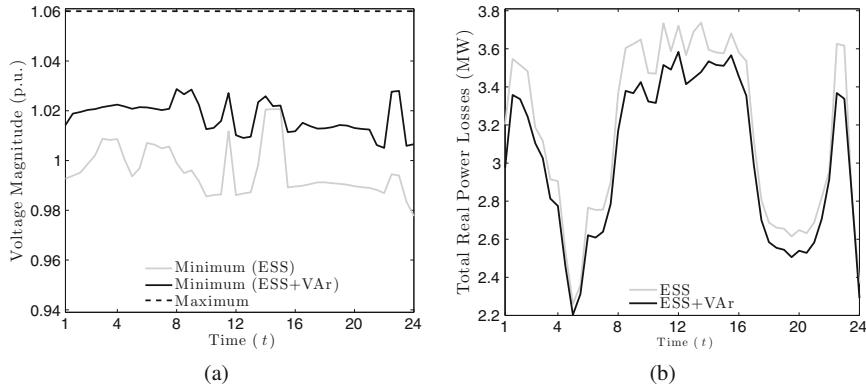


Fig. 3 (a) The maximum and minimum system voltages for the network in Figure 1 for cases 1 and 2. In both cases, the maximum network voltage is 1.06 p.u. (b) The total real power network losses corresponding to the voltage profiles in panel (a).

Table 4 The effect of optimal ESS integration on the costs to consumers.

Bus	Consumer Costs		Generator Marginal Profits	
	ESS (\$)	ESS + VAr Support (\$)	ESS (\$)	ESS + VAr Support (\$)
1	-	-	52,101.34	52,191.70
2	17,600.00	17,661.68	32,932.91	33,008.10
3	154,579.53	154,558.66	1,117.54	1,110.97
4	66,983.00	66,872.91	-	-
5	9,916.90	9,926.44	-	-
6	15,022.78	14,994.88	47.85	10.40
7	-	-	-	-
8	-	-	310.45	354.33
9	38,918.46	38,943.43	-	-
10	11,836.50	11,840.61	-	-
11	5,231.05	5,227.22	-	-
12	8,123.65	8,107.98	-	-
13	16,138.23	16,111.45	-	-
14	19,492.92	19,479.95	-	-
Energy Services (\$)	363,728.02	363,671.68	86,461.39	86,611.41
VAr Support (\$)	114.99	53.54	48.70	64.06
Total (\$)	363,843.02	363,725.22	86,510.09	86,675.47

buses 2, 5, 9, and 10 experience a *slight increase* in costs and generators at buses 3 and 4 experience a *slight decrease* in profits.

The positive effects related to the provision of VAr by ESSs demonstrated in Tables 3 and 4 are not unexpected as allowing the storage to provide VAr support increases overall VAr availability in the system and increases voltage control points.

Table 5 Reactive power capability (power rating) and maximum dispatch, including reactive power produced or absorbed, for the given operating cycle.

Bus	VAr Capability (MVAr)	$\max_t z_{n,t} $ Dispatch (MVAr)	Total VAr Dispatch (MVAr)
1	0	0	0
2	3.0	0.08	1.69
3	0	0	0
4	27.05	27.05	1,298.79
5	18.3	18.3	879.56
6–13	0	0	0
14	1.65	1.63	74.70
Total	50	47.06	2,254.74

Table 6 The minimum and maximum Q-LMPs, the load weighted mean Q-LMPs total Q-LMP payments to storage.

Bus	Nodal Q-LMP			ESS + VAr Support	
	Minimum (\$/MVAr)	Maximum (\$/MVAr)	Load Weighted Mean (\$/MVAr)	Total VAr Dispatch (MVAr)	Total Q-LMP Payments (\$)
2	0	0	0	1.69	0
4	0.058	0.255	0.176	1,298.79	213.61
5	0.086	0.424	0.285	879.56	233.10
14	0	0.321	0.202	74.70	13.22

Prior to ESS integration, the only dynamic reactive power capability is provided by generating units at buses 1, 2, 3, 6, and 8. The ESS units increase the dynamic reactive power capability on the network by 33.8% and introduce new VAr support at buses 4, 5, and 14. The local availability of VAr support is particularly important because reactive power losses across lines tend to be large due to the high reactance characteristic of transmission line conductors. The efficiency of having VAr support at more buses is illustrated in Table 5 which shows that in Case 2 the reactive power capability of the ESS at bus 2 is 97.3% underutilized, whereas the ESS units at buses 4 and 5 are dispatching VAr support at full capacity for the duration of the operating cycle. The variation in how ESS units are best utilized in Case 2 demonstrates that optimal ESS integration is driven by profitability in both real and reactive power services. Table 6 illustrates that Q-LMP payments can be quite substantial depending on the magnitude of the Q-LMP and the amount of VAr dispatched from the ESS (as summarized in Table 5).

For the Q-LMP-based profits used here, the payments for VAr support increase the marginal profits of the storage operator by 50.6%, which indicates that Q-LMP-based payments create substantial incentives for ESS units to provide VAr support in addition to traditional power and energy services. However, the positive system effects illustrated in Table 3 are realized regardless as to whether or not the storage operator is compensated for the VAr support. In fact, it is worth noting that the ESS units are not typically employed to provide VAr support and may not be

Table 7 The total revenues, costs, and marginal profits to ESSs without and with VAr support; the total Q-LMP payments to the storage operator is \$459.93.

Energy Services	ESS	ESS + VAr Support
Revenues (\$)	5,304.55	4,798.65
Costs (\$)	4,506.17	4,056.19
Profits (\$)	798.37	742.46
Reactive Power Compensation		
Revenues (\$)	0	459.93
Costs (\$)	0	0
Profits (\$)	0	459.93
Total Marginal Profits (\$)	798.37	1,202.39

compensated using the Q-LMP as assumed here. However we use this approach for consistency with real power compensation mechanisms. Proposing a payment scheme of some type is motivated by the fact that changes to the makeup of electric power systems is making system operators consider requiring nontraditional resources [16] to provide ancillary services such as VAr support to the grid. Related concerns have made VAr markets an area of growing interest [2, 18, 23, 29, 40, 41], and our results demonstrate that a market that reflects the value of the VAr support provided by ESSs would have a substantial impact on the financial viability of energy storage projects.

The impact of remuneration is illustrated in Table 7, which reports equivalent results to Table 2 for the same case study but breaks out the revenues, costs, and profits based on energy services and reactive power compensation. The results in Table 7 show that the introduction of a market mechanism such as Q-LMP can benefit the storage operator and therefore incentivize the storage operator to help decrease system costs. For the current market design with nodal pricing on the real power dispatch only, i.e., the LMP, the marginal profits to a storage operator without VAr support would be \$798.37, which would then decrease to \$742.46 if VAr support were also provided *but Q-LMP payments are not included*. However with nodal pricing also on the reactive power dispatch, i.e., Q-LMP, the overall marginal profits to the storage operator with VAr support would increase to \$1,202.39. This simple example illustrates that the current market design disincentivizes storage operators from providing VAr support because marginal profits decrease unless the VAr dispatch is remunerated.

In the next section, we take a step toward addressing the question of VAr compensation mechanisms by comparing the Q-LMP-based payment mechanism to the reactive power capability payments used by two Independent System Operators (ISOs) in the Eastern United States: the New York ISO (NYISO) and ISO New England (ISO-NE).

4.3 Payments for VAr Support: Capability or Dispatch Rates?

The case study in Section 4.2 demonstrates that having ESS units which can provide VAr support is beneficial to grid operations. Here we compare our proposed nodal Q-LMP payment mechanism, which pays for reactive power dispatch, to current market mechanisms, which pay for reactive power capability. In this analysis we consider OATT Schedule 2 Rates for NYISO and ISO-NE [14]. Both NYISO and ISO-NE pay qualified units for VAr capability, where the cost is typically allocated to customers based on a load ratio that is measured in terms of real power. The NYISO and ISO-NE VAr capability rates are \$3,919/MVar – year and \$2,190/MVar – year, respectively [14]. Both system operators also make lost opportunity cost (LOC) payments when the qualified unit's real power output is dispatched down for the purpose of providing VAr support. LOC payments are not considered in this study since it is assumed that there is no trade-off between real and reactive power outputs from the power electronics on the ESS unit; however, such trade-offs do occur in practice for conventional generation operating along the border of its capability curve. This trade-off in ESS operations is largely dependent on the configuration of the system and its storage technology. A growing trend for asynchronous generation is to oversize the power converters in order to increase reactive power capacity [6, 52], which may result in little to no trade-off depending upon the local reactive power needs of the system.

Table 8 compares the Q-LMP payments to the NYISO and ISO-NE capability payments. The results show that the NYISO and ISO-NE type VAr capability rates are inconsistent between the system operators. This inconsistency is also observed for the other ISOs including MISO, PJM, and CAISO [14]. The disconnect between these rates and nodal prices can lead to significant oversupply of VAr in some areas and scarcity in others. This can be thought of as a misalignment of system costs, investment and dispatch incentives, and operational needs of the electric power grid. Through pricing the locational and temporal needs for VAr support on the system, the Q-LMP mechanism reflects the marginal value of VAr support as a paid service and is incentive compatible and revenue adequate at a global optimum of the OPF+S problem.

Table 8 Q-LMP payments compared to the effective daily reactive power capability rates for reactive capability in NYISO and ISO-NE.

Bus	Q-LMP (\$/Day)	NYISO Capability Rate (\$/Day)	ISO-NE Capability Rate (\$/Day)
1	0	0	0
2	0.04	32.21	18.00
3	0	0	0
4	213.61	290.43	162.30
5	233.06	196.49	109.80
6–13	0	0	0
14	13.22	17.72	9.90
Total	459.93	536.85	300.00

5 Conclusions

This paper uses an OPF+S framework to study how the provision of VAr support by ESS units affects optimal storage siting and sizing decisions as well as overall system performance. We solve the nonconvex problem through a semidefinite relaxation (SDR) that allows us to exploit strong duality to form a storage operator subproblem from its Lagrangian dual. This storage operator subproblem optimizes profits for the storage operator. Our theoretical results provide an analytical relationship between maximizing storage operator profits and the minimum cost allocation and operation of distributed energy storage systems for the system operator. In particular, we show that there is a natural duality between these problems in a purely competitive market.

We illustrate the theoretical relationships and discuss their impact on siting decisions through a numerical study. Our results show that the inclusion of VAr support greatly changes the optimal allocation of resources in the network. The provision of VAr support also makes the power system operations more efficient, resulting in market settlements that clear at a lower operating cost. The combination of theoretical results and case studies highlight the tight connections between market design and the financial viability of large-scale storage integration in AC power systems.

A direct extension to this work includes sensitivity analysis to better understand the effects of wind and load uncertainty on optimal storage siting and sizing. The optimal storage allocation problem and solution approach discussed herein can also be modified to investigate optimal allocation of renewables and other grid resources, although new theory including proofs of strong duality for the associated subproblem would be required.

Two important directions for future work include developing methods to scale the results to larger test systems and the inclusion of uncertainty. Scaling the approach to large power systems can be accomplished through the use of newly proposed decompositions and scaling techniques for the SDR of the OPF problem, e.g., those discussed in [34, 35], although these would first need to be extended to the multi-period OPF with storage problem. Reformulation as a stochastic optimization model would allow the inclusion of demand and wind uncertainty, as well as include other flexible resources such as price-responsive demand. We could then frame our results in terms of either the expectation or the conditional value-at-risk (CVaR), which would enable us to quantify the effects of resource variability and uncertainty for a range of scenarios. As the stochastic formulation may not provide an exact relaxation, application to problems where strong duality does not hold is an important direction for ongoing work; these cases would require more advanced solution techniques (and extensions of these techniques for multi-period problems with storage) to determine the global optimum; see, e.g., [28, 34].

6 Appendix

6.1 Primal Relaxation

First we introduce parameters to construct the SDR. For the nodal admittance matrix Y we define $\bar{Y}_n := e_n e_n^T Y$ for each bus $n \in \mathcal{N}$, and $\bar{Y}_{k(n,i)}^\ell := (y_n^s + y_{ni}) e_n e_n^T - (y_{ni}) e_n e_i^T$ for each line $k \in \mathcal{K}$, where y_n^s is the shunt component of the admittance y_n and $e_n \in \mathbb{R}^N$ is the standard basis vector. We then define $\Phi_n := h(\bar{Y}_n)$ for $h : \mathbb{C}^{N \times N} \rightarrow \mathbb{R}^{2N \times 2N}$ and $\Phi_k^\ell := h(\bar{Y}_k^\ell)$ for $h : \mathbb{C}^{K \times N} \rightarrow \mathbb{R}^{2N \times 2N}$ where

$$h(\Omega) := \frac{1}{2} \begin{bmatrix} \operatorname{Re}\{\Omega + \Omega^T\} & \operatorname{Im}\{\Omega - \Omega^T\} \\ \operatorname{Im}\{\Omega - \Omega^T\} & \operatorname{Re}\{\Omega + \Omega^T\} \end{bmatrix},$$

and $\operatorname{Re}\{\cdot\}$ and $\operatorname{Im}\{\cdot\}$ denote the real and imaginary parts of their arguments, respectively. Similarly, we define $\Psi_n := \tilde{h}(\bar{Y}_n)$ for $\tilde{h} : \mathbb{C}^{N \times N} \rightarrow \mathbb{R}^{2N \times 2N}$ and $\Psi_k^\ell := \tilde{h}(\bar{Y}_k^\ell)$ for $\tilde{h} : \mathbb{C}^{K \times N} \rightarrow \mathbb{R}^{2N \times 2N}$ where

$$\tilde{h}(\Omega) := -\frac{1}{2} \begin{bmatrix} \operatorname{Im}\{\Omega + \Omega^T\} & \operatorname{Re}\{\Omega^T - \Omega\} \\ \operatorname{Re}\{\Omega^T - \Omega\} & \operatorname{Im}\{\Omega + \Omega^T\} \end{bmatrix}.$$

Using the trace operator $\operatorname{tr}\{\cdot\}$, we determine the active and reactive power injections for each bus $n \in \mathcal{N}$, i.e., $\operatorname{tr}\{\Phi_n W(t)\}$ and $\operatorname{tr}\{\Psi_n W(t)\}$, and the active and reactive power flows on each line $k \in \mathcal{K}$, i.e. $\operatorname{tr}\{\Phi_k^\ell W(t)\}$ and $\operatorname{tr}\{\Psi_k^\ell W(t)\}$, respectively. We also define the coefficient matrix $M_n := e_n e_n^T \oplus e_n e_n^T$ to reference with $\operatorname{tr}\{M_n W(t)\}$ the equivalent terms to the real component of the nodal voltage squared and the imaginary component of the nodal voltage squared for each bus $n \in \mathcal{N}$.

Let \mathbf{W} and $\boldsymbol{\alpha}^g$ denote the decision variable sets $\{W(t)\}_{t \in \mathcal{T}}$ and $\{\alpha_n^g(t)\}_{g \times \mathcal{T}}$, respectively. We refer to the relaxation of OPF+S in (16) and (17) as the SDR-OPF+S:

$$\hat{\mathbf{p}}^* := \min_{\mathbf{W}, \boldsymbol{\alpha}^g, \mathbf{r}^c, \mathbf{r}^d, \mathbf{z}, \mathbf{P}^g, \mathbf{Q}^g, \mathbf{C}} \sum_{n \in \mathcal{N}} \left(\sum_{t \in \mathcal{T}} \alpha_n^g(t) + f_n^s(\cdot) \right) \quad (34)$$

subject to

$$P_n^{\min} \leq P_n^g(t) \leq P_n^{\max} : \lambda_n^{\min}(t), \lambda_n^{\max}(t) \quad (35a)$$

$$Q_n^{\min} \leq Q_n^g(t) \leq Q_n^{\max} : \varphi_n^{\min}(t), \varphi_n^{\max}(t) \quad (35b)$$

$$-RR_n \leq P_n^g(t) - P_n^g(t-1) \leq RR_n : \delta_n^{\min}(t), \delta_n^{\max}(t) \quad (35c)$$

$$\begin{aligned}
0 \leq P_n^w(t) \leq C_n^w(t) & : \xi_n^{min}(t), \xi_n^{max}(t) \quad (35d) \\
C_n^{min} \leq s_n(t) \leq C_n & : \beta_n^{min}(t), \beta_n^{max}(t) \quad (35e) \\
s_n(t) = s_n(t-1) + \eta_n^c r_n^c(t) - r_n^d(t) / \eta_n^d & : \gamma_n(t) \quad (35f) \\
0 \leq r_n^c(t) \leq R_n^c & : \rho_n^{min}(t), \rho_n^{max}(t) \quad (35g) \\
0 \leq r_n^d(t) \leq R_n^d & : \sigma_n^{min}(t), \sigma_n^{max}(t) \quad (35h) \\
\tilde{s}_n(T) - s_n(T) = 0 & : \varkappa_n \quad (35i) \\
Z_n^{min} \leq z_n(t) \leq Z_n^{max} & : \psi_n^{min}(t), \psi_n^{max}(t) \quad (35j) \\
\sum_{n \in \mathcal{N}} C_n \leq h & : \phi \quad (35k) \\
\text{tr}\{\Phi_n W(t)\} = P_n^w(t) + P_n^g(t) - P_n^d(t) - r_n^c(t) + r_n^d(t) & : \lambda_n(t) \quad (35l) \\
\text{tr}\{\Psi_n W(t)\} = Q_n^g(t) - Q_n^d(t) + z_n(t) & : \varphi_n(t) \quad (35m) \\
(V_n^{min})^2 \leq \text{tr}\{M_n W(t)\} \leq (V_n^{max})^2 & : \vartheta_n^{min}(t), \vartheta_n^{max}(t) \quad (35n) \\
\begin{bmatrix} -(S_k^{max})^2 & \text{tr}\{\Phi_k^\ell W(t)\} & \text{tr}\{\Psi_k^\ell W(t)\} \\ \text{tr}\{\Phi_k^\ell W(t)\} & -1 & 0 \\ \text{tr}\{\Psi_k^\ell W(t)\} & 0 & -1 \end{bmatrix} \leq 0 & : \zeta_k(t) \quad (35o) \\
\begin{bmatrix} c_{n1}^g(t) P_n^g(t) - \alpha_n^g(t) \sqrt{c_{n2}^g(t)} P_n^g(t) & \\ \sqrt{c_{n2}^g(t)} P_n^g(t) & -1 \end{bmatrix} \leq 0 & : \nu_n(t) \quad (35p) \\
W(t) \succeq 0 & : \mu(t) \quad (35q)
\end{aligned}$$

with its corresponding dual variables for all $n \in \mathcal{N}$, $k \in \mathcal{K}$, and $t \in \mathcal{T}$. For the reformulated constraints, equations (35l)–(35n) and (35p) apply to each bus $n \in \mathcal{N}$ and time $t \in \mathcal{T}$; equation (35o) applies to each line $k \in \mathcal{K}$ and time $t \in \mathcal{T}$; equation (35q) applies to all $t \in \mathcal{T}$. Note that SDR-OPF+S has a linear cost function (34), linear equality and inequality constraints in equations (35l)–(35n), and linear matrix inequality (LMI) constraints in (35o) and (35p).

The change of variables that transforms (16) and (17) into (34) and (35) follows from the fact that a symmetric matrix $W(t) \in \mathbb{R}^{2N \times 2N}$ is positive semidefinite and rank one if and only if there exists $\omega(t) \in \mathbb{R}^{2N}$ such that $W(t) = \omega(t) \omega(t)^T$ for all $t \in \mathcal{T}$. This condition can be enforced by including the following constraint to the constraint set in SDR-OPF+S:

$$\text{rank}(W(t)) = 1 \quad (36)$$

for all $t \in \mathcal{T}$. However equation (36) is a nonconvex constraint, and therefore we can exclude (36) by analytically determining if the solution to SDR-OPF+S in (34)

and (35) meets the rank one criteria. In the case when \mathbf{W}^* has a rank one solution, $\hat{\mathbf{p}}^*$ is the global optimum of the OPF+S problem, and SDR-OPF+S is equivalent to OPF+S where $\hat{\mathbf{p}}^* = \mathbf{p}^*$, i.e., the SDR is exact.

6.2 Lagrangian Dual

We present the Lagrangian dual formulation, which is the basis for the storage subproblem in (21) and (22) as presented in Section 3. The Lagrangian dual for optimization of (34) and (35) excluding the rank constraint (36) is

$$\hat{\mathbf{d}}^* := \max_{x \geq 0, \lambda, \varphi, \kappa, v, \zeta} g(\cdot) \quad (37)$$

subject to

$$\begin{aligned} & \sum_{n \in \mathcal{N}} \left\{ \lambda_n(t) \Phi_n + \left(\vartheta_n^{max}(t) - \vartheta_n^{min}(t) \right) M_n + \varphi_n(t) \Psi_n \right\} \\ & + \sum_{k \in \mathcal{K}} \left\{ 2\zeta_{k2}(t) \Phi_k + 2\zeta_{k3}(t) \Psi_k \right\} \succeq 0 \end{aligned} \quad (38a)$$

$$\rho_n^{max}(t) - \rho_n^{min}(t) + \lambda_n(t) + \gamma_n(t) \eta^c = 0 \quad (38b)$$

$$\sigma_n^{max}(t) - \sigma_n^{min}(t) - \lambda_n(t) - \gamma_n(t) \left(\eta^d \right)^{-1} + c_{n1}^s(t) = 0 \quad (38c)$$

$$\begin{aligned} & \lambda_n^{max}(t) - \lambda_n^{min}(t) - \lambda_n(t) + \delta_n^{max}(t) - \delta_n^{min}(t) \\ & - \delta_n^{max}(t-1) + \delta_n^{min}(t-1) + c_{n1}^g(t) + 2\sqrt{c_{n2}^g(t)}v_{n1}(t) = 0 \end{aligned} \quad (38d)$$

$$\varphi_n^{max}(t) - \varphi_n^{min}(t) - \varphi_n(t) = 0 \quad (38e)$$

$$\psi_n^{max}(t) - \psi_n^{min}(t) - \varphi_n(t) = 0 \quad (38f)$$

$$\xi_n^{max}(t) - \lambda_n(t) = 0 \quad (38g)$$

$$\gamma_n(t) - \gamma_n(t-1) + \beta_n^{max}(t) - \beta_n^{min}(t) + \kappa_n = 0 \quad (38h)$$

$$\phi - \beta_n^{max}(t) = 0 \quad (38i)$$

$$\xi_k(t) := \begin{bmatrix} \zeta_{k1}(t) & \zeta_{k2}(t) & \zeta_{k3}(t) \\ \zeta_{k2}(t) & \zeta_{k4}(t) & \zeta_{k5}(t) \\ \zeta_{k3}(t) & \zeta_{k5}(t) & \zeta_{k6}(t) \end{bmatrix} \succeq 0 \quad (38j)$$

$$v_n(t) := \begin{bmatrix} 1 & v_{n1}(t) \\ v_{n1}(t) & v_{n2}(t) \end{bmatrix} \succeq 0 \quad (38k)$$

where the cost function in (37) is

$$\begin{aligned}
g(\cdot) := & -\phi h - \sum_{n \in \mathcal{N}} \varkappa_n \tilde{s}_n(T) - \sum_{t=2}^T \sum_{n \in \mathcal{N}} R R_n \left(\delta_n^{min}(t) + \delta_n^{max}(t) \right) \\
& + \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{N}} \left\{ \lambda_n(t) P_n^d(t) + \varphi_n(t) Q_n^d(t) + c_{n1}^s(t) + \lambda_n^{min}(t) P_n^{min} - \lambda_n^{max}(t) P_n^{max} \right. \\
& + \varphi_n^{min}(t) Q_n^{min} - \varphi_n^{max}(t) Q_n^{max} + \vartheta_n^{min}(t) (V_n^{min})^2 \\
& - \vartheta_n^{max}(t) (V_n^{max})^2 + \beta_n^{min}(t) C_n^{min} \\
& \left. - \xi_n^{max}(t) C_n^w(t) - \rho_n^{max}(t) R^c - \sigma_n^{max}(t) R^d + \psi_n^{min}(t) Z^{min} - \psi_n^{max}(t) Z^{max} \right\} \\
& + \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{G}} \left(c_{n1}^g(t) + 2\sqrt{c_{n2}^g(t)} v_{n1}(t) - v_{n2}(t) \right) \\
& + \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{K}} \left(\zeta_{k1}(t) (S_k^{max})^2 + \zeta_{k4}(t) + \zeta_{k6}(t) \right)
\end{aligned} \tag{39}$$

with $n \in \mathcal{N}$, $k \in \mathcal{K}$, and $t \in \mathcal{T}$. For equation (38d), we drop the $\delta_n^{min}(t-1)$ and $\delta_n^{max}(t-1)$ terms when $t = 1$. For equation (38h), $\gamma_n(0)$ and $\gamma_n(T)$ are unrestricted.

Accordingly, let $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\varkappa}$ denote the respective sign indefinite dual variable sets $\{\lambda_n(t)\}_{\mathcal{N} \times \mathcal{T}}$, $\{\varphi_n(t)\}_{\mathcal{N} \times \mathcal{T}}$, $\{\gamma_n(t)\}_{\mathcal{N} \times \mathcal{T}}$, and $\{\varkappa_n\}_{\mathcal{N}}$. Also let $\boldsymbol{\xi}$ and \boldsymbol{v} denote the respective Lagrange multiplier matrix sets $\{\zeta_k(t)\}_{\mathcal{K} \times \mathcal{T}}$ and $\{v_n(t)\}_{\mathcal{G} \times \mathcal{T}}$ that are associated with the LMI constraints in (35o) and (35p), respectively. Without loss of generality let \boldsymbol{x} denote the remaining nonnegative decision variable sets.

Therefore we define the Lagrangian dual problem LD-OPF+S, which is a convex problem, by equations (37) and (38). By solving for $\hat{\mathbf{d}}^*$, we can determine the tightest lower bound for all the primal variables in SDP-OPF+S. Moreover since the primal problem SDP-OPF+S is also convex, the KKT conditions are sufficient for the solution to be both primal and dual optimal. Gayme and Topcu prove that, excluding the rank constraint, strong duality holds by Slater's condition [20].

7 Appendix

7.1 Proof of Lemma 1

Proof At the KKT point for the LD-OPF+S in (37) and (38), the storage variables C_n , $s_n(t)$, $r_n^c(t)$, $r_n^d(t)$, and $z_n(t)$ provide a feasible solution for every bus $n \in \mathcal{N}$.

By construction, a rank one solution to the SDR-OPF+S problem satisfies the constraints (5)–(12), the associated complementary slackness conditions:

$$\gamma_n(t) [s_n(t-1) - s_n(t) + \eta_n^c r_n^c(t) - r_n^d(t)/\eta_n^d] = 0, \quad (40)$$

$$\rho_n^{min}(t) r_n^c(t) = 0, \quad (41)$$

$$\rho_n^{max}(t) [r_n^c(t) - R_n^c] = 0, \quad (42)$$

$$\sigma_n^{min}(t) r_n^d(t) = 0, \quad (43)$$

$$\sigma_n^{max}(t) [r_n^d(t) - R_n^d] = 0, \quad (44)$$

$$\psi_n^{min}(t) [Z_n^{min} - z_n(t)] = 0, \quad (45)$$

$$\psi_n^{max}(t) [z_n(t) - Z_n^{max}] = 0, \quad (46)$$

$$\beta_n^{min}(t) [C_n^{min} - s_n(t)] = 0, \quad (47)$$

$$\beta_n^{max}(t) [s_n(t) - C_n] = 0, \quad (48)$$

$$\phi [\sum_{n \in \mathcal{N}} C_n - h] = 0, \quad (49)$$

$$\varkappa_n [\tilde{s}_n(T) - s_n(T)] = 0, \quad (50)$$

$$\lambda_n(t) [\tilde{P}_n(t) + r_n^c(t) - r_n^d(t)] = 0, \quad (51)$$

for all $t \in \mathcal{T}$ where $\tilde{P}_n(t) = P_n^g(t) + P_n^w(t) - \text{tr}\{\Phi_n W(t)\} - P_n^d(t)$, and the respective zero gradient conditions of C_n , $s_n(t)$, $r_n^c(t)$, $r_n^d(t)$, and $z_n(t)$ which are

$$\phi - \beta_n^{max}(t) = 0 \quad (52)$$

$$\gamma_n(t) - \gamma_n(t-1) + \beta_n^{max}(t) - \beta_n^{min}(t) + \varkappa_n = 0 \quad (53)$$

$$\rho_n^{max}(t) - \rho_n^{min}(t) + \lambda_n(t) + \eta_n^c \gamma_n(t) = 0 \quad (54)$$

$$\sigma_n^{max}(t) - \sigma_n^{min}(t) - \lambda_n(t) - \gamma_n(t)/\eta_n^d + c_{n,1}^s = 0 \quad (55)$$

$$\psi_n^{max}(t) - \psi_n^{min}(t) - \varphi_n(t) = 0 \quad (56)$$

for all $t \in \mathcal{T}$. From the zero gradient conditions, when $\lambda_n(t) \geq 0$ the ESS dynamics at bus n are given by

$$\begin{aligned} & -c_{n,1}^s + (\rho_n^{min}(t) - \rho_n^{max}(t)) \\ & + (\sigma_n^{min}(t) - \sigma_n^{max}(t)) + (1/\eta_n^d - \eta_n^c) \gamma_n(t) \geq 0. \end{aligned} \quad (57)$$

Finally we note that for any storage technology with efficiencies $\eta_n^c, \eta_n^d \in (0, 1]$, $(1/\eta_n^d - \eta_n^c) \geq 0$.

We now demonstrate that whenever $\lambda_n(t) \geq 0$, either $r_n^c(t)$ or $r_n^d(t)$ is zero, so $r_n^c(t)r_n^d(t) = 0$ implicitly holds for every time interval $t \in \mathcal{T}$.

This fact can be seen by observing that the following properties hold for all $t \in \mathcal{T}$ at the KKT optimal point when $\lambda_n(t) \geq 0$:

1. If $\gamma_n(t) > 0$, then from (41) and (54) we have that $\rho_n^{min}(t) > \lambda_n(t) \geq 0$. As a result $r_n^c(t) = 0$, i.e., the storage is not charging.
2. If $\gamma_n(t) \leq 0$, then either $\rho_n^{min}(t) > 0$, $\sigma_n^{min}(t) > 0$, or both for the condition in (57) to hold. Then (41) and (43), respectively, imply that either $r_n^c(t) = 0$ or $r_n^d(t) = 0$, i.e., the storage is not charging, the storage is not discharging, or the storage is idle if both criteria hold.

7.2 Proof of Lemma 2

Proof (\Leftarrow) The real power balance constraint in (12) for $P_n(t) = \text{tr}\{\Phi_n W(t)\}$ is equivalent to the following two inequalities:

$$P_n(t) + P_n^d(t) + r_n^c(t) \leq P_n^g(t) + P_n^w(t) + r_n^d(t) \quad (58)$$

$$P_n(t) + P_n^d(t) + r_n^c(t) \geq P_n^g(t) + P_n^w(t) + r_n^d(t) \quad (59)$$

for all $t \in \mathcal{T}$, $n \in \mathcal{N}$ where $\lambda_n^+(t)$ is the dual variable to (58), $\lambda_n^-(t)$ is the dual variable to (59), and $\lambda_n^+(t), \lambda_n^-(t) \geq 0$. The dual variable $\lambda_n(t)$ to (12) is sign indefinite where $\lambda_n(t) = \lambda_n^+(t) - \lambda_n^-(t)$.

Dual feasibility of equations (58) and (59) requires that

$$\lambda_n^+(t) - \lambda_n^-(t) \geq 0. \quad (60)$$

This implies that if $\lambda_n^+(t) \geq \lambda_n^-(t)$, then (58) holds.

(\Rightarrow) Assuming we replace the equality constraint in (12) by the inequality

$$P_n(t) + P_n^d(t) + r_n^c(t) \leq P_n^g(t) + P_n^w(t) + r_n^d(t), \quad (61)$$

and suppose $\hat{P}_n^g(t)$, $\hat{r}_n^c(t)$, $\hat{r}_n^d(t)$, and $\hat{P}_n(t)$ results in a feasible solution to the SDR-OPF+S primal problem in (34) and (35) and LD-OPF+S dual problem in (37) and (38). By dual feasibility, the dual problem gives a nontrivial lower bound on the OPF+S only when (61) is binding, which implies $\lambda_n(t) \geq 0$.

Acknowledgment This work was partially supported through the National Science Foundation (grant numbers ECCS 1230788 and OISE 1243482) and Sandia National Laboratories' Laboratory Directed Research and Development (LDRD) program. The authors would also like to thank Benjamin Hobbs for useful discussions.

References

1. Abbaspour M et al (2013) Optimal operation scheduling of wind power integrated with compressed air energy storage. *Renew Energy* 51:53–59
2. Ahmadi H, Foroud AA (2016) Improvement of the simultaneous active and reactive power markets pricing and structure. *IET Gener Transm Distrib* 10(1):81–92
3. Akhil AA, Huff G, Currier AB, Kaun BC, Rastler DM, Chen SB, Cotter AL, Bradshaw DT, Gauntlett WD (2013) DOE/EPRI 2013 electricity storage handbook in collaboration with NRECA. Tech. Rep. SAND2013-5131, Sandia National Laboratories
4. Atwa YM, El-Saadany EF (2013) Optimal allocation of ESS in distribution systems with a high penetration of wind energy. *IEEE Trans Power Syst* 25(4):1815–1822
5. Bai X, Wei H, Fujisawa K, Wang Y (2008) Semidefinite programming for optimal power flow problems. *Int J Electr Power Energy Syst* 30(6–7):383–392
6. Boldea I (2015) Variable speed generators. CRC Press, Boca Raton, FL
7. Bose S, Gayme DF, Topcu U, Chandy KM (2012) Optimal placement of energy storage in the grid. In: Proceeding of the 51st IEEE conference on decision and control, Maui, HI, pp 5605–5612
8. Bose S, Gayme D, Chandy K, Low S (2015) Quadratically constrained quadratic programs on acyclic graphs with application to power flow. *IEEE Trans Control Netw Syst* 2(3):278–287. <https://doi.org/10.1109/TCNS.2015.2401172>
9. Bragard M, Soltau N, Thomas S, Doncker RWD (2010) The balance of renewable sources and user demands in grids: power electronics for modular battery energy storage systems. *IEEE Trans Power Electron* 25(12):3049–3056
10. Castillo A, Gayme DF (2013) Profit maximizing storage allocation in power grids. In: Proceeding of the 52nd IEEE conference on decision and control, Firenze, pp 429–435
11. Castillo A, Gayme DF (2014) Grid-scale energy storage applications in renewable energy integration: a survey. *Energy Convers Manag* 87:885–894
12. Castillo A, Gayme DF (2017) Evaluating the effects of real power losses in optimal power flow based storage integration. *IEEE Trans Control Netw Syst*. <https://doi.org/10.1109/TCNS.2017.2687819>
13. Chandy KM, Low S, Topcu U, Xu H (2010) A simple optimal power flow model with energy storage. In: Proceeding of the 49th IEEE conference on decision and control, pp 1051–1057
14. Commission Staff Report: Payment for reactive power. Tech. Rep. AD14-7, Federal Energy Regulatory Commission (2014)
15. Dvijotham K, Backhaus S, Chertkov M (2011) Operations-based planning for placement and sizing of energy storage in a grid with a high penetration of renewables. Tech. rep., Los Alamos National Lab
16. Ellis A, Nelson R, Engeln EV, Walling R, MacDowell J, Casey L, Seymour E, Peter W, Barker C, Kirby B, Williams JR: Review of existing reactive power requirements for variable generation. In: IEEE power and energy society general meeting, pp 1–7 (2012)
17. Eyer J, Corey G (2010) Energy storage for the electricity grid: benefits and market potential assessment guide. Tech. rep., Sandia National Laboratories
18. Farahani HF, Shayanfar HA, Ghazizadeh MS (2014) Modeling of stochastic behavior of plug-in hybrid electric vehicle in a reactive power market. *Electr Eng* 96(1):1–13
19. Gabash A, Li P (2012) Active-reactive optimal power flow in distribution networks with embedded generation and battery storage. *IEEE Trans Power Syst* 27(4):2026–2035
20. Gayme DF, Topcu U (2013) Optimal power flow with large-scale storage integration. *IEEE Trans Power Syst* 28(2):709–717
21. Ghofrani M, Arabali A, Etezadi-Amoli M, Fadali MS (2013) A framework for optimal placement of energy storage units within a power system with high wind penetration. *IEEE Trans Sustainable Energy* 4(2):434–442

22. Gopalakrishnan A, Raghunathan AU, Nikovski D, Biegler LT (2013) Global optimization of multi-period optimal power flow. In: Proceeding of the American control conference, Washington, DC, pp 1157–1164
23. Homaee O, Jadid S (2014) Investigation of synchronous generator in reactive power market – an accurate view. *IET Gener Transm Distrib* 8(11):1881–1890
24. Hu Z, Jewell WT (2011) Optimal power flow analysis of energy storage for congestion relief, emissions reduction, and cost savings. In: 2011 IEEE/PES systems Conference and exposition. Phoenix, AZ
25. Irving MR, Sterling MJH (1985) Economic dispatch of active power by quadratic programming using a sparse linear complementary algorithm. *Electr Power Energy Syst* 7:2–6
26. Jabr RA, Coonick AH, Cory BJ (2002) A primal-dual interior point method for optimal power flow dispatching. *IEEE Trans Power Syst* 17(3):654–662
27. Jenkins N, Allan R, Crossley P, Kirschen D, Strbac G (2000) Technical impacts of embedded generation on the distribution system. In: Embedded generation, pp 11–12. The Institution of Electrical Engineers, London
28. Josz C, Maeght J, Panciatici P, Gilbert JC (2015) Application of the moment-SOS approach to global optimization of the OPF problem. *IEEE Trans Power Syst* 30(1):463–470
29. Kargarian A, Raooofat M, Mohammadi M (2011) Reactive power market management considering voltage control area reserve and system security. *Appl Energy* 88(11):3832–3840
30. Lavaei J, Low SH (2012) Zero duality gap in optimal power flow problem. *IEEE Trans Power Syst* 27(1):92–107
31. Low SH (2014) Convex relaxation of optimal power flow; Part I: formulations and equivalence. *IEEE Trans Control Netw Syst* 1(1):15–27
32. Madani R, Sojoudi S, Lavaei J (2015) Convex relaxation for optimal power flow problem: mesh networks. *IEEE Trans Power Syst* 30(1):199–211
33. MATLAB (2014) version 8.3.0.532 (R2014a). The MathWorks Inc.
34. Molzahn DK, Hiskens IA (2015) Sparsity-exploiting moment-based relaxations of the optimal power flow problem. *IEEE Trans Power Syst* 30(6):3168–3180
35. Molzahn DK, Holzer J, Lesieutre B, DeMarco C (2013) Implementation of a large-scale optimal power flow solver based on semidefinite programming. *IEEE Trans Power Syst* 28(4):3987–3998
36. Momoh JA (2001) Electric power system applications of optimization. Marcel Dekker, New York
37. MOSEK ApS (2015) The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28). <http://docs.mosek.com/7.1/toolbox/index.html>
38. National Renewable Energy Laboratory (2006) Western wind resources dataset. http://wind.nrel.gov/Web_nrel/
39. Rebennack S, Kallrath J, Pardalos PM (2011) Optimal storage design for a multi-product plant: a non-convex MINLP formulation. *Comput Chem Eng* 35:255–271
40. Rueda-Medina AC, Padilha-Feltrin A (2013) Distributed generators as providers of reactive power support—a market approach. *IEEE Trans Power Syst* 28(1), 347–370
41. Saraswat A, Saini A, Saxena AK (2013) A novel multi-zone reactive power market settlement model: a pareto-optimization approach. *Energy* 51(1):85–100
42. Scaini V (2012) Grid support stability for reliable, renewable power. Tech. Rep., WP083002EN, Eaton Corporation
43. Sioshansi R, Denholm P, Jenkin T (2012) Market and policy barriers to deployment of energy storage. *Econ Energy Environ Policy* 1(2):47–63
44. Smith SC, Sen PK, Kroposki B, Malmedal K (2010) Renewable energy and energy storage systems in rural electrical power systems: issues, challenges and application guidelines. In: Proceeding of the IEEE Rural Electric Power Conference (REPC), pp B4-1–B4-7
45. Sojoudi S, Lavaei J (2012) Physics of power networks makes hard optimization problems easy to solve. In: Proceeding of the IEEE PES general meeting

46. Sojoudi S, Lavaei J (2013) Convexification of generalized network flow problem with application to power systems. In: Proceeding of 52nd IEEE conference on decision and control, pp 7552–7559
47. Thrampoulidis C, Bose S, Hassibi B (2013) Optimal placement of distributed energy storage in power networks. Tech. rep., Caltech
48. Tripathy S (1997) Improved load-frequency control with capacitive energy storage. *Energy Convers Manag* 38(6):551–562
49. University of Washington (1993) Power systems test case archive. <http://www.ee.washington.edu/research/pstca>
50. Warrington J, Goulart P, Mariéthoz S, Morari M (2012) A market mechanism for solving multi-period optimal power flow exactly on AC networks with mixed participants. In: Proceeding of the American control conference, Montreal, QC, pp 3101–3107
51. Wogrin S, Gayme DF (2014) Optimizing storage siting, sizing and technology portfolios in transmission-constrained networks. Preprint
52. Wu B, Lang Y, Zargari N, Kouro S (2011) Power conversion and control of wind energy systems. Wiley, New York
53. Tan Z et al (2014) A two-stage scheduling optimization model and solution algorithm for wind power and energy storage system considering uncertainty and demand response. *Electr Power Energy Syst* 63:1057–1069
54. Zimmerman RD, Murillo-Sánchez CE, Thomas RJ (2011) MATPOWER: Steady-state operations, planning and analysis tools for power systems research and education. *IEEE Trans Power Syst* 26(1):12–19

Virtual Inertia Placement in Electric Power Grids



Bala Kameshwar Poolla, Dominic Groß, Theodor Borsche, Saverio Bolognani, and Florian Dörfler

Abstract The past few years have witnessed a steady shift in the nature of power generation worldwide. While the share of renewable-based distributed generation has been on the rise, there has also been a decline in the conventional synchronous-based generation. The renewable-based power generation interfaced to the grid via power-electronic converters, however, does not provide rotational inertia, an inherent feature of synchronous machines. This absence of inertia has been highlighted as the prime source for the increasing frequency violations and severely impacting grid stability. As a countermeasure, virtual or synthetic inertia and damping emulated by advanced control techniques have been proposed. In this chapter, we study the optimal placement and tuning of these devices. We discuss two approaches based on the control notion of \mathcal{H}_2 system gain characterizing the amplification of a disturbance and the spectral notion of pole-placement. A comprehensive analysis accompanied by iterative gradient-based algorithms is presented for both the approaches and validated on a three-area test case for comparison.

1 Introduction

Modern power systems are experiencing many transformations and are facing unprecedented challenges in order to accommodate the shift from classical synchronous generation to power electronic-based nonsynchronous generation (typically from renewable power sources).

In this scenario, the most pressing issue that grid operators need to tackle is the loss of frequency stability of the grid [17, 29]. In fact, by retiring synchronous machines, the locally available system inertia decreases. System inertia, as a global parameter, represents the capability to store and inject kinetic energy to the grid.

B. K. Poolla · D. Groß · T. Borsche · S. Bolognani · F. Dörfler (✉)

ETH Zurich, 8092 Zurich, Switzerland

e-mail: bpoolla@control.ee.ethz.ch; gross@control.ee.ethz.ch; tborsche@gmx.net;
bsaverio@control.ee.ethz.ch; dorfler@ethz.ch

Lower inertia means larger frequency fluctuations following a disturbance [33], i.e., any event that causes power imbalance in the grid: disconnection of a generator, sudden drop in power injection from a renewable (uncontrollable) source, tie line faults, grid splits, etc. Even as of today, the deteriorating effects of low-inertia levels on the system frequency and related incidents are being observed by transmission system operators worldwide [1, 27, 31].

1.1 Quantification of Frequency Stability

The amount of inertia available in the network directly affects the rate of change of frequency (RoCoF) at the grid buses in the instants that immediately follow a large disturbance, as depicted in upper panel of Figure 1 (see, e.g., the numerical investigation in [27] for a quantification of this relation). Based on this understanding, RoCoF is typically adopted as the main metric to evaluate the robustness of the system in terms of frequency stability, for the following reasons.

- Steep changes in generator frequencies (i.e., large RoCoF at the bus where the generators are connected) are poorly tolerated by the prime movers of power plants, leading to a higher probability of further disconnections and ultimately to cascading events.
- Immediately after a fault, the grid frequency will be substantially different at different buses. Therefore, larger RoCoF translates into potentially larger voltage angle differences across power lines and therefore higher probability of protection tripping and even network splitting.
- Moreover, as the governor response of the generators (primary frequency regulation) does not act until seconds after an incident, the RoCoF directly affects the lowest frequency reached by the grid (the frequency *nadir*, in Figure 1). A low-frequency peak can lead to the disconnection of generators, load shedding intervention, and is therefore dangerous for system stability.

Based on historical observations of exceptional events in the continental European grid, operators have derived recommendations regarding the maximum allowed values for the RoCoF (typically of the order of 500 mHz/s up to 1 Hz/s).

However, the robustness of the system frequency against power imbalance disturbances can be also assessed and quantified via different metrics. For example, if faster primary control mechanisms are deployed (e.g., by exploiting the flexibility of the power converters, or the smart loads available in the grid, or battery storage as in [16]), then the time scale separation between primary control and inertial response of the grid may become less sharp. In such a scenario, the frequency nadir should be explicitly evaluated (and not indirectly, via the RoCoF, which falls short in describing the entire response curve).

Another approach for the assessment of frequency stability consists of evaluating a *signal norm* for the post-fault frequency response. As depicted in Figure 1, the

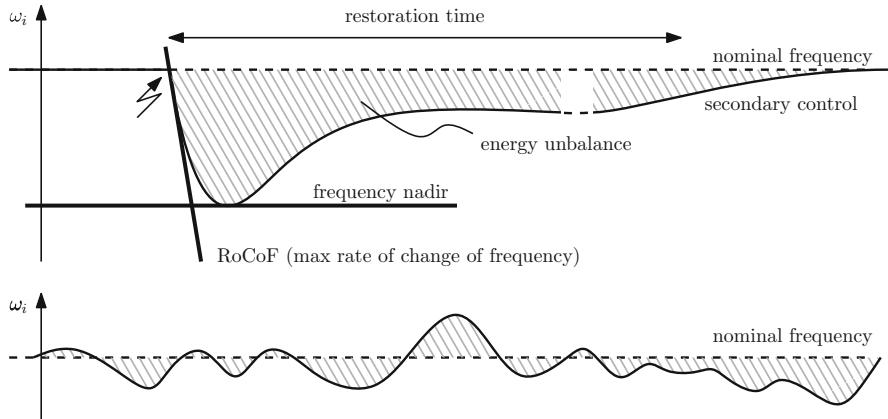


Fig. 1 Schematic representation of the frequency response at a generic grid bus i , caused by a large disturbances, such as generator faults and network splits (upper panel), and by smaller persistent disturbances, such as fluctuating power generation from renewables (lower panel)

total area between the frequency evolution and the steady-state post-disturbance frequency describes how promptly (and efficiently) the bus frequency returns to its nominal operating conditions after a disturbance. By selecting a proper signal norm, we can evaluate the norm of these transients for all the grid buses, therefore obtaining a meaningful aggregate performance metric that can be interpreted as the total *energy unbalance* caused by the disturbance.

Interestingly, the same signal norm is informative also in another scenario. When a large number of renewable sources are connected to the grid, power imbalance will not only be caused by large (although rare) events, such as the loss of a generator, but also by sudden unpredicted fluctuations of uncontrollable sources such as wind and solar. The size of this latter class of events is expected to be smaller, but their occurrence is quite more frequent. For analysis purposes, one should consider a persistent disturbance on the power infeed of the buses where renewable sources are connected. In this case, the aforementioned signal norm would describe the amplification gain between these disturbances and the resulting fluctuations in the grid frequencies at all the buses, as depicted in the bottom panel of Figure 1 for a generic grid bus i .

In this chapter, we consider all these metrics, in order to build a deeper understanding of the effects of lower grid inertia and of the related frequency stability phenomena. We present their mathematical formulation in Section 3. The resulting analysis and methodologies have the advantage of remaining valid also for the future scenarios that we identified, namely, grids with faster primary frequency regulation loops and systems that host large amounts of fluctuating power sources.

1.2 Virtual Inertia and Damping

Considering the role played by system inertia in the stability of the grid, it is not surprising that inertia has been recognized as a key ancillary service. To overcome the lack of inertia provided by synchronous generators, different technologies have been proposed to provide virtual (or synthetic) inertia.

A number of control schemes have been designed in order to make power converters behave as closely as possible to synchronous machines [15, 37]. These schemes range from simple proportional-derivative droop control laws [30] up to the more complex control schemes reviewed in [5] and [8] under the name of *virtual synchronous generators*. All these strategies require some amount of energy storage (to play the role of the missing rotor kinetic energy), which could be batteries [35], super-capacitors, or flywheels.

On the other hand, specialized control schemes have been proposed for those power sources in which some kinetic energy is available, although not synchronously with the grid frequency, notably wind turbines [21] and diesel generators [32]. For these sources, and in particular for wind turbines interconnected to the grid via doubly fed induction generators, the power converters can be controlled in order to mimic the inertial and damping response of a synchronous machine, i.e., proportional to the rate of change of frequency and to the frequency observed by a PLL, respectively. See, for example, the solutions proposed in [2, 10, 11, 14]. Interestingly, it is also possible to control these power converters so that an inertial response is induced, without relying on (possibly destabilizing) PLL measurements [36].

Given the maturity of these solutions, in this work we assume that synthetic inertia and damping can be deployed in the grid, and we are agnostic with respect to the specific technology and energy storage solution that is adopted.

1.3 Where to Deploy Virtual Inertia and Damping?

The authors in [33] recognized that the detrimental effects of reduced system inertia are worsened by spatially heterogeneous inertia profiles. In other words, not only is the total amount of system inertia directly connected to the frequency stability and robustness of the grid but also its specific location.

Given the opportunity of placing synthetic inertia and damping in the grid, and therefore of deciding its location, in this chapter we focus on the fundamental problem of “where to optimally place synthetic inertia and damping”. Different authors considered this problem of optimally placing and tuning of virtual inertia controllers based on either spectral performance metrics [6, 7, 13, 25] or system norms [12, 19, 23, 24].

In this chapter, we consider and compare two algorithms toward this goal. Both algorithms aim at tuning the parameters and the location of the synthetic inertia and damping devices available in the grid, in order to optimize some performance metric of the frequency stability of the system. In the first algorithm, presented in Section 4.1, we consider the amplification gain from disturbances to frequency fluctuations of the synchronous machines and control effort of the synthetic inertia devices. In the second algorithm, presented in Section 4.2, the performance metric is a weighted combination of time-domain indices (RoCoF and frequency nadir) and spectral parameters (damping ratio).

We analyze the performance of the two algorithms for a three-area test system in Section 5, where we compare the resulting spatial allocations of synthetic inertia and damping, the post-fault time-domain response of the frequency at generator buses, and a set of system performance indices.

2 Model

2.1 Synchronous Machines

A common, simple model to assess dynamic phenomena in power systems is the swing equation. It models each generator i with two dynamic states, angle θ_i and frequency ω_i . The dynamics of generators are assumed to be dominated by the rotational inertia, and voltages are assumed constant; see [18, 28] for a detailed derivation. We further linearize over the current steady state and assume that the mechanical input to the generator stays constant over the time scale of interest. The differential equation describing the dynamics of the phase angles at each generator bus is then

$$m_i \dot{\omega}_i = -d_i \omega_i + p_{\text{mech},i} + p_{\text{el},i}, \quad (1)$$

with m_i the inertia and d_i the damping of the generator. The term $p_{\text{mech},i}$ represents changes in the mechanical torque on the machine, while $p_{\text{el},i}$ represents changes in the electrical torque, including line flows to neighboring buses, bus power injections, and local disturbances.

Moreover, we also use (1) with $p_{\text{mech},i} = 0$ to model dynamics of the voltage phase angles of a load bus with index i . Specifically, a small inertia constant m_i at the load buses is used to model fast initial transients in the angle and frequency of the load buses after a disturbance, and d_i represents the typical load damping. Using $m_i = 0$ for the load buses results in the well-known frequency-dependent load model with damping [3]. The differences between the two models become negligible for a small enough m_i at the load buses. Moreover, letting both $m_i = 0$ and $d_i = 0$ for the load buses results in an implicit formulation of the frequency divider [20]. It should also be noted that, in some cases, using the load inertia $m_i = 0$

will result in a more involved formulation of the optimization problems presented in Section 4. In the remainder, we consider the more general case of a small load bus inertia constants $m_i > 0$.

The load buses and generator buses are connected via power lines, described by the graph Laplacian L (the bus susceptance matrix of the grid). Under small-signal DC power flow assumptions, the electric torque term p_{el} can then be linearized as

$$p_{\text{el}} = -L\theta + p,$$

where the i -th element p_i of p represents the change in electric power injection at bus i .

The system dynamics can be written as

$$\begin{bmatrix} \dot{\theta} \\ \dot{\omega} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & I \\ -M^{-1}L & -M^{-1}D \end{bmatrix}}_{A_0} \underbrace{\begin{bmatrix} \theta \\ \omega \end{bmatrix}}_{x_0} + \underbrace{\begin{bmatrix} 0 \\ M^{-1} \end{bmatrix}}_{B_0} (p_{\text{mech}} + p), \quad (2)$$

where the diagonal matrices M and D contain the inertia and damping coefficients m_i and d_i , respectively. We denote the state of system (2), which contains all generator angles and frequencies, by x_0 .

2.2 Governors

The swing equation model is best suited for small-signal disturbance analysis. As we investigate constant disturbances that bring the system to a new steady state, we need to extend the model of each synchronous machine with a governor model. Conventionally, the turbine and governor are modeled by a first-order low-pass filter [18] of the form

$$p_{\text{gov},i} = -\frac{K_{\text{gov},i}}{T_{\text{gov},i} s + 1} \omega_i. \quad (3)$$

As a state-space representation for each governor, we adopt the form

$$A_{\text{gov},i} = -\frac{1}{T_{\text{gov},i}}, \quad B_{\text{gov},i} = \frac{1}{T_{\text{gov},i}}, \quad C_{\text{gov},i} = -K_{\text{gov},i}, \quad (4)$$

which constitute the diagonal elements of the aggregate state-space representation matrices A_{gov} , B_{gov} , and C_{gov} . The state x_{gov} of the aggregated system is a low-pass-filtered version of the generator frequencies (the input of the system), while the output p_{gov} is fed into the mechanical power control term p_{mech} in (2).

2.3 Virtual Inertia and Damping as a Feedback Control Loop

Virtual inertia and damping devices are abstracted as local feedback control loops. Each such device receives bus frequency ω_i as input and feeds power $p_{v,i}$ into the system according to the transfer function

$$p_{v,i} = \frac{\tilde{m}_i s + \tilde{d}_i}{(T_{1i}s + 1)(T_{2i}s + 1)} \omega_i. \quad (5)$$

We call \tilde{m}_i virtual inertia, as it reacts proportional to the derivative of the frequency, and \tilde{d}_i virtual damping, as it reacts proportional to the frequency itself.

The transfer function has two poles—one is needed for causality of the PD control and the other can be interpreted as time constant of the PLL. In fact, ω_i is the physical frequency at bus i , which cannot be measured without a time delay.

One possible state-space realization of the controller (5) is

$$\tilde{A}_i = \begin{bmatrix} -\frac{T_{1i}+T_{2i}}{T_{1i}T_{2i}} & -\frac{1}{T_{1i}T_{2i}} \\ 1 & 0 \end{bmatrix}, \quad \tilde{B}_i = \begin{bmatrix} \frac{1}{T_{1i}T_{2i}} \\ 0 \end{bmatrix}, \quad \tilde{C}_i = [\tilde{m}_i \ \tilde{d}_i]. \quad (6)$$

The control system (6) has two states. The second state can be interpreted as low-pass-filtered measurement of the frequency ω_i , while the first state can be seen as low-pass-filtered derivative of ω_i . The output of the controller (6) is the electric power $p_{v,i}$ injected by the virtual inertia and damping device at bus i and therefore acts via the term $p_{el,i}$ in (2). We finally define the aggregate representation matrices \tilde{A} , \tilde{B} , and \tilde{C} , which have the blocks \tilde{A}_i , \tilde{B}_i , and \tilde{C}_i , respectively, on their diagonal. We denote the state of the aggregated virtual inertia devices by \tilde{x} .

2.4 Interconnected Closed-Loop Power System

The interconnection of the dynamical models of synchronous machines, governors, and virtual inertia and damping devices is schematically represented in Figure 2. Notice that in the interconnected diagram, we have also considered a disturbance input η in the electric power injection. Convenient state-space representations of the interconnected systems will be introduced for each of the virtual inertia and damping placement and tuning methods in Section 4.

2.5 Assumptions and Limitations

As stated throughout the model description, the swing equation model is well suited for small-signal analysis, as it is a linearization around the current operating

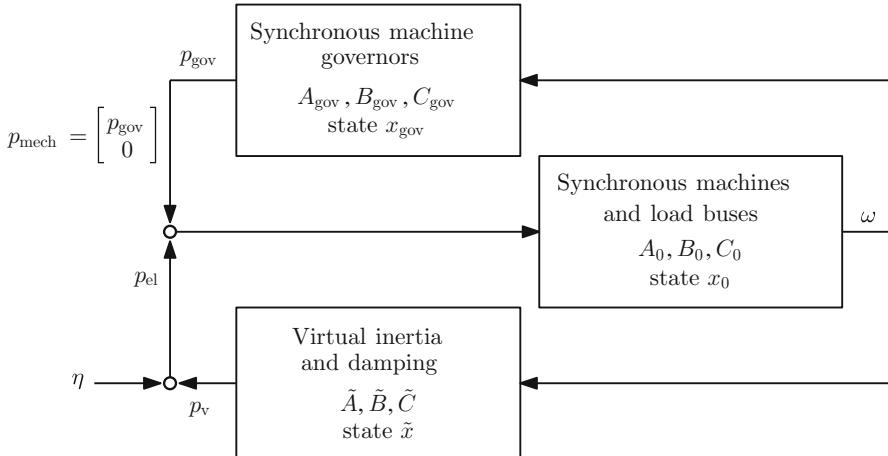


Fig. 2 Schematic representation of the interconnection of the dynamical models for synchronous machines, governors, and virtual inertia and damping devices

point. Adding governors accounts for the change in operating point, as it correctly describes the lasting frequency deviation after a fault and before secondary control mechanisms are activated. The interconnection of multiple swing equation models and governor models via linearized power flow equations yields a high-dimensional linear system which exhibits complex and coupled transient behaviors in the bus voltage angles and frequencies. It does not, however, account for the nonlinearities in the power flows which become increasingly significant for larger disturbances. Nonetheless, adopting a linear model enables us to use efficient tools from linear system theory while capturing the main phenomena in the post-fault response of a power system.

The second simplification we make relates to the model detail and granularity. We ignore all voltage dynamics and related automatic voltage regulation and power system stabilizers. Adding these would not interfere with the algorithms proposed in the next sections, as long as they are modeled as linear controllers, but for the sake of simplicity, we do not consider them in this chapter.

3 Performance Metrics and Design Constraints

Based on the model presented in Section 2, here we formally define a set of performance metrics that we shall use to assess the frequency stability of the grid, when subject to power disturbances.

As discussed in Section 1.1 and schematically represented in Figure 1, different metrics can be defined on the time-domain response of the system following a specified disturbance input η_0 .

In particular, given a step disturbance η_0 at time $t = 0$, we define the following indices on the time-domain evolution of bus frequencies.

- **Rate of change of frequency (RoCoF):**

$$\max_i \left(\max_{t \geq 0} |\dot{\omega}_i(t)| \right). \quad (7)$$

- **Frequency nadir:**

$$\max_i \left(\max_{t \geq 0} |\omega_i(t)| \right). \quad (8)$$

- **Total energy unbalance:**

$$\int_0^\infty \sum_i q_i \omega_i^2 dt = \int_0^\infty \omega^\top Q \omega dt, \quad (9)$$

where q_i are positive, possibly bus-dependent, weights on the different bus frequencies and are collected in the diagonal matrix Q .

- **Damping ratio of a power system:**

Independently of the particular disturbance, the damping ratio describes how fast oscillations in the power system are vanishing. The damping ratio of a power system is defined as the smallest damping ratio of its eigenvalues λ_k . A higher numeric value hence corresponds to better performance.

$$\min_k \frac{-\sigma_k}{\sqrt{(\sigma_k)^2 + (\omega_k)^2}}, \quad (10)$$

where $\lambda_k = \sigma_k + i\omega_k$ is the k -th eigenvalue of the closed-loop power system model.

For the same step disturbance as considered above, we also define the following indices to quantify the control effort that the governors and the virtual inertia devices need to exert.

- **Total governor effort:**

$$\int_0^\infty \sum_i r_{\text{gov},i} p_{\text{gov},i}^2 dt = \int_0^\infty p_{\text{gov}}^\top R_{\text{gov}} p_{\text{gov}} dt, \quad (11)$$

where $r_{\text{gov},i}$ are positive, possibly bus-dependent, weights on the control effort of different governors and are collected in the diagonal matrix R_{gov} .

- **Total virtual inertia and damping effort:**

$$\int_0^\infty \sum_i r_i p_{v,i}^2 dt = \int_0^\infty p_v^\top R p_v dt, \quad (12)$$

where r_i are positive, possibly bus-dependent, weights on the control effort of different virtual inertia and damping devices and are collected in the diagonal matrix R .

- **Peak virtual inertia and damping power injection:**

$$\max_i \left(\max_{t \geq 0} |p_{v,i}(t)| \right). \quad (13)$$

The following two system norms provide a measure of the system output in response to a disturbance η (see Section 1.1 and [38]).

- **\mathcal{H}_2 -norm:** The \mathcal{H}_2 -norm can be interpreted as the energy of the response to impulsive faults or the expected energy of the response to white noise. By defining a suitable performance output, the energy metrics (11), (12), and (9) can be directly considered in this framework.
- **\mathcal{H}_∞ -norm:** The \mathcal{H}_∞ -norm corresponds to the RMS gain from the disturbance to the performance output, which may include the frequencies ω as well as the control inputs p_{gov} and p_v .

The performance indices (12) and (13) quantify the control effort of the virtual inertia and damping devices, either in terms of total energy or peak power injection. They need to be considered when tuning the virtual inertia gains \tilde{m}_i and virtual damping gains \tilde{d}_i , as these devices will necessarily have bounds on both their energy and the maximum instantaneous power that their power converters can inject in the grid.

In order to convert bounds on the peak power injection into bounds on the values of \tilde{m}_i and \tilde{d}_i , we examined some real frequency measurements from two different grids (Figure 3). In both cases, it is evident that maximum frequency deviations and maximum RoCoF do not happen simultaneously. Therefore, the derivative control term $\tilde{m}_i \dot{\omega}_i$ and the proportional term $\tilde{d}_i \omega_i$ in (5) will not reach their peak value at the same time, and it is recommended to consider a joint constraint on the two gains. In this chapter we adopt the constraint

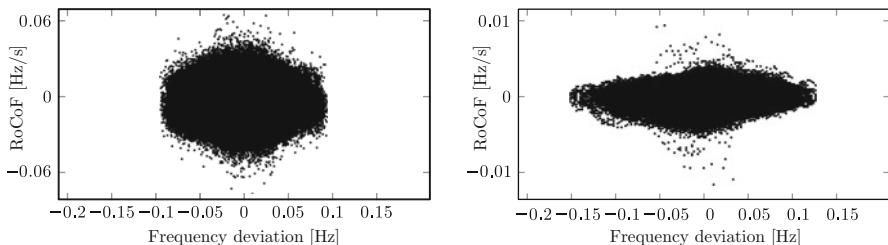


Fig. 3 Scatter plot of frequency data measurements from Ireland (left panel, courtesy of F. Milano, University College Dublin) and from continental Europe (right panel, courtesy of RTE France)

$$\max\{\alpha \tilde{m}_i, \beta \tilde{d}_i\} \leq p_{v,i}^{\max},$$

where $p_{v,i}^{\max}$ depends on the power inverter size, while α and β depend on the typical observed ranges in the RoCoF and the frequency deviation, respectively.

When considering the problem of allocating virtual inertia and damping in a grid, from an economic perspective, we expect to have a budget on the maximum total size of the power converters in these devices. We shall therefore consider a system-wide constraint of the form

$$\sum_i \max\{\alpha \tilde{m}_i, \beta \tilde{d}_i\} \leq p_v^{\text{budget}}. \quad (14)$$

4 Methods

In the following, we present two algorithms to optimally place and tune virtual inertia and damping throughout the power system. The two algorithms are iterative gradient-based schemes targeting the different performance metrics presented in Section 3.

4.1 Control Design via \mathcal{H}_2 Optimization

A first approach to answer the question of “where to optimally place virtual inertia and damping” consists of recasting the problem as that of minimizing an input-output gain. Here, the disturbances acting on the system form the set of inputs, and the frequency deviations of the synchronous machines and control energy used by the virtual inertia and damping devices and governors are the outputs. Such an input-output gain is also referred to as the \mathcal{H}_2 system gain as discussed in Section 3.

To this end, we consider the interconnected grid model presented in Figure 2. More precisely, we combine the state-space representations of the synchronous machines (2), the governors (4), and the virtual inertia filters (6) but keep the virtual inertia and damping gains as explicit feedback inputs. While the frequencies ω of the synchronous machines (2) are stable, their angles θ are not. However, the input-output behavior of (2) with output ω can be equivalently expressed in terms of the state vector $x_\delta = (\delta, \omega)$, where $\delta_i = \theta_1 - \theta_i$ corresponds to the angle relative to the angle of bus 1. After applying a corresponding similarity transformation and removing the remaining unstable mode corresponding to the absolute angle θ_1 , we obtain a stable system $(A_\delta, B_\delta, C_\delta)$. The overall system is then given by

$$\begin{bmatrix} \dot{x}_\delta \\ \dot{x}_g \\ \ddot{x} \end{bmatrix} = \underbrace{\begin{bmatrix} A_\delta & B_{\delta,\text{gov}} & 0 \\ B_{\text{gov},\delta} & A_{\text{gov}} & 0 \\ \tilde{B}C_\delta & 0 & \tilde{A} \end{bmatrix}}_A \underbrace{\begin{bmatrix} x_\delta \\ x_{\text{gov}} \\ \ddot{x} \end{bmatrix}}_x + \underbrace{\begin{bmatrix} B_\delta \\ 0 \\ 0 \end{bmatrix}}_B p_v + \underbrace{\begin{bmatrix} B_\delta \Pi \\ 0 \\ 0 \end{bmatrix}}_G \eta, \quad (15)$$

where η is the vector of disturbances and G is the disturbance gain matrix. The matrix Π encodes the a priori information about the location of the sources of these disturbances, such as large synchronous generators and HVDC terminal points, among others. Finally, the matrix Π_G is used to obtain the generator frequencies ω_G from the system state, i.e., $\omega_G = \Pi_G C_\delta x_\delta$, and the matrices $B_{\delta,\text{gov}} = B_\delta \Pi_G^\top C_{\text{gov}}$ and $B_{\text{gov},\delta} = B_{\text{gov}} \Pi_G C_\delta$ are used to model the interconnection between the synchronous machines and the governors.

The output of the virtual inertia and damping devices fed into the interconnected system as in Figure 2 is given by

$$p_v = \underbrace{\begin{bmatrix} 0 & 0 & [\tilde{M} \tilde{D}] \end{bmatrix}}_{\tilde{K}} \underbrace{\begin{bmatrix} x_\delta \\ x_{\text{gov}} \\ \tilde{x} \end{bmatrix}}_x, \quad (16)$$

where \tilde{K} is the matrix of virtual inertia and damping parameters (proportional and derivative gains). \tilde{M} and \tilde{D} are diagonal matrices collecting the inertia constants \tilde{m}_i and damping constants \tilde{d}_i of the individual emulation devices (5).

We then introduce a performance output y which contains the signals that we wish to include in the \mathcal{H}_2 gain analysis. This output can be constructed as

$$y = \underbrace{\begin{bmatrix} Q^{\frac{1}{2}} C_\delta & 0 & 0 \\ 0 & R_{\text{gov}}^{\frac{1}{2}} C_{\text{gov}} & 0 \\ 0 & 0 & 0 \end{bmatrix}}_C \underbrace{\begin{bmatrix} x_\delta \\ x_{\text{gov}} \\ \tilde{x} \end{bmatrix}}_x + \underbrace{\begin{bmatrix} 0 \\ 0 \\ R^{\frac{1}{2}} \end{bmatrix}}_F p_v, \quad (17)$$

where Q penalizes the frequency deviations ω , R_{gov} penalizes the governor control effort p_{gov} , and R is a penalty on the control energy used by the virtual inertia and damping devices p_v , as discussed in Section 3.

The performance output (17) is then used to formulate the following performance metric which combines the metrics (9), (11), and (12):

$$\int_0^\infty y^\top y \, dt = \int_0^\infty \omega^\top Q \omega + p_{\text{gov}}^\top R_{\text{gov}} p_{\text{gov}} + p_v^\top R p_v \, dt. \quad (18)$$

By explicitly closing the loop, this results in the following dynamic system \mathcal{G} :

$$\begin{aligned} \dot{x} &= (A + B\tilde{K})x + G\eta, \\ y &= (C + F\tilde{K})x. \end{aligned} \quad (19)$$

To compute the norm $\|\mathcal{G}\|_2^2$ between the disturbance input η and the performance output y of system (19), let $P_{\tilde{K}}$ denote the solution of the Lyapunov equation

$$P(A + B\tilde{K}) + (A + B\tilde{K})^\top P + C^\top C + \tilde{K}^\top F^\top F\tilde{K} = 0, \quad (20)$$

parameterized in \tilde{K} for the given system matrices A , B , C , and F . Based on the observability Gramian $P_{\tilde{K}}$, the norm $\|\mathcal{G}\|_2^2$ is given by [38]

$$\|\mathcal{G}\|_2^2 = \text{trace}(G^\top P_{\tilde{K}} G). \quad (21)$$

Thus, the optimization problem to compute the optimal allocation with respect to the \mathcal{H}_2 -norm $\|\mathcal{G}\|_2^2$ is obtained as

$$\begin{aligned} \min_{\tilde{K}} \quad & \text{trace}(G^\top P_{\tilde{K}} G) \\ \text{s.t.} \quad & \tilde{K} \in \mathcal{S} \cap \mathcal{C} \end{aligned} \quad (22)$$

where \mathcal{C} is a convex constraint for the magnitudes of \tilde{m}_i and \tilde{d}_i . Furthermore, \mathcal{S} encodes the structural constraint on \tilde{K} , i.e., the purely local feedback structure of the virtual inertia and damping control in (5). Note that evaluating the cost function requires solving the Lyapunov equation (20).

In general, the optimization problem (22) is non-convex and may be very large-scale, but its structure can be exploited to obtain efficient solution methods. By using the implicit linearization technique from [26], the gradient of the norm $\|\mathcal{G}\|_2^2(\tilde{K})$ with respect to \tilde{K} is given by

$$\nabla_{\tilde{K}} \|\mathcal{G}\|_2^2 = \begin{bmatrix} \frac{\partial}{\partial \tilde{K}_{1,1}} \|\mathcal{G}\|_2^2 & \cdots & \frac{\partial}{\partial \tilde{K}_{1,n}} \|\mathcal{G}\|_2^2 \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \tilde{K}_{m,1}} \|\mathcal{G}\|_2^2 & \cdots & \frac{\partial}{\partial \tilde{K}_{m,n}} \|\mathcal{G}\|_2^2 \end{bmatrix} = 2(B^\top P_{\tilde{K}} + R\tilde{K})L_{\tilde{K}}, \quad (23)$$

where $L_{\tilde{K}}$ is the controllability Gramian obtained as solution to the Lyapunov equation

$$L(A + B\tilde{K})^\top + (A + B\tilde{K})L + GG^\top = 0, \quad (24)$$

parameterized in \tilde{K} for the given system matrices A , B , and G . Thus, computing the norm $\|\mathcal{G}\|_2^2$ and its gradient $\nabla_{\tilde{K}} \|\mathcal{G}\|_2^2$ for a given \tilde{K} mainly requires solving the Lyapunov equations (20) and (24). Moreover, the number of decision variables of the optimization problem (22) can be reduced by projecting the gradient $\nabla_{\tilde{K}} \|\mathcal{G}\|_2^2$ on the sparsity constraint \mathcal{S} . Using the vector of nonzero parameters $\phi = [\tilde{m}_1, \tilde{d}_1, \dots, \tilde{m}_m, \tilde{d}_m]$, the projected gradient is given by

$$\nabla_{\phi} \|\mathcal{G}\|_2^2 = \begin{bmatrix} \frac{\partial}{\partial \tilde{m}_1} \|\mathcal{G}\|_2^2 \\ \frac{\partial}{\partial \tilde{d}_1} \|\mathcal{G}\|_2^2 \\ \vdots \\ \frac{\partial}{\partial \tilde{m}_m} \|\mathcal{G}\|_2^2 \\ \frac{\partial}{\partial \tilde{d}_m} \|\mathcal{G}\|_2^2 \end{bmatrix}. \quad (25)$$

Because the \mathcal{H}_2 norm is infinite for unstable systems, both the system norm $\|\mathcal{G}\|_2^2$ and its gradient (23) are only well defined for a stable closed-loop system (19). Thus, to optimize the control gain \tilde{K} , an initial guess for \tilde{K} is required that stabilizes (19) and satisfies the sparsity constraint \mathcal{S} . As the plant is stable, an initial guess which satisfies these conditions is given by $\tilde{m}_i = 0$ and $\tilde{d}_i = 0$.

Assuming that the projections onto \mathcal{C} can be efficiently computed via the projected gradient method [4], the gradient computation outlined above can be used to solve the optimization problem (22) even for systems of very large dimension, e.g., when \mathcal{C} encodes upper and lower bounds on \tilde{m}_i and \tilde{d}_i . However, if the projection onto \mathcal{C} cannot be computed efficiently, the above gradient computation can still be used to speed up the computation's higher-order methods.

4.2 Control Design via Optimization of Spectral and Time-Domain Criteria

This subsection outlines a placement algorithm (colloquially referred to as *spectral algorithm* in the following) that uses a combination of spectral analysis and time-domain limits on the step response. We refer to [6, 7] for a more in-depth presentation and analysis. Previously, spectral optimization has been extensively used in multi-machine PSS gain scheduling [34]; the main extension here is the inclusion of time-domain limits.

The damping ratio of a power system introduced in (10) describes how fast power system oscillations are declining and often is a classical proxy to quantify system stability [34]. For notational convenience, we define the damping ratio of the k -th eigenvalue $\lambda_k = \sigma_k + i\omega_k$ of the power system as

$$\zeta_k = \frac{-\sigma_k}{\sqrt{(\sigma_k)^2 + (\omega_k)^2}}. \quad (26)$$

In addition, we are interested in the magnitude of the overshoot after a given disturbance. This can be computed from the step response of the transfer function from a disturbance to a performance output $y(t)$ (e.g., a generator or load bus frequency) as

$$y(t) = \sum_k \frac{r_k}{\lambda_k} (1 - e^{\lambda_k t}), \quad (27)$$

where the residues r_k are defined via the right and left eigenvectors u_k and v_k^\top as

$$r_k = C u_k v_k^\top B. \quad (28)$$

In general, we consider all input-output step responses, indexed by y_{db} , where d corresponds to the index of the disturbance and b to the index of the performance output.

Next, we define the overshoot as in (8) by the largest value of each step response

$$S_{db} = \max_t |y_{db}(t)| = |y_{db}(t_{S,db})|, \quad (29)$$

and the RoCoF as the largest rate of change (time derivative) of the step response

$$R_{db} = \max_t \left| \frac{dy_{db}(t)}{dt} \right| = \left| \frac{dy_{db}(t)}{dt} \right|_{t=t_{R,db}}, \quad (30)$$

where $t_{S,db}$ and $t_{R,db}$ are the time instants when the largest amplitude and slope of the signal $y_{db}(t)$ occur.

4.2.1 Objectives: Maximizing Damping Ratio or Minimizing Overshoot and RoCoF

In the following, we formalize the optimization objectives according to the performance metrics laid out in Section 3. To maximize the worst-case damping ratio ζ , we define the cost term

$$-c^\zeta \zeta^{\min} \quad (31)$$

with a positive coefficient c^ζ and the variable ζ^{\min} lower-bounding the damping ratio, i.e.,

$$\zeta^{\min} \leq \zeta_k \forall \{k | \omega_k > 0\}. \quad (32)$$

Similarly, one can minimize the worst-case absolute RoCoF $|R|_\infty = \max_{d,b} |R_{db}|$ by defining the cost function

$$c^R |R|_\infty \quad (33)$$

with some positive coefficient c^R and subject to the constraint

$$|R|_\infty \geq |R_{db}| \forall d, b. \quad (34)$$

An analogous approach can be used for the worst-case overshoot $|S|_\infty = \max_{d,b} |S_{db}|$.

4.2.2 Virtual Inertia and Damping Placement Algorithm

The vector of system parameters ϕ (e.g., the control gains in (5)) can be obtained by solving a multi-objective optimization problem of the form

$$\begin{aligned} \min_{\phi \in \mathcal{C}} \quad & -c^\zeta \zeta^{\min} + c^R |R|_\infty + c^S |S|_\infty \\ \text{s.t.} \quad & \underline{\zeta} \leq \zeta_k \leq \bar{\zeta} \forall k, \\ & \underline{S} \leq S_{db} \leq \bar{S} \forall d, b, \\ & \underline{R} \leq R_{db} \leq \bar{R} \forall d, b, \end{aligned} \quad (35)$$

where the coefficients $c^\zeta > 0$, $c^R > 0$, and $c^S > 0$ penalize the damping ratio, the overshoot, and the RoCoF. In addition, we consider limits on the worst-case values of the performance metrics while confining the control gains to a set \mathcal{C} as in Section 3.

Because the underlying eigenvalue problem is heavily nonlinear and non-convex, the optimization problem (35) is very difficult to solve. Here, we adopt a sequential linear programming approach that iteratively optimizes ϕ and updates the eigenvalues as well as eigenvectors to ultimately arrive at a local optimum. In the remainder we shall use linear approximations of (35). To this end, we define the sensitivity operator \mathcal{D} which characterizes a first-order approximation (not necessarily a gradient) of the change of a function f with respect to a parameter x :

$$\mathcal{D}_x f \approx \frac{\partial f}{\partial x}. \quad (36)$$

4.2.3 Sensitivity of Damping Ratio

The sensitivity of the damping ratio (26) with respect to a vector of system parameters ϕ (e.g., a set of control gains as in (5)) is given by

$$\mathcal{D}_\phi \zeta_k = \mathcal{D}_\phi \left(\frac{-\sigma_k}{\sqrt{(\sigma_k)^2 + (\omega_k)^2}} \right) = \omega_k \frac{(\sigma_k \mathcal{D}_\phi \omega_k - \omega_k \mathcal{D}_\phi \sigma_k)}{\left((\sigma_k)^2 + (\omega_k)^2 \right)^{\frac{3}{2}}}. \quad (37)$$

Observe that in order to compute (37), we need the eigenvalue derivatives, which can be obtained as in [22]

$$\mathcal{D}_\phi \lambda_k = v_k^\top (\mathcal{D}_\phi A) u_k . \quad (38)$$

4.2.4 Sensitivities of Time-Domain Indices: Overshoot and RoCoF

Because the time instants $t_{S,db}$ and $t_{R,db}$ implicitly depend on the parameters ϕ , the sensitivities of the overshoot S_{db} and the largest RoCoF R_{db} with respect to ϕ are more involved. The full sensitivity of S is obtained as

$$\mathcal{D}_\phi S = \sum_k \left[\left(\mathcal{D}_\phi \frac{r_k}{\lambda_k} \right) (1 - e^{\lambda_k t_S}) - \frac{r_k}{\lambda_k} ((\mathcal{D}_\phi \lambda_k) t_S + \lambda_k \mathcal{D}_\phi t_S) e^{\lambda_k t_S} \right], \quad (39)$$

and makes use of the derivatives of the residues r_k , the derivative of the eigenvalues λ_k , and the sensitivity of the peak time $t_{S,db}$ (see [6, 7] for details). The sensitivity of the RoCoF $\mathcal{D}_\phi R$ can be found in a similar fashion as

$$\mathcal{D}_\phi R = - \sum_k \left[\mathcal{D}_\phi r_k + r_k (\mathcal{D}_\phi \lambda_k) t_R + r_k \lambda_k \mathcal{D}_\phi t_R \right] e^{\lambda_k t_R} . \quad (40)$$

4.2.5 Sequential Linear Programming and Updates of Performance Metrics

We adopt a sequential linear programming approach to solve (35). Given a sequence of parameters ϕ^ν with iteration index ν , the performance indices are approximated at each iteration ν according to

$$\tilde{\zeta}_k^\nu = \zeta_k^\nu + (\mathcal{D}_\phi \zeta_k^\nu) (\phi - \phi^\nu) \forall \{k | \omega_k > 0\}, \quad (41)$$

$$\tilde{R}_{db}^\nu = R_{db}^\nu + (\mathcal{D}_\phi R_{db}^\nu) (\phi - \phi^\nu) \forall d, b, \quad (42)$$

$$\tilde{S}_{db}^\nu = S_{db}^\nu + (\mathcal{D}_\phi S_{db}^\nu) (\phi - \phi^\nu) \forall d, b, \quad (43)$$

where $S_{db}^\nu = S_{db}(\phi^\nu)$, $R_{db}^\nu = R_{db}(\phi^\nu)$, and $\zeta_k^\nu = \zeta_k(\phi^\nu)$ are the exact values at ϕ^ν . Letting $\zeta_k = \tilde{\zeta}_k^\nu$, $R_{db} = \tilde{R}_{db}^\nu$, $S_{db} = \tilde{S}_{db}^\nu$ in (35) a linear program is obtained. This linear program is solved at every iteration ν to obtain a new parameter vector $\phi^{\nu+1}$. Based on $\phi^{\nu+1}$, new linear approximations are constructed according to (41)–(43) resulting in a linear program to be solved in iteration $\nu+1$. This iterative scheme is repeated until convergence. To ensure convergence of the sequential linear programming scheme, the step size is limited by a constraint $\|\phi - \phi^\nu\| \leq \varepsilon^\nu$ and an appropriate step size rule.

5 Test Case

In this section, we investigate the virtual inertia and damping placement problem for a case study described in Section 5.1. The optimal virtual inertia and virtual damping allocations are derived via the \mathcal{H}_2 -based and the spectral optimization-based approaches presented in Section 4. In Section 5.2, we compare the performance of these two allocations based on different system metrics and draw suitable interpretations. Finally, in Section 5.3, we conclude by illustrating the time-domain evolution of some of the performance metrics.

5.1 Description of the Test System

The three-area test system is adapted from [7], and its topology is shown in Figure 4. The system is an aggregation of twelve buses, which are classified as load or generator buses. As the bus labeled 11 does not belong to either of these categories, we can effectively remove it and reduce the system without any loss of generality to an eleven-bus (i.e. buses labeled 1–10 and 12) system via Kron reduction [9]. We do not initially constrain ourselves concerning possible sites for virtual inertia and damping placement. Rather, we assume that each of the remaining eleven buses can be assigned a virtual inertia and damping device (with identical time constants $T_1 = 0.1$ and $T_2 = 0.3$), and we find the optimal placement through the algorithms presented in Section 4.

To set up the test case, we consider a disturbance input at every load bus (i.e., the buses labeled 3, 4, 7, 8, 12). A disturbance of 1 p.u. corresponds to 100 MVA. We consider the optimal placement and tuning of virtual inertia

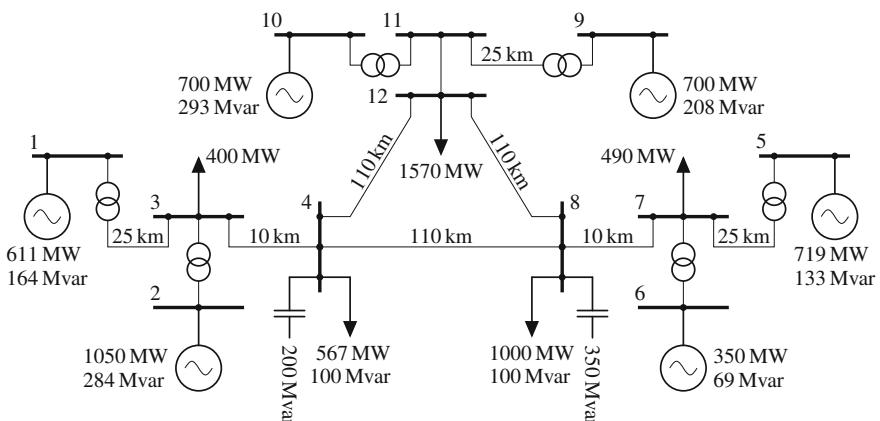


Fig. 4 Topology of the twelve-bus test system with six generators and five load buses

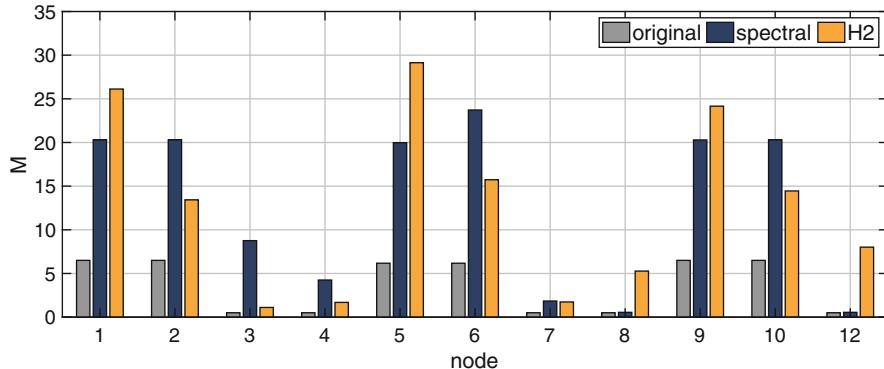


Fig. 5 Optimal inertial allocations for \mathcal{H}_2 optimized, spectral optimized algorithms, and original allocation

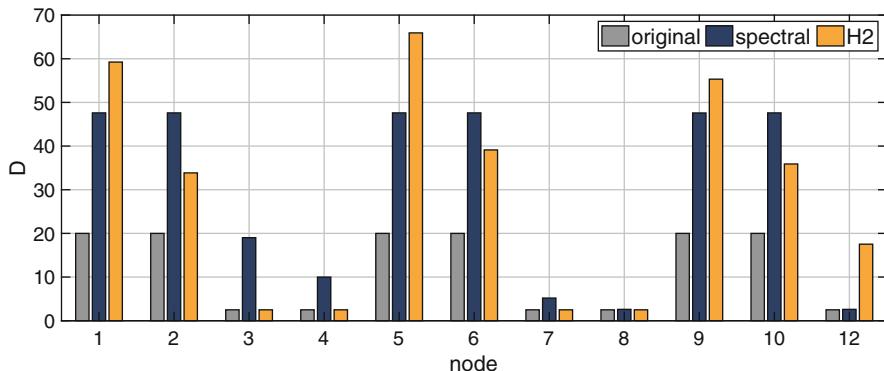


Fig. 6 Optimal damping allocations for \mathcal{H}_2 optimized, spectral optimized algorithms, and original allocation

and damping according to the performance metrics and algorithms laid out in Sections 3 and 4, respectively. Specifically, the spectral optimization presented in Section 4.2 considers the step response of the performance outputs with respect to every disturbance input, and the \mathcal{H}_2 -based optimization presented in Section 4.1 optimizes the gains from all disturbance inputs to all disturbance outputs. The allocation of the \mathcal{H}_2 -based optimization considers a cost identically penalizing the generator frequency deviations, the governor control inputs, and the virtual inertia power injections. This is a feasible choice as all the states are normalized in the p.u. scale. The spectral algorithm on the other hand considers identical penalties on frequency overshoot and RoCoF, while ensuring a minimum limit on damping ratio of 0.07. The constraints on inertia and damping are considered as described in (14) with $\alpha = 1$, $\beta = 0.5$, and $p_v^{\text{budget}} = 100$.

Based on the two algorithms, the inertia and damping allocations are obtained as depicted in Figures 5 and 6, respectively. Observe that both optimization algorithms are essentially in agreement concerning the allocation of virtual inertia and damping

Table 1 Comparison of pure frequency metrics

	Original system	Spectral optimization	\mathcal{H}_2 optimization
System norms (from all disturbance inputs to all performance outputs)			
\mathcal{H}_2 gain ^a	0.4606	0.3551	0.3590
\mathcal{H}_∞ gain ^a	0.2222	0.1729	0.1600
Localized disturbance at bus 4			
Damping ratio	0.0648	0.0700	0.0701
Max RoCoF	0.2874 (at node 2)	0.2489 (at node 2)	0.2682 (at node 2)
Frequency nadir	0.0419 (at node 2)	0.0334 (at node 2)	0.0374 (at node 2)

^aFrom disturbance to frequency

Table 2 Comparison of system metrics

	Original system	Spectral optimization	\mathcal{H}_2 optimization
System norms (from all disturbance inputs to all performance outputs)			
\mathcal{H}_2 gain ^b	0.5574	0.4720	0.3956
\mathcal{H}_∞ gain ^b	0.6632	0.4276	0.4259
Localized disturbance at bus 4			
Peak power injection	0	0.0103 (at node 4)	0.0029 (at node 4)

^bFrom disturbance to frequency and control effort

devices dominantly placed at the periphery buses of the grid (see Figures 5 and 6). The differences between the two allocations are subtle but lead to quite distinct characteristics, which we shall discuss below.

5.2 Comparison of Performance

For the purpose of comparing the efficacy of these two allocations vis-à-vis the original allocations, we compare in this subsection a few performance metrics that have been introduced in Section 3. In particular, we consider the damping ratio, maximum RoCoF (rate of change of frequency), frequency overshoot, peak power injection, and \mathcal{H}_2 and \mathcal{H}_∞ system input-output gains, which can effectively capture system-wide performance in response to a disturbance. A comparison of the aforementioned metrics across different allocations is tabulated forthwith based on the classification of each metric either as a pure frequency performance measure (Table 1) or a system performance measure (Table 2) that includes the control effort as well.

In Tables 1 and 2, we consider the system norms (i.e., the gains from all disturbance inputs to all performance outputs) as well as a spectral and time-domain criteria for a single localized fault. We compare the \mathcal{H}_2 and \mathcal{H}_∞ norms for the system with optimized control gains when (a) the generator frequency violations alone are penalized (Table 1), i.e., $R = 0$, $R_{gov} = 0$, and (b) the generator frequency violations and the control input are penalized (Table 2).

To further illustrate the performance of the optimized allocations and compare the spectral and time-domain criteria which are only well defined for a single disturbance input, we inspect the response when the test system is subjected to a localized disturbance of 1 p.u. at the load bus labeled 4 in Figure 4.

The following inferences can be drawn from the above results:

- The system input-output gains, \mathcal{H}_2 and \mathcal{H}_∞ , are significantly reduced for the two optimal allocations in comparison to the original allocation. This holds for both scenarios—with and without control input penalties.
- The spectral optimization algorithm marginally outperforms the \mathcal{H}_2 optimization for the time-domain criteria of maximum RoCoF and frequency overshoot.
- For both the algorithms, the peak control input occurs at the bus where the disturbance strikes. However, for obtaining a similar performance, the spectral optimization roughly expends four times the effort of what is required by the \mathcal{H}_2 optimization. Indeed, compared to \mathcal{H}_2 -optimal algorithm, the spectral algorithm as presented in Section 4.2 does not penalize the control effort.

5.3 Time-Domain and Spectral Simulations

In this subsection, we present a few simulation plots which enable a better understanding of the post-fault system behavior. The time-domain plots in Figure 7 suggest that all generators experience an improved transient frequency response behavior post-fault for both the \mathcal{H}_2 and the spectral-optimized algorithms over the original allocation. However, the control effort displayed in Figure 8 reveals a significantly higher peak control effort for the spectral optimization than for the \mathcal{H}_2 optimization. Finally, the plot of the closed-loop eigenvalues in Figure 9 does not shed much light on the effectiveness of the optimization techniques, as all methods achieve similar worst-case damping ratios and damping asymptotes. We conclude that the spectrum itself is not very insightful, which points toward the need to investigate other meaningful metrics as discussed above in Section 3.

6 Summary and Conclusions

In this chapter we presented performance metrics for low-inertia power systems and studied the optimal placement and tuning of control devices providing virtual inertia and damping subject to power constraints. We introduced two different algorithms that formalized this control problem: the first algorithm was based on the control-system notion of an \mathcal{H}_2 system gain characterizing the amplification of a disturbance, whereas the second algorithm targeted spectral and time-domain performance indices that are of immediate concern to system operators. For both optimization criteria, we presented iterative and gradient-based strategies leading

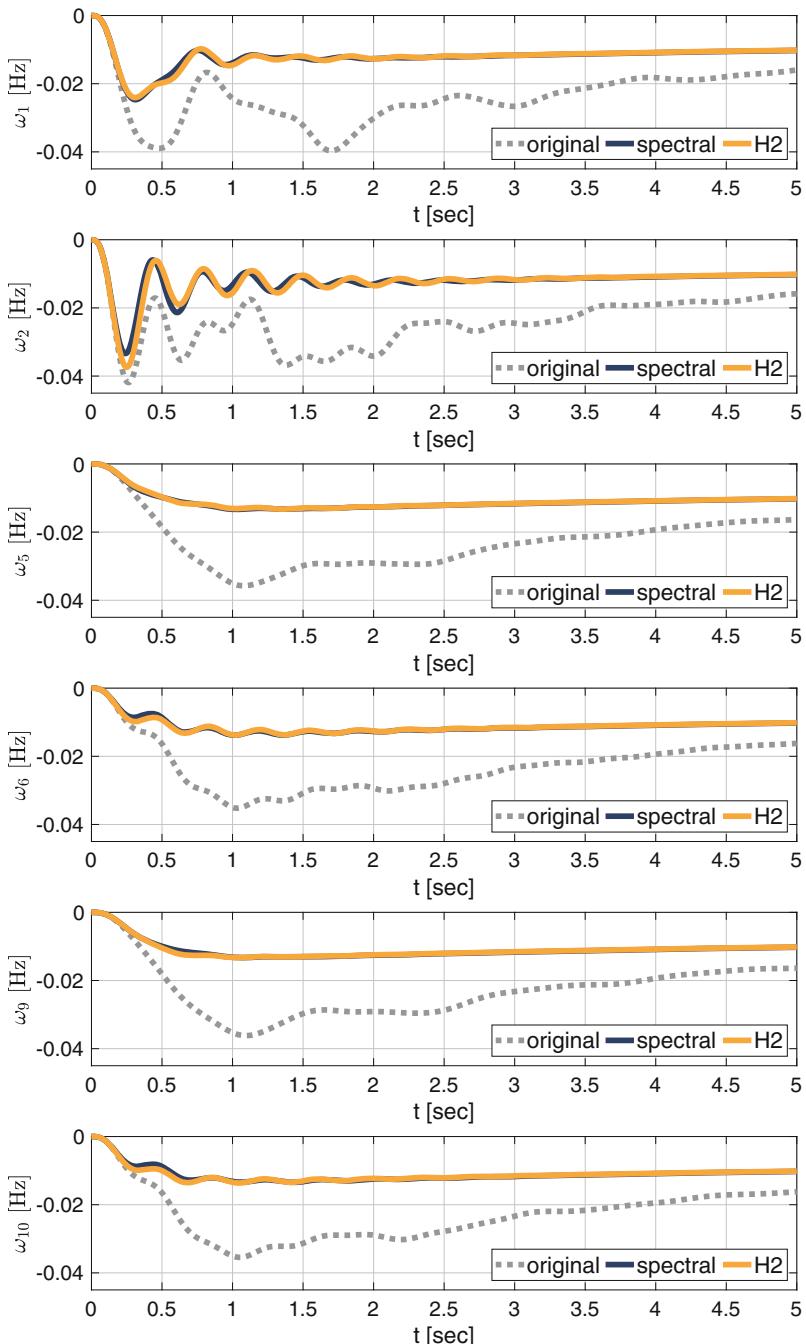


Fig. 7 Time-domain plots for frequency variation at different generator nodes post a step fault at load bus labeled 4

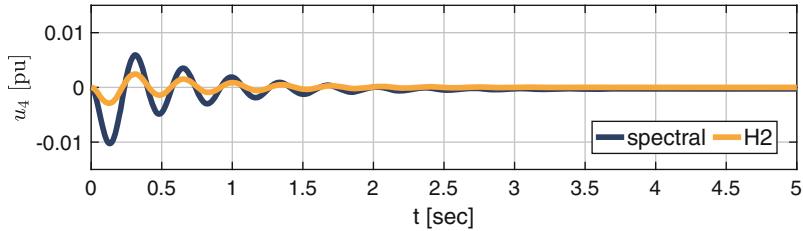


Fig. 8 Time-domain plots for control input at node 4 post a step fault at load bus labeled 4

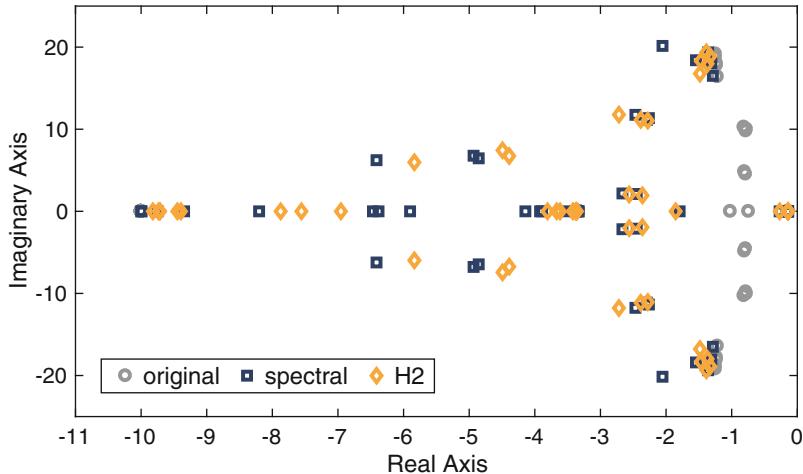


Fig. 9 Spectrum of the system matrices for the original, \mathcal{H}_2 , and the spectral optimized algorithms

to locally optimal controllers. Finally, we compared the performance of the two approaches for a three-area power system test case.

Our results revealed at first glance only subtle differences, as both algorithms led to a fairly similar allocation of inertia and damping (see Figures 5 and 6), nearly identical transient frequency performance (see Figure 7), and almost perfectly identical damping ratios and damping asymptotes (see Figure 9), and likewise most performance indices were in the same range (see Tables 1 and 2). However, Table 2 and Figure 8 revealed one important distinction: namely, for obtaining a similar performance, the spectral optimization roughly expends four times the effort of what is required by the \mathcal{H}_2 optimization. This is due to the fact that the \mathcal{H}_2 -optimal algorithm explicitly penalizes the control effort, whereas the spectral algorithm targets only frequency performance. Thus, as an immediate conclusion, we recommend that the spectral algorithm presented in Section 4.2 needs to be extended to explicitly penalize or limit the effort expended by the virtual inertia and damping control devices. Finally, it would also be interesting to investigate other performance indices, derive constructive optimization algorithms, and contrast their performance.

References

1. AEMO (2016) Update report—black system event in South Australia on 28 September 2016. Tech. rep.
2. Anaya-Lara O, Hughes F, Jenkins N, Strbac G (2006) Contribution of DFIG-based wind farms to power system short-term frequency regulation. *IEE Proc Gener Transm Distrib* 153(2):164–170
3. Bergen AR, Hill DJ (1981) A structure preserving model for power system stability analysis. *IEEE Trans Power Apparatus Syst* 100(1):25–35
4. Bertsekas D (1995) Nonlinear programming. Athena Scientific, Nashua
5. Bevrani H, Ise T, Miura Y (2014) Virtual synchronous generators: a survey and new perspectives. *Int J Electr Power Energy Syst* 54:244–254
6. Borsche TS, Dörfler F (2017) On placement of synthetic inertia with explicit time-domain constraints. *IEEE Trans Power Syst*, Submitted. Available at <https://arxiv.org/abs/1705.03244>
7. Borsche TS, Liu T, Hill DJ (2015) Effects of rotational inertia on power system damping and frequency transients. In: 54th IEEE conference on decision and control
8. D'Arco S, Suul JA (2013) Virtual synchronous machines—classification of implementations and analysis of equivalence to droop controllers for microgrids. Proceedings of the IEEE Powertech conference, pp 1–7
9. Dörfler F, Bullo F (2013) Kron reduction of graphs with applications to electrical networks. *IEEE Trans Circuits Syst Regul Pap* 60(1):150–163
10. Ekanayake J, Jenkins N (2004) Comparison of the response of doubly fed and fixed-speed induction generator wind turbines to changes in network frequency. *IEEE Trans Energy Convers* 19(4):800–802
11. Ekanayake J, Holdsworth L, Jenkins N (2003) Control of DFIG wind turbines. *Power Eng* 17(1):28–32
12. Groß D, Bolognani S, Poolla BK, Dörfler F (2017) Increasing the resilience of low-inertia power systems by virtual inertia and damping. In: Bulk power systems dynamics and control symposium (iREP)
13. Guggilam SS, Zhao C, Dall'Anese E, Chen YC, Dhople SV (2017) Engineering inertial and primary-frequency response for distributed energy resources. arXiv preprint arXiv:170603612
14. Hughes F, Anaya-Lara O, Jenkins N, Strbac G (2005) Control of DFIG-based wind generation for power network support. *IEEE Trans Power Syst* 20(4):1958–1966
15. Jouini T, Arghir C, Dörfler F (2017, Submitted) Grid-forming control for power converters based on matching of synchronous machines. Available at <https://arxiv.org/abs/1706.09495>
16. Koller M, Borsche T, Ulbig A, Andersson G (2015) Review of grid applications with the Zurich 1 MW battery energy storage system. *Electr Power Syst Res* 120:128–135
17. Lalor G, Ritchie J, Rourke S, Flynn D, O'Malley M (2004) Dynamic frequency control with increasing wind generation. In: IEEE power engineering society general meeting
18. Machowski J, Bialek J, Bumby J (2011) Power system dynamics: stability and control. Wiley, Hoboken
19. Mešanović A, Münz U, Heyde C (2016) Comparison of H_{∞} , H_2 , and pole optimization for power system oscillation damping with remote renewable generation. In: IFAC workshop on control of transmission and distribution smart grids, pp 103–108
20. Milano F, Ortega A (2017) Frequency divider. *IEEE Trans Power Syst* 32(2):1493–1501
21. Morren J, de Haan S, Kling W, Ferreira J (2006) Wind turbines emulating inertia and supporting primary frequency control. *IEEE Trans Power Syst* 21(1):433–434
22. Murthy DV, Haftka RT (1988) Derivatives of eigenvalues and eigenvectors of a general complex matrix. *Int J Numer Methods Eng* 26:293–311
23. Pirani M, Hashemi E, Fidan B, Simpson-Porco JW (2016) H_{∞} robustness in mechanical and power networks. *IFAC-PapersOnLine* 50(1):5196–5201
24. Poolla BK, Bolognani S, Dörfler F (2017) Optimal placement of virtual inertia in power grids. *IEEE Trans Autom Control* 62(12):6209–6220

25. Rakhshani E, Remon D, Cantarellas AM, Rodriguez P (2016) Analysis of derivative control based virtual inertia in multi-area high-voltage direct current interconnected power systems. *IET Gener Transm Distrib* 10(6):1458–1469
26. Rautert T, Sachs EW (1997) Computational design of optimal output feedback controllers. *SIAM J Optim* 7(3):837–852
27. RG-CE System Protection & Dynamics Sub Group (2016) Frequency stability evaluation criteria for the synchronous zone of continental Europe. Tech. rep., ENTSO-E
28. Sauer PW, Pai M (1997) Power system dynamics and stability. Urbana 51:61,801
29. Slootweg J, Kling W (2002) Impacts of distributed generation on power system transient stability. In: Proceedings of IEEE power engineering society summer meeting
30. Soni N, Doolla S, Chandorkar MC (2013) Improvement of transient response in microgrids using virtual inertia. *IEEE Trans Power Delivery* 28(3):1830–1838
31. Svenska kraftnät, Statnett, Fingrid and Energinetdk (2016) Challenges and opportunities for the nordic power system. Tech. rep.
32. Torres M, Lopes LA (2009) Virtual synchronous generator control in autonomous wind-diesel power systems. In: Proceedings of IEEE electrical power & energy conference
33. Ulbig A, Borsche TS, Andersson G (2014) Impact of low rotational inertia on power system stability and operation. In: Proceedings of 19th IFAC world congress
34. Vournas CD, Papadis BC (1987) Power system stabilization via parameter optimization-application to the Hellenic interconnected system. *IEEE Trans Power Syst* 2(3):615–622
35. Vu Van T, Visscher K, Diaz J, Karapanos V, Woyte A, Albu M, Bozelie J, Loix T, Federenciu D (2010) Virtual synchronous generator: an element of future grids. In: IEEE PES innovative smart grid technologies conference Europe
36. Wang S, Hu J, Yuan X (2015) Virtual synchronous control for grid-connected DFIG-based wind turbines. *IEEE J Emerging Sel Top Power Electron* 3(4):932–944
37. Zhong QC, Weiss G (2011) Synchronverters: inverters that mimic synchronous generators. *IEEE Trans Ind Electron* 58(4):1259–1267
38. Zhou K, Doyle JC, Glover K (1996) Robust and optimal control. Prentice-Hall, Upper Saddle River

A Hierarchy of Models for Inverter-Based Microgrids



Olaoluwapo Ajala, Alejandro D. Domínguez-García, and Peter W. Sauer

Abstract This chapter develops a timescale-based hierarchy of microgrid models that can be utilized in analysis and control design tasks. The focus is on microgrids with distributed generation interfaced via grid-forming inverters. The process of developing the model hierarchy involves two key stages: (1) the formulation of a microgrid high-order model using circuit and control laws, and (2) the systematic reduction of this high-order model to several reduced-order models using singular perturbation techniques. The timescale-based hierarchy of models is comprised of the aforementioned microgrid high-order model (μHOM), along with three reduced-order models: microgrid reduced-order model 1 (μROM1), microgrid reduced-order model 2 (μROM2), and microgrid reduced-order model 3 (μROM3). A numerical validation of all the models is also presented.

1 Introduction

A microgrid may be defined as a collection of loads and distributed energy resources (DERs), interconnected via an electrical network with a small physical footprint, which is capable of operating in (1) grid-connected mode, as part of a large power system; or (2) islanded mode, as an autonomous power system. The DERs that constitute a microgrid are often interfaced to the electrical network via a grid-feeding inverter, where the output real and reactive powers are controlled to track a given reference, or via a grid-forming inverter, where the output voltage magnitude and frequency are controlled to track a given reference.

As the popularity and adoption of the microgrid concept in electricity systems increases, it becomes necessary to develop comprehensive mathematical models. Models are tools that control engineers, scientists, mathematicians, and other

O. Ajala · A. D. Domínguez-García (✉) · P. W. Sauer

Department of Electrical and Computer Engineering, University of Illinois at Urbana Champaign, 306 N Wright St, Urbana, IL 61801, USA

e-mail: ooajala2@illinois.edu; aledan@illinois.edu; psauer@illinois.edu

nonexperts in the field of microgrids, require for the different analysis and control design tasks necessary for development of innovative microgrid technologies. For example, to design and test microgrid frequency controllers, models that capture phenomena in the same timescale as the frequency, while neglecting phenomena in faster timescales, are required. Otherwise, the design of such a controller could prove difficult. Accurate mathematical models may be developed for inverter-based microgrids by utilizing concepts from circuit and control theory. However, the resulting models are often highly complex and too detailed for the particular application. It therefore becomes necessary to simplify these models to less detailed ones which, though less accurate, can represent the phenomena relevant to the application of interest.

The main contribution of this chapter is the development of a timescale-based hierarchy of models for inverter-based microgrids. Specifically, the focus is on microgrids with grid-forming inverter-interfaced power supplies interconnected to loads through an electrical network. Using Kirchhoff's laws and the inverter control laws, a microgrid high-order model (μHOM) is developed. Afterward three reduced-order models (microgrid reduced-order model 1 (μROM1), microgrid reduced-order model 2 (μROM2), and microgrid reduced-order model 3 (μROM3)) are formulated from the μHOM using singular perturbation techniques for model order reduction—the Kuramoto-type model developed in [5] can be derived from μROM3 . The time resolution, or timescale, for which the reduced-order models are valid is also identified, and all four models are explicitly presented, with the small parameters used for singular perturbation analysis identified. Finally, a comparison of the model responses, for a given test case, is presented.

The development of high-order and reduced-order models for inverter-based microgrids has received significant attention in the literature recently. More specifically, Pogaku et al. [10] present a high-order model for grid-forming inverter-based microgrids but exclude a discussion on model order reduction. Anand and Fernandes [2] and Rasheduzzaman et al. [11] present reduced-order models for microgrids, but the models are obtained using small-signal analysis, which is only valid within certain operating regions. Kodra et al. [6] discuss the model order reduction of an islanded microgrid using singular perturbation analysis. However, the electrical network dynamics are not included in the high-order model presented, and a simple linear model, which does not fully capture the dynamics of the islanded microgrid, is used for the singular perturbation analysis. Dörfler and Bullo [5] present a Kuramoto-type model for a grid-forming inverter developed using singular perturbation analysis. The electrical network is considered in the analysis and sufficient conditions for which the reduced-order Kuramoto-type model is valid are presented. However, the analysis is not as detailed as that presented in this chapter. More specifically, the timescale resolution associated with the Kuramoto-type is not discussed, the analysis is performed for a lossless electrical network, and the high-order model, on which singular perturbation analysis is performed, is not rigorously developed. Schiffer et al. [13] develop a detailed high-order model for grid-forming inverter-based microgrids. Singular perturbation analysis is then

employed to perform timescale separation and model order reduction, as done in this chapter with underlying assumptions stated. However, though the authors claim that the model order reduction can be performed, the small parameters used for singular perturbation analysis are not explicitly identified, and details of the singular perturbation analysis are not presented. Also, the time resolution associated with the reduced-order model developed is not identified. [Luo and Dhople \[9\]](#) present three models for a grid-forming inverter-based microgrid, which are obtained by performing successive model reduction steps on a high-order model, using singular perturbation analysis. However, the singular perturbation analysis is presented in a much less detailed form than that in this chapter, the timescales associated with each reduced model are not identified, and the high-order model from which all other models are derived is not explicitly stated with all the small parameters used for singular perturbation analysis identified.

The remainder of this chapter is organized as follows. In Section 2, the relevant concepts, to be used in later developments, are introduced. In Section 3, the microgrid high-order model (μHOM) is developed. In Sections 4–6, by using singular perturbation techniques, we obtain three reduced-order models that we refer to as μROM_1 , μROM_2 , and μROM_3 , respectively. Finally, in Section 7, the time resolutions of μROM_1 , μROM_2 , and μROM_3 are identified, and a comparison between the models responses, for a given test case, is presented.

2 Preliminaries

In this section, we first introduce the $qd0$ transformation of three-phase variables to arbitrary and synchronous reference frames. Next, we introduce graph-theoretic notions used in later developments to develop the network model. Finally, a primer on singular perturbation analysis for timescale modeling and model order reduction is presented.

2.1 The $qd0$ Transformation

Let $\alpha(t)$ denote the angular position of a reference frame rotating at an arbitrary angular velocity, $\omega(t)$, and let $\mathbf{f}_{qd0[\alpha(t)]}(t) = [f_{q[\alpha(t)]}(t) \ f_{d[\alpha(t)]}(t) \ f_{0[\alpha(t)]}(t)]^T$ denote the $qd0$ transform of a vector of three-phase variables, $\mathbf{f}_{abc}(t) = [f_a(t) \ f_b(t) \ f_c(t)]^T$, to the reference frame. The general form of the non-power-invariant $qd0$ transformation is given by

$$\mathbf{f}_{qd0[\alpha(t)]}(t) = \mathbf{K}_1(\alpha(t))\mathbf{f}_{abc}(t), \quad (1)$$

where

$$\mathbf{K}_1(\alpha(t)) = \frac{2}{3} \begin{bmatrix} \cos(\alpha(t)) \cos(\alpha(t) - \frac{2\pi}{3}) \cos(\alpha(t) + \frac{2\pi}{3}) \\ \sin(\alpha(t)) \sin(\alpha(t) - \frac{2\pi}{3}) \sin(\alpha(t) + \frac{2\pi}{3}) \\ \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \end{bmatrix},$$

$$\alpha(t) = \int_0^t \omega(\tau) d\tau + \alpha(0).$$

The $qd0$ reference frame in Eq. 1 is referred to as the arbitrary reference frame, but when $\alpha(t) = \omega_0 t$, where ω_0 denotes the synchronous frequency, it is referred to as the synchronously rotating reference frame [8].

Assume that $f_a(t)$, $f_b(t)$, and $f_c(t)$ are a balanced three-phase set. Let $\overrightarrow{\mathbf{f}}_{qd0[\omega_0 t]}(t)$ and $\overrightarrow{\mathbf{f}}_{qd0[\alpha(t)]}(t)$ denote the complex representation of $\mathbf{f}_{abc}(t)$ in the synchronously rotating reference frame and the arbitrary reference frame, respectively. Then, by using Eq. 1, we have that for

$$\overrightarrow{\mathbf{f}}_{qd0[\cdot]}(t) := f_{q[\cdot]}(t) - j f_{d[\cdot]}(t), \quad (2)$$

where j denotes the complex variable, i.e., $j = \sqrt{-1}$,

$$\overrightarrow{\mathbf{f}}_{qd0[\alpha(t)]}(t) = \overrightarrow{\mathbf{f}}_{qd0[\omega_0 t]}(t) \exp(-j\delta(t)), \quad (3)$$

with

$$\delta(t) := \alpha(t) - \omega_0 t.$$

[Note that because of the balanced assumption on $f_a(t)$, $f_b(t)$, and $f_c(t)$, $f_{0[\alpha(t)]}(t) = 0$.]

Let $\hat{\mathbf{f}}_{qd0[\alpha(t)]}(t) = [f_{q[\alpha(t)]}(t) \ f_{d[\alpha(t)]}(t)]^T$, and $\hat{\mathbf{f}}_{qd0[\omega_0 t]}(t) = [f_{q[\omega_0 t]}(t) \ f_{d[\omega_0 t]}(t)]^T$; then from Eq. 2–Eq. 3, it follows that:

$$\hat{\mathbf{f}}_{qd0[\alpha(t)]}(t) = \mathbf{K}_2(\delta(t)) \hat{\mathbf{f}}_{qd0[\omega_0 t]}(t), \quad (4)$$

with

$$\mathbf{K}_2(\delta(t)) = \begin{bmatrix} \cos(\delta(t)) - \sin(\delta(t)) \\ \sin(\delta(t)) \cos(\delta(t)) \end{bmatrix},$$

and the evolution of $\delta(t)$ is governed by

$$\frac{d\delta(t)}{dt} = \omega(t) - \omega_0. \quad (5)$$

2.2 Graph-Theoretic Network Model

The topology of the microgrid electrical network can be described by a connected undirected graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with \mathcal{V} denoting the set of buses in the network, so that $\mathcal{V} := \{1, 2, \dots, |\mathcal{V}|\}$, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, so that $\{j, k\} \in \mathcal{E}$ if buses j and k are electrically connected. Choose an arbitrary orientation for each of the elements in \mathcal{E} ; then we can define an incidence matrix, $\mathbf{M} = [m_{ie}] \in \mathbb{R}^{n \times |\mathcal{E}|}$, associated with this orientation as follows:

$$\begin{aligned} m_{ie} &= 1 && \text{if edge } e \text{ is directed away from node } i, \\ m_{ie} &= -1 && \text{if edge } e \text{ is directed into node } i, \\ m_{ie} &= 0 && \text{if edge } e \text{ is not incident on node } i. \end{aligned}$$

Connected to some buses, we assume that there is an inverter-interfaced source, the dynamics of which are described in Section 3.1; and at each bus, we assume there is another element, the dynamics of which are described by a generic dynamical model satisfying some properties, as described in Section 3.3.

Let $\mathcal{V}_J \subseteq \mathcal{V}$ denote the set of buses with an inverter-interfaced source. For $j = 1, 2, \dots, |\mathcal{V}_J|$, let s_j be used to identify parameters or variables associated with the inverter-interfaced source connected to bus j . As a result, we can represent the resistance, inductance, and current injection of the source as $R^{(s_j)}$, $L^{(s_j)}$, and $I^{(s_j)}(t)$, respectively.

For $j = 1, 2, \dots, |\mathcal{V}|$, let l_j be used to identify parameters or variables associated with an element connected to bus j . As a result, we can represent the resistance, inductance, and current injection of the element as $R^{(l_j)}$, $L^{(l_j)}$, and $I^{(l_j)}(t)$, respectively.

For $m = 1, 2, \dots, |\mathcal{E}|$, let $e_m := \{j, k\}$, $\{j, k\} \in \mathcal{E}$. As a result, we can represent the resistance, inductance, and current across a line extending from bus j to bus k as $R^{(e_m)}$, $L^{(e_m)}$, and $I^{(e_m)}(t)$, respectively.

2.3 A Primer on Singular Perturbation Analysis

Definition 1 (Big O notation) Consider a function $f(\epsilon)$, defined on some subset of the real numbers. We write $f(\epsilon) = \mathbf{O}(\epsilon^k)$ if and only if there exists a positive real number C , such that

$$|f(\epsilon)| \leq C\epsilon^k, \text{ as } \epsilon \rightarrow 0.$$

The material in this section follows closely from the developments in ([7], pp. 1–12) and ([4], pp. 7–11). Consider the following two-timescale dynamical model:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= f(\mathbf{x}(t), \mathbf{z}(t), \mathbf{w}(t), \epsilon), & \mathbf{x}(0) &= \mathbf{x}^0, \\ \epsilon \dot{\mathbf{z}}(t) &= g(\mathbf{x}(t), \mathbf{z}(t), \mathbf{w}(t), \epsilon), & \mathbf{z}(0) &= \mathbf{z}^0, \\ \mathbf{0} &= h(\mathbf{x}(t), \mathbf{z}(t), \mathbf{w}(t), \epsilon), & \mathbf{w}(0) &= \mathbf{w}^0,\end{aligned}\quad (6)$$

with slow and fast timescales, t and τ , respectively, where $\tau = \frac{t}{\epsilon}$, $f(\cdot, \cdot, \cdot, \epsilon) = \mathbf{O}(1)$, $g(\cdot, \cdot, \cdot, \epsilon) = \mathbf{O}(1)$, and $h(\cdot, \cdot, \cdot, \epsilon) = \mathbf{O}(1)$.

Assumption 2.1 Let the bar ($\bar{\cdot}$) and tilde ($\tilde{\cdot}$) notations be used to describe the slow t -scale and fast τ -scale variables, respectively. \mathbf{x} , \mathbf{z} , and \mathbf{w} can be decoupled to

$$\begin{aligned}\mathbf{x}(t) &= \bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}(\tau), \\ \mathbf{z}(t) &= \bar{\mathbf{z}}(t) + \tilde{\mathbf{z}}(\tau), \\ \mathbf{w}(t) &= \bar{\mathbf{w}}(t) + \tilde{\mathbf{w}}(\tau),\end{aligned}$$

where

$$\begin{aligned}\bar{\mathbf{x}}(t) &= \bar{\mathbf{x}}_0(t) + \epsilon \bar{\mathbf{x}}_1(t) + \epsilon^2 \bar{\mathbf{x}}_2(t) + \dots, \\ \tilde{\mathbf{x}}(\tau) &= \tilde{\mathbf{x}}_0(\tau) + \epsilon \tilde{\mathbf{x}}_1(\tau) + \epsilon^2 \tilde{\mathbf{x}}_2(\tau) + \dots, \\ \bar{\mathbf{z}}(t) &= \bar{\mathbf{z}}_0(t) + \epsilon \bar{\mathbf{z}}_1(t) + \epsilon^2 \bar{\mathbf{z}}_2(t) + \dots, \\ \tilde{\mathbf{z}}(\tau) &= \tilde{\mathbf{z}}_0(\tau) + \epsilon \tilde{\mathbf{z}}_1(\tau) + \epsilon^2 \tilde{\mathbf{z}}_2(\tau) + \dots, \\ \bar{\mathbf{w}}(t) &= \bar{\mathbf{w}}_0(t) + \epsilon \bar{\mathbf{w}}_1(t) + \epsilon^2 \bar{\mathbf{w}}_2(t) + \dots, \\ \tilde{\mathbf{w}}(\tau) &= \tilde{\mathbf{w}}_0(\tau) + \epsilon \tilde{\mathbf{w}}_1(\tau) + \epsilon^2 \tilde{\mathbf{w}}_2(\tau) + \dots.\end{aligned}$$

The dynamical model in Eq. 6 may be rewritten in terms of t and τ as

$$\begin{aligned}\dot{\bar{\mathbf{x}}}(t) + \frac{1}{\epsilon} \frac{d\tilde{\mathbf{x}}(\tau)}{d\tau} &= f(\bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}(\tau), \bar{\mathbf{z}}(t) + \tilde{\mathbf{z}}(\tau), \bar{\mathbf{w}}(t) + \tilde{\mathbf{w}}(\tau), \epsilon), \\ \epsilon \dot{\bar{\mathbf{z}}}(t) + \frac{d\tilde{\mathbf{z}}(\tau)}{d\tau} &= g(\bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}(\tau), \bar{\mathbf{z}}(t) + \tilde{\mathbf{z}}(\tau), \bar{\mathbf{w}}(t) + \tilde{\mathbf{w}}(\tau), \epsilon), \\ \mathbf{0} &= h(\bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}(\tau), \bar{\mathbf{z}}(t) + \tilde{\mathbf{z}}(\tau), \bar{\mathbf{w}}(t) + \tilde{\mathbf{w}}(\tau), \epsilon),\end{aligned}$$

and by setting $\epsilon = 0$, it follows that

$$\begin{aligned}\frac{d\tilde{\mathbf{x}}_0(\tau)}{d\tau} &= 0, \\ \dot{\bar{\mathbf{x}}}_0(t) &= f(\bar{\mathbf{x}}_0(t) + \tilde{\mathbf{x}}_0(\infty), \bar{\mathbf{z}}_0(t) + \tilde{\mathbf{z}}_0(\infty), \bar{\mathbf{w}}_0(t) + \tilde{\mathbf{w}}_0(\infty), 0), \\ \frac{d\tilde{\mathbf{z}}_0(\tau)}{d\tau} &= g(\bar{\mathbf{x}}_0(0) + \tilde{\mathbf{x}}_0(\tau), \bar{\mathbf{z}}_0(0) + \tilde{\mathbf{z}}_0(\tau), \bar{\mathbf{w}}_0(0) + \tilde{\mathbf{w}}_0(\tau), 0),\end{aligned}$$

and

$$\begin{aligned}\mathbf{0} &= h(\bar{\mathbf{x}}_0(0) + \tilde{\mathbf{x}}_0(\tau), \bar{\mathbf{z}}_0(0) + \tilde{\mathbf{z}}_0(\tau), \bar{\mathbf{w}}_0(0) + \tilde{\mathbf{w}}_0(\tau), 0), \\ \mathbf{0} &= h(\bar{\mathbf{x}}_0(t) + \tilde{\mathbf{x}}_0(\infty), \bar{\mathbf{z}}_0(t) + \tilde{\mathbf{z}}_0(\infty), \bar{\mathbf{w}}_0(t) + \tilde{\mathbf{w}}_0(\infty), 0).\end{aligned}\quad (7)$$

Assumption 2.2 *Equation 7 has distinct real roots, one of which is*

$$\begin{aligned}\bar{\mathbf{w}}_0(0) + \tilde{\mathbf{w}}_0(\tau) &= v(\bar{\mathbf{x}}_0(0) + \tilde{\mathbf{x}}_0(\tau), \bar{\mathbf{z}}_0(0) + \tilde{\mathbf{z}}_0(\tau)), \\ \bar{\mathbf{w}}_0(t) + \tilde{\mathbf{w}}_0(\infty) &= v(\bar{\mathbf{x}}_0(t) + \tilde{\mathbf{x}}_0(\infty), \bar{\mathbf{z}}_0(t) + \tilde{\mathbf{z}}_0(\infty)).\end{aligned}$$

Choosing initial conditions $\tilde{\mathbf{x}}_0(0) = 0$ and $\bar{\mathbf{x}}_0(0) = \mathbf{x}^0$, let $\bar{\mathbf{z}}_0(t) = \zeta(\bar{\mathbf{x}}_0(t))$ be a root of

$$\mathbf{0} = g(\bar{\mathbf{x}}_0(t), \bar{\mathbf{z}}_0(t), v(\bar{\mathbf{x}}_0(t), \bar{\mathbf{z}}_0(t)), 0). \quad (8)$$

As a result, the two-timescale dynamical model in Eq. 6 may be expressed in the approximate form

$$\dot{\bar{\mathbf{x}}}_0(t) = f(\bar{\mathbf{x}}_0(t), \zeta(\bar{\mathbf{x}}_0(t)), v(\bar{\mathbf{x}}_0(t), \zeta(\bar{\mathbf{x}}_0(t))), 0), \quad (9)$$

and

$$\frac{d\tilde{\mathbf{z}}_0(\tau)}{d\tau} = g\left(\mathbf{x}^0, \zeta(\mathbf{x}^0) + \tilde{\mathbf{z}}_0(\tau), v(\mathbf{x}^0, \zeta(\mathbf{x}^0) + \tilde{\mathbf{z}}_0(\tau)), 0\right), \quad (10)$$

where $\bar{\mathbf{x}}_0(0) = \mathbf{x}^0$ and $\tilde{\mathbf{z}}_0(0) = \mathbf{z}^0 - \zeta(\mathbf{x}^0)$.

Assumption 2.3 *The equilibrium $\tilde{\mathbf{z}}_0(\tau) = \mathbf{0}$ of Eq. 10 is asymptotically stable in \mathbf{x}^0 , and $\tilde{\mathbf{z}}_0(0)$ belongs to its domain of attraction.*

Assumption 2.4 *The eigenvalues of $\frac{\partial g}{\partial \mathbf{z}}$ (the Jacobian of Eq. 8) evaluated, for $\epsilon = 0$, along $\bar{\mathbf{x}}_0(t)$, $\bar{\mathbf{z}}_0(t)$, have real parts smaller than a fixed negative number.*

Theorem 1 (Tikhonov's theorem) *Let f and g in Eq. 6 be sufficiently many times continuously differentiable functions of their arguments, and let the root $\bar{\mathbf{z}}_0(t) = \zeta(\bar{\mathbf{x}}_0(t))$ of Eq. 8 be distinct and real, in the domain of interest (it follows from the implicit function theorem that the Jacobian of Eq. 8 must be invertible). Then, if assumptions 2.1, 2.2, 2.3, and 2.4 are satisfied, Eq. 6 can be approximated by Eq. 9 and Eq. 10, where*

$$\begin{aligned}\mathbf{x}(t) &= \bar{\mathbf{x}}_0(t) + \mathbf{O}(\epsilon), \\ \mathbf{z}(t) &= \zeta(\bar{\mathbf{x}}_0(t)) + \tilde{\mathbf{z}}_0(\tau) + \mathbf{O}(\epsilon), \\ \mathbf{w}(t) &= v(\bar{\mathbf{x}}_0(t), \zeta(\bar{\mathbf{x}}_0(t)) + \tilde{\mathbf{z}}_0(\tau)) + \mathbf{O}(\epsilon),\end{aligned}$$

and there exists $t_0 > 0$ such that

$$\mathbf{z}(t) = \zeta(\bar{\mathbf{x}}_0(t)) + \mathbf{O}(\epsilon),$$

$$\mathbf{w}(t) = v(\bar{\mathbf{x}}_0(t), \zeta(\bar{\mathbf{x}}_0(t))) + \mathbf{O}(\epsilon),$$

for all $t > t_0$.

In this work, we refer to the approximate slow component in Eq. 9 as the reduced-order model.

Definition 2 (Time resolution) The time resolution of the reduced-order model in Eq. 9 is the time it takes the approximate fast component in Eq. 10 to reach the equilibrium $\tilde{\mathbf{z}}_0(\tau) = \mathbf{0}$ from an initial state $\tilde{\mathbf{z}}_0(0) = \mathbf{z}^0 - \zeta(\mathbf{x}^0)$.

3 Microgrid High-Order Model (μ HOM)

In this section, basic circuit laws are used in conjunction with notions introduced in Section 2 to develop a high-order model for a grid-forming inverter-based AC microgrid operating in islanded mode. First, a model is developed for an inverter-interfaced source, which comprises a battery, a three-phase inverter, an LCL filter, and a voltage magnitude controller model. Next, a three-phase model for the electrical network is developed, along with a generic model for an element (typically a load) connected between each bus and the ground. The microgrid high-order model (μ HOM) is developed by combining the inverter-interfaced source model, the network model, and the generic element model. In this work, the models developed are expressed using the per-unit representation to ease analysis in later developments.

3.1 Inverter-Interfaced Source Model

The structure of the inverter-interfaced source adopted in this work is comprised of a three-phase inverter coupled with a battery, an LCL filter, and a voltage magnitude controller. An averaged model, as opposed to a switched model, is used to describe the three-phase inverter dynamics (see [14], pp. 27–38, for more details).

For the inverter connected to bus j of the microgrid network, let $V_{DC}^{(s_j)}$ denote the DC voltage at the inverter input. Let $U^{(s_j)}(t)$, $E^{(s_j)}(t)$, $\hat{E}^{(s_j)}(t)$, and $V^{(s_j)}(t)$ denote the pulse-width modulation (PWM) output voltage of the inverter, the internal voltage of the inverter, the LCL filter capacitor voltage, and the voltage at bus j , in per-unit representation, respectively. Let $\Xi^{(s_j)}(t)$ and $I^{(s_j)}(t)$ denote

the inverter output current and the filtered inverter output current, in per-unit representation, respectively; let $\Phi^{(s_j)}(t)$ and $\Gamma^{(s_j)}(t)$ denote the state variables for the voltage and current proportional-integral (PI) controllers, in per-unit representation, respectively; let $E_r^{(s_j)}(t)$ denote the voltage magnitude controller reference, in per-unit representation; and let $E_r^{(s_j)}(t) = E_{rq}^{(s_j)}(t) - jE_{rd}^{(s_j)}(t)$, where $E_{rd}^{(s_j)}(t) = 0$; let $\Xi_r^{(s_j)}(t)$ denote the current controller reference, in per-unit representation; let $\Xi_r^{(s_j)}(t) = \Xi_{rq}^{(s_j)}(t) - j\xi_{rd}^{(s_j)}(t)$; and let $P_f^{(s_j)}(t)$ and $Q_f^{(s_j)}(t)$ denote the filtered real and reactive power measurements, respectively. Then, using the $qd0$ transformation discussed in Section 2, the dynamics of the inverter-interfaced source connected to bus j of the microgrid electrical network can be described by

$$\begin{aligned}
D_\omega^{(s_j)} \frac{d\delta^{(s_j)}(t)}{dt} &= P_r^{(s_j)} - P_f^{(s_j)}(t), \\
\frac{1}{\omega_c^{(s_j)}} \frac{dQ_f^{(s_j)}(t)}{dt} &= -Q_f^{(s_j)}(t) + E_{q[\omega_0 t]}^{(s_j)}(t)I_{d[\omega_0 t]}^{(s_j)}(t) - E_{d[\omega_0 t]}^{(s_j)}(t)I_{q[\omega_0 t]}^{(s_j)}(t), \\
\frac{1}{\omega_c^{(s_j)}} \frac{dP_f^{(s_j)}(t)}{dt} &= -P_f^{(s_j)}(t) + E_{q[\omega_0 t]}^{(s_j)}(t)I_{q[\omega_0 t]}^{(s_j)}(t) + E_{d[\omega_0 t]}^{(s_j)}(t)I_{d[\omega_0 t]}^{(s_j)}(t), \\
\frac{L^{(s_j)}}{\omega_0} \frac{dI_{q[\omega_0 t]}^{(s_j)}(t)}{dt} &= -R^{(s_j)}I_{q[\omega_0 t]}^{(s_j)}(t) - L^{(s_j)}I_{d[\omega_0 t]}^{(s_j)}(t) + E_{q[\omega_0 t]}^{(s_j)}(t) - V_{q[\omega_0 t]}^{(l_j)}(t), \\
\frac{L^{(s_j)}}{\omega_0} \frac{dI_{d[\omega_0 t]}^{(s_j)}(t)}{dt} &= L^{(s_j)}I_{q[\omega_0 t]}^{(s_j)}(t) - R^{(s_j)}I_{d[\omega_0 t]}^{(s_j)}(t) + E_{d[\omega_0 t]}^{(s_j)}(t) - V_{d[\omega_0 t]}^{(l_j)}(t), \\
\frac{1}{\omega_0} \frac{d\Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} &= -\hat{E}_{q[\omega_0 t]}^{(s_j)}(t)\cos(\delta^{(s_j)}(t)) + \hat{E}_{d[\omega_0 t]}^{(s_j)}(t)\sin(\delta^{(s_j)}(t)) - \hat{R}_0^{(s_j)}\Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
&\quad + \hat{R}_0^{(s_j)}I_{q[\omega_0 t]}^{(s_j)}(t)\cos(\delta^{(s_j)}(t)) - \hat{R}_0^{(s_j)}I_{d[\omega_0 t]}^{(s_j)}(t)\sin(\delta^{(s_j)}(t)) \\
&\quad - \frac{1}{D_E^{(s_j)}}Q_f^{(s_j)}(t) + E_0^{(s_j)} + \frac{1}{D_E^{(s_j)}}Q_r^{(s_j)}, \\
\frac{1}{\omega_0} \frac{d\Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} &= -\hat{E}_{q[\omega_0 t]}^{(s_j)}(t)\sin(\delta^{(s_j)}(t)) - \hat{E}_{d[\omega_0 t]}^{(s_j)}(t)\cos(\delta^{(s_j)}(t)) - \hat{R}_0^{(s_j)}\Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
&\quad + \hat{R}_0^{(s_j)}I_{q[\omega_0 t]}^{(s_j)}(t)\sin(\delta^{(s_j)}(t)) + \hat{R}_0^{(s_j)}I_{d[\omega_0 t]}^{(s_j)}(t)\cos(\delta^{(s_j)}(t)), \\
C^{(s_j)} \frac{d\hat{E}_{q[\omega_0 t]}^{(s_j)}(t)}{dt} &= -I_{q[\omega_0 t]}^{(s_j)}(t) - C^{(s_j)}\hat{E}_{d[\omega_0 t]}^{(s_j)}(t) + \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t)\cos(\delta^{(s_j)}(t)) \\
&\quad + \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t)\sin(\delta^{(s_j)}(t)),
\end{aligned}$$

(continued)

$$\begin{aligned}
\frac{C^{(s_j)}}{\omega_0} \frac{d\hat{E}_{d[\omega_0 t]}^{(s_j)}(t)}{dt} &= -I_{d[\omega_0 t]}^{(s_j)}(t) + C^{(s_j)} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) - \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
&\quad + \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)), \\
L_0^{(s_j)} \frac{d\Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} &= K_{P\gamma}^{(s_j)} \left(1 + \frac{V_{DC}^{(s_j)} K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}}{2} \right) I_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
&\quad - K_{P\gamma}^{(s_j)} \left(1 + \frac{V_{DC}^{(s_j)} K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}}{2} \right) I_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
&\quad - \left(R_0^{(s_j)} + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)}}{2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) \right) \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
&\quad - \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)}}{2} \Gamma_{q[\alpha^{(j)}(t)]}^{(s_j)} \\
&\quad + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2} E_0^{(s_j)} \\
&\quad + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2} \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2 D_E^{(s_j)}} Q_r^{(s_j)} \\
&\quad - \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2 D_E^{(s_j)}} Q_f^{(s_j)}(t) + K_{P\gamma}^{(s_j)} C^{(s_j)} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
&\quad + \frac{K_{P\gamma}^{(s_j)} C^{(s_j)}}{D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
&\quad + K_{P\gamma}^{(s_j)} C^{(s_j)} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
&\quad + \frac{K_{P\gamma}^{(s_j)} C^{(s_j)}}{D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)), \\
L_0^{(s_j)} \frac{d\Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} &= K_{P\gamma}^{(s_j)} \left(1 + \frac{V_{DC}^{(s_j)} K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}}{2} \right) I_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
&\quad + K_{P\gamma}^{(s_j)} \left(1 + \frac{V_{DC}^{(s_j)} K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}}{2} \right) I_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
&\quad - \left(R_0^{(s_j)} + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)}}{2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) \right) \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t)
\end{aligned}$$

(continued)

$$\begin{aligned}
& - \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) + \frac{V_{DC}^{(s_j)} K_{I\gamma}^{(s_j)}}{2} \Gamma_{d[\alpha^{(j)}(t)]}^{(s_j)} \\
& - \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{P\phi}^{(s_j)}}{2} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& - \frac{K_{P\gamma}^{(s_j)} C^{(s_j)}}{D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& + \frac{K_{P\gamma}^{(s_j)} C^{(s_j)}}{D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& - K_{P\gamma}^{(s_j)} C^{(s_j)} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& + K_{P\gamma}^{(s_j)} C^{(s_j)} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& + \frac{V_{DC}^{(s_j)} K_{P\gamma}^{(s_j)} K_{I\phi}^{(s_j)}}{2} \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t), \\
\frac{1}{\omega_0} \frac{d\Gamma_{q[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} & = \left(\frac{2}{V_{DC}^{(s_j)}} + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) I_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& - \left(\frac{2}{V_{DC}^{(s_j)}} + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) I_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) + K_{P\phi}^{(s_j)} E_0^{(s_j)} \\
& - K_{P\phi}^{(s_j)} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) + K_{P\phi}^{(s_j)} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& - \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) + \frac{K_{P\phi}^{(s_j)}}{D_E^{(s_j)}} \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t) \right) \\
& + \frac{2C^{(s_j)}}{V_{DC}^{(s_j)}} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& + \frac{2C^{(s_j)}}{V_{DC}^{(s_j)} D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& + \frac{2C^{(s_j)}}{V_{DC}^{(s_j)} D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& + K_{I\phi}^{(s_j)} \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t),
\end{aligned}$$

(continued)

$$\begin{aligned}
\frac{1}{\omega_0} \frac{d\Gamma_{d[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} = & \left(\frac{2}{V_{DC}^{(s_j)}} + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) I_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& + \left(\frac{2}{V_{DC}^{(s_j)}} + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right) I_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& - K_{P\phi}^{(s_j)} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) - K_{P\phi}^{(s_j)} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& - \left(1 + K_{P\phi}^{(s_j)} R_0^{(s_j)} \right) \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) + K_{I\phi}^{(s_j)} \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
& - \frac{2C^{(s_j)}}{V_{DC}^{(s_j)}} \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) + \frac{2C^{(s_j)}}{V_{DC}^{(s_j)}} \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& - \frac{2C^{(s_j)}}{V_{DC}^{(s_j)} D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& + \frac{2C^{(s_j)}}{V_{DC}^{(s_j)} D_\omega^{(s_j)} \omega_0} \left(P_r^{(s_j)} - P_f^{(s_j)}(t) \right) \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)), \\
E_{q[\omega_0 t]}^{(s_j)}(t) = & - \hat{R}_0^{(s_j)} I_{q[\omega_0 t]}^{(s_j)}(t) + \hat{R}_0^{(s_j)} \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) \\
& + \hat{R}_0^{(s_j)} \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) + \hat{E}_{q[\omega_0 t]}^{(s_j)}(t), \\
E_{d[\omega_0 t]}^{(s_j)}(t) = & - \hat{R}_0^{(s_j)} I_{d[\omega_0 t]}^{(s_j)}(t) - \hat{R}_0^{(s_j)} \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) \\
& + \hat{R}_0^{(s_j)} \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) + \hat{E}_{d[\omega_0 t]}^{(s_j)}(t),
\end{aligned}$$

where $L_0^{(s_j)}$, $L^{(s_j)}$, and $C^{(s_j)}$ denote the inductances and capacitance of the *LCL* filter, in per-unit representation, respectively; $R_0^{(s_j)}$, $\hat{R}_0^{(s_j)}$, and $R^{(s_j)}$ denote the inverter and filter resistances, in per-unit representation, respectively; $K_{P\phi}^{(s_j)}$ and $K_{P\gamma}^{(s_j)}$ denote the proportional controller gains for the voltage and current controllers, in per-unit representation, respectively; $K_{I\phi}^{(s_j)}$ and $K_{I\gamma}^{(s_j)}$ denote the corresponding integral controller gains; $D_E^{(s_j)}$ and $D_\omega^{(s_j)}$ denote the voltage and frequency droop coefficients, respectively; $E_0^{(s_j)}$ denotes the voltage droop law constant; $\omega_c^{(s_j)}$ denotes the filter cut-off frequency; $P_r^{(s_j)}$; and $Q_r^{(s_j)}$ denote real and reactive power set points, respectively. See [1], pp. 10–13 for details of this result.

3.2 Network Model

Assumption 3.1 All lines connecting the network buses can be represented using the short transmission line model [3].

Let $V_{q[\omega_0 t]}^{(l_j)}(t) - j V_{d[\omega_0 t]}^{(l_j)}(t)$ denote the per-unit voltage at bus j , and let $R^{(e_m)}$, $L^{(e_m)}$, and $I_{q[\omega_0 t]}^{(e_m)}(t) - j I_{d[\omega_0 t]}^{(e_m)}(t)$ denote the per-unit resistance, inductance, and current across line (j, k) , respectively, as introduced in Section 2.2. Then, the voltage across a line connecting bus j and bus k of the network can be described by

$$\begin{aligned} V_{q[\omega_0 t]}^{(l_j)}(t) - V_{q[\omega_0 t]}^{(l_k)}(t) &= \frac{L^{(e_m)}}{\omega_0} \frac{d I_{q[\omega_0 t]}^{(e_m)}(t)}{dt} + R^{(e_m)} I_{q[\omega_0 t]}^{(e_m)}(t) + L^{(e_m)} I_{d[\omega_0 t]}^{(e_m)}(t), \\ V_{d[\omega_0 t]}^{(l_j)}(t) - V_{d[\omega_0 t]}^{(l_k)}(t) &= \frac{L^{(e_m)}}{\omega_0} \frac{d I_{d[\omega_0 t]}^{(e_m)}(t)}{dt} + R^{(e_m)} I_{d[\omega_0 t]}^{(e_m)}(t) - L^{(e_m)} I_{q[\omega_0 t]}^{(e_m)}(t). \end{aligned}$$

Let

$$\begin{aligned} \mathbf{V}_{q[\omega_0 t]}^{(\mathcal{V})}(t) &= \left[V_{q[\omega_0 t]}^{(l_1)}(t) \ V_{q[\omega_0 t]}^{(l_2)}(t) \ \dots \ V_{q[\omega_0 t]}^{(l_{|\mathcal{V}|})}(t) \right]^T, \\ \mathbf{V}_{d[\omega_0 t]}^{(\mathcal{V})}(t) &= \left[V_{d[\omega_0 t]}^{(l_1)}(t) \ V_{d[\omega_0 t]}^{(l_2)}(t) \ \dots \ V_{d[\omega_0 t]}^{(l_{|\mathcal{V}|})}(t) \right]^T, \\ \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) &= \left[I_{q[\omega_0 t]}^{(e_1)}(t) \ I_{q[\omega_0 t]}^{(e_2)}(t) \ \dots \ I_{q[\omega_0 t]}^{(e_{|\mathcal{E}|})}(t) \right]^T, \\ \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) &= \left[I_{d[\omega_0 t]}^{(e_1)}(t) \ I_{d[\omega_0 t]}^{(e_2)}(t) \ \dots \ I_{d[\omega_0 t]}^{(e_{|\mathcal{E}|})}(t) \right]^T. \end{aligned}$$

Then the network dynamics are described by

$$\begin{aligned} \frac{1}{\omega_0} \mathbf{L}^{(\mathcal{E})} \frac{d \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t)}{dt} &= -\mathbf{R}^{(\mathcal{E})} \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) - \mathbf{L}^{(\mathcal{E})} \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) + \mathbf{M}^T \mathbf{V}_{q[\omega_0 t]}^{(\mathcal{V})}(t), \\ \frac{1}{\omega_0} \mathbf{L}^{(\mathcal{E})} \frac{d \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t)}{dt} &= -\mathbf{R}^{(\mathcal{E})} \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) + \mathbf{L}^{(\mathcal{E})} \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) + \mathbf{M}^T \mathbf{V}_{d[\omega_0 t]}^{(\mathcal{V})}(t), \end{aligned} \quad (11)$$

with

$$\begin{aligned} \mathbf{R}^{(\mathcal{E})} &= \text{diag} \left(R^{(e_1)}, R^{(e_2)}, \dots, R^{(e_{|\mathcal{E}|})} \right), \\ \mathbf{L}^{(\mathcal{E})} &= \text{diag} \left(L^{(e_1)}, L^{(e_2)}, \dots, L^{(e_{|\mathcal{E}|})} \right), \end{aligned}$$

where $\text{diag} (d^{(1)}, d^{(2)}, \dots, d^{(n)})$ is a diagonal matrix with diagonal entries $d^{(1)}, d^{(2)}, \dots, d^{(n)}$; and \mathbf{M} denotes the network incidence matrix as defined in Section 2.

3.3 Generic Element Model

Let $V_{q[\omega_0 t]}^{(l_j)}(t) - j V_{d[\omega_0 t]}^{(l_j)}(t)$ denote the per-unit voltage at bus j , and let $I_{q[\omega_0 t]}^{(l_j)}(t) - j I_{d[\omega_0 t]}^{(l_j)}(t)$ denote the per-unit current injection by an element (typically a load) at bus j . The dynamics can be described by a generic nonlinear system of differential equations which we assume to be of the form

$$\begin{aligned} \mu_V^{(l_j)} \dot{V}_{q[\omega_0 t]}^{(l_j)}(t) &= q_V(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\ \mu_V^{(l_j)} \dot{V}_{d[\omega_0 t]}^{(l_j)}(t) &= d_V(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\ \mu_I^{(l_j)} \dot{I}_{q[\omega_0 t]}^{(l_j)}(t) &= q_I(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\ \mu_I^{(l_j)} \dot{I}_{d[\omega_0 t]}^{(l_j)}(t) &= d_I(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \end{aligned} \quad (12)$$

where $\mu_V^{(l_j)}$ and $\mu_I^{(l_j)}$ represent time constants of the generic element at bus j ; and $q_V(\cdot, \cdot, \cdot, \cdot)$, $d_V(\cdot, \cdot, \cdot, \cdot)$, $q_I(\cdot, \cdot, \cdot, \cdot)$, and $d_I(\cdot, \cdot, \cdot, \cdot)$ are nonlinear functions of its state variables.

4 Microgrid Reduced-Order Model 1 (μROM1)

In this section, the singular perturbation techniques discussed in Section 2.3 are used to reduce the order (state-space dimension) of the μHOM to obtain μROM1 .

Assumption 4.1 For $\epsilon_1 = 1 \times 10^{-5}$, the dynamic properties of the μHOM are such that at each bus j :

$$\begin{aligned} \mathbf{x}_1(t) &= \left[\delta^{(s_j)}(t) \ Q_f^{(s_j)}(t) \ P_f^{(s_j)}(t) \ I_{q[\omega_0 t]}^{(s_j)}(t) \ I_{d[\omega_0 t]}^{(s_j)}(t) \ \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Phi_{d[\alpha^{(s_j)}(t)]}^{(j)}(t) \right. \\ &\quad \left. \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) \ \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) \ I_{q[\omega_0 t]}^{(l_j)}(t) \ I_{d[\omega_0 t]}^{(l_j)}(t) \ V_{q[\omega_0 t]}^{(l_j)}(t) \ V_{d[\omega_0 t]}^{(l_j)}(t) \right]^T, \\ \mathbf{z}_1(t) &= \left[\Gamma_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Gamma_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \ \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \right]^T, \\ \text{and } \mathbf{w}_1(t) &= \left[E_{q[\omega_0 t]}^{(s_j)}(t) \ E_{d[\omega_0 t]}^{(s_j)}(t) \right]^T, \text{ the dynamics of } \mathbf{z}_1(t) \text{ are faster than those of } \mathbf{x}_1(t), \text{ and the } \mu\text{HOM} \text{ can be expressed compactly as follows:} \end{aligned}$$

$$\begin{aligned} \dot{\mathbf{x}}_1(t) &= f_1(\mathbf{x}_1(t), \mathbf{z}_1(t), \mathbf{w}_1(t), \epsilon_1), \\ \epsilon_1 \dot{\mathbf{z}}_1(t) &= g_1(\mathbf{x}_1(t), \mathbf{z}_1(t), \mathbf{w}_1(t), \epsilon_1), \\ \mathbf{0} &= h_1(\mathbf{x}_1(t), \mathbf{z}_1(t), \mathbf{w}_1(t), \epsilon_1). \end{aligned} \quad (13)$$

Assumption 4.2 Equation 13 satisfies the conditions for Tikhonov's theorem, as presented in Section 2.3.

Given Assumptions 4.1–4.2, the μ HOM can be reduced to the so-called microgrid reduced-order model 1 (μ ROM1).

The explicit ordinary differential equations (ODEs) that constitute μ ROM1 are as follows (see [1] pp. 25–28 for a detailed derivation of this result):

$$\begin{aligned}
D_{\omega}^{(s_j)} \frac{d\delta^{(s_j)}(t)}{dt} &= P_r^{(s_j)} - P_f^{(s_j)}(t), \\
\frac{1}{\omega_c^{(s_j)}} \frac{dQ_f^{(s_j)}(t)}{dt} &= -Q_f^{(s_j)}(t) + E_{q[\omega_0 t]}^{(s_j)}(t)I_{d[\omega_0 t]}^{(s_j)}(t) - E_{d[\omega_0 t]}^{(s_j)}(t)I_{q[\omega_0 t]}^{(s_j)}(t), \\
\frac{1}{\omega_c^{(s_j)}} \frac{dP_f^{(s_j)}(t)}{dt} &= -P_f^{(s_j)}(t) + E_{q[\omega_0 t]}^{(s_j)}(t)I_{q[\omega_0 t]}^{(s_j)}(t) + E_{d[\omega_0 t]}^{(s_j)}(t)I_{d[\omega_0 t]}^{(s_j)}(t), \\
\mu_V^{(l_j)} \dot{V}_{q[\omega_0 t]}^{(l_j)}(t) &= q_V(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\
\mu_V^{(l_j)} \dot{V}_{d[\omega_0 t]}^{(l_j)}(t) &= d_V(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\
\mu_I^{(l_j)} \dot{I}_{q[\omega_0 t]}^{(l_j)}(t) &= q_I(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\
\mu_I^{(l_j)} \dot{I}_{d[\omega_0 t]}^{(l_j)}(t) &= d_I(V_{q[\omega_0 t]}^{(l_j)}(t), V_{d[\omega_0 t]}^{(l_j)}(t), I_{q[\omega_0 t]}^{(l_j)}(t), I_{d[\omega_0 t]}^{(l_j)}(t)), \\
\frac{1}{\omega_0} \mathbf{L}^{(\mathcal{E})} \frac{d\mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t)}{dt} &= -\mathbf{R}^{(\mathcal{E})}\mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) - \mathbf{L}^{(\mathcal{E})}\mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) + \mathbf{M}^T \mathbf{V}_{q[\omega_0 t]}^{(\mathcal{V})}(t), \\
\frac{1}{\omega_0} \mathbf{L}^{(\mathcal{E})} \frac{d\mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t)}{dt} &= -\mathbf{R}^{(\mathcal{E})}\mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) + \mathbf{L}^{(\mathcal{E})}\mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) + \mathbf{M}^T \mathbf{V}_{d[\omega_0 t]}^{(\mathcal{V})}(t), \\
\frac{L^{(s_j)}}{\omega_0} \frac{dI_{q[\omega_0 t]}^{(s_j)}(t)}{dt} &= -R^{(s_j)}I_{q[\omega_0 t]}^{(s_j)}(t) - L^{(s_j)}I_{d[\omega_0 t]}^{(s_j)}(t) + E_{q[\omega_0 t]}^{(s_j)}(t) - V_{q[\omega_0 t]}^{(l_j)}(t), \\
\frac{L^{(s_j)}}{\omega_0} \frac{dI_{d[\omega_0 t]}^{(s_j)}(t)}{dt} &= L^{(s_j)}I_{q[\omega_0 t]}^{(s_j)}(t) - R^{(s_j)}I_{d[\omega_0 t]}^{(s_j)}(t) + E_{d[\omega_0 t]}^{(s_j)}(t) - V_{d[\omega_0 t]}^{(l_j)}(t), \\
\frac{1}{\omega_0} \frac{d\Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} &= -\frac{K_{I\phi}^{(s_j)} K_{P\phi}^{(s_j)}}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
&\quad + \frac{C^{(s_j)} K_{I\phi}^{(s_j)}}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
&\quad + \frac{K_{P\phi}^{(s_j)} \left(I_{q[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) - I_{d[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t))\right)}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2}
\end{aligned}$$

(continued)

$$\begin{aligned}
& + \frac{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 E_0^{(s_j)}}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& + \frac{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t)\right)}{D_E^{(s_j)} C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + D_E^{(s_j)} \left(K_{P\phi}^{(s_j)}\right)^2}, \\
\omega_0 \frac{d\Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t)}{dt} = & - \frac{K_{I\phi}^{(s_j)} K_{P\phi}^{(s_j)}}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
& + \frac{C^{(s_j)} K_{I\phi}^{(s_j)}}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \\
& + \frac{K_{P\phi}^{(s_j)} \left(I_{q[\omega_0 t]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) + I_{d[\omega_0 t]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t))\right)}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& - \frac{C^{(s_j)} K_{P\phi}^{(s_j)} \left(E_0^{(s_j)} + \frac{1}{D_E^{(s_j)}} \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t)\right)\right)}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2},
\end{aligned}$$

where

$$\begin{aligned}
E_{q[\omega_0 t]}^{(s_j)}(t) = & - \frac{K_{P\phi}^{(s_j)}}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} I_{q[\omega_0 t]}^{(s_j)}(t) \\
& + \frac{C^{(s_j)} K_{I\phi}^{(s_j)} \left(\Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) - \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t))\right)}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& + \frac{K_{I\phi}^{(s_j)} K_{P\phi}^{(s_j)} \left(\Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) + \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t))\right)}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& + \frac{E_0^{(s_j)} C^{(s_j)} K_{P\phi}^{(s_j)} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right) \sin(\delta^{(s_j)}(t))}{C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& + \frac{C^{(s_j)} K_{P\phi}^{(s_j)} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right) \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t)\right) \sin(\delta^{(s_j)}(t))}{D_E^{(s_j)} C^{(s_j)}^2 \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + D_E^{(s_j)} \left(K_{P\phi}^{(s_j)}\right)^2}
\end{aligned}$$

(continued)

$$\begin{aligned}
& + \frac{\left(K_{P\phi}^{(s_j)}\right)^2 \cos(\delta^{(s_j)}(t)) \left(E_0^{(s_j)} + \frac{1}{D_E^{(s_j)}} \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t)\right)\right)}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2}, \\
E_{d[\omega_0 t]}^{(s_j)}(t) = & - \frac{K_{P\phi}^{(s_j)}}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} I_{d[\omega_0 t]}^{(s_j)}(t) \\
& + \frac{C^{(s_j)} K_{I\phi}^{(s_j)} \left(\Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t)) + \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t))\right)}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& - \frac{K_{I\phi}^{(s_j)} K_{P\phi}^{(s_j)} \left(\Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \sin(\delta^{(s_j)}(t)) - \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \cos(\delta^{(s_j)}(t))\right)}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& + \frac{E_0^{(s_j)} C^{(s_j)} K_{P\phi}^{(s_j)} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right) \cos(\delta^{(s_j)}(t))}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2} \\
& + \frac{C^{(s_j)} K_{P\phi}^{(s_j)} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right) \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t)\right) \cos(\delta^{(s_j)}(t))}{D_E^{(s_j)} C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + D_E^{(s_j)} \left(K_{P\phi}^{(s_j)}\right)^2} \\
& - \frac{\left(K_{P\phi}^{(s_j)}\right)^2 \sin(\delta^{(s_j)}(t)) \left(E_0^{(s_j)} + \frac{1}{D_E^{(s_j)}} \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t)\right)\right)}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)}\right)^2 + \left(K_{P\phi}^{(s_j)}\right)^2}.
\end{aligned}$$

5 Microgrid Reduced-Order Model 2 (μ ROM2)

In this section, the singular perturbation techniques discussed in Section 2.3 are used to reduce the order (state-space dimension) of the μ HOM to obtain μ ROM2.

Assumption 5.1 For $\epsilon_2 = 1 \times 10^{-3}$, the dynamic properties of the μ HOM are such that at each bus j :

$$\begin{aligned}
\mathbf{z}_2(t) = & \left[I_{q[\omega_0 t]}^{(l_j)}(t) \ I_{d[\omega_0 t]}^{(l_j)}(t) \ V_{q[\omega_0 t]}^{(l_j)}(t) \ V_{d[\omega_0 t]}^{(l_j)}(t) \ \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) \ \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) \ I_{q[\omega_0 t]}^{(s_j)}(t) \ I_{d[\omega_0 t]}^{(s_j)}(t) \right. \\
& \left. \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Gamma_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Gamma_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) \ \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) \right. \\
& \left. \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) \ \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) \right]^T,
\end{aligned}$$

$\mathbf{x}_2(t) = \left[\delta^{(s_j)}(t) \ Q_f^{(s_j)}(t) \ P_f^{(s_j)}(t) \right]^T$, and $\mathbf{w}_2(t) = \left[E_{q[\omega_0 t]}^{(s_j)}(t) \ E_{d[\omega_0 t]}^{(s_j)}(t) \right]^T$, the dynamics of $\mathbf{z}_2(t)$ are faster than those of $\mathbf{x}_2(t)$, and the μHOM can be expressed compactly as follows:

$$\begin{aligned} \dot{\mathbf{x}}_2(t) &= f_2(\mathbf{x}_2(t), \mathbf{z}_2(t), \mathbf{w}_2(t), \epsilon_2), \\ \epsilon_2 \dot{\mathbf{z}}_2(t) &= g_2(\mathbf{x}_2(t), \mathbf{z}_2(t), \mathbf{w}_2(t), \epsilon_2), \\ \mathbf{0} &= h_2(\mathbf{x}_2(t), \mathbf{z}_2(t), \mathbf{w}_2(t), \epsilon_2). \end{aligned} \quad (14)$$

Assumption 5.2 Equation 14 satisfies the conditions for Tikhonov's theorem, as presented in Section 2.3.

Given Assumptions 5.1–5.2, the μHOM can be reduced to the so-called microgrid reduced-order model 2 (μROM2) using the developments in Section 2.3.

Let $\theta^{(s_j)}(t) := \arctan\left(\frac{-E_{d[\omega_0 t]}^{(s_j)}(t)}{E_{q[\omega_0 t]}^{(s_j)}(t)}\right)$, and $\theta^{(l_j)}(t) := \arctan\left(\frac{-V_{d[\omega_0 t]}^{(l_j)}(t)}{V_{q[\omega_0 t]}^{(l_j)}(t)}\right)$. Let $\beta^{(j)} \in \{0, 1\}$ be a constant such that $\beta^{(j)} = 1$ if bus $j \in \mathcal{V}_{\mathcal{J}}$, and $\beta^{(j)} = 0$ otherwise. Also, let $|\vec{\mathbf{V}}^{(e_m)}(t)| = |\vec{\mathbf{V}}^{(l_j)}(t)| |\vec{\mathbf{V}}^{(l_k)}(t)|$, and

$$\text{at bus } j : \quad \theta^{(e_m)}(t) = \theta^{(l_j)}(t) - \theta^{(l_k)}(t),$$

$$\text{at bus } k : \quad \theta^{(e_m)}(t) = \theta^{(l_k)}(t) - \theta^{(l_j)}(t).$$

Assumption 5.3 The generic model in (12) can be reduced to the so-called ZIP model (see e.g. [12]), given by

$$V_{q[\omega_0 t]}^{(l_j)}(t) I_{q[\omega_0 t]}^{(l_j)}(t) + V_{d[\omega_0 t]}^{(l_j)}(t) I_{d[\omega_0 t]}^{(l_j)}(t) = -P_0^{(l_j)} - |\vec{\mathbf{V}}^{(l_j)}(t)| P_1^{(l_j)} - |\vec{\mathbf{V}}^{(l_j)}(t)|^2 P_2^{(l_j)},$$

$$V_{q[\omega_0 t]}^{(l_j)}(t) I_{d[\omega_0 t]}^{(l_j)}(t) - V_{d[\omega_0 t]}^{(l_j)}(t) I_{q[\omega_0 t]}^{(l_j)}(t) = -Q_0^{(l_j)} - |\vec{\mathbf{V}}^{(l_j)}(t)| Q_1^{(l_j)} - |\vec{\mathbf{V}}^{(l_j)}(t)|^2 Q_2^{(l_j)},$$

where $P_0^{(l_j)}$, $P_1^{(l_j)}$, $P_2^{(l_j)}$, $Q_0^{(l_j)}$, $Q_1^{(l_j)}$ and $Q_2^{(l_j)}$ denote constants for the element at bus j , and $|\vec{\mathbf{V}}^{(l_j)}(t)|$ denotes the phasor magnitude of $V_{q[\omega_0 t]}^{(l_j)}(t) - jV_{d[\omega_0 t]}^{(l_j)}(t)$.

The explicit ordinary differential equations (ODEs) that constitute μROM2 are as follows (see [1] pp. 31–36 for a detailed derivation of this result):

$$\begin{aligned}
D_{\omega}^{(s_j)} \frac{d\theta^{(s_j)}(t)}{dt} &= P_r^{(s_j)} - P_f^{(s_j)}(t), \\
\frac{1}{\omega_c^{(s_j)}} \frac{dQ_f^{(s_j)}(t)}{dt} &= -B^{(s_j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 - \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \sin(\theta^{(s_j)}(t) \right. \\
&\quad \left. - \theta^{(l_j)}(t)) - B^{(s_j)} \cos(\theta^{(s_j)}(t) - \theta^{(l_j)}(t)) \right) - Q_f^{(s_j)}(t), \\
\frac{1}{\omega_c^{(s_j)}} \frac{dP_f^{(s_j)}(t)}{dt} &= G^{(s_j)} \left| \vec{\mathbf{E}}^{(s_j)}(t) \right|^2 - \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \cos(\theta^{(s_j)}(t) \right. \\
&\quad \left. - \theta^{(l_j)}(t)) + B^{(s_j)} \sin(\theta^{(s_j)}(t) - \theta^{(l_j)}(t)) \right) - P_f^{(s_j)}(t),
\end{aligned}$$

and for $\mathbf{E}(j)$ representing the set of edges incident to node j , such that $e_m \in \mathbf{E}(j)$ if and only if the edge e_m is incident to node j , the power balance equations at bus $j \in \mathcal{V}$ are given by

$$\begin{aligned}
0 &= P_0^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| P_1^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 P_2^{(l_j)} + \beta^{(j)} G^{(s_j)} \left| V^{(l_j)}(t) \right|^2 \\
&\quad - \beta^{(j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \cos(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right. \\
&\quad \left. + B^{(s_j)} \sin(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right) + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 \sum_{e_m \in \mathbf{E}(j)} G^{(e_m)} \\
&\quad - \sum_{e_m \in \mathbf{E}(j)} \left| \vec{\mathbf{V}}^{(e_m)}(t) \right| \left(G^{(e_m)} \cos(\theta^{(e_m)}(t)) + B^{(e_m)} \sin(\theta^{(e_m)}(t)) \right), \\
0 &= Q_0^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| Q_1^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 Q_2^{(l_j)} - \beta^{(j)} B^{(l_j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 \\
&\quad - \beta^{(j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \sin(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right. \\
&\quad \left. - B^{(s_j)} \cos(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right) - \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 \sum_{e_m \in \mathbf{E}(j)} B^{(e_m)} \\
&\quad - \sum_{e_m \in \mathbf{E}(j)} \left| \vec{\mathbf{V}}^{(e_m)}(t) \right| \left(G^{(e_m)} \sin(\theta^{(e_m)}(t)) - B^{(e_m)} \cos(\theta^{(e_m)}(t)) \right).
\end{aligned}$$

(continued)

and

$$\left| \overrightarrow{\mathbf{E}}^{(s_j)}(t) \right| = \frac{\left(K_{P\phi}^{(s_j)} \right)^2 \left(E_0^{(s_j)} + \frac{1}{D_E^{(s_j)}} \left(Q_r^{(s_j)} - Q_f^{(s_j)}(t) \right) \right)}{C^{(s_j)2} \left(1 + K_{P\phi}^{(s_j)} \hat{R}_0^{(s_j)} \right)^2 + \left(K_{P\phi}^{(s_j)} \right)^2}.$$

$$\text{where } \hat{G}^{(s_j)} = \frac{\hat{R}_0^{(s_j)}}{\left(\hat{R}_0^{(s_j)} \right)^2 + \left(\frac{1}{C^{(s_j)}} \right)^2}, \quad \hat{B}^{(s_j)} = \frac{C^{(s_j)}}{\left(C^{(s_j)} \hat{R}_0^{(s_j)} \right)^2 + 1}, \quad G^{(s_j)} = \frac{R^{(s_j)}}{\left(R^{(s_j)} \right)^2 + \left(L^{(s_j)} \right)^2}, \quad B^{(s_j)} = \frac{-L^{(s_j)}}{\left(R^{(s_j)} \right)^2 + \left(L^{(s_j)} \right)^2}, \quad G^{(e_m)} = \frac{R^{(e_m)}}{\left(R^{(e_m)} \right)^2 + \left(L^{(e_m)} \right)^2}, \text{ and } B^{(e_m)} = \frac{-L^{(e_m)}}{\left(R^{(e_m)} \right)^2 + \left(L^{(e_m)} \right)^2}.$$

6 Microgrid Reduced-Order Model 3 (μROM3)

In this section, the singular perturbation techniques discussed in Section 2.3 are used to reduce the order (state-space dimension) of the μHOM to obtain μROM3 .

Assumption 6.1 For $\epsilon_3 = 1 \times 10^{-1}$, the dynamic properties of the μHOM are such that at each bus j :

$$\mathbf{z}_3(t) = \begin{bmatrix} Q_f^{(s_j)}(t) & P_f^{(s_j)}(t) & I_{q[\omega_0 t]}^{(l_j)}(t) & I_{d[\omega_0 t]}^{(l_j)}(t) & V_{q[\omega_0 t]}^{(l_j)}(t) & V_{d[\omega_0 t]}^{(l_j)}(t) & \mathbf{I}_{q[\omega_0 t]}^{(\mathcal{E})}(t) & \mathbf{I}_{d[\omega_0 t]}^{(\mathcal{E})}(t) \\ I_{q[\omega_0 t]}^{(s_j)}(t) & I_{d[\omega_0 t]}^{(s_j)}(t) & \Phi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) & \Phi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) & \Gamma_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) & \Gamma_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) & & \\ \Xi_{q[\alpha^{(j)}(t)]}^{(s_j)}(t) & \Xi_{d[\alpha^{(j)}(t)]}^{(s_j)}(t) & \hat{E}_{q[\omega_0 t]}^{(s_j)}(t) & \hat{E}_{d[\omega_0 t]}^{(s_j)}(t) & & & & \end{bmatrix}^T,$$

$\mathbf{x}_3(t) = \delta^{(s_j)}(t)$, and $\mathbf{w}_3(t) = \begin{bmatrix} E_{q[\omega_0 t]}^{(s_j)}(t) & E_{d[\omega_0 t]}^{(s_j)}(t) \end{bmatrix}^T$, the dynamics of $\mathbf{z}_3(t)$ are faster than those of $\mathbf{x}_3(t)$, and the μHOM can be expressed compactly as follows:

$$\begin{aligned} \dot{\mathbf{x}}_3(t) &= f_3(\mathbf{x}_3(t), \mathbf{z}_3(t), \mathbf{w}_3(t), \epsilon_3), \\ \epsilon_3 \dot{\mathbf{z}}_3(t) &= g_3(\mathbf{x}_3(t), \mathbf{z}_3(t), \mathbf{w}_3(t), \epsilon_3), \\ \mathbf{0} &= h_3(\mathbf{x}_3(t), \mathbf{z}_3(t), \mathbf{w}_3(t), \epsilon_3). \end{aligned} \tag{15}$$

Assumption 6.2 Equation 15 satisfies the conditions for Tikhonov's theorem, as presented in Section 2.3.

Given Assumptions 6.1–6.2, the μHOM can be reduced to the so-called microgrid reduced-order model 3 (μROM3).

Using Assumption 5.3 and the definitions in Section 5, the explicit ordinary differential equations (ODEs) that constitute μROm3 are as follows (see [1] pp. 37–41 for a detailed derivation of this result):

$$\begin{aligned} D_{\omega}^{(s_j)} \frac{d\theta^{(s_j)}(t)}{dt} &= P_r^{(s_j)} - G^{(s_j)} \left| \vec{\mathbf{E}}^{(s_j)}(t) \right|^2 + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \\ &\quad \left(G^{(s_j)} \cos(\theta^{(s_j)}(t) - \theta^{(l_j)}(t)) + B^{(s_j)} \sin(\theta^{(s_j)}(t) - \theta^{(l_j)}(t)) \right), \\ \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| &= \frac{1}{D_E^{(s_j)}} \left(Q_r^{(s_j)} + B^{(s_j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \right. \right. \\ &\quad \left. \left. \sin(\theta^{(s_j)}(t) - \theta^{(l_j)}(t)) - B^{(s_j)} \cos(\theta^{(s_j)}(t) - \theta^{(l_j)}(t)) \right) \right) + E_0^{(s_j)}, \end{aligned}$$

and for $E(j)$ representing the set of edges incident to node j , such that $e_m \in E(j)$ if and only if the edge e_m is incident to node j , the power balance equations at bus $j \in \mathcal{V}$ are given by

$$\begin{aligned} 0 &= P_0^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| P_1^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 P_2^{(l_j)} + \beta^{(j)} G^{(s_j)} \left| V^{(l_j)}(t) \right|^2 \\ &\quad - \beta^{(j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \cos(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right. \\ &\quad \left. + B^{(s_j)} \sin(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right) + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 \sum_{e_m \in E(j)} G^{(e_m)} \\ &\quad - \sum_{e_m \in E(j)} \left| \vec{\mathbf{V}}^{(e_m)}(t) \right| \left(G^{(e_m)} \cos(\theta^{(e_m)}(t)) + B^{(e_m)} \sin(\theta^{(e_m)}(t)) \right), \\ 0 &= Q_0^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| Q_1^{(l_j)} + \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 Q_2^{(l_j)} - \beta^{(j)} B^{(l_j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 \\ &\quad - \beta^{(j)} \left| \vec{\mathbf{V}}^{(l_j)}(t) \right| \left| \vec{\mathbf{E}}^{(s_j)}(t) \right| \left(G^{(s_j)} \sin(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right. \\ &\quad \left. - B^{(s_j)} \cos(\theta^{(l_j)}(t) - \theta^{(s_j)}(t)) \right) - \left| \vec{\mathbf{V}}^{(l_j)}(t) \right|^2 \sum_{e_m \in E(j)} B^{(e_m)} \\ &\quad - \sum_{e_m \in E(j)} \left| \vec{\mathbf{V}}^{(e_m)}(t) \right| \left(G^{(e_m)} \sin(\theta^{(e_m)}(t)) - B^{(e_m)} \cos(\theta^{(e_m)}(t)) \right). \end{aligned}$$

(continued)

$$\text{where } \hat{G}^{(s_j)} = \frac{\hat{R}_0^{(s_j)}}{\left(\hat{R}_0^{(s_j)}\right)^2 + \left(\frac{1}{C^{(s_j)}}\right)^2}, \quad \hat{B}^{(s_j)} = \frac{C^{(s_j)}}{\left(C^{(s_j)}\hat{R}_0^{(s_j)}\right)^2 + 1}, \quad G^{(s_j)} = \frac{R^{(s_j)}}{\left(R^{(s_j)}\right)^2 + \left(L^{(s_j)}\right)^2}, \quad B^{(s_j)} = \frac{-L^{(s_j)}}{\left(R^{(s_j)}\right)^2 + \left(L^{(s_j)}\right)^2}, \quad G^{(e_m)} = \frac{R^{(e_m)}}{\left(R^{(e_m)}\right)^2 + \left(L^{(e_m)}\right)^2}, \text{ and } B^{(e_m)} = \frac{-L^{(e_m)}}{\left(R^{(e_m)}\right)^2 + \left(L^{(e_m)}\right)^2}.$$

7 Comparison of μHOM and μROM

In this section, the time resolutions, or timescales, of the reduced-order models are discussed and for the given test cases, the responses of μHOM , μROM1 , μROM2 , and μROM3 are compared and validated.

7.1 Reduced-Model Time Resolution

For the formulation of μROM_i , $i = 1, 2, 3$, the value for ϵ_i was chosen such that $\frac{1}{10\epsilon_i}$ represents the largest eigenvalues of the system, associated with the fast states $\mathbf{z}_i(t)$. Consequently, the fast-varying terms in the system response reach steady state in approximately $50\epsilon_i$ seconds, and the time resolution of μROM_i is $50\epsilon_i$ seconds. Table 1 shows the time resolution for the reduced-order models.

7.2 Model Validation

To validate μHOM , μROM1 , μROM2 , and μROM3 , the following test case is employed: two grid-forming inverter-based battery sources connected to a three-bus microgrid electrical network with an RLC load. One inverter-interfaced source is connected to bus 1 and the other to bus 2, and the load is connected to bus 3. Using μHOM , μROM1 , μROM2 , and μROM3 , we developed the test case in MATLAB/Simulink. The model parameters are shown in Table 2.

Table 1 Reduced-model time resolution

	Small parameter	Timescale
μROM1	$\epsilon_1 = 1 \times 10^{-5}$	$500 \mu\text{s}$
μROM2	$\epsilon_2 = 1 \times 10^{-3}$	50 ms
μROM3	$\epsilon_3 = 0.1$	5 s

Table 2 System parameters

	<i>Parameter</i>	s_1	s_2	$e_1 = \{1, 3\}$	$e_2 = \{2, 3\}$	l_3
Battery	$V_{DC}^{(s_j)}$	900V	900V	n/a	n/a	n/a
Three-phase inverter	$S^{(s_j)}$	10kVA	12kVA	n/a	n/a	n/a
	$V_{DQ}^{(s_j)}$	321.0265V	321.0265V	n/a	n/a	n/a
<i>LCL</i> filter	$r_0^{(s_j)}$	0.1Ω	0.15Ω	n/a	n/a	n/a
	$l_0^{(s_j)}$	1.35mH	1.5mH	n/a	n/a	n/a
	$r^{(s_j)}$	0.03Ω	0.04Ω	n/a	n/a	n/a
	$l^{(s_j)}$	0.35mH	0.33mH	n/a	n/a	n/a
	$\tilde{r}_0^{(s_j)}$	15mΩ	16mΩ	n/a	n/a	n/a
	$c^{(s_j)}$	50μF	60μF	n/a	n/a	n/a
Inner current control	$\kappa_{P\gamma}^{(s_j)}$	10.4479	10.4479	n/a	n/a	n/a
	$\kappa_{I\gamma}^{(s_j)}$	6.374×10^5	6.374×10^5	n/a	n/a	n/a
Outer voltage control	$\kappa_{P\phi}^{(s_j)}$	6.1825	6.1825	n/a	n/a	n/a
	$\kappa_{I\phi}^{(s_j)}$	1.364×10^4	1.364×10^4	n/a	n/a	n/a
Droop control	$D_\omega^{(s_j)}$	13.2629	13.2629	n/a	n/a	n/a
	$D_E^{(s_j)}$	2.3368	2.3368	n/a	n/a	n/a
Network	$r^{(e_m)}$	n/a	n/a	0.35Ω	0.4Ω	n/a
	$l^{(e_m)}$	n/a	n/a	1.5mH	2mH	n/a
Load	$r^{(l_j)}$	n/a	n/a	n/a	n/a	0.2kΩ
	$l^{(l_j)}$	n/a	n/a	n/a	n/a	11mH
	$c^{(l_j)}$	n/a	n/a	n/a	n/a	64μF

Let $I_{q[\omega_0 t]}^{(l_3)}(t) - j I_{d[\omega_0 t]}^{(l_3)}(t)$ denote the current across the load inductance. The load model used in μHOM and μROM1 is given by

$$\frac{C^{(l_3)}}{\omega_0} \frac{dV_{q[\omega_0 t]}^{(l_3)}(t)}{dt} = -\frac{1}{R^{(l_3)}} V_{q[\omega_0 t]}^{(l_3)}(t) - C^{(l_3)} V_{d[\omega_0 t]}^{(l_3)}(t) - I_{lq[\omega_0 t]}^{(l_3)}(t) + I_{dq[\omega_0 t]}^{(l_3)}(t), \quad (16)$$

$$\frac{C^{(l_3)}}{\omega_0} \frac{dV_{d[\omega_0 t]}^{(l_3)}(t)}{dt} = -\frac{1}{R^{(l_3)}} V_{d[\omega_0 t]}^{(l_3)}(t) + C^{(l_3)} V_{q[\omega_0 t]}^{(l_3)}(t) - I_{ld[\omega_0 t]}^{(l_3)}(t) + I_{qd[\omega_0 t]}^{(l_3)}(t), \quad (17)$$

$$\frac{L^{(l_3)}}{\omega_0} \frac{dI_{lq[\omega_0 t]}^{(l_3)}(t)}{dt} = -L^{(l_3)} I_{ld[\omega_0 t]}^{(l_3)}(t) + V_{q[\omega_0 t]}^{(l_3)}(t), \quad (18)$$

$$\frac{L^{(l_3)}}{\omega_0} \frac{dI_{ld[\omega_0 t]}^{(l_3)}(t)}{dt} = L^{(l_3)} I_{lq[\omega_0 t]}^{(l_3)}(t) + V_{d[\omega_0 t]}^{(l_3)}(t), \quad (19)$$

The load model used in $\mu\text{ROm}2$ and $\mu\text{ROm}3$ is given by

$$\begin{aligned} V_{q[\omega_0 t]}^{(l_3)}(t) &= \frac{R^{(l_3)}}{1 + (R^{(l_3)} C^{(l_3)})^2} \left(I_{q[\omega_0 t]}^{(l_3)}(t) + \frac{V_{d[\omega_0 t]}^{(l_3)}(t)}{L^{(l_3)}} \right) \\ &\quad - \frac{(R^{(l_3)})^2 C^{(l_3)}}{1 + (R^{(l_3)} C^{(l_3)})^2} \left(I_{d[\omega_0 t]}^{(l_3)}(t) - \frac{V_{q[\omega_0 t]}^{(l_3)}(t)}{L^{(l_3)}} \right), \end{aligned} \quad (20)$$

$$\begin{aligned} V_{d[\omega_0 t]}^{(l_3)}(t) &= \frac{(R^{(l_3)})^2 C^{(l_3)}}{1 + (R^{(l_3)} C^{(l_3)})^2} \left(I_{q[\omega_0 t]}^{(l_3)}(t) + \frac{V_{d[\omega_0 t]}^{(l_3)}(t)}{L^{(l_3)}} \right) \\ &\quad + \frac{R^{(l_3)}}{1 + (R^{(l_3)} C^{(l_3)})^2} \left(I_{d[\omega_0 t]}^{(l_3)}(t) - \frac{V_{q[\omega_0 t]}^{(l_3)}(t)}{L^{(l_3)}} \right). \end{aligned} \quad (21)$$

7.3 Results

A test case is considered where all four models have the same initial conditions, but at $t = 20$ s, the load resistance changes to $0.1 \text{ k}\Omega$, the load inductance changes to 10 ms , and the capacitance changes to $70 \mu\text{F}$. The comparison between the models is captured in Figures 1, 2, 3, and 4.

The model responses are depicted with timescale resolutions of 5 s , 50 ms , and $500 \mu\text{s}$ microseconds. We observe that at these time resolutions, the accuracies of $\mu\text{ROm}3$, $\mu\text{ROm}2$, and $\mu\text{ROm}1$, respectively, are visible. This is consistent with the result in Table 1.

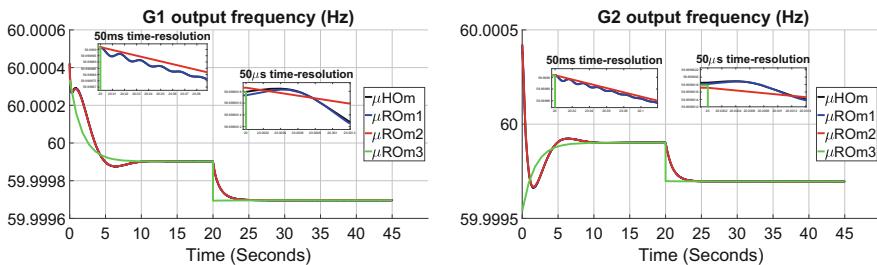


Fig. 1 Generator output frequency (Hz)

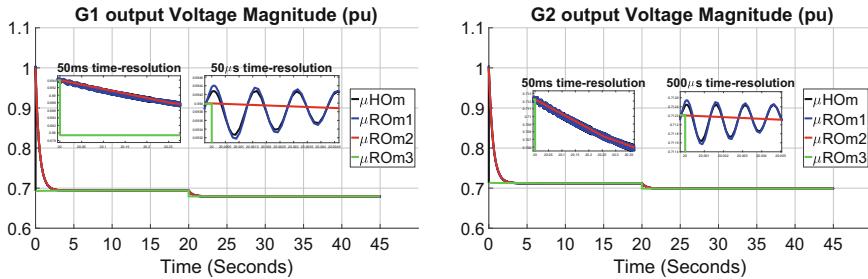


Fig. 2 Generator output voltage magnitude (pu)

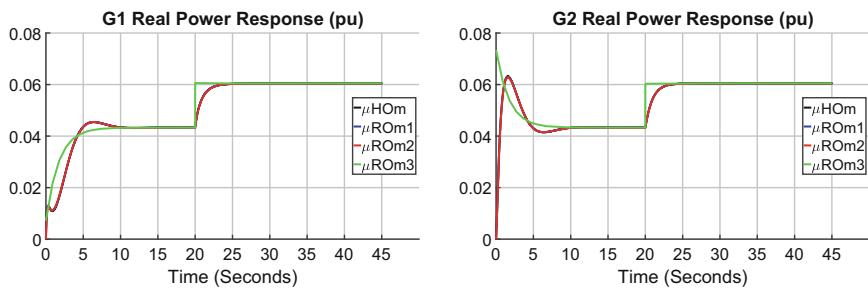


Fig. 3 Generator output real power (pu)

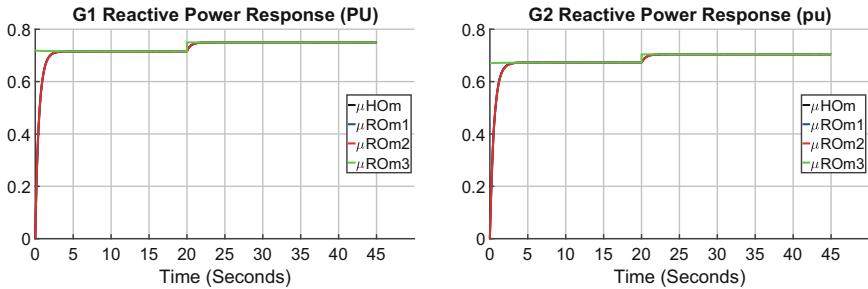


Fig. 4 Generator output reactive power (pu)

8 Conclusion

In this work, we developed a microgrid high-order model (μ HOM) by using circuit-theoretic and control laws. We used singular perturbation techniques for model order reduction of the μ HOM, which allowed us to obtain three reduced-order models, μ ROM1, μ ROM2, and μ ROM3, with the time resolution of each reduced model identified.

For a given test case, we compared the responses of all three models to that of the μ HOM. Using the time resolutions to analyze each model response, we observed that the responses of μ ROM1, μ ROM2, and μ ROM3 track the response of the μ HOM with errors $\mathbf{O}(\epsilon_1)$, $\mathbf{O}(\epsilon_2)$, and $\mathbf{O}(\epsilon_3)$, respectively, where $\epsilon_1 = 10^{-5}$, $\epsilon_2 = 10^{-3}$, and $\epsilon_3 = 10^{-1}$.

Acknowledgment The information, data, or work presented herein was supported by the Advanced Research Projects Agency-Energy (ARPA-E), US Department of Energy, within the NODES program, under Award DE-AR0000695, and by the MidAmerica Regional Microgrid Education and Training (MARMET) Consortium, US Department of Energy.

References

1. Ajala OO, Domínguez-García AD, Sauer PW (2017) A hierarchy of models for inverter-based microgrids. Coordinated science laboratory technical report UILU-ENG-17-2201, University of Illinois at Urbana-Champaign, May 2017. <http://hdl.handle.net/2142/96001>
2. Anand S, Fernandes BG (2013) Reduced-order model and stability analysis of low-voltage dc microgrid. *IEEE Trans Ind Electron* 60(11):5040–5049
3. Bergen AR, Vittal V (2000) Power systems analysis. Prentice Hall, Upper Saddle River
4. Chow JH (1982) Time-scale modeling of dynamic networks with applications to power systems. Springer, Berlin
5. Dörfler F, Bullo F (2010) Synchronization and transient stability in power networks and non-uniform kuramoto oscillators. In: Proceedings of the 2010 American control conference, June 2010, pp 930–937
6. Kodra K, Zhong N, Gajić Z (2016) Model order reduction of an islanded microgrid using singular perturbations. In: Proceedings of American control conference, Chicago, IL, pp 3650–3655
7. Kokotović P, Khalil HK, O'Reilly J (1986) Singular perturbation methods in control: analysis and design. Classics in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia
8. Krause PC, Wasenczuk O, Sudhoff SD, Pekarek S (2013) Institute of Electrical, and Electronics Engineers. Analysis of electric machinery and drive systems. IEEE press series on power engineering. Wiley, Hoboken
9. Luo L, Dhople SV (2014) Spatiotemporal model reduction of inverter-based islanded microgrids. *IEEE Trans Energy Convers* 29(4):823–832. ISSN 0885–8969
10. Pogaku N, Prodanovic M (2007) Green TC Modeling, analysis and testing of autonomous operation of an inverter-based microgrid. *IEEE Trans Power Electron* 22(2):613–625
11. Rasheduzzaman M, Mueller JA, Kimball JW (2015) Reduced-order small-signal model of microgrid systems. *IEEE Trans Sustainable Energy* 6(4):1292–1305
12. Sauer PW, Pai A (2006) Power system dynamics and stability. Stipes, Champaign
13. Schiffer J, Zonetti D, Ortega R, Stankovic AM, Sezi T, Raisch J (2015) Modeling of microgrids—from fundamental physics to phasors and voltage sources. CoRR, abs/1505.00136, May 2015
14. Yazdani A, Iravani R (2010) Voltage-sourced converters in power systems. Wiley, Hoboken

Asynchronous Coordination of Distributed Energy Resources with Packetized Energy Management



Mads Almassalkhi, Luis Duffaut Espinosa, Paul D. H. Hines, Jeff Frolik,
Sumit Paudyal, and Mahraz Amini

Abstract To enable greater penetration of renewable energy, there is a need to move away from the traditional form of ensuring electric grid reliability through fast-ramping generators and instead consider an active role for flexible and controllable distributed energy resources (DERs), e.g., plug-in electric vehicles (PEVs), thermostatically controlled loads (TCLs), and energy storage systems (ESSs) at the consumer level. However, in order to facilitate consumer acceptance of this type of load coordination, DERs need to be managed in a way that avoids degrading the consumers' quality of service (QoS), autonomy, and privacy. This work leverages a probabilistic packetized approach to energy delivery that draws inspiration from random access, digital communications. Packetized energy management (PEM) is an asynchronous, bottom-up coordination scheme for DERs that both abides by the constraints of the transmission and distribution grids and does not require explicit knowledge of specific DER's local states or schedules. We present a novel macro-model that approximates the aggregate behavior of packetized DERs and is suitable for estimation and control of available flexible DERs to closely track a time-varying regulation signal. PEM is then implemented in a transmission/distribution system setting, validated with realistic numerical simulations, and compared against state-of-the-art load coordination schemes from industry.

M. Almassalkhi (✉) · L. Duffaut Espinosa · P. D. H. Hines · J. Frolik · M. Amini
University of Vermont, 210 Colchester Avenue, Farrell Hall 200A, Burlington, VT 05405, USA
e-mail: malmassa@uvm.edu; lduffaut@uvm.edu; phines@uvm.edu; jfrolik@uvm.edu;
mamini2@uvm.edu

S. Paudyal
Department of Electrical and Computer Engineering, Michigan Technological University, 1400
Townsend Drive, Houghton, MI 49931, USA

1 What Is Packetized Energy Management (PEM)?

PEM leverages the packet-based strategies from random access communication channels, which have previously been applied to the distributed management of wireless sensor networks [1]. In particular, PEM may be thought of as a multichannel, multi-receiver version of ALOHA or RTS/CTS (*request/clear to send*) [2, 3]. Under PEM, the delivery of energy to a flexible load (e.g., electric vehicle, water heater, battery, air conditioner, refrigerator, etc.) is accomplished by having the load stochastically request “energy packets,” just as digital communication networks break data into packets. An energy packet represents a fixed duration/fixed power block of demand consumed (or delivered) by the flexible load. For example, a 5 minute/ 5 kW energy packet consumes 5 kW for 5 minutes (i.e., 0.417 kWh of energy). PEM engenders the following technical advantages:

- **Local decision-making:** devices offer their own flexibility to the grid operator in bottom-up fashion based on unique local energy demands, which ensures customers’ quality of service (QoS).
- **Privacy:** individual energy usage information is not needed, which ameliorates privacy concerns.
- **Fairness:** all devices have equitable access to the grid resources
- **Responsiveness:** the aggregation of devices can adapt to rapid changes in supply and demand.
- **Scalability:** asynchronous control enables plug-and-play capability and scales to millions of devices.

Furthermore, with the proposed PEM coordination architecture, the grid operator (or load aggregator) only requires two scalar measurements from the collection of loads: aggregate power consumption and the loads’ requests for packets. This represents a significant advantage over many other load coordination methods that can require up to an entire histogram of states from the population of loads. This distinction is further elaborated upon in Section 2.

This chapter will (1) present the PEM scheme within the context of existing approaches from literature, (2) develop and validate an aggregate and homogeneous macro-model, and (3) illustrate coordination of heterogeneous DERs at the consumer level under PEM. PEM is suitable for a large class of deferrable loads and is illustrated with residential electric water heaters, vehicles, and batteries, with an emphasis of content on electric water heaters. Since this chapter focuses on the closed-loop control performance of PEM, the underlying communication network is assumed ideal (i.e., no delays or lost requests or responses). This assumption does not detract from the results presented herein since realistic communication delays and losses are related to an individual device, which is coupled to the system in an asynchronous and randomized manner. Specifically, communication delays are on the order of seconds while packet durations are on the order of minutes. Of course, widespread disruption to the communication infrastructure will affect

PEM, which is why implementation of cybersecurity and validation against realistic communication parameters are critical topics to consider but are outside the scope of this chapter.

2 The Need for PEM Coordination of DERs

Fast-ramping generators have provided the electric grid reliable operating reserves for decades. However, power systems with high penetrations of renewable energy challenge this operating paradigm. At high levels of renewable penetration, current approaches to manage the variability in wind or solar generation would require having more fast-ramping conventional generators online. However, that leads to more generators idling, burning fuel, and increasing harmful air emissions. Therefore, there is a need to move away from the traditional form of ensuring reliability to consider an active role for flexible and controllable net-load distributed energy resources (DERs), e.g., plug-in electric vehicles (PEVs), thermostatically controlled loads (TCLs), energy storage systems (ESSs), and distributed generation at the consumer level [4]. While the core concepts underlying modern demand-side management (DSM) have existed for decades [5, 6], the technology for coordinating the activities of DERs is nascent but maturing rapidly. Indeed, there is a growing consensus that balancing supply and demand in power systems with large amounts of variable renewable energy will require an active role for flexible DERs in addition to balancing services from conventional power plants [4].

With the proposed PEM architecture, the grid operator or aggregator only requires two scalar measurements from the collection of loads: the aggregate power consumption and the aggregate request rate. This limited data requirement represents a significant advantage over other aggregate model-estimator-controller state-space approaches, e.g., [7], which require an entire histogram of states from the collection of loads to update a state bin transition model. To generate these states for control, an observer is designed to estimate the histogram based on aggregated power consumption; however, in some cases, the model may not be observable [8]. Note that in addition to aggregated power consumption, which only informs the observer about devices that are ON, PEM also receives packet requests from the loads that are OFF, which supplies information about the OFF population and offers a valuable mechanism for observability and state feedback.

Recently, the work in [7] has been extended to include higher-order dynamic models and end user and compressor delay constraints [9] and stochastic dynamical performance bounds [10]. Specifically, the modeling of packet duration in this chapter was inspired by the compressor lockout method utilized in [9]. A mean-field approach to direct load control is developed for heterogeneous TCL populations in [11]. Similar to the PEM paradigm, the mean-field approach developed in [8, 12] maintains the quality of service (QoS) through opt-out mechanisms and also employs local randomization, which reduces the effect of synchronization in the population of loads. That is, prior work uses either direct load control [4] or employs

local randomization at the device level for the ON/OFF transitions in a population of flexible loads [7, 12]. In the latter case, stochastic device behavior is regulated by either broadcasting (i.e., sending in top-down fashion to the entire population of loads) an updated probability density function [7] or broadcasting an updated scalar variable, which perturbs probability density functions defined at each device [12]. In contrast to those prior works, PEM does not perturb the probability density function at the device level nor broadcasts the control signal to the entire population. Instead, PEM listens to each load’s individual and stochastic request (i.e., in a locally driven, bottom-up fashion). The coordinator then responds in real time to each packet request based on grid and/or market conditions.

The most closely related work on energy packets is found in references [13, 14], where an omniscient centralized *packetized direct load controller* (PDLC) is developed for TCLs. The average controller performance and consumer QoS are analytically investigated, and queuing theory is employed by the authors to quantify the centralized controller’s performance. In [15], a distributed (binary information) version of PDLC is proposed that requires only (binary) packet request information from the loads. Unlike the proposed PEM scheme, the distributed PDLC assumes that the exact number of participating packetized loads at any given time is known, the allocation of packet requests from the queue is synchronized, and the queue stores packet requests if the packets cannot be allocated, which creates delays in service. Instead, this chapter extends the authors’ previous packetized energy results for managing PEVs [16, 17] to consider TCLs and bidirectional residential batteries. In addition, it is shown how PEM uses local randomization and packetization to overcome the challenges surrounding synchronization of devices during extended peak reduction events. Furthermore, PEM does not require the storing of packet requests. More precisely, this chapter focuses on results on PEM coordination of TCLs (specifically, electric water heaters) from [18] and [19] and presents a macro-model of TCLs under PEM. PEM coordination is then extended with a case study that considers *diverse* heterogeneous loads (TCLs, PEVs, and ESSs). The chapter concludes with a discussion of future directions for PEM in transmission and distribution system operations.

3 PEM Fundamentals

Figure 1 illustrates an example of the cyber-physical interactions needed to realize PEM in distribution system operations, such as managing power constraints, peak demand, and variability from and balancing of renewable generation. We will separately describe the functions of the grid operator (e.g., a utility or ISO), the coordinator (e.g., DER management system or a “virtual power plant” or VPP), and the packetized energy controller (PEC). The PEC connects a single flexible load to a VPP and can directly interact with the load to engender the “packetized” response. Owing to the proposed bottom-up approach, the concept of a packetized load is introduced next.

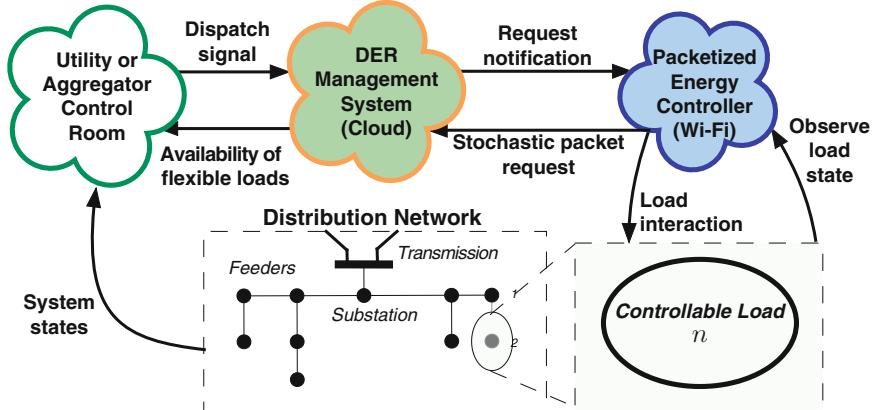


Fig. 1 Cyber-physical infrastructure to realize PEM.

3.1 Packetized Loads

As noted, PEM has previously been proposed for coordinated charging of electric vehicles subject to transformer constraint in a distribution feeder [16, 20, 21]. In this earlier work, PEVs asynchronously request the authority to charge with a specific probability according to their logic state in a probabilistic automaton. For example, consider a three-state finite-state machine (e.g., see Figure 2(right)). The probability that a packetized load in logic state i requests access to the grid during period Δt is P_i where $P_1 > P_2 > P_3$. If there is capacity available in the grid, a PEV's request for a (charging) packet is accepted, and the PEV is granted authority to charge, but only for a predetermined fixed duration of time (e.g., 15mins), referred to as the *control epoch*. Upon having the packet request accepted, a logic state transition takes place, $P_i \rightarrow P_{i-1}$, which reduces the mean time to request (MTTR). In contrast, if the PEV is denied authority to charge due to insufficient capacity (or overload), the MTTR increases with transition $P_i \rightarrow P_{i+1}$. In this chapter, the PEM concept is adapted for the purpose of managing a set of diverse DER types, TCLs, PEVs, and ESSs, by specifically coupling a device's local dynamic state (e.g., temperature or state of charge) to the device's MTTR (i.e., the stochastic request rate). With the inclusion of packetized ESS devices, bidirectional power exchanges are considered by distinguishing between charge (consume) and discharge (inject) requests. The diverse DER types are combined under a single VPP, and the closed-loop performance is presented in Section 5.

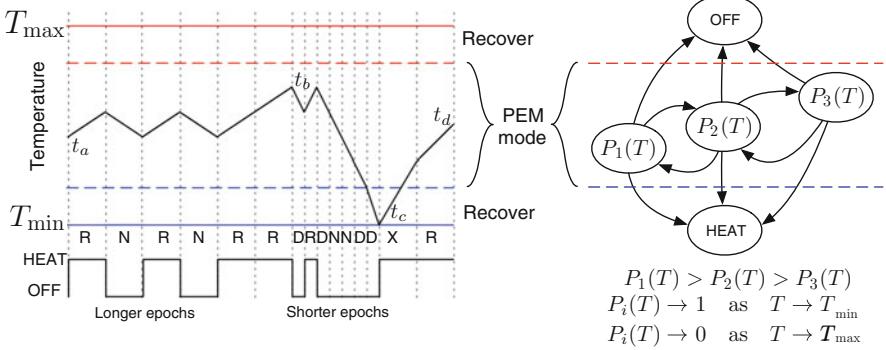


Fig. 2 Water heater managed by PEM. The left figure shows a sequence of events. At time t_a , when grid resources are unconstrained, loads stochastically request (R) or do not request (N) energy. At t_b , the system approaches a period of constrained supply, in which the system aggregator mostly denies requests (D) and reduces the epoch length. As a result, the automaton transitions to a lower probability state (e.g., $P_1 \rightarrow P_2$). At t_c , the temperature hits the QoS bound, and the load exits (X) from PEM and rapidly seeks to recover temperature to within the QoS bounds, which occurs at t_d . The right figure shows the state machine that changes its request probabilities ($P_i(T)$) and its epoch lengths, based on responses the local temperature state. Also embedded in the automaton is the epoch lengths between state transitions and making requests.

3.1.1 Dynamic Modeling of DERs

We will summarize the three DERs models in this section. After developing the necessary models, the stochastic request rate of packetized loads is defined as a function of the dynamic state.

PEV and ESS Models

The dynamic models for PEV and ESS are nearly identical, except that a PEV is inherently mobile and, during time periods when they are away from home (or a charging station), the PEV battery's state of charge decreases at a rate corresponding to the driving pattern. In addition, unlike an ESS, it is assumed that a PEV cannot inject power to the grid and, therefore, represents a unidirectional energy storage device. A general discrete-time model with sampling time Δt of PEV or ESS battery n 's state of charge (SOC [kWh], x_n) is summarized by the following:

$$x_n[k+1] = x_n[k] + \Delta t (-\eta_{sl,n} x_n[k] + \eta_{ch,n} u_{dis,n}[k] - \eta_{dis,n} u_{dis,n}[k]), \quad (1)$$

where $\eta_{sl,n}, \eta_{ch/dis,n}$ represents parameters associated with standing losses and charging/discharging efficiency, respectively. If standing losses are not considered, $\eta_{sl,n} \equiv 0$. The control inputs are charge and discharge rates [kW], which are each bounded. For a PEV, $u_{dis,n}$ is uncontrollable and reflects the vehicle's away-from-home driving pattern. The SOC is also bounded by battery capacity bounds:

$x_n \in \{\underline{x}_n, \bar{x}_n\}$. The ESS is assumed to be subject to an uncontrollable background net-demand process (charging and/or discharging), but that simultaneous charging and discharging are feasible.

TCL Model for Electric Water Heater

Generic models of TCLs can be found in [4, 7]; however, this chapter presents a simple model of an electric water heater that uses a stochastic residential hot-water withdrawal rate to describe the temperature dynamics. Therefore, this section focuses on the electric water heater (EWH) model, which is modeled as a first-order single-heating-element thermodynamic model motivated by [22, 23] but modified to consider a uniform thermal mass and hot-water withdrawal by the consumer in liters/min rather than as a fixed energy loss. Detailed discussions on parameters and hot-water withdrawal rates and events can be found in [18] and [19] but are summarized below:

The temperature at time-step $k + 1$ is given by:

$$T_n[k + 1] = T_n[k] + \Delta t \left(\frac{\eta_n P_n^{\text{rate}}}{c\rho L} z_n[k] - \frac{(T_n[k] - T_{\text{amb}}[k])}{\tau_n} - \frac{(T_n[k] - T_{\text{inlet}}[k])}{60L} w_n[k] \right) \quad (2)$$

where $c = 4.186$ [kJ/kg·°C] and $\rho = 0.990$ [kg/liters] represent specific heat capacity and density of water close to 50°C¹. L [liters] represents the total capacity of the EWH. Note that P_n^{rate} , η_n , and z_n are the heating element power transfer rate [kW], the heat transfer efficiency, and the binary ON/OFF logic state ($\Rightarrow z_n \equiv 1/0$) of EWH n , respectively². The terms T_{amb} , τ_n , T_{inlet} , and w_n are the ambient temperature [°C], time constant due to ambient insulation losses [s], inlet temperature [°C], and hot-water withdrawal rate by consumer n in [liters/min], respectively. The hot-water withdrawal rate, w_n , represents the uncontrollable background demand for hot water and is modeled as a Poisson rectangular pulse (PRP) random process as discussed briefly in Section 4.1 and [18, 19]. To ensure numerical stability, all simulations use $\Delta t \leq 60$ s and are presented in Section 5.

3.1.2 Conventional Control of DERs

The vast majority of existing traditional DERs operate in a binary (ON/OFF) manner and are already controlled by simple state machines. For example, a PEV (when charging) will charge continuously at maximum rating until SOC reaches

¹Physically, c and ρ vary with water temperature, but this relationship is ignored herein as it does not affect the results or conclusion of PEM's local decision-making.

²The binary z_n implies that (2) is a hybrid dynamic model.

upper limit and then switch to OFF, while a TCL will change logic state based on temperature deadbands. Specifically, a TCL is controlled to maintain a desired temperature set point, T_{set}^n , within a temperature deadband, $T_{\text{set}}^n \pm T_{\text{DB}}^n/2$. The local discrete-time control logic can then be described by the following for TCLs:

$$z_n[k] = \begin{cases} 1, & T_n[k] \leq T_{\text{set}}^n - T_{\text{DB}}^n/2 \\ 0, & T_n[k] \geq T_{\text{set}}^n + T_{\text{DB}}^n/2 \\ z_n[k-1], & \text{otherwise} \end{cases}, \quad (3)$$

and for PEVs:

$$z_n[k] = \begin{cases} 1, & x_n[k] < \bar{x}_n \\ 0, & x_n[k] \geq \bar{x}_n \\ z_n[k-1], & \text{otherwise.} \end{cases} \quad (4)$$

The battery control logic for an ESS device can be described by similar local logic depending on the operating mode (e.g., peak reduction or arbitrage) but is omitted herein. Thus, the proposed PEM scheme requires only the replacement of the existing state machine with a more sophisticated one (i.e., the equivalent of a firmware upgrade) that interacts with a coordinator/aggregator.

3.1.3 Adaptation of PEM for DERs

As discussed previously, the key to enable PEM is the local decision-making of the packetized energy controller (PEC), which observes the physical load's local dynamic state. This state is the temperature for a TCL and the state of charge (SOC) for PEV and ESS devices. By coupling a device's dynamic state to a stochastic request rate for accessing the grid, PEM effectively perturbs the ON/OFF transition rate of the device, which, in the aggregate, begets flexibility for the VPP operator. The description below described the PEM adaptation for an electric water heater (i.e., a TCL) but is straightforward to extend to the other DER types.

Figure 2(right) illustrates a TCL automaton under PEM. When the local temperature of the TCL, T , is between its upper and lower temperature limits for PEM operation, the TCL's mean time to request (MTTR) is driven by an exponential distribution whose mean is inversely proportional to T relative to the upper limit. That is, TCLs with temperatures very close to the lower threshold will make requests with near certainty (i.e., $P_i(T \rightarrow T_{\min}) \approx 1$), and those near the upper limit in temperature will make requests with low probability (i.e., $P_i(T \rightarrow T_{\max}) \approx 0$). Upon transmitting a request and if there is capacity in the grid, the TCL will be given the authority to turn ON for a fixed control epoch length δ (i.e., $z_n(t) = 1$ for $t \in (t_0, t_0 + \delta)$), and a state transition occurs: $P_i(T) \rightarrow P_{i-1}(T)$. If the request is denied, the TCL finite state machine transitions to a state with lower MTTR, $P_i(T) \rightarrow P_{i+1}(T)$, but will immediately resume requesting with temperature-dependent probability. If access is denied repeatedly, T reaches T_{\min} , which causes

the TCL to exit the PEM scheme (*exit-ON*) to satisfy quality-of-service (QoS) constraints. An illustrative ON/OFF cycle of a packetized water heater is illustrated in Figure 2(left).

In addition to the TCL receiving an “Yes/No” response to a request, the TCL may also receive an updated (global) control epoch length, δ , thus enabling tighter tracking in the aggregate, which is helpful during ramping events. Clearly, while the TCL is ON, it does not make requests. Furthermore, we require $\delta \geq \Delta t$.

Remark 1 Since all packetized loads operate in this manner, the DER aggregator granting (“yes”) or denying (“no”) the authority to turn ON does not require any knowledge of a particular load. Furthermore, the aggregator does not even need to track which load is making a particular request. As each load type runs the same automaton logic and its ability to turn ON depends only on the (present and past) system capacity (and, potentially, past VPP decisions), any load making a request at the same point in time will be treated the same by the aggregator. As such, the PEM approach inherently maintains privacy while still offering equitable access to the grid.

3.1.4 The Stochastic Request Rate with PEM

This section explicitly defines the stochastic request rate for a packetized load as a function of the device’s local dynamic state. Consider a TCL, PEV, or ESS packetized load with just one automaton state. Then, in the discrete-time implementation of PEM, the probability that the packetized load n with local dynamic state $x_n[k] \in [\underline{x}_n, \bar{x}_n]$ and desired set-point $x_n^{\text{set}} \in (\underline{x}_n, \bar{x}_n)$ requests access to the grid during time-step k (over interval Δt) is defined by the cumulative exponential distribution function:

$$P(x_n[k]) := 1 - e^{-\mu(x_n[k])\Delta t}$$

where rate parameter $\mu(x_n[k]) > 0$ is dependent on the local dynamic state. For a consume (or charge) request, this dependence is established by considering the following boundary conditions:

- $P(\text{load } n \text{ requests to consume packet during time } k | x_n[k] \leq \underline{x}_n) = 1$
- $P(\text{load } n \text{ requests to consume packet during time } k | x_n[k] \geq \bar{x}_n) = 0$,

which permits a simple functional form for the rate parameter that ensures the boundary conditions are met:

$$\mu(x_n[k]) = \begin{cases} 0, & \text{if } x_n[k] \geq \bar{x}_n \\ m_R \left(\frac{\bar{x}_n - x_n[k]}{x_n[k] - \underline{x}_n} \right) \cdot \left(\frac{x_n^{\text{set}} - \underline{x}_n}{\bar{x}_n - x_n^{\text{set}}} \right), & \text{if } x_n[k] \in (\underline{x}_n, \bar{x}_n) \\ \infty, & \text{if } x_n[k] \leq \underline{x}_n \end{cases} \quad (5)$$

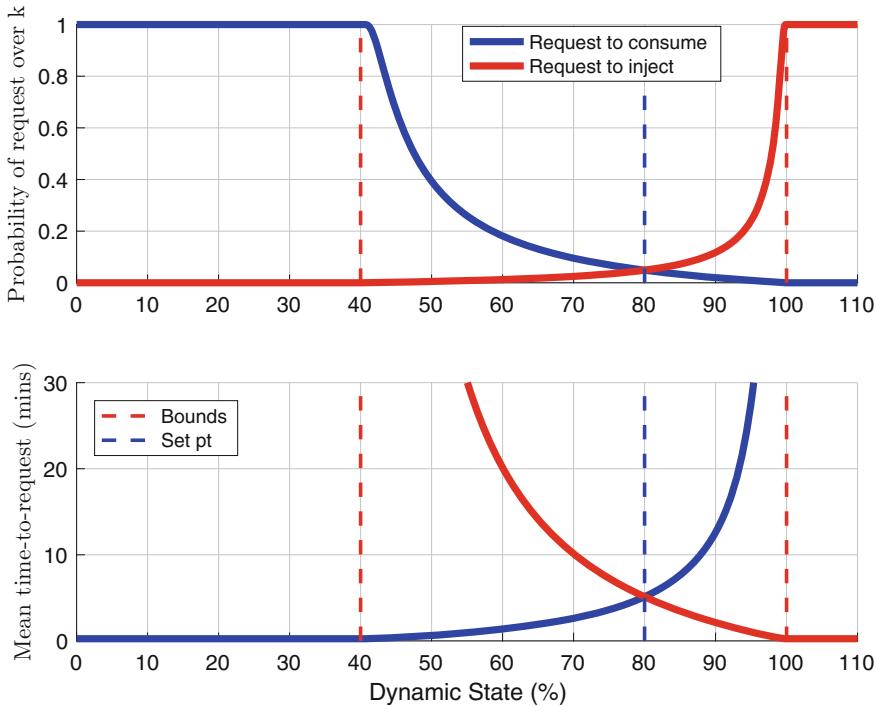


Fig. 3 Illustrating the effect of the local state x_n (e.g., temperature or state of charge) on the access request probabilities (top) and MTTR (bottom) of a controllable load under PEM. Note that if the state exceeds bounds (x_n, \bar{x}_n), the probability of request is 0 or 1 depending on the type of request (consume or inject). While TCLs can only consume power from the grid, controllable ESS batteries can discharge and inject power into the grid. The ability of a device to request either form of participation is captured with the consume and inject packets. We assume that if both packets are requested simultaneously, the requests cancel each other out and no request is made.

where $m_R > 0$ [Hz] is a design parameter that defines the mean time to request (MTTR). For example, if one desires a MTTR of 5 minutes when $x_n[k] \equiv x_n^{\text{set}}$, then $m_R = \frac{1}{600}$ Hz.

Figure 3 illustrates the bidirectional stochastic request rates and MTTR for a generic packetized load that can request to consume power from and inject power into the grid. Note that (5) is represented by the blue lines (left to right).

3.1.5 Quality of Service Under PEM

With the stochastic nature of DERs under PEM, it is entirely possible that a local disturbance (e.g., a large hot-water withdrawal rate for TCLs) can drive $x_n[k]$ below \underline{x}_n . Therefore, to maximize QoS to the consumer (e.g., avoid cold

showers), a DER under PEM can temporarily exit (i.e., opt out of) PEM and operate under traditional control (e.g., turn ON and stay ON). This is illustrated for TCLs in Figure 2(left) at event t_c and with automaton states HEAT and OFF in Figure 2(right). That is, once a TCL under PEM exceeds temperature bounds, the traditional control logic is employed temporarily to bring the local temperature within PEM “recovery bounds” $T_{\text{set}}^n \pm T_{\text{PEM}}^n / 2$ with $T_{\text{PEM}}^n < T_{\text{DB}}^n$ where PEM control logic is reinstated (i.e., TCL opts back into PEM). The recovery bounds are helpful to avoid excessive exit/reentry cycling at the boundary.

Remark 2 Clearly, if packetized loads exit PEM en masse, the available flexibility can be greatly reduced and, therefore, will impact the ability of a coordinator to track a given reference balancing signal. Hence, the macro-model effort in Section 4 is focused on developing a population model of homogeneous TCLs that will permit analysis to leverage incoming requests (modulating yes/no response rates) to reduce the need for opting out without sacrificing controllability.

3.2 *Closing the Loop on PEM with the Virtual Power Plant*

As shown in Figure 1, a packetized energy controller enables bidirectional Wi-Fi communication between a load and the virtual power plant (VPP). The VPP receives balancing dispatch signals akin to automatic generation control (AGC) from a grid operator and coordinates flexible energy resources to track the balancing command³. Within the proposed PEM scheme, the VPP tracks the balancing signal by responding to individual, asynchronous, and stochastic load access requests with “yes” or “no” notifications based on real-time output error between actual aggregate output, $y(t)$, and the VPP reference signal, $r(t)$: $e(t) := r(t) - y(t)$. This simple closed-loop controller is illustrated in Figure 4. The VPP is similar to a relay controller in the sense that it accepts a request (“yes”) if $e(t) > 0$, otherwise, “no.” However, unlike standard relay control of a single plant, the VPP responds to asynchronous, stochastic requests from N plants, which overcomes common drawbacks associated with relay control (e.g., switching leading to oscillations) and permits accurate tracking. While the above describes a simple control scheme for VPP, more advanced approaches can leverage past load requests rates, VPP responses, and aggregate net demand of the VPP to further improve upon performance. The VPP is described by the following inputs and outputs:

Input: Balancing reference signal, asynchronous requests;

Output: Yes/no access notification to individual load.

³While the VPP needs to estimate and predict the aggregate flexibility from available loads, these results focus on the tracking control problem as the estimation problem represents ongoing research

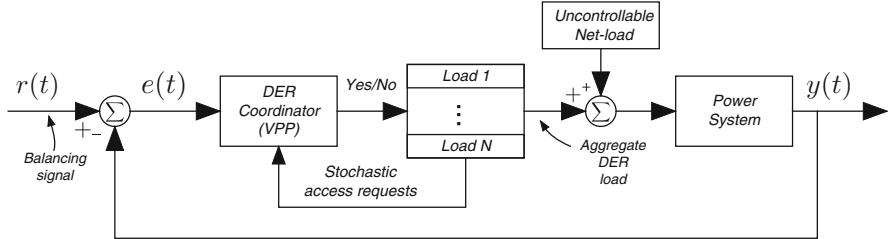


Fig. 4 The closed-loop feedback system for PEM with the reference $r(t)$ provided by the grid operator and the aggregate packetized loads' output response $y(t)$ measured by VPP.

3.3 Providing Grid-Level Service with PEM

The transmission (e.g., ISO New England) or distribution utility system operator (e.g., the DSO control room in Figure 1) is able to measure or estimate the states of the grid, such as voltage, frequency, and power flows. Under scenarios with high penetration of renewable energy, the grid operator will find it ever more difficult to balance demand and supply while satisfying network conditions and, therefore, seeks to leverage the flexible packetized DERs sitting in customer homes and industrial/commercial facilities. This is achieved by signaling individual balancing requests to VPPs across the network in near real time akin to automatic generation control (AGC; secondary frequency control) signals, which are transmitted every 4–5 seconds today. The signaling may be computed via solving an optimal power flow (OPF) problem that dispatches VPPs optimally with respect to network constraints and available net-load resources. Thus, the grid operator is summarized by the following inputs and outputs:

Input: Grid voltages, frequencies, net-load forecasts, and price forecasts;

Output: Dispatch balancing signal.

Remark 3 By managing the anonymous, fair, and asynchronous pings of packetized loads via a VPP that receives grid or market-based balancing signals from the grid operator, PEM represents a bottom-up distributed control scheme that has been adapted for TCLs, PEVs, and ESS devices in this section.

The next section develops and validates a macro-model that captures the aggregate behavior of a population of homogeneous EWHs. This is particularly valuable since the complexity of a large-scale VPP increases exponentially when augmenting the hybrid dynamics in (1) and (2) for thousands of flexible loads under PEM. Devising suitable control techniques, therefore, becomes intractable for the proposed VPP, and a simpler, scalable, lower-dimensional, aggregate model is needed.

4 Macro-model for Homogeneous Packetized EWHs

This section presents a state bin transition (macro-) model for a large homogeneous population of TCLs. The aggregate energy use of these TCLs is coordinated with PEM. First, a macro-model for a population of TCLs is developed and then augmented with a timer to capture the duration and consumption of energy packets and with exit-ON/exit-OFF dynamics to ensure consumer quality of service. This permits a virtual power plant (VPP) operator to interact with TCL population through the stochastic packet request mechanism. The VPP regulates the proportion of accepted packet requests to allow tight tracking of balancing signals. The developed macro-model compares well with (agent-based) micro-simulations of TCLs under PEM and can be represented by a controlled Markov chain. Details on the macro-model development can be found in [19].

4.1 Conventional Thermostatic Control

A macro-model for a large population of TCLs is developed in this section as an abstraction of the augmented (agent-based) dynamic micro-models. Specifically, consider the TCL population dynamics over a discretization of the temperature state space, and employ a state bin transition model, such that the macro-model approaches the behavior of the micro-model as the number of devices increases [24]. The transitions between these bins are determined by the dynamical system equations of the homogeneous TCLs as discussed below. The macro-model utilizes a finite set $\mathcal{X} = \{x_1, \dots, x_N\}$, where each element is called a state. Assume that there exists an appropriate probability space (Ω, P, \mathcal{F}) , where Ω is the set of events, \mathcal{F} a filtration, and P the probability measure of elements in \mathcal{F} . Then, random variables $\{X_k\}_{k \geq 0}$ are defined as $X_k : \Omega \rightarrow \mathcal{X}$. Let $x_j \in \mathcal{X}$, and denote $q_j[k] = P(X_k = x_j)$ as the probability of $X_k = x_j$, $k \geq 0$. The column vector $q[k] := (q_1, \dots, q_N)^T$ then gives the probability mass function of the random variable X_k . Also, if one denotes the transition probability of an homogeneous Markov chain as $p_{ij} = P(X_{k+1} = x_i | X_k = x_j)$, it then follows that

$$q[k+1] = Mq[k], \quad (6)$$

where $M = \{p_{ij}\}_{1 \leq i, j \leq N}$ [25]. Given an initial distribution $q[0]$, one can solve for (6) and find the distribution at time k as $q[k] = M^k q[0]$.

Conventional thermostatic control of an EWH is based on keeping the local state variable (e.g., temperature) within a deadband $[T_{\min}, T_{\max}]$ by switching the device ON (when $T \leq T_{\min}$) or OFF (when $T \geq T_{\max}$), where $[T_{\min}, T_{\max}] = [T_{\text{set}} - T_{\text{DB}}, T_{\text{set}} + T_{\text{DB}}]$ as in Section 3. The interval $[T_{\min}, T_{\max}]$ is divided into N consecutive bins each corresponding to a *bin state* in \mathcal{X} . Since (2) includes hybrid ON/OFF dynamics, the state space for the system consists of two discrete state

spaces: \mathcal{X}_{on} and \mathcal{X}_{off} . That is, the full state space is given by $\mathcal{X} = \mathcal{X}_{\text{on}} \cup \mathcal{X}_{\text{off}}$. At time k , the probability mass function of the system is $q^\top = (q_{\text{on}}^\top, q_{\text{off}}^\top)$ with $q_{\text{on}} = (q_{\text{on}}^1, \dots, q_{\text{on}}^N)^\top$ and $q_{\text{off}} = (q_{\text{off}}^1, \dots, q_{\text{off}}^N)^\top$. Note that q contains the percentage of the population in each state of \mathcal{X} . For instance, if R is the total number of EWHs and R_{on}^i is the number in state x_{on}^i , then $R_{\text{on}}^i = q_{\text{on}}^i R$. Similarly, the total ON population is given by

$$y = c^\top q \quad \text{for } c = (\mathbf{1}_N^\top, 0 \cdots 0)^\top \in \mathbb{R}^{2N}, \quad (7)$$

and $\mathbf{1}_N = (1, \dots, 1)^\top \in \mathbb{R}^N$. The transition rates are computed by considering how the temperature bin corresponding to a particular state is altered by the hybrid dynamics in (2).

Together with discrete sampling time and temperature bin widths, the hot-water withdrawal rate w_n in (2) is one of the main factors affecting these transition rates. For this purpose, a *Poisson rectangular pulse* stochastic differential model is employed to model the individual water withdrawal rates w , including the stochastic duration and intensities (e.g., l/min) of hot-water events [26]. Upon aggregating w across the entire population, it is shown in [19] that a constant steady-state water consumption rate can be derived, \bar{w}_{sst} . For example, consider two realizations a, b of the water profiles with identical parameters except for the water withdrawal intensities of the random variable w ($\lambda_a \neq \lambda_b$). An EWH in ON state with λ_a ($< \lambda_b$) at temperature T_i reaches temperature T_{i+1} faster than realization b , which draws more hot water on average and increases the time required to reach T_{i+1} . Nevertheless, since the hot-water draw profiles in the population are assumed to be statistically identical, the average of these profiles reaches \bar{w}_{sst} for $t \rightarrow \infty$. Thus, the state transition rates for the large population are calculated considering the evolution of (2) with respect to the average hot-water draw of the population. The transition rates for the ON and OFF populations are computed next. Dropping the subscript n in (2), it follows that the solution with steady-state consumption $w = \bar{w}_{\text{sst}}$ and $T(0) = T_0$ is

$$T(t) = \Phi_{T_0}(t) = e^{-at} \left(T_0 - \frac{b(z)}{a} \right) + \frac{b(z)}{a}, \quad (8)$$

where $a = \frac{1}{\tau} + \frac{\bar{w}_{\text{sst}}}{60L}$ and $b(z) = \frac{T_{\text{amb}}}{\tau} + \frac{T_{\text{in}}}{60L} \bar{w}_{\text{sst}} + \frac{P_{\text{rate}}}{c\rho L\eta} z$. In particular, define $\Phi_{T_0}^{\text{on}}(t) = \Phi_{T_0}(t)|_{z=1}$ and $\Phi_{T_0}^{\text{off}}(t) = \Phi_{T_0}(t)|_{z=0}$. For the ON population, the dynamics imply forward transitions, i.e., from x_{on}^i to x_{on}^{i+1} as shown in Figure 5. First, take the boundaries of the temperature bin T_{i-1} and T_i corresponding to state x_{on}^i , and compute $T'_{i-1} = \Phi_{\Delta t}^{\text{on}}(T_{i-1})$ and $T'_i = \Phi_{\Delta t}^{\text{on}}(T_i)$. Note that in this case $T_i < T'_i$. Thus, the percentage of water heaters that remain in x_{on}^i and move to x_{on}^{i+1} , respectively, is given by

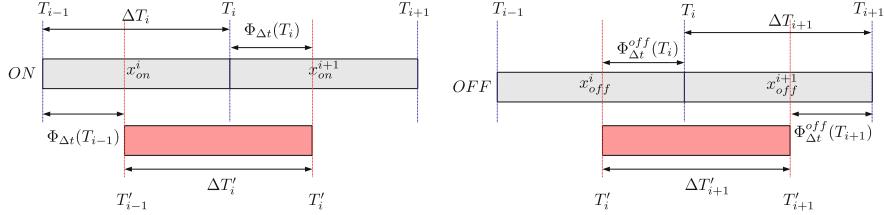


Fig. 5 Transition rate calculation for ON and OFF populations.

$$p_{ii}^{\text{on}} = \frac{T_i - T'_{i-1}}{T'_i - T'_{i-1}} \quad \text{and} \quad p_{i(i+1)}^{\text{on}} = \frac{T'_i - T_i}{T'_i - T'_{i-1}}.$$

Note that $p_{ii}^{\text{on}} + p_{i(i+1)}^{\text{on}} = 1$, as expected. Transition rates for the OFF dynamics are determined similarly but are reversed, i.e., from x_{off}^{i+1} to x_{off}^i since $T'_i = \Phi_{\Delta t}^{\text{off}}(T_i) < T_i$. Thus,

$$p_{ii}^{\text{off}} = \frac{T'_{i+1} - T_i}{T'_{i+1} - T'_i} \quad \text{and} \quad p_{i(i-1)}^{\text{off}} = \frac{T_i - T'_i}{T'_{i+1} - T'_i}.$$

The previous analysis was purposely restricted to state transitions between contiguous states. Using (8), one can compute an upper bound for Δt such that any EWH in state x_{on}^i only transitions to x_{on}^{i+1} and any EWH in x_{off}^{i+1} only transition to x_{on}^i for all i . Define

$$t_i^{\text{on}} = a^{-1} \log \left(\frac{T_i - \frac{b(z)}{a}}{T_{i+1} - \frac{b(z)}{a}} \right) \Big|_{z=1} \quad (9)$$

as the time that an EWH takes to go from T_i to T_{i+1} . Observe that if an EWH at T_i is kept ON for $t > t_i^{\text{on}}$ seconds, then $T(t) > T_{i+1}$. This implies that some EWHs in x_{on}^i will transition to x_{on}^{i+2} and skip x_{on}^{i+1} . Similarly, t_i^{off} is defined as in (9) but $z = 0$, and the transitions are in a reverse direction. The condition on the discretization time-step Δt for *contiguous transitions* is then formulated as $\Delta t < \min_i \{t_i^{\text{on}}, t_i^{\text{off}}\}$. For example, a state space partitioning having $N = 30$ temperature bins for each of the ON and OFF populations implies that $\Delta t < 60.27$ seconds in order to keep transitions between contiguous states.

In addition, the OFF-to-ON ($p_{\text{off}}^{\text{on}}$) and ON-to-OFF ($p_{\text{on}}^{\text{off}}$) transition rates must be computed to capture the jump to a transitory state that automatically transitions to x_{on}^1 and x_{off}^N , respectively. The complete Markov chain for conventional thermostatic control is shown in Figure 6a. It is important to observe that the transient effects on temperature caused by stochastic hot-water withdrawals are not captured since

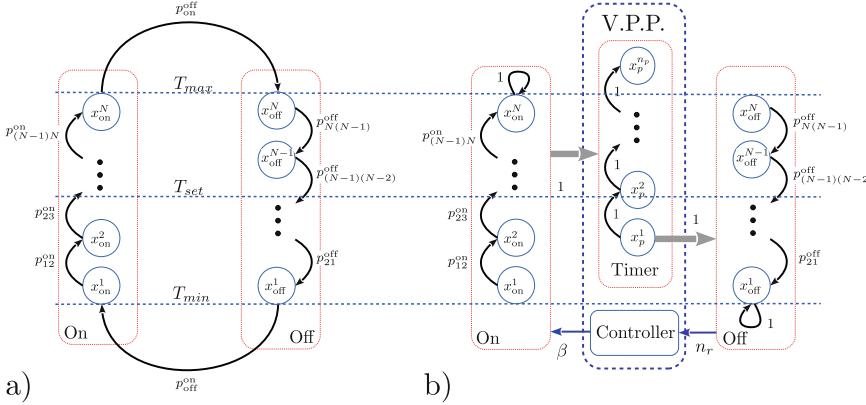


Fig. 6 Abstraction for (a) conventional thermostatic control and (b) PEM control, where self-loops are not visualized.

the transition rates assume a steady-state (mean) consumption of hot water. The Markov transition matrix M associated with conventional thermostatic control is then given as

$$M = \begin{pmatrix} p_{11}^{on} & 0 & \cdots & 0 & p_{on}^{off} & \cdots & 0 & 0 & 0 \\ p_{12}^{on} & p_{22}^{on} & \ddots & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & p_{23}^{on} & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{NN}^{on} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{11}^{off} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & p_{(N-1)(N-2)}^{off} & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & p_{(N-1)(N-1)}^{off} & p_{N(N-1)}^{off} & 0 \\ 0 & 0 & \cdots & p_{off}^{on} & 0 & \cdots & 0 & 0 & p_{NN}^{off} \end{pmatrix}. \quad (10)$$

Observe that the Markov chain associated to M is irreducible since one can reach any state from any arbitrarily initial state. It follows then that this abstraction possesses a unique invariant distribution as well since \mathcal{X} is finite dimensional. Nonetheless the conventional model lacks the flexibility inherent to PEM.

4.2 PEM Markov Model

Recall from Section 3 that, under PEM, an EWH can only switch ON for an epoch if its packet request is accepted by the VPP coordinator. That is, given the aggregate request rate, the VPP selects the proportion of EWHs that will receive a packet and

automatically switch them ON. To capture the unique nature of PEM's fixed packet duration and VPP's role, we leverage prior literature on fault tolerant recovery logic [27] and TCL modeling with compressor lockout periods [9]. Specifically, a fixed *timer* is added to the state bin transition model to track the population with accepted packet requests. The objective of this section is to present a macro-model describing PEM control as a *controlled Markov chain*.

Definition 1 (Controlled Markov Chain [25]) Let $\{u_k\}_{k \geq 0}$ be a sequence of real valued functions taking values on a set U . A Markov chain $\{X_k\}_{k \geq 0}$ is said to be a *controlled Markov chain* (CMC) if its transition matrix $M = M(u) := \{q_{ij}(u)\}_{1 \leq i, j \leq N}$ satisfies

$$\begin{aligned} P(X_{n+1} = x_{i_{n+1}} | X_n = x_{i_n}, \dots, X_0 = x_{i_0}, u_n, \dots, u_0) \\ = P(X_{n+1} = x_{i_{n+1}} | X_n = x_{i_n}, u_n) = p_{i_{n+1}i_n}(u_n). \end{aligned}$$

The definition also implies that $M(u)$ for a CMC must be a (column) stochastic matrix for any choice of $u \in U$. Assuming that all states of the CMC are observed, one can define a control policy: $u = \mathcal{X} \rightarrow U$, and, thus, $M = M(u(x))$. The probability mass function of a CMC is computed similarly using $q[k+1] = M(u[k])q[k]$ given an initial distribution $q[0]$ and control input $u[0]$.

In what follows, PEM control will be introduced in the context of CMCs. The underlying Markov transition matrix over which PEM is implemented is given by (10), but with $p_{\text{off}}^{\text{on}} = p_{\text{on}}^{\text{off}} = 0$ and $p_{NN}^{\text{on}} = p_{11}^{\text{off}} = 1$. That is, x_{off}^1 and x_{on}^N are absorbent states indicating that ON states cannot be reached from OFF states and vice versa. VPP control, therefore, becomes the interface between these two modes of operation. The mechanics of switching EWHs ON and OFF under PEM control are described next.

Suppose $q[k] \in \mathbb{R}^{2N}$ is the probability distribution of the PEM macro-model population at time k , $\beta_{\text{on}} = \text{diag}\{\beta_{\text{on}}^1, \dots, \beta_{\text{on}}^N\}$ with $\beta_{\text{on}}^i \in [0, 1]$ the percentage of the OFF population in state x_{off}^i that is switched ON by the VPP, and $\beta_{\text{off}} = \text{diag}\{\beta_{\text{off}}^1, \dots, \beta_{\text{off}}^N\}$ with $\beta_{\text{off}}^i \in [0, 1]$ the percentage of the ON population in state x_{on}^i that is switched OFF. The action of instantaneously switching ON and OFF proportion of devices in q is given by the transformation

$$\bar{q}[k] = \bar{M}(\beta_{\text{on}}[k], \beta_{\text{off}}[k]) q[k], \quad (11)$$

where

$$\bar{M}(\beta_{\text{on}}, \beta_{\text{off}}) = \left(\begin{array}{c|c} I_N - \beta_{\text{off}} & \beta_{\text{on}} \\ \hline \beta_{\text{off}} & I_N - \beta_{\text{on}} \end{array} \right), \quad (12)$$

and I_N denotes the N -dimensional identity matrix. Once $\bar{M}(\beta_{\text{on}}, \beta_{\text{off}})$ has switched some EWHs ON and some other OFF, the underlying transition matrix M acts on \bar{q} . This provides the dynamics

$$q[k+1] = M\bar{M}(\beta_{\text{on}}, \beta_{\text{off}}) q[k]. \quad (13)$$

The next theorem simply says that the sequence $\{X_k\}_{k \geq 0}$ associated (13) is a CMC.

Theorem 1 Let $\beta_{\text{on}}[k], \beta_{\text{off}}[k] \in \mathbb{R}^{N \times N}$ be defined as in (11) for all $k \geq 0$. The sequence $\{X_k\}_{k \geq 0}$ of random variables X_k taking values in \mathcal{X} and probability distribution satisfying (13) is a controlled Markov chain as described by Definition 1 with input $u[k] = (\mathbf{1}_N^\top \beta_{\text{on}}[k], \mathbf{1}_N^\top \beta_{\text{off}}[k])^\top \in \mathbb{R}^{2N}$.

Proof The proof is straightforward since (10) and (12) are stochastic matrices for arbitrary $\beta_{\text{on}}^i, \beta_{\text{off}}^i \in [0, 1]$ for all i , and the product of stochastic matrices is a stochastic matrix. ■

An important aspect of PEM control is that only EWHs that are OFF request a packet and do so as a function of the (local temperature) bin, which implies that not all OFF EWHs will turn ON. Therefore, define

$$\hat{q}[k] := \hat{M}q[k] = \begin{pmatrix} I_N & 0_N \\ 0_N & T_{\text{req}} \end{pmatrix} q[k] = \begin{pmatrix} q_{\text{on}}[k] \\ \hat{q}_{\text{off}}[k] \end{pmatrix},$$

where $T_{\text{req}} = \text{diag}\{p_1^{\text{req}}, \dots, p_N^{\text{req}}\}$ and $p_i^{\text{req}} := 1 - e^{-\mu(T_i^m)\Delta t}$ is the request probability assigned to x_{off}^i by (5) with respect to the midpoint of temperature bin i , T_i^m . Note that \hat{q} is not a probability mass function since $\mathbf{1}_N^\top (q_{\text{on}} + \hat{q}_{\text{off}}) < 1$, which means that the aggregate request rate, i.e., the population that can be switched ON, is given by

$$n_r[k] := \mathbf{1}_N^\top \hat{q}_{\text{off}}[k]. \quad (14)$$

Under PEM, the VPP determines the rate of accepting packets, $\beta[k]$. The resulting EWHs instantly switch ON when packet requests are accepted. The population of devices that switch from OFF to ON, q^+ , is a function of β and q_{off} . That is,

$$q^+[k] := \begin{pmatrix} 0_N & \beta[k]T_{\text{req}} \\ 0_N & -\beta[k]T_{\text{req}} \end{pmatrix} q[k] = M_{\beta[k]}^+ q[k] \quad (15)$$

In contrast, to model the population of EWHs that switch from ON to OFF requires information on the rate of expiring packets. In other words, let δ [secs] be the duration of a packet epoch, and then the EWHs that have been ON for δ seconds will turn OFF. This requires keeping track of how many EWHs were turned ON δ seconds ago and, essentially, constitutes a delayed system. However, one can augment states to the system dynamics to account for the needed memory, which is equivalent to having a timer. That is, given δ , the time-step Δt , and the vector of augmented (timer) states $x_p \in \mathbb{R}^{n_p}$ with $n_p = \lfloor \delta/\Delta t \rfloor$, the *timer dynamics* is given by

$$x_p[k+1] = M_p x_p[k] + C_p q_{\text{on}}^+[k] \quad (16a)$$

$$y_p[k] = x_p[k], \quad (16b)$$

where $M_p \in \mathbb{R}^{n_p \times n_p}$ is a zero matrix except for its first lower diagonal whose components are 1 and $C_p \in \mathbb{R}^{n_p \times N}$ is responsible for allocating the newly switched ON population into the timer states. Note that there exists a temperature T_p such that $\Phi_{T_p}^{\text{on}}(\delta) = T_{\max}$. Therefore it is necessary for C_p to interrupt packets to prevent exceeding temperature limits and thus wasting resources. Specifically, if $T_{i+1} < T_p$, C_p allocates all requesting EWHs from bin $[T_i, T_{i+1}]$ into the timer state x_p^1 . Otherwise, it allocates EWHs with $T_j > T_p$ in the timer state x_p^j with $j = \lfloor (\delta - t_j)/\Delta t \rfloor$, and t_j is the time it will take to increase the EWH's temperature from T_j to T_{\max} . Note that since the macro-model considers only binned (rather than exact) temperatures, the allocation of requests assumes that the mass function in each state is uniformly distributed.

The timer states are internal states of the VPP and provide information of the distribution of total ON population in PEM (i.e., $\mathbf{1}_N^\top q_{\text{on}}$) across all packet intervals, x_p . As in (15), one can define the population of EWHs that just completed their δ -second packet and will turn OFF instantly as

$$q^-[k] := \begin{pmatrix} \beta^-[k] I_N & | & 0_N \\ -\beta^-[k] I_N & | & 0_N \end{pmatrix} q[k] = M_{\beta^-[k]}^- q[k], \quad (17)$$

where $\beta^-[k] := y_p^{n_p}[k]/(\mathbf{1}_{n_p}^\top y_p[k])$. One can now formulate the ON/OFF switching events for the entire population as:

$$\bar{q}[k] := q[k] + q^+[k] - q^-[k] = (I + M_{\beta^-[k]}^+ - M_{\beta^-[k]}^-) q[k],$$

which yields the EWH population dynamics:

$$\begin{aligned} q[k+1] &= M(I + M_{\beta^-[k]}^+ - M_{\beta^-[k]}^-) q[k] \\ &= \bar{M}(\beta_{\text{on}}[k], \beta_{\text{off}}[k]) q[k], \end{aligned} \quad (18)$$

where $\beta_{\text{on}}[k] = \beta^-[k] T_{\text{req}}$ and $\beta_{\text{off}}[k] = \beta^-[k] I_N$. Note that there is no order in which EWHs are switched ON or OFF since both happen simultaneously. Figure 6b shows the state diagram of the population model under PEM control.

The next corollary follows directly from Theorem 1.

Corollary 1 *The sequence $\{X_k\}_{k \geq 0}$ of random variables X_k taking values in \mathcal{X} and probability distribution satisfying (18) is a controlled Markov chain with input $u[k] = (\mathbf{1}_N^\top \beta^-[k] T_{\text{req}}, \mathbf{1}_N^\top \beta^-[k] I_N)^\top$.*

4.3 Tracking with PEM Macro-model

In PEM, the input β is exogenous. Recall $P^{\text{rate}} := \frac{1}{n} \sum_{n=1}^N P_n^{\text{rate}}$, P_{ref} , and P_{dem} (see [18] for a list of the system parameters values) denote the average, reference, and demand power for the large-scale water heater system. In particular, $P_{\text{dem}}[k] := P_{\text{ON}}[k] - P_{\text{OFF}}[k]$, where $P_{\text{ON}}[k]$ is the power drawn by all EWHs that are ON at time k and $P_{\text{OFF}}[k]$ is the power released by all EWHs that were ON at time $k-1$ and subsequently were switched OFF at time k . Given n_r in (14) and that PEM tracking is activated (per Figure 4), the input $\beta[k]$ in Figure 6b is designed, using information generated by the VPP's macro-model at each instant of time k , to be

$$\beta[k] = \min \left\{ 1, \frac{P_{\text{ref}}[k] - P_{\text{dem}}[k]}{P^{\text{rate}} n_r[k]} \right\}$$

when $P_{\text{ref}} > P_{\text{dem}}$ and 0, otherwise. Observe in the diagram that the timer dynamics automatically releases the population in $x_p^{n_p}$ and transitions them all to the OFF states. Also, note that if $\beta[k] = 0$ for all k , then the state diagram becomes reducible since there the states cannot transition from ON to OFF. This last fact is undesirable given that x_{off}^1 ends up accumulating the entire population when k goes to infinity, which implies that every EWH becomes synchronized. This shortcoming is addressed by additional states that will allow cold EWHs to turn ON even when the VPP sets $\beta[k] = 0$. This *exit-ON/exit-OFF* mechanism is augmented to the PEM macro-model to ensure QoS as described next.

Remark 4 The above design of input β is convenient, yet valuable, but can be improved by considering β in an optimal control setting, which is outside the scope of this chapter.

4.4 Exit-ON/Exit-OFF Dynamics

As mentioned previously, the end-consumer QoS is of paramount importance when controlling a large-scale system of DERs. Specifically, *no one likes to take a cold shower*. Therefore, whenever an EWH's temperature falls outside the deadband $[T_{\min}, T_{\max}]$, it will exit the packetized scheme and revert to conventional control until a prespecified PEM *exit-ON* set point is reached. Once the *exit-ON* set point is reached, the EWH is allowed to reenter the packetized scheme.

The population of EWHs that are too cold and exit PEM (to turn ON) join the *exit-ON* mode dynamics (denoted by \oplus). On the other hand, if a water heater is too hot and has to turn OFF, then it joins *exit-OFF* mode dynamics (denoted by \ominus) at state x_{\ominus}^0 after which EWHs transition under M naturally to the requesting states. These two PEM exit modes of operations were introduced in Section 3. Adding these modes of operation to the PEM macro-model only requires

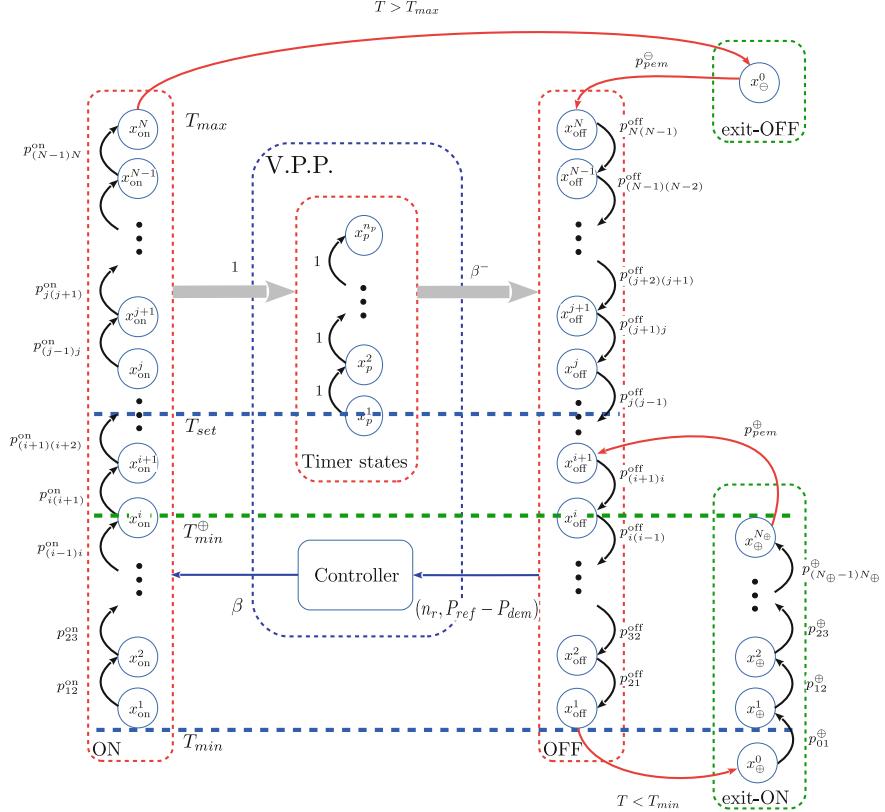


Fig. 7 PEM macro-model with exit-ON (\oplus) dynamics. ON/OFF state transitions are controlled by VPP and illustrated with gray and blue arrows.

a simple augmentation of states with their corresponding transition rates as shown in Figure 7. In the same figure, $T_{\min}^\oplus = T_{\text{set}} - T_{\text{PEM}}$ as explained in Section 3. The updated full population dynamics is given by (16) and

$$q[k+1] = M_{\text{exit}} \left(I + M_{\beta[k]}^+ + M_{\beta-[k]}^- \right) q[k], \quad y[k] = c^\top q[k],$$

where $M_{\text{exit}} := \text{diag}\{M_{\text{exit-ON}}, \bar{M}, M_{\text{exit-OFF}}\}$, $M_{\text{exit-ON}}$ is a matrix of zeros except for the main diagonal ($p_{11}^\oplus, \dots, p_{N_\oplus N_\oplus}^\oplus$) and the first lower diagonal ($p_{12}^\oplus, p_{23}^\oplus, \dots, p_{(N_\oplus-1)N_\oplus}^\oplus$), $M_{\text{exit-OFF}}$ introduces the probabilities to reenter PEM from x_{on}^N to x_Θ^0 and from x_{on}^N to x_{on}^N with p_{pem}^\oplus corresponding to the *exit-OFF* mode, and \bar{M} is such that $\bar{M}_{ij} = M_{ij}$ except for $M_{(N+N_\oplus+1)N_\oplus} = p_{pem}^\oplus$, which describe the transition probabilities to reenter PEM from the *exit-ON* mode.

4.5 Validating the EWH Macro-model Against the Micro-model

The internal temperature state and aggregate power output from a simulation of the macro-model are compared to those of a homogeneous agent-based (micro-) simulation with $N = 1000$ packetized EWHs. Specifically, the parameters for the homogeneous collection of 1000 EWHs are chosen as $P^{\text{rate}} = 4.5$; $T_{\text{inlet}} = T_{\text{amb}} = 14$; $L = 250$; and $\eta = 1.0$. The simulation parameters are $\delta = 300$ secs and $\Delta t = 5$ secs. The errors for the power signals are computed as

$$E_{\text{pow}} = \mathbb{E}[|y_{\text{micro}} - y_{\text{macro}^{\text{avg}}}||/y_{\text{micro}}] \times 100,$$

which is a percentage with respect to the micro-model simulation's output power. Also, the average temperatures profiles obtained from micro/macro-models are compared using $\text{RMSE}_{\text{temp}} = \sqrt{\mathbb{E}[(T_{\text{micro}}^{\text{avg}} - T_{\text{macro}}^{\text{avg}})^2]}$.

Figure 8 is a 6-hour simulation of the homogeneous micro- and macro-models. Both simulations start by accepting all requests ($\beta = 1$) for the first 2 hours, which illustrates “control-free” PEM (i.e., only local stochastic access driving the

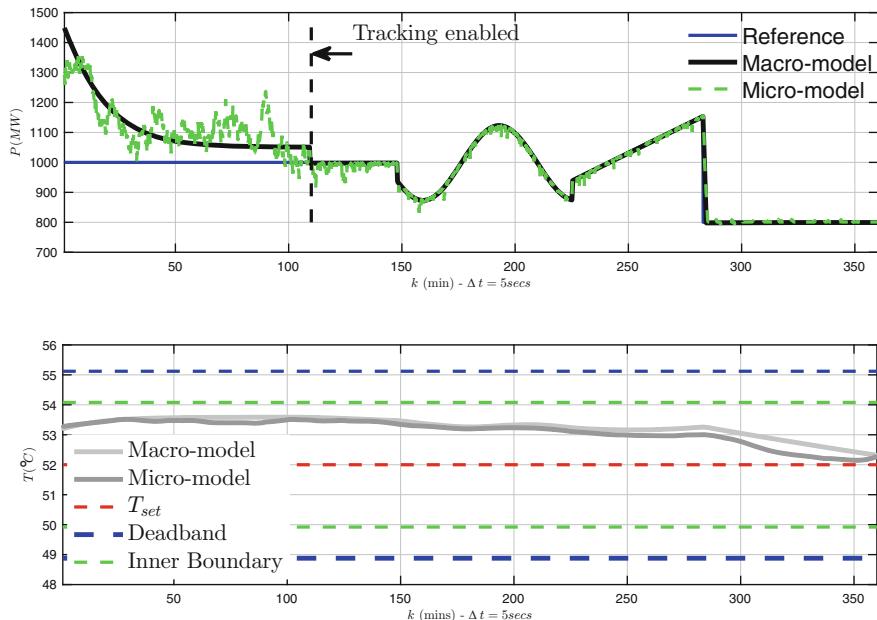


Fig. 8 Validating the homogeneous macro-model by comparing the aggregate power (top) and internal average temperature (bottom) against those of a homogeneous micro-simulation with $N = 1000$ packetized electric water heaters.

Table 1 Capacity parameter for micro-model with $L = 250 + \sigma v$ and $v \sim N(0, 1)$.

σ	$E_{\text{pow}} (\beta = 1)$	E_{pow}	$\text{RMSE}_{\text{temp}}$
10	1.5244	0.0646	0.2085
20	1.6054	0.0947	0.2118
30	1.7056	0.0699	0.2132
40	1.5968	0.0655	0.1960
50	1.6268	0.0619	0.2062

Table 2 Efficiency parameter for micro-model with $\eta = 1 - \sigma(1 - v)$ and $v \sim U(0, 1)$.

σ	$E_{\text{pow}} (\beta = 1)$	E_{pow}	$\text{RMSE}_{\text{temp}}$
0.1	2.8933	0.1763	0.5685
0.2	6.7277	0.3715	0.9730
0.3	10.778	1.2613	1.3740
0.4	15.432	4.4017	1.7924
0.5	18.872	8.0183	2.1642

system). After 2 hours, PEM tracking is enabled, and it is observed that the average population temperature and aggregate power output of both simulations agree during most of the tracking period. Also, note that the tracking errors stay within $\pm 5\%$, which highlights PEM's ability to extract flexibility from the packetized population of loads and accurately track provided reference.

The accuracy of the macro-model under increasing levels of heterogeneity is shown in Tables 1 and 2 for a few salient parameters. Heterogeneity in parameters T_{amb} and T_{inlet} is omitted here as their impact on errors is similar to parameter L (Table 1). The efficiency parameter η has a strong effect on micro-simulation's aggregate power output and average temperature; however, it is expected that, in practice, η is uniform across the population (and close to 1.0). Overall, the macro-model is accurate and captures the dominant behavior of a large population of packetized EWHs, which makes it viable to be used for control and estimation to develop further insight into the capabilities of PEM.

5 Numerical Case Study of PEM with Diverse DERs

While the previous section focused on homogeneous packetized TCLs, this section investigates how a single VPP, under PEM, can operate a diverse fleet of heterogeneous DERs. Specifically, the following case study will illustrate how 1500 heterogeneous packetized TCL (1000), PEV (250), and ESS (250) devices can all be coordinated under with single VPP and simultaneously track a reference signal (in the aggregate) and satisfy (local) QoS constraints. For an in-depth case study on just heterogeneous TCLs under PEM, please see [18], where a ramp-rate limit is introduced to the VPP to counteract the synchronization effects related to temperatures during long periods of reject all.

The uncontrollable background demand for each load type describes the random perturbations to the local dynamic state.

- **TCL:** for the 1000 residential electric water heaters, the uncontrollable demand represents the use of hot water in the home, such as a shower and running the washing machine or dishwasher. To model these hot-water events, we employ the following stochastic process:
 1. Choose the average number of hot-water (HW) events per hour, $H W_{avg}^{hr}$.
 2. For each TCL, uniformly distribute the total number of HW events with mean $2T_{sim} H W_{avg}^{hr}$.
 3. Randomly select hot-water events' starting times from available times, k_0^{HW} .
 4. For each HW event, choose duration Δk^{HW} from normally distributed $\min\{\max\{N(700, 300)/\Delta t, 1\}, 3600/\Delta t\}$.
 5. From the duration of each HW event, choose a constant hot-water withdrawal rate $v_n[k]$ [liters/min] based on the exponential distribution with mean $1200/(\Delta t \Delta k^{HW})$, which is inversely proportional to duration. A capacity of 30 liters/min is imposed on $v_n[k]$, which represents a high residential flow rate [28].
- **PEV:** the background demand in the case of the 250 plug-in electric vehicle batteries represents the driving patterns that discharge the battery. The PEV travel patterns were randomly sampled from travel survey data [29] for New England, as described in [30], which provides the stochastic model for residential arrival and departure times, as well as miles driven. From an assumed electric driving range of 150 miles and an electric driving efficiency of 6.7 miles per kWh, the total reduction in SOC is computed upon arriving home (to charge).
- **ESS:** the 250 home batteries represent Tesla's Powerwalls (2.0), which have battery capacity of 13.5 kWh, charge and discharge efficiency of around 95% (round trip of 92%), and a maximum (continuous) power rating of 5.0 kW. It is assumed that the battery owner stochastically charges or discharges the battery based on a Gaussian random walk with a minimum power draw of 1.5 kW in either direction. This could be representative of excess or deficit residential solar PV production or short-term islanding conditions.

The $N = 1500$ diverse DER devices are then packetized, and over an 8-hour period (16:00 to 24:00), the VPP will interact with the loads, and from 18:40 to 24:00, the VPP tracks a mean-reverting random signal that represents a balancing signal from the ISO. The tracking is achieved by denying or accepting packet requests based on real-time error between reference and aggregated VPP power output as described in Section 3. The tracking errors are less than 5% for packet epochs of $\delta = 5$ minutes. Figures 9 and 10 illustrate the tracking performance of the VPP and that QOS requirements are satisfied as well. Table 3 outlines the errors and other metrics of the diverse VPP during its tracking period, as a function of packet length. With increasing packet epoch δ , the flexibility of the VPP is reduced, which will increase tracking error metrics (MAPE, RMSE). However,

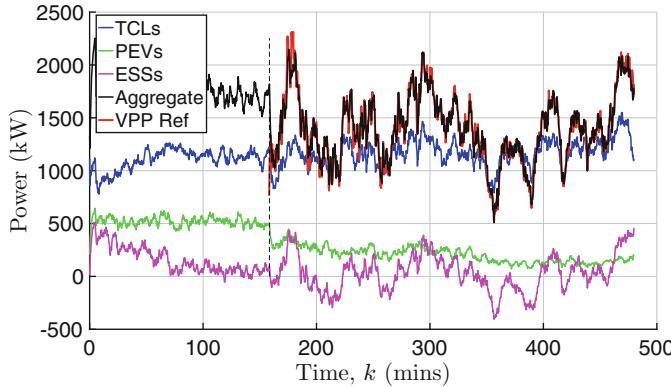


Fig. 9 The 8-hour case study of a diverse VPP under PEM (with devices=1000 TCLs, 250 PEVs, and 250 ESS), where the initial time (0) represents 4:00 PM, while the end time is midnight, which affects the arrival/departure rates of the PEVs. The aggregate power produced by the VPP is shown for an initial accept-all phase and a later tracking phase (after minute 160). The aggregate VPP output power is shown and has mean average percent tracking error (MAPE) and RMSE of 2.03% and 4.51%, respectively. The aggregate power from each of the three DER types is provided as well. Note that the ESS devices operate about 0 kW, while the PEVs can charge only when home.

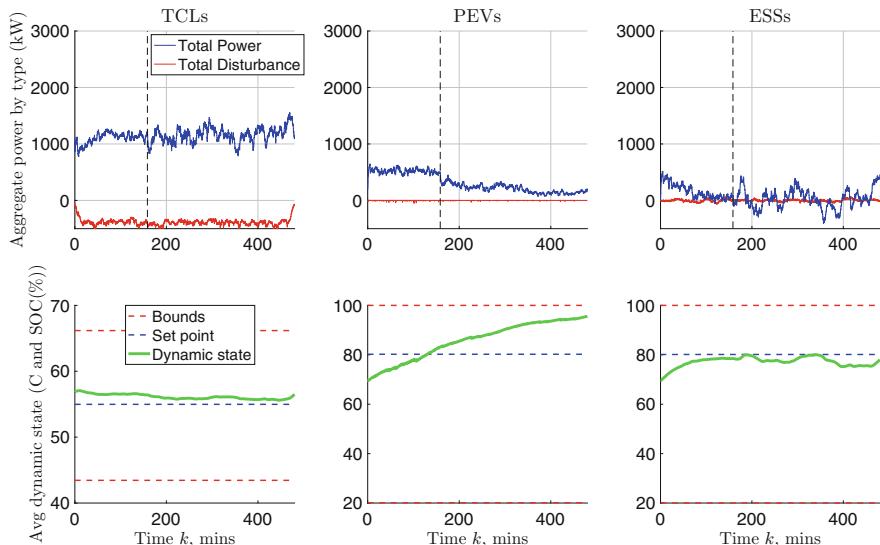


Fig. 10 Aggregate power and average dynamic state for each DER type. Despite the VPP tracking the reference for over 5 hours, the individual devices are able to both provide flexibility to VPP and satisfy QoS requirements. The left-most figures (top and bottom) are the TCLs, the center are PEVs, and the far-right are the ESS fleet. The TCL water heater and ESS battery should be close to their set point, while the objective of a PEV is to a) cross set-point barrier and b) aim to be fully charged.

Table 3 Comparing tracking performance of diverse VPP for different packet lengths.

Metric	$\delta = 5$ mins	$\delta = 10$	$\delta = 15$	$\delta = 20$	$\delta = 30$
MAPE (%)	2.03	3.08	4.11	4.46	5.84
RMSE (%)	4.51	6.07	7.44	8.10	10.35
RMSE (kW)	58.93	76.76	86.93	98.13	126.01
Avg TCL ON/OFF cycles per hour	4.12	2.58	2.08	1.83	1.64
Avg device state deviation from set point (%) ^a	4.69	4.96	5.28	5.51	5.91

^aFor the PEVs, the set point has been defined as being fully charged (or \bar{x}_n)

with increasing packet lengths, the devices will cycle less often, which can help preserve the mechanical integrity of relays in electric water heaters (but may not have a significant effect on battery inverters). Therefore, there is a tradeoff between tracking performance and mechanical device degradation. In addition, as the packets become longer, the individual loads deviate further from their set points, which implies that the VPP requires greater control effort despite the reduced tracking performance. Thus, there is a need to develop the analysis and optimal controller design for a VPP, which will be achieved by extending the macro-model in Section 4 to include not just TCLs but also PEVs and ESS. The final simulation further illustrates the value in managing a diverse fleet of DERs.

Consider two VPPs: one is comprised of 500 TCLs, 250 PEVs, and 250 ESS batteries (i.e., diverse VPP), while the other contains 1000 TCLs (i.e., *TCL-only VPP*). Figure 11 illustrates how these two VPPs perform in tracking a signal composed of step, periodic, and ramp changes. It is clear that the diverse VPP outperforms the TCL-only VPP. In fact, the tracking RMSE for the diverse VPP is four times smaller than the TCL-only VPP (54 kW vs. 220 kW), while the MAPE is 30% lower (2% vs. 36%). Moreover, observe that this gain in performance comes without sacrificing QoS as the TCLs in both VPPs experience nearly identical mean absolute deviation from the temperature set point: 2.4°C vs. 2.5°C (with similar standard deviations). To further illustrate the value of a diverse fleet of resources, Figure 12 provides the ON/OFF statuses for each device in the respective VPPs. Careful comparison of the VPP illustrates that the TCL-only VPP fails to track the lower parts of the reference signal due to many TCLs opting out (i.e., transitions to *exit-ON*) as signified by very long continuous ON periods for the TCL-only VPP in Figure 12. That is, diversity in resources not only improves tracking ability but also improves QoS delivered to end consumer.

6 Conclusion and Next Steps for PEM

The focus of this chapter has been on the coordination of diverse DERs under PEM, including the validation of a macro-model for homogeneous TCLs and a case study to support PEM under heterogeneity. These results lay the ground work for

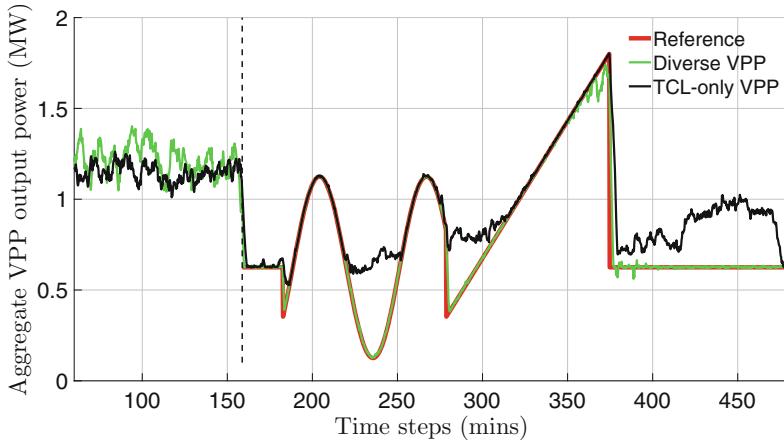


Fig. 11 Two VPPs are tracking a multimode reference signal with different sets of DERs. The diverse VPP (with 500 TCLs, 250 PEVs, and 250 ESS batteries) significantly outperforms the 1000 TCL-only VPP by leveraging the bidirectional capability of the batteries while maintaining QOS across all DER types. The TCL-only VPP is unable to track due to a large number of TCLs the enter exit-ON and opt out of PEM.

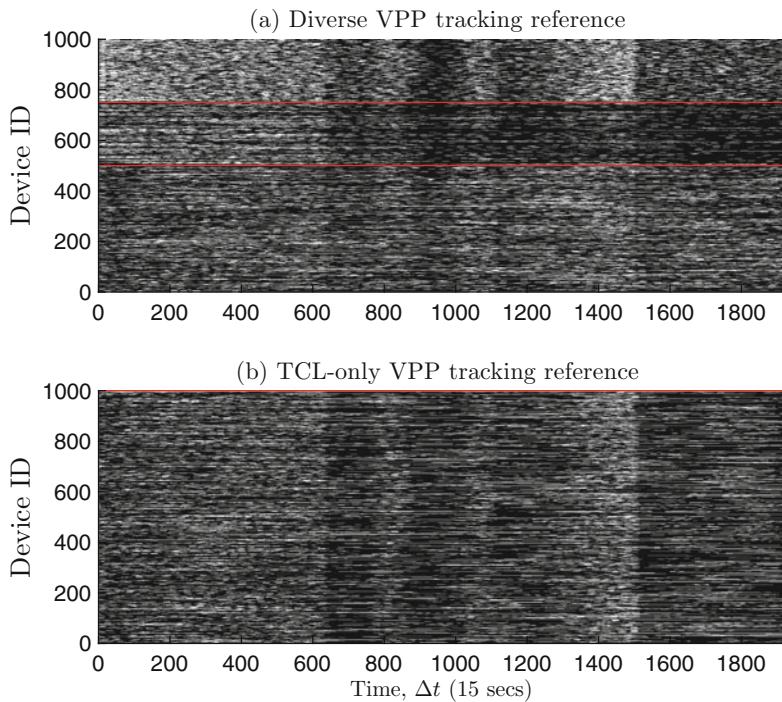


Fig. 12 The ON/OFF status of all devices in each VPP. The red lines indicate different DER types available to the diverse VPP in (a): 250 PEVs in the middle and 250 ESS batteries in the top band. The batteries are important during the downward phases of the signal as they offer the VPP added flexibility. In addition, the diversity helps avoid large-scale opting out of TCLs after the final step change (in minute 380).

analysis to quantify macro-model uncertainty, forecast VPP flexibility capabilities and uncertainty bounds on performance, and enable the development of optimal control techniques for managing the VPP resources in order to improve tracking performance and QoS. Specifically, we are interested in embedding multiple VPPs into optimal power flow problems for transmission and/or distribution to develop grid-optimized reference signals that can be used in conjunction with real-time balancing between VPPs to improve resilience, reliability, and economics of power system operations.

This chapter has presented a novel and innovative paradigm for coordinating DERs: packetized energy management (PEM). PEM has numerous advantages over many of today's state-of-the-art coordination algorithms. At the core of PEM is local decision-making that randomizes the requests rates, which promotes asynchronous coordination across the population and protects the system from unwanted effects of synchronization. That is, PEM is truly a bottom-up approach that protects the privacy of the consumer. In addition, since a VPP only requires a measurement of the aggregate power output and the packet request rate, PEM offers a simple framework for modeling and control that avoids having to rely on entire histograms to be transmitted to devices for control.

These unique properties of PEM have been explored in this chapter through simulations and modeling. On the modeling front, a state bin transition population model is augmented with a timer that tracks the completion of a packet and informs the VPP of expiring packets (i.e., device that soon will switch OFF). This information will be valuable as we build up the optimal control framework for PEM. In addition, the model captures the opt-out mechanism, which provides a mechanism for improving QOS (albeit it by reducing available flexibility).

Acknowledgements This work was supported by the US Department of Energy's Advanced Research Projects Agency-Energy (ARPA-E) grant DE-AR0000694.

References

1. Frolik J (2004) Qos control for random access wireless sensor networks. In: Proceedings of 2004 wireless communications and networking conference (WCNC04), Greenville, SC
2. Abramson N (1977) Throughput of packet broadcasting channels. IEEE Trans Commun 25(1):117–128
3. Bianchi G (1998) IEEE 802.11-saturation throughput analysis. IEEE Commun Lett 2(12):318–320
4. Callaway DS, Hiskens IA (2011) Achieving controllability of electric loads. Proc IEEE 99(1):184–199
5. Morgan M, Talukdar S (1979) Electric power load management: some technical, economic, regulatory and social issues. Proc IEEE 67(2):241–312
6. Schwepp FC, Tabors RD, Kirtley JL, Outhred HR, Pickel FH, Cox AJ (1980) Homeostatic utility control. IEEE Trans Power Syst 99(3):1151–1163
7. Mathieu JL, Koch S, Callaway DS (2013) State estimation and control of electric loads to manage real-time energy imbalance. IEEE Trans Power Syst 28(1):430–440

8. Chen Y, Basic A, Meyn SP (2015) State estimation and mean field control with application to demand dispatch. In: 2015 54th IEEE conference on decision and control. IEEE, New York, pp 6548–6555
9. Zhang W, Lian J, Chang C-Y, Kalsi K (2013) Aggregated modeling and control of air conditioning loads for demand response. *IEEE Trans Power Syst* 28(4):4655–4664
10. Esmaeil Zadeh Soudjani S, Abate A (2015) Aggregation and control of populations of thermostatically controlled loads by formal abstractions. *IEEE Trans Control Syst Technol* 23(3):975–990
11. Mahdavi N, Braslavsky J, Perfumo C (2016) Mapping the effect of ambient temperature on the power demand of populations of air conditioners. *IEEE Trans Smart Grid* 99:1–11
12. Meyn SP, Barooah P, Basic A, Chen Y, Ehren J (2015) Ancillary service to the grid using intelligent deferrable loads. *IEEE Trans Autom Control* 60(11):2847–2862
13. Zhang B, Baillieul J (2012) A packetized direct load control mechanism for demand side management. In: IEEE conference on decision and control, pp 3658–3665
14. Zhang B, Baillieul J (2013) A novel packet switching framework with binary information in demand side management. In: IEEE conference on decision and control. IEEE, New York, pp 4957–4963
15. Zhang B, Baillieul J (2016) Control and communication protocols based on packetized direct load control in smart building microgrids. *Proc IEEE* 104(4):837–857
16. Frolik J, Hines P (2012) Urgency-driven, plug-in electric vehicle charging. In: Proceedings of IEEE PES Innovative Smart Grid Technology (ISGT-Europe), Berlin
17. Rezaei P, Frolik J, Hines PDH (2014) Packetized plug-in electric vehicle charge management. *IEEE Trans Smart Grid* 5(2):642–650
18. Almassalkhi M, Frolik J, Hines P (2017) Packetized energy management: asynchronous and anonymous coordination of thermostatically controlled loads. In: American control conference
19. Duffaut Espinosa L, Almassalkhi M, Hines P, Heydari S, Frolik J (2017) Towards a macro-model for packetized energy management of resistive water heaters. In: IEEE conference on information sciences and systems
20. Frolik J, Hines P (2012) Random access, electric vehicle charge management. In: 1st IEEE international electric vehicle conference (IEVC), Greenville
21. Rezaei P, Frolik J, Hines P (2014) Packetized plug-in electric vehicle charge management. *IEEE Trans Smart Grid* 5(2):642–650
22. Goh C, Apt J (2004) Consumer strategies for controlling electric water heaters under dynamic pricing. In: Proceeding of Carnegie Mellon Electricity Industry Center
23. Kondoh J, Lu N, Hammerstrom DJ (2011) An evaluation of the water heater load potential for providing regulation service. *IEEE Trans Power Syst* 26(3):1309–1316
24. Lu N, Chassin DP (2004) A state-queueing model of thermostatically controlled appliances. *IEEE Trans Power Syst* 19(3):1666–1673
25. Kumar PR, Varaiya P (1986) Stochastic systems: estimation, identification and adaptive control. Prentice Hall, Upper Saddle River
26. Buchberger SG, Wu L (1995) Model for instantaneous residential water demands. *J Hydraul Eng* 121(3):232–246
27. Zhang H, Gray WS, Gonzalez OR (2008) Performance analysis of digital flight control systems with rollback error recovery subject to simulated neutron-induced upsets. *IEEE Trans Control Syst Technol* 16(1):46–59
28. ASHRAE (2002) Chapter 49: service water heating. In: ASHRAE applications handbook. ASHRAE, New York, pp 49.1–49.22
29. U.S. Department of Transportation, Federal Highway Administration (2009) National Household Travel Survey (NHTS) [Online]. Available: <http://nhts.ornl.gov/download.shtml>
30. Hilshey AD, Rezaei P, Hines P, Frolik J (2012) Electric vehicle charging: transformer impacts and smart, decentralized solutions. In: IEEE Power and Energy Society General Meeting. IEEE, New York, pp 1–8

Ensemble Control of Cycling Energy Loads: Markov Decision Approach



Michael Chertkov, Vladimir Y. Chernyak, and Deepjyoti Deka

Abstract A Markov decision process (MDP) framework is adopted to represent ensemble control of devices with cyclic energy consumption patterns, e.g., thermostatically controlled loads. Specifically we utilize and develop the class of MDP models previously coined linearly solvable MDPs, that describe optimal dynamics of the probability distribution of an ensemble of many cycling devices. Two principally different settings are discussed. First, we consider optimal strategy of the ensemble aggregator balancing between minimization of the cost of operations and minimization of the ensemble welfare penalty, where the latter is represented as a KL-divergence between actual and normal probability distributions of the ensemble. Then, second, we shift to the demand response setting modeling the aggregator's task to minimize the welfare penalty under the condition that the aggregated consumption matches the targeted time-varying consumption requested by the system operator. We discuss a modification of both settings aimed at encouraging or constraining the transitions between different states. The dynamic programming feature of the resulting modified MDPs is always preserved; however, "linear solvability" is lost fully or partially, depending on the type of modification. We also conducted some (limited in scope) numerical experimentation using the formulations of the first setting. We conclude by discussing future generalizations and applications.

M. Chertkov (✉)

Theoretical Division, T-4 & CNLS, Los Alamos National Laboratory, Los Alamos, NM, USA

Energy System Center, Skoltech, Moscow, Russia

e-mail: chertkov@lanl.gov

V. Y. Chernyak

Department of Chemistry, Wayne State University, Detroit, MI, USA

e-mail: chernyak@chem.wayne.edu

D. Deka

Theoretical Division, T-4 & CNLS, Los Alamos National Laboratory, Los Alamos, NM, USA

e-mail: deepjyoti@lanl.gov

1 General Motivation/Introduction

Power systems, as well as other energy systems, have undergone a transition from traditional device-oriented and deterministic approaches to a variety of novel approaches to account for

- stochasticity and uncertainty in how devices, especially new devices such as wind farms, are operated;
- network aspects, e.g., with respect to optimization, control, and design/planning; and
- utilization of massive amounts of newly available measurements/data for the aforementioned settings.

Such approaches, in particular Demand Response (DR), have become one important component of smart grid development, see [7, 32] and the references therein. Novel DR architectures break the traditional paradigm where only generators are flexible, and hence suggest that participation of flexible loads in control can benefit the grid at large without compromising load/consumer comforts significantly.

DR assumes that the loads are capable of following operational commands from the system operator (SO). Using DR in the range from tens of seconds to minutes is a potential attractive niche for frequency control that maintains the balance between production and consumption [1, 13, 31, 38]. Traditional frequency control is achieved by adjusting the generators, whereas DR (when developed) helps to achieve the balance additionally by adjusting the loads. The control task for loads in the DR setting is set by the SO as a temporal consumption request. Such requests can be formulated ahead of time (e.g., for the next 10 min) or in real time, using frequent updates (e.g., every 10 seconds).

This type of frequency control, namely DR services, is already provided by big consumers such as aluminum smelters [35] and desalination plants [33]. However, the potential effect would be an order of magnitude larger if small loads are available to provide DR services. Nevertheless, direct involvement of small-scale consumers is expensive because of associated communication and control costs. A viable solution to this problem is to control many small-scale consumers indirectly via an intermediate entity, also called the aggregator of an ensemble that includes many (e.g., thousands or tens of thousands) small-scale consumers [2, 6, 7, 11, 12, 21, 24–27]. Therefore, this novel DR architecture assumes, action-wise, separation into the following three levels:

- (a) SO sends control request/signal to an aggregator. The SO request is stated as a temporal profile (possibly binned into short intervals, e.g., 15 seconds) that is expected from the ensemble for an upcoming duration of length, e.g., 10 minutes.
- (b) Aggregator (A) processes the SO signal by solving an optimization/control problem and broadcasts the same A-signal, which is the output/solution of the optimization/control problem, to all members of the ensemble (end-point consumers). The broadcast of identical signal to end-point devices/consumers makes the communication step cheap.

- (c) End-point devices receive and implement the A-signal into control action. The implementation is assumed to be straightforward, and at most require only light device-level computations.

A number of challenges are associated with this novel aggregator-based DR architecture. The challenges are of both formulation (conceptual) and solution (implementation) type. We will mention a couple of these challenges most relevant to level (b) of the control architecture, which is the focus of this manuscript.

In posing the A-optimization, one desires a simple/minimal, yet sufficiently accurate, way of modeling individual devices. In particular, an acceptable framework needs to model the devices in terms of their achievable states and spatio-temporal resolution. To address this challenge, we will state the aggregator-level problems through the language of Markov decision process (MDP), and therefore utilize the transition probabilities between states in MDP as control variables. More accurately, we will assume that each device can find itself in all (or a subset of all) of the finite number of states. Then we describe the probabilistic state of the ensemble in terms of a vector of non-negative numbers that represent the fraction of all consumers observed in different states at any given moment of time. At each time, the probabilistic state vector thus sums to unity. The optimization/control degree of freedom for the A-optimization is represented by the set of stochastic transition probability matrices between the states, defined at each time slot over the discretized and finite time horizon.

In addition to the frequency control formulation, we will also discuss another setting that is of interest by itself but also less challenging in terms of finding efficient computational solutions. In this algorithmically simplified case we will look for an optimal balance between the cost of the ensemble operations when the price of electricity changes in time and the deviation of the probabilistic state of the ensemble from its natural/normal behavior under uniform prices or without the price bias.

The discrete-time, discrete-space ensemble modeling has a continuous-time, mixed-space counterpart known as thermostatically controlled loads (TCLs), or even more generally cycling loads, that are characterized by periodic (or quasi-periodic) evolution in the phase space. See [7] and our recent paper [9] for detailed discussions of TCL modeling and relevant references. In this manuscript we choose to work with discrete-time, discrete-space Markov process (MP) models because of their universality and flexibility. Indeed, an MP model can be viewed as a macro-model that follows from its micro-model counterpart—the TCL, after spatio-temporal model reduction (coarse-graining), see, e.g., [29]. However, MPs can represent a much broader class of models than those obtained via coarse-graining of TCLs or even than a combination (average) of different TCL models. The more general class of MDP models is especially useful in the context of machine learning (ML), where an MP is reconstructed from actual measurements/samples of the underlying ensemble.

From MP, which represents stochastic dynamics, we transition to MDP, which represents stochastic optimization with the MP/ensemble state constrained to being within a specified class of state dynamics. Many choices of MDP are reasonable from the perspective of practical engineering. In this manuscript we adopt and develop an approach pioneered in [18] and further developed in [15, 16, 22, 23, 28, 36]. This relatively unexplored formulation of the stochastic optimal control is known as “path integral control” in the case of continuous-time, continuous-space formulations [22] and as “Linearly Solvable MDP” (LS-MDP) in the case of a discrete-time, discrete-space setting [15, 16, 36]. Linear solvability is advantageous because it allows us to determine analytic expressions for the optimal solution in settings that go one step further than possibly through Dynamic Programming (DP) approaches. (We remind the reader that DP involves recursive solutions of the Hamilton–Jacobi–Bellman [HJB] equations.) One extra advantage of the LS-MDP, which holds even when reduction of the HJB equation to a linear equation is not possible, is related to the fact that the penalty term associated with deviation of the optimized transition probability from its ideal shape (not perturbed by the SO signal) has a very natural form of the generalized Kullback–Leibler distance between the two probability measures.

When discussing MPs and MDPs, we will utilize the following (rather standard) notations and terminology:

- The dynamic state of the ensemble is described in terms of a vector, $\rho(t) = (\rho_\alpha(t)|\forall\alpha)$, where $\rho_\alpha(t) \geq 0$ is the probability to find a device (from the ensemble) in state α at time t . The vector is normalized (no probability loss) at all times considered, i.e., $\sum_\alpha \rho_\alpha(t) = 1$, $\forall t = 0, \dots, T$, $\forall\alpha$.
- $p(t) = (p_{\alpha\beta}(t)|\forall t, \forall\alpha, \beta)$ is the transition probability matrix that describes an MP. It is set to be an optimization variable within the MDP framework. The matrix element, $p_{\alpha\beta}(t)$, represents the transition probability, i.e. the probability for a device which was at state β at time t to transition to state α at $t + 1$. The matrix is stochastic, i.e.

$$\sum_\alpha p_{\alpha\beta}(t) = 1, \quad \forall t = 0, \dots, T - 1, \quad \forall\beta, \quad (1)$$

- The evolution of the MP/ensemble state is described by the master equation (ME):

$$\rho_\alpha(t + 1) = \sum_\beta p_{\alpha\beta}(t)\rho_\beta(t), \quad \forall t, \quad \forall\alpha. \quad (2)$$

The ME should be supplemented by the initial condition for the state of the ensemble:

$$\rho_\alpha(0) = \rho_{in;\alpha}, \quad \forall\alpha \quad (3)$$

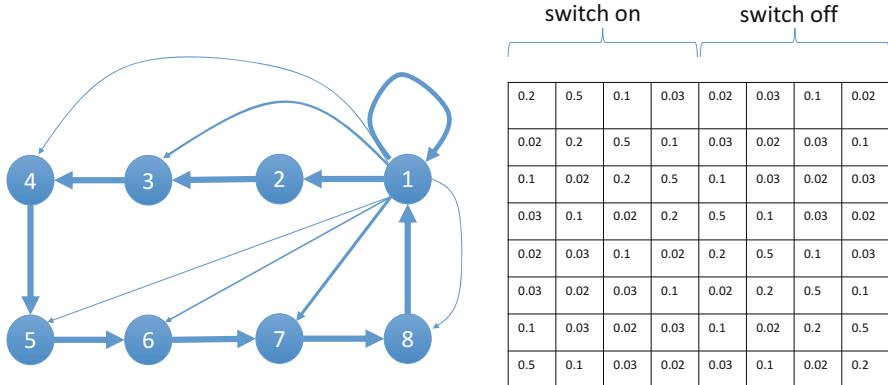


Fig. 1 An example MDP resulting from space-binning/discretization and time-discretization of a two-level mixed-state TCL model of the type discussed in [9]. This model is used in Section 4 for illustrative experiments. The directed graph of the “natural” transitions and respective stochastic \bar{p} matrix are shown in the left and right subfigures, respectively.

Given the initial condition (3), one needs to run ME (2) forward in time to find the states of the MP ensemble at all future times. The direction of time is important because it reflects the physical causality of the setting.

- MDP formulations discussed in the manuscript are stated as optimizations over the transition probability matrix p . All of the formulations also depend on the target \bar{p} , which corresponds to the optimal transition matrix if the ensemble is “left alone” for a sufficiently long time. In the case of frequency control, discussed in Section 5, \bar{p} corresponds to the steady state of the ensemble when the frequency tracking guidance is ignored. In the formulations of Sections 2 and 3, \bar{p} corresponds to the steady state of the ensemble operating for a sufficiently long time in the case of a flat, i.e., time- and state-independent, cost. An example MP and respective “natural” \bar{p} , introduced in the result of coarse-graining of a TCL, is shown in Figure 1.

The material in the remainder of the manuscript is organized as follows. MDPs aimed at finding the profit optimal transition probability vector for an ensemble of devices are defined and analyzed in Section 2. MDPs discussed in this section measure the deviation of the “optimal” p from its “normal” counterpart, \bar{p} , in terms of the standard KL divergence. This formulation is an LS-MDP; we have placed a detailed technical discussion of the MDP’s linear solvability and related features and properties in Section 8. Next, in Section 3, a modified MDP is discussed that differentiates between state transitions, i.e., discounts some transitions and encourages others. This formulation is richer in comparison to the differentiation-neutral formulation of Section 2. Numerical simulations for the optimization objectives discussed in Sections 2 and 3 are presented in Section 4. In Section 5 we describe and discuss the solution of an MDP, seeking to minimize the welfare deviation, stated in terms of the KL distance (or weighted KL distance) between

the “optimal” and “normal” transition probabilities while also matching the SO objective exactly. Preliminary discussion on integrating the MDP approaches into the optimum power flow formulations, e.g., to jointly control grid voltages and to minimize the power losses, is given in Section 6. Section 7 is reserved for the conclusions and discussions of the path forward.

2 Cost-vs-Welfare Optimal

Our main MDP formulation of interest, which we term “cost vs welfare,” is stated as the following optimization:

$$\min_{p, \rho} \mathbb{E}_\rho \left[\sum_{t=0}^{T-1} \sum_{\alpha} \left(\underbrace{U_\alpha(t+1)}_{\text{Cost of Electricity}} + \underbrace{\sum_{\beta} \log \frac{p_{\alpha\beta}(t)}{\bar{p}_{\alpha\beta}}}_{\text{welfare penalty}} \right) \right] , \quad (4)$$

Equations (1,2,3)

$$= \min_{p, \rho} \sum_{t=0}^{T-1} \sum_{\beta} \rho_{\beta}(t) \left(\sum_{\alpha} p_{\alpha\beta}(t) \left(U_\alpha(t+1) + \log \frac{p_{\alpha\beta}(t)}{\bar{p}_{\alpha\beta}} \right) \right) , \quad (5)$$

Equations (1,2,3)

where the matrix p is the optimization/control variable, which is stochastic according to Equation (1). Here in Equation (4), \bar{p} is an exogenously known stochastic matrix describing the transition probabilities corresponding to normal dynamics/mixing within the ensemble, i.e., \bar{p} explains the dynamics that the ensemble would show in the case of “cost ignored” objective ($U = 0$). The stochasticity of \bar{p} means that \bar{p} satisfies Equation (1), when p is replaced by \bar{p} .

We assume that when following \bar{p} , the ensemble mixes sufficiently fast to reach statistical steady state, $\rho^{(\text{st})}$, i.e., $\bar{p}\rho^{(\text{st})} = \rho^{(\text{st})}$.

The essence of the optimization (4) is a compromise that an aggregator aims to achieve between cost savings for the ensemble and keeping the level of discomfort (welfare penalty) to its minimum. The two conflicting objectives are represented by the two terms in Equation (4).

Optimization (4) is rather general, whereas for the example of a specific discrete-time–discrete-space TCL, one sets U_α to non-zero for only the states α that represent the “switch-on” states of the TCL.

Equation (4) belongs to the family of the so-called LS-MDPs introduced in [36], discussed in [15, 16, 28], and briefly described as a special case in Section 8. Solution of Equations (4) is fully described by Equations (24, 27, 28).

According to the general description part of Section 8, the problem is solved in two DP steps:

- Backward in time step. Compute p recursively by advancing backward in time according to Equations (24, 27), where $\gamma(t)$ is substituted by 1, and starting from the final condition Equation (28).
- Forward in time step. Reconstruct ρ by running Equation (2) forward in time with the initial condition Equation (3).

3 Differentiating States Through a Penalty/Encouragement

The KL welfare penalty term in Equation (4) is restrictive in terms of how the transitions between different states, and also those observed at different moments of time, compare to each other. To encourage/discourage or generally differentiate the transitions, one may weight the terms in the KL sum differently, thus introducing the $\gamma_{\alpha\beta}(t)$ factors:

$$\min_{p, \rho} \mathbb{E}_\rho \left[\sum_{t=0}^{T-1} \sum_{\alpha} \left(\underbrace{U_\alpha(t+1)}_{\text{Cost of Electricity}} + \underbrace{\sum_{\beta} \gamma_{\alpha\beta}(t) \log \frac{p_{\alpha\beta}(t)}{\bar{p}_{\alpha\beta}}}_{\text{welfare penalty}} \right) \right] \quad . \quad (6)$$

Equations (1,2,3)

This generalization of Equation (4) aims to ease the implementation of the optimal decision, e.g., emphasizing or downplaying the controllability of transitions between the states and at different moments of time.

Solution of the optimization (6) is described in Section 8. Even though linear solvability of the state-uniform formulation, discussed in Section 2, does not extend to the non-uniform formulation of Equation (6), the basic DP approach still holds and the problem is solved via the following backward-forward algorithm:

- Backward in time step. Starting with the final conditions Equation (17), one solves Equation (18) for φ recursively backward in time. Solving Equation (18) requires, at each time step, execution of inner loop iterations (until convergence), according to Equation (22), to find the Lagrangian multiplier λ . Following it, p at that time is reconstructed according to Equation (19). Alternatively, p can be determined by minimizing convex function Equation (18) directly with a linear constraint reflecting the stochasticity of p .
- Forward in time step. Reconstruct ρ running Equation (2) forward in time with the initial condition Equation (3).

4 Computational Experiments

In this section we describe some (preliminary) computational experiments conducted for MDP settings discussed in the two preceding sections. We choose the example case with 8 states (four “on,” four “off”) and target transition probability,

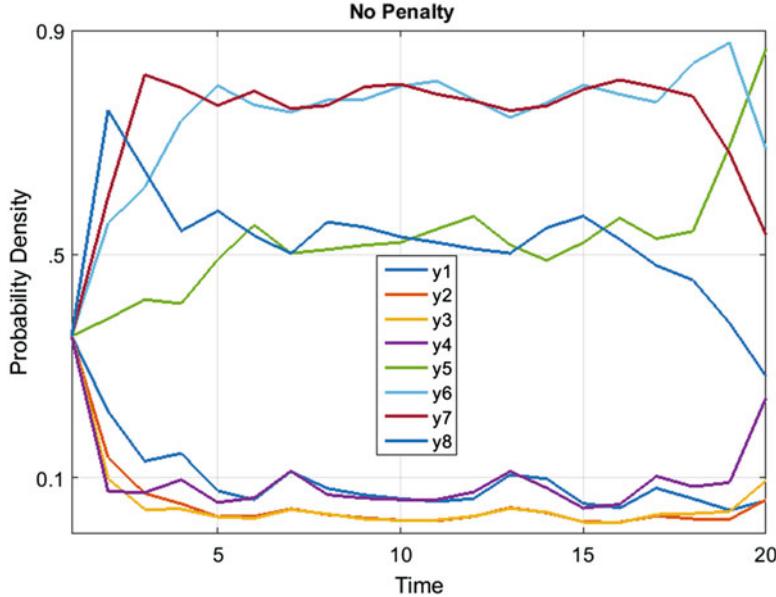


Fig. 2 Example solution of the MDP problem for the case without penalty, i.e., with the objective represented by Equation (4). The initial probability distribution corresponds to the steady state of the “target” MC with the transition probabilities shown in Figure 1. Eight curves show optimal dynamics of the respective, $\rho_\alpha(t)$, $\alpha = 1, \dots, 8$.

\bar{p} , in the form shown in Figure 1. We conduct the experiments in the regime with a non-stationary, time-dependent, and random cost term, $\sim 1 + \text{rand}(t)$, that is nonzero for only the first four (“on”) states. For the same target \bar{p} , we show solutions of two optimization formulations that correspond to the setting of Equation (4) and Equation (6), respectively. In the latter case, we choose a time-independent penalty factor $\gamma_{\alpha\beta} = 10$ for all transitions except for those that correspond to advancing one (counterclockwise) step along the cycle. The special “along the cycle” transitions are not penalized and given a penalty factor equal to unity, $\gamma_{\text{mod}[\alpha+1],\alpha} = 1$, $\alpha = 1, \dots, 7$. We use the algorithms described at the end of Section 2 and Section 3, respectively, to solve the MDPs.

The results are shown in Figures 2 and 3. Comparing the figures corresponding to the two regimes (with and without penalty), one observes that imposing penalty leads to a more homogeneous distribution ρ_α over the states α .

5 Ancillary Services-vs-Welfare Optimal

Another viable business model for an aggregator of an ensemble of cycling devices is to provide ancillary (frequency control) services to a regional SO. The ancillary services consist of adjusting the energy consumption of the ensemble to an

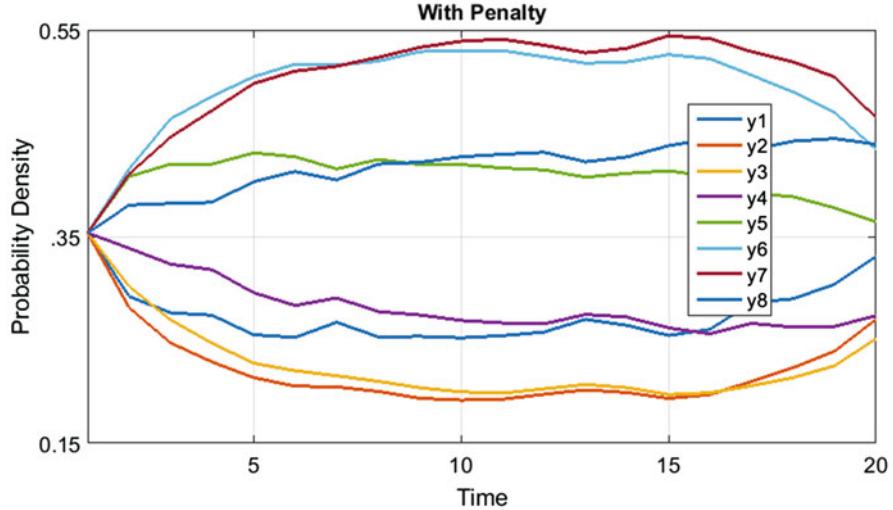


Fig. 3 Example solution of the MDP problem for the case with penalty, i.e., with the objective represented by Equation (6). The initial probability distribution corresponds to the steady state of the “target” MC with the transition probabilities shown in Figure 1. Eight curves show optimal dynamics of the respective, $\rho_\alpha(t)$, $\alpha = 1, \dots, 8$ for the setting, equivalent to the one used in Figure 2.

exogenous signal. Tracking the signal may require some (or all) participants of the ensemble to sacrifice their natural cycling behavior. In this section we aim to study whether a perfect tracking of a predefined (exogenous) signal is feasible and then, in the case of feasibility, we would like to find the optimal solution causing least discomfort to the ensemble. Putting it formally, the aggregator solves the following optimization problem

$$\min_p \mathbb{E}_\rho \left[\underbrace{\sum_{t=0}^{T-1} \sum_{\alpha, \beta} \gamma_{\alpha\beta}(t) \log \frac{p_{\alpha\beta}(t)}{\bar{p}_{\alpha\beta}}}_{\text{welfare penalty}} \right] \quad (7)$$

s.t. Equations (1)

$$s(t) = \sum_{\alpha} \varepsilon_{\alpha} \rho_{\alpha}(t), \quad \forall \alpha, \quad \forall t = 1, \dots, T \quad (8)$$

$\underbrace{\qquad\qquad\qquad}_{\text{energy tracking constraint}}$

where \bar{p} is the “target” distribution represented by a stochastic matrix, e.g., one that leads to the standard steady state when $U = 0$. p is the stochastic matrix which is

constrained by Equation (1). $s(t)$ is the amount of energy requested by the system operator to balance the (transmission level) grid, and ε_α is the amount of energy consumed by a device when it stays in the state α for a unit time slot. The setting of Equation (8) assumes perfect tracking, that is, the total consumption of the ensemble is exactly equal to the amount requested by the SO.

Introducing the Lagrangian multiplier for the energy tracking constraint, one restates Equation (7) as the following min-max (max-min according to normal Lagrangian formulation) optimization:

$$\max_{\xi} \min_{p, \rho} \left(\mathbb{E}_p \left[\sum_{t=0}^{T-1} \sum_{\alpha, \beta} \gamma_{\alpha \beta}(t) \log \frac{p_{\alpha \beta}(t)}{\bar{p}_{\alpha \beta}} + \sum_{t=1}^T \sum_{\alpha} \xi(t) \varepsilon_{\alpha} \right] - \sum_{t=1}^T \xi(t) s(t) \right) \quad , \quad (9)$$

We observe that optimization over p and ρ in the resulting expression becomes equivalent under a substitution

$$U_\alpha(t) = \xi(t) \varepsilon_\alpha, \quad \forall \alpha, \quad t = 1, \dots, T. \quad (10)$$

to the penalized KL welfare penalty optimization (6), discussed in Section 3.

In spite of the close relation between the energy tracking problem Equation (8) and the profit optimality problem Equation (6), the former is more difficult to implement. Indeed one can solve Equation (6) in only one backward-forward run. On the other hand, we are not aware of the existence of a similar efficient algorithm for solving Equation (8). The difficulty is related to the fact that $\xi(t)$ itself should be derived as the result of a KKT condition that reinforces the energy tracking constraint Equation (8). A natural resolution of this problem is through an outer-loop iteration including the following two substeps:

- Run the backward-forward penalty optimization algorithm described at the end of Section 3 using the current $\xi(t)$ (outer-step-specific) profile.
- Update current $\xi(t)$ according to Equation (8), utilizing the current $\rho(t)$ derived from the previous substep.

The outer-loop iterative scheme is initiated with $\xi(t)$ derived from Equation (8), with $\rho(t)$ corresponding to the “normal” MP, $p \rightarrow \bar{p}$. Iterations are run until a preset tolerance is achieved. We plan to experiment with and analyze convergence of this iterative scheme in the future.

6 Hybrid Modeling: Toward Voltage-Aware and Hierarchical Ensemble Control

In this section we describe one possible scheme of an MDP approach for integration into control of power distribution networks. The description here is preliminary and meant to motivate a formulation for further exploration.

The aggregation of many loads discussed in the manuscript so far has ignored details of power flows (PF) as well as related voltage and line flow constraints.

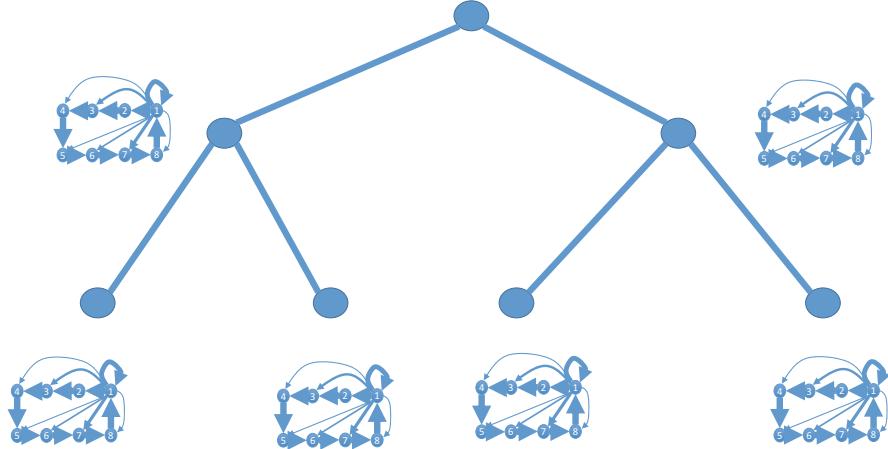


Fig. 4 Scheme illustrating hybrid/integrated and power-grid-aware MDP formulation. Nodes of the distribution level power system are each characterized by a sub-ensemble modeled as an MP. Nodes are connected in a tree-like power system.

We have assumed that any of the discussed solutions is PF-feasible, i.e. solution of PF equations is realizable for any of the consumption configurations with voltages and line flows staying within the prescribed bounds. For these assumptions to hold, either the aggregated devices should be in immediate geographical proximity of one another or the power system should be operated with a significant safety margin. In general, satisfying either of the two conditions globally for a large ensemble is impractical. This motivates a discussion of the following hybrid model, stated in terms of many geographically localized and different sub-ensembles connected into a power distribution network (see also Figure 4 for illustration):

$$\min_{p, \rho, s, V} \mathbb{E}_\rho \left[\sum_{t=0}^{T-1} \sum_i \sum_\alpha \left(U_\alpha^{(i)}(t+1) + \sum_\beta \gamma_{\alpha\beta}^{(i)}(t) \log \frac{p_{\alpha\beta}^{(i)}(t)}{\bar{p}_{\alpha\beta}^{(i)}} \right) \right] \quad (11)$$

s.t.

$$\rho_\alpha^{(i)}(t+1) = \sum_\beta p_{\alpha\beta}^{(i)}(t) \rho_\beta^{(i)}(t), \quad \forall t, \quad \forall i, \quad \forall \alpha \quad (12)$$

$$\underbrace{s^{(i)}(t) \doteq \sum_\alpha \varepsilon_\alpha^{(i)} \rho_\alpha^{(i)}(t) = V^{(i)}(t) \sum_{j \sim i} \left(\frac{V^{(i)}(t) - V^{(j)}(t)}{z_{ij}} \right)^*}_{\text{Nodal PF relations}}, \quad \forall i, \forall t, \quad (13)$$

$$\underbrace{v^{(i)} \leq |V^{(i)}|(t) \leq \bar{v}^{(i)}}_{\text{voltage constraints}}, \quad \forall i, \quad \forall t. \quad (14)$$

where each sub-ensemble i , representing, for example, stochastic/fluctuating consumption of a large apartment complex, is modeled as an aggregated MP that includes devices and consumers of different types. The objective (11) generalizes Equation (6), thus accounting for the cost of services and welfare quality (deviation from normal) for multiple ensembles. As done earlier in the manuscript, we introduce the additional penalty factor, $\gamma_{\alpha\beta}^{(i)}(t)$, in front of the KL (welfare penalty) term. This allows us to reflect the significance of different ensembles, transitions, and times. Equation (12) generalizes the ME (2) to sub-ensembles. Equations (13, 14) introduce PF and voltage constraints into the MDP setting. We thus seek for an optimum (11) over three vectors: the vector $p = \left(p_{\alpha\beta}^{(i)}(t) \mid \forall i, \forall \alpha, \beta, \forall t\right)$, constructed from the stochastic transition-probability matrices, where each component represents a node in the network at all moments of time; the vector $\rho = \left(\rho_{\alpha}^{(i)}(t) \mid \forall i, \forall \alpha, \forall t\right)$, reconstructed from p according to Equations (12) with the proper initial conditions provided; and the vector of voltages $V = \left(V^{(i)}(t) \mid \forall i, \forall t\right)$.

Notice that other objectives (such as a contribution enforcing minimization of power losses in the distribution systems), other constraints (such as imposing bounds on line flows), as well as other controls (such as voltage/ position of tap-changers and/or reactive consumption at the nodes containing inverters) can be incorporated into the scheme in the spirit of [34, 37].

Because the MDP problem is stochastic by nature, it also allows incorporation of other stochastic sources, e.g., solar or wind renewables, which can be done either via modeling the stochastic sources as MP (with no control) or by extending the model by adding so-called chance-constrained descriptions in the spirit of [3].

An efficient solution of the hybrid problem (11) can be built by combining the techniques of temporal DP, developed in this manuscript for individual MDPs, with spatial (tree-graph) DP techniques, developed recently for the Optimal PF (OPF) in power distributions [17] and taking advantage of the tree-graph operational layout of power distribution networks. We plan to work on practical implementation of these and other components of the hybrid model in the future.

7 Conclusions and Path Forward

In this manuscript we review and develop a computationally scalable approach for optimization of an ensemble of devices modeled via finite-space, finite-time MP. This approach builds upon earlier publications [19, 28, 29] addressing DR applications in power systems and beyond. A particularly practical and popular example of MP, relevant for power systems, is the ensemble of cycling loads, such as air conditioners, water heaters, or residential water pumps [7, 32].

We have developed a number of useful MDP formulations aiming to achieve optimality for a diverse set of objectives of interest for an aggregator (of the ensemble) under different circumstances. We started the manuscript by describing MDP that balances overall expenses of the ensemble acquired when the cost of electricity varies in time, with a welfare penalty that measures ensemble operational

deviation from its normal behavior. Then we proceeded to discuss MDP built to test feasibility of the ensemble to provide ancillary frequency control services to SO. Finally, we addressed the future challenge of building a hybrid model that incorporates MP and MDP modeling into deterministic and stochastic frameworks of OPFs (operational dispatch) for power distribution.

The most important technical achievement of this paper is the development of an appropriate DP framework for stochastic optimization including DR. For the cost-vs-welfare optimization we also tested the developed methodology and algorithms numerically (on a small-scale example).

This manuscript does not offer a concluding fine-tuned summary of a completed project. Instead, we have focused here on presenting a new open-ended paradigm. In other words, we expect to see emergence of many more future extensions and generalizations of the approach that we have started to develop here. Some of these proposed future developments have been already discussed in the preceding sections, especially in Section 6. Others are briefly highlighted below.

- *Utilizing and Extending Lin-Solvability.* Intuitively, it is clear that lin-solvability is a rich property that should be advantageous for both analysis and algorithms. However, the use of this strong property in this manuscript was rather limited. We expect that the lin-solvability will provide an actual computational/algorithmic benefit not just for solving MDP per se but also for solving more-complex multilevel optimization/control problems where an individual MDP represents an element of a richer model. (Some examples are mentioned below.) On the other hand, the lin-solvability described in Section 8.1.1 is a rather delicate property of the model and it is easy to lose either partially or completely, as shown for the setting/formulation discussed in Section 8.1 and the main part of Section 8, respectively. This observation motivates further investigation of other settings/formulations amenable for either full or partial lin-solvability. One promising direction for future analysis is to consider a generalization from the KL-divergence to Renyi-divergence, building upon and extending the approach of [15, 16].
- *Model Reduction.* Many models of power systems are too large and detailed for computations. This applies even to the routine PF computation, which is a routing subtask for many power system problems of high level involving optimization, control, and generalizations accounting for stochasticity and robustness. Seeking to represent large-scale, long-time behavior, one is interested in building a reduced model in order to aggregate the small-scale and/or short-time details in a compact way. The reduction may be lossless or lossy, with or without the ability to reconstruct the small-scale/short-time details. The intrinsically stochastic MP framework developed here is appropriate for the lossy case, where stochasticity represents the loss of insignificant details. We envision building reduced models capable of more efficient but still accurate computations of, e.g., PF, in the format similar to that discussed in Section 6. The reduced model may consist of power lines with effective impedances and stochastic load/generation elements represented by MPs. The level of coarse-graining may be predefined by

a geography-preserving procedure, e.g., of the type discussed in [20], but it may also be left flexible/adaptive, where the number of states and allowable transitions for an individual MP is the subject of optimization.

- *Hierarchical Control.* When the reduced model with MPs representing aggregated loads/generators is to be used in the context of optimization and control, changing from MP to MDP may be understood as allowing the optimization/control to be split into two levels. Optimization with many MDPs modeling aggregated end-users, as discussed in Section 6, will produce optimal transition probability matrices for each individual MDP. Then, the task of implementing this policy is delegated to aggregators responsible for individual MDPs. In this case, implementation means the global optimality/control is substituted by a two-level hierarchical control. The scheme may be extended to represent more levels of control.
- *Derivation of MDP from a Detailed Microscopic Model.* Consider an ensemble of continuous-time TCLs, each represented by a microscopic state in terms of temperature (continuous variable) and switch on/off status (discrete variable). In the case of an inhomogeneous ensemble, where each TCL or a group of TCLs may be parameterized differently, e.g., in terms of the allowed temperature rates and/or transition rates between on/off states, one is posing the question of representing this inhomogeneous ensemble as a single discrete time and discrete space (binned) MP or MDP. Developing a methodology for constructing a representative MP/MDP for an aggregated ensemble is an important task for future research.
- *Accounting for Risk-Metric.* The MP/MDP methodology is sufficiently rich and flexible to account for and mitigate risks of different types, e.g., overloading. Indeed, incorporating additional constraints into MDP optimization by limiting some elements of ρ or p is a straightforward way of limiting the risk in probabilistic terms, natural for MDP.
- *Stochastic MDP.* MDP models already account for the intrinsic stochasticity of an ensemble of devices (or coarse-grained areas) that a model represents. However, exogenous effects, such as those representing the cost of electricity or the frequency signal, are modeled in this manuscript deterministically, even though it is often more appropriate to represent generically uncertain and stochastic exogenous signals probabilistically. Formulating appropriate “second-order” statistical models represents an interesting challenge for future research. Interesting recent studies dealing with exogenous fluctuations and uncertainty added to the MP/MDP setting are presented in [4, 5].
- *From MDP to Reinforcement Learning (RL).* MDP is a standard tool used in the field of RL. One of the data-driven RL approaches [36] relevant for our discussion suggests considering “ideal” transition probabilities, \bar{p} , as unknown/uncertain and then attempting to learn \bar{p} from the data in parallel with solving the MDP. The approach was coined in [36] as Z-learning, and it is also closely related to the so-called approximate DP (see [30] and the references therein for many books and reviews). We anticipate that this general approach, when applied to the models introduced in the manuscript, will allow

us to build a data-driven and MDP-based framework for controlling an ensemble whose normal behavior is known only through limited samples of representative behavior.

- *MDP for Supervised Learning (SL).* Recently a SL methodology was applied as a real-time proxy to solve difficult power system problems, such as finding an efficient description of the feasibility domain to solve OPF or power system reliability management problems [14]. The main idea in this line of research rests on replacing an expensive power system computation with an ML black box trained to evaluate a sufficient number of samples labeled by relevant output characteristics. One interesting option consists of using properly designed MPs and/or MDPs for opening up the black box and turning it into (at least partially) a physics-informed ML. Another possible direction would be to use MPs/MDPs as labels.
- *MDP as an Element of a Graphical Model (GM) framework.* It was argued recently in [10, 17] that the approach of GM offers a flexible framework and efficient solutions/algorithms for a variety of optimization and control problems in energy systems (power systems and beyond). It is of interest to extend the GM framework and build into it MDP methodology, formulations, and solutions.
- *Applications to Systems, e.g., Energy Systems.* The MDP models discussed in this manuscript are of interest well beyond representing ensembles (and coarse regions) in power systems. Similar methods and approaches are relevant for problems representing behavior and DR capabilities of consumers'/producers' ensembles in other energy infrastructures, such as natural gas systems and district heating systems. In fact, the models discussed above fit practically “as is” to describe aggregation of many consumers of district heating systems residing in a big apartment complex or a densely populated residential area. The language of MDP is also universal enough to optimize joint energy consumption through multiple energy infrastructures of a “residential” ensemble (including electricity, heat, and gas—possibly used as an alternative to central heating in local boilers).

8 Dynamic/Bellman Programming

In this section we discuss the DP solution for the most general of the formulations considered in this paper, the profit-vs-welfare optimal formulations. Specifically we consider Equation (6) that introduces in-homogeneity in the inter-state transitions, expressed through the state and time-dependent $\gamma_{\beta\alpha}(\tau)$ factors.

Let us restate Equation (6) as

$$\min_p C \left(p|_0^{T-1}, \rho(0) \right) \Big|_{\sum_\beta p_{\beta\alpha}(\tau)=1, \forall \alpha}, \quad (15)$$

where $p|_0^{T-1}$ is the shortcut notation for the vector constructed of the transition probability matrices evaluated at $t = 0, \dots, T - 1$, i.e., $(p(t)| \forall t = 0, \dots, T - 1)$. The so-called value function in Equation (15) can be decomposed according to

$$C(p|_0^{T-1}, \rho(0)) = C(p|_0^{\tau-1}, \rho(0)) + \sum_{\alpha} \varphi_{\alpha}(\tau, p|_{\tau}^{T-1}) \rho_{\alpha}(\tau, p|_0^{\tau-1}), \quad (16)$$

where the notation $\varphi_{\alpha}(\tau, p|_0^{\tau-1})$ is introduced temporarily (just for the purpose of this derivation) to emphasize that ρ_{α} computed at the moment of time τ also depends, according to Equation (2), on the transition probability matrices computed at all the preceding times. Here in Equation (16) φ_{α} is defined at the final moment of time according to

$$\varphi_{\alpha}(T) = U_{\alpha}(T), \quad (17)$$

and then solved backward in time by the following recursive equations

$$\begin{aligned} \forall \alpha, \quad \tau = T - 1, \dots, 0 : \quad \varphi_{\alpha}(\tau, p|_{\tau}^{T-1}) &= \sum_{\beta} \varphi_{\beta}(\tau + 1, p|_{\tau+1}^{T-1}) p_{\beta\alpha}(\tau) \\ &+ \sum_{\beta} \gamma_{\beta\alpha}(\tau) p_{\beta\alpha}(\tau) \log \left(\frac{p_{\beta\alpha}(\tau)}{\bar{p}_{\beta\alpha}} \right) + U_{\alpha}(\tau), \end{aligned} \quad (18)$$

where, the notation $\varphi_{\alpha}(\tau, p|_{\tau}^{T-1})$ emphasizes that (by construction) $\varphi_{\alpha}(\tau)$ depends only on $p|_{\tau}^{T-1}$.

The DP-decomposed (recursive) structure of Equations (16, 18) allows us to evaluate optimization over p in Equation (15) greedily as Karush–Kuhn–Tucker (KKT) first-order conditions—first over $p(T - 1)$ and then over $p(T - 2)$ all the way to $p(0)$. The KKT condition for minimization of (18) with linear constraint $\sum_{\beta} p_{\beta\alpha}(\tau) = 1$ gives the following DP relation for the optimal p :

$$p_{\beta\alpha}(\tau) = \bar{p}_{\beta\alpha} \exp \left(-1 - \frac{\varphi_{\beta}(\tau + 1) - \lambda_{\alpha}(\tau)}{\gamma_{\beta\alpha}(\tau)} \right), \quad \forall \alpha, \beta, \quad \forall \tau = T - 1, \dots, 0, \quad (19)$$

where the Lagrange multipliers, $\lambda_{\alpha}(\tau)$, are determined implicitly from the stochasticity of $p(\tau)$

$$\sum_{\beta} \bar{p}_{\beta\alpha} \exp \left(-1 - \frac{\varphi_{\beta}(\tau + 1) - \lambda_{\alpha}(\tau)}{\gamma_{\beta\alpha}(\tau)} \right) = 1, \quad \forall \alpha, \quad \forall \tau = T - 1, \dots, 0. \quad (20)$$

The Lagrange multipliers λ can also be extracted from one-dimensional convex optimizations one gets substituting (19) into the corresponding Lagrangian relaxation of (18)

$$\begin{aligned} \lambda_{\alpha}(\tau) &= \arg \min_{\mu} \left(\sum_{\beta} \bar{p}_{\beta\alpha} \gamma_{\beta\alpha}(\tau) \exp \left(-1 - \frac{\varphi_{\beta}(\tau + 1) - \mu}{\gamma_{\beta\alpha}(\tau)} \right) - \mu \right), \quad (21) \\ \forall \alpha, \quad \forall \tau &= T - 1, \dots, 0. \end{aligned}$$

Computationally, one may solve Equation (21) via gradient descent

$$\lambda_\alpha(\tau) \leftarrow \lambda_\alpha(\tau) - \delta \left(\sum_\beta \bar{p}_{\beta\alpha} \exp \left(-1 - \frac{\varphi_\beta(\tau+1) - \lambda_\alpha(\tau)}{\gamma_{\beta\alpha}(\tau)} \right) - 1 \right), \quad (22)$$

$$\forall \alpha, \forall \tau = T-1, \dots, 0.$$

choosing an appropriate step, δ , and iterating Equation (22) until the target tolerance of the solution's accuracy is reached. Then Equations (19, 18) can be used to determine $p(\tau)$ and $\varphi(\tau)$.

Another way to determine $p(\tau), \varphi(\tau)$ is to minimize (18) directly resolving the p -stochasticity constraint. As $x \log x$ is a convex function, minimizing (18) is a convex problem that can be solved with standard solvers, e.g. Ipopt or cvx.

8.1 The Case Which Allows Explicit Normalization

In the case of a general $\gamma(\tau)$ in Equation (19), the Lagrange multipliers, $\lambda(t)$, cannot be expressed via p in a closed form. An exception is the case when $\gamma_{\beta\alpha}(\tau)$ is independent of β , i.e.

$$\gamma_{\beta\alpha}(\tau) \Rightarrow \gamma_\alpha(\tau). \quad (23)$$

Then, Equation (19) results in

$$p_{\beta\alpha}(\tau) = \frac{\exp \left(-\frac{\varphi_\beta(\tau+1)}{\gamma_\alpha(\tau)} \right) \bar{p}_{\beta\alpha}}{\sum_v \exp \left(-\frac{\varphi_v(\tau+1)}{\gamma_\alpha(\tau)} \right) \bar{p}_{v\alpha}}, \quad \forall \alpha, \beta, \quad \forall \tau. \quad (24)$$

Substituting Equation (24) into Equation (18), one derives

$$\varphi_\alpha(\tau) = -\gamma_\alpha(\tau) \log \left(\sum_\beta \exp \left(-\frac{\varphi_\beta(\tau+1)}{\gamma_\alpha(\tau)} \right) \bar{p}_{\beta\alpha} \right) + U_\alpha(\tau), \quad \forall \alpha, \quad \forall \tau. \quad (25)$$

Notice that there are also other special cases that may allow for analytic expressions for the normalization. In particular, the case when $\gamma_{\alpha\beta}(\tau), \forall \tau, \alpha, \beta$ take values from a finite alphabet. We leave discussion of this and other interesting cases to future studies.

8.1.1 Linearly Solvable Case

Reduction (23) allows us to map Equation (15) to solution of DP equations (25) that are, however, nonlinear. We now make an additional reduction to limit Equation (23) even further to the state independent case

$$\gamma_{\alpha\beta}(\tau) \Rightarrow \gamma(\tau), \quad (26)$$

Equation (25) now reduces to what was coined in [15, 16, 36] as the linearly solvable MDP

$$\exp\left(-\frac{\varphi_\alpha(\tau)}{\gamma(\tau)}\right) \doteq u_\alpha(\tau) = \sum_\beta u_\beta(\tau+1) \bar{p}_{\beta\alpha} \exp\left(-\frac{U_\alpha(\tau)}{\gamma(\tau)}\right). \quad (27)$$

Equation (27) is solved backward in time with the final condition

$$u_\alpha(T) = \exp\left(-\frac{U_\alpha(T)}{\gamma(T)}\right), \quad \forall \alpha. \quad (28)$$

Two remarks are in order. First, we note that Equation (26) is of practical interest when one aims to change the relative importance of the welfare reinforcement vs. price balance in the optimization. Second, other linearly solvable cases, in addition to those described by Equation (26), may exist. We postpone a more general discussion of a broader class of the linearly solvable cases as well as their practical utility, to future publications.

Acknowledgements The authors are grateful to S. Backhaus, I. Hiskens, and D. Calloway for fruitful discussions, guidance, and valuable comments. The work at LANL was carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy under Contract No. DE-AC52-06NA25396.

References

1. Aunedi M, Kountouriotis PA, Calderon JEO, Angelis D, Strbac G (2013) Economic and environmental benefits of dynamic demand in providing frequency regulation. *IEEE Trans Smart Grid* 4(4):2036–2048. <https://doi.org/10.1109/TSG.2013.2258047>
2. Bashash S, Fathy HK (2011) Modeling and control insights into demand-side energy management through setpoint control of thermostatic loads. In: Proceedings of the 2011 American control conference, pp 4546–4553. <https://doi.org/10.1109/ACC.2011.5990939>
3. Bienstock D, Chertkov M, Harnett S (2014) Chance-constrained optimal power flow: risk-aware network control under uncertainty. *SIAM Rev* 56(3):461–495. <https://doi.org/10.1137/130910312>
4. Bušić A, Meyn S (2016) Distributed randomized control for demand dispatch. ArXiv e-prints
5. Bušić A, Meyn S (2016) Ordinary differential equation methods for Markov decision processes and application to Kullback-Leibler control cost. ArXiv e-prints
6. Callaway DS (2009) Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. *Energy Convers Manag* 50(5):1389–1400. <https://doi.org/10.1016/j.enconman.2008.12.012>. <http://www.sciencedirect.com/science/article/pii/S0196890408004780>
7. Callaway D, Hiskens I (2011) Achieving controllability of electric loads. *Proc IEEE* 99(1):184–199. <https://doi.org/10.1109/JPROC.2010.2081652>
8. Canyasse R, Dalal G, Mannor S (2016) Supervised learning for optimal power flow as a real-time proxy. ArXiv e-prints
9. Chertkov M, Chernyak V (2017) Ensemble of thermostatically controlled loads: statistical physics approach. *Sci Rep* 17:8673

10. Chertkov M, Krishnamurthy D, Misra S, Hentenryck PV, Vuffray M (2017) Graphical models and belief propagation-hierarchy for optimal physics-constrained network flows (2017). arXiv:1702.01890
11. Chong CY, Debs AS (1979) Statistical synthesis of power system functional load models. In: 18th IEEE conference on decision and control including the symposium on adaptive processes, vol 2, pp 264–269. <https://doi.org/10.1109/CDC.1979.270177>
12. Chong CY, Malhami RP (1984) Statistical synthesis of physically based load models with applications to cold load pickup. IEEE Trans Power Apparatus Syst PAS-103(7):1621–1628. <https://doi.org/10.1109/TPAS.1984.318643>
13. Donnelly M, Harvey D, Munson R, Trudnowski D (2010) Frequency and stability control using decentralized intelligent loads: benefits and pitfalls. In: IEEE PES general meeting, pp 1–6. <https://doi.org/10.1109/PES.2010.5589835>
14. Duchesne L (2016) Machine learning of proxies for power systems reliability management. Master's thesis, Faculty of Applied Sciences, Department of Electrical Engineering and Computer Science, University of Liege, Belgium. https://matheo.ulg.ac.be/bitstream/2268.2/1374/4/master_thesis_laurine_duchesne.pdf
15. Dvijotham K, Todorov E (2012) A unifying framework for linearly solvable control. arxiv:1202.3715
16. Dvijotham K, Todorov E (2013) Linearly solvable optimal control. Wiley, New York, pp 119–141. <https://doi.org/10.1002/9781118453988.ch6>
17. Dvijotham K, Chertkov M, Van Hentenryck P, Vuffray M, Misra S (2016) Graphical models for optimal power flow. Constraints 1–26. <https://doi.org/10.1007/s10601-016-9253-y>
18. Fleming WH, Mitter SK (1982) Optimal control and nonlinear filtering for nondegenerate diffusion processes. Stochastics 8(1):63–77. <https://doi.org/10.1080/17442508208833228>
19. Ghaffari A, Moura S, Krstic M (2015) Pde-based modeling, control, and stability analysis of heterogeneous thermostatically controlled load populations. J Dyn Syst Meas Control 137(10):101001
20. Grudzien C, Deka D, Chertkov M, Backhaus S (2017) Structure- & physics-preserving reductions of power grid models. arXiv:1707.03672
21. Ihara S, Schweppe F (1981) Physically based modeling of cold load pickup. IEEE Trans Power Syst PAS-100(9):4142–4150. <https://doi.org/10.1109/TPAS.1981.316965>
22. Kappen HJ (2005) Linear theory for control of nonlinear stochastic systems. Phys Rev Lett 95:200201. <https://doi.org/10.1103/PhysRevLett.95.200201>
23. Kontoyiannis I, Meyn S (2005) Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. Electron J Probab 10(3):61–123. <https://doi.org/10.1214/EJP.v10-231>. <http://ejp.ejpecp.org/article/view/231>
24. Lu N, Chassin D (2004) A state-queueing model of thermostatically controlled appliances. IEEE Trans Power Syst 19(3):1666–1673. <https://doi.org/10.1109/TPWRS.2004.831700>
25. Lu N, Chassin D, Widergren S (2005) Modeling uncertainties in aggregated thermostatically controlled loads using a state queueing model. IEEE Trans Power Syst 20(2):725–733. <https://doi.org/10.1109/TPWRS.2005.846072>
26. Malhame R, Chong CY (1985) Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system. IEEE Trans Autom Control 30(9):854–860. <https://doi.org/10.1109/TAC.1985.1104071>
27. Malhame R, Chong CY (1988) On the statistical properties of a cyclic diffusion process arising in the modeling of thermostat-controlled electric power system loads. SIAM J Appl Math 48(2):465–480. <https://doi.org/10.1137/0148026>
28. Meyn S, Barooah P, Basic A, Chen Y, Ehren J (2015) Ancillary service to the grid using intelligent deferrable loads. IEEE Trans Autom Control 60(11):2847–2862. <https://doi.org/10.1109/TAC.2015.2414772>
29. Paccagnan D, Kamgarpour M, Lygeros J (2015) On the range of feasible power trajectories for a population of thermostatically controlled loads. In: 2015 54th IEEE conference on decision and control (CDC), pp 5883–5888. <https://doi.org/10.1109/CDC.2015.7403144>
30. Reinforcement learning. https://en.wikipedia.org/wiki/reinforcement_learning. Accessed 07 Jan 2017

31. Schwepp FC, Tabors RD, Kirtley JL, Outhred HR, Pickel FH, Cox AJ (1980) Homeostatic utility control. *IEEE Trans Power Apparatus Syst PAS-99(3):1151–1163*. <https://doi.org/10.1109/TPAS.1980.319745>
32. Siano P (2014) Demand response and smart grids—a survey. *Renew Sust Energ Rev* 30:461–478. <https://doi.org/http://dx.doi.org/10.1016/j.rser.2013.10.022>. <http://www.sciencedirect.com/science/article/pii/S1364032113007211>
33. Study on the demand response potential for seawater desalination projects (2016). http://www.ercot.com/content/wcm/lists/89476/Demand_Response_Potential_for_Seawater_Desalination_Projects_11_18_2016.pdf
34. Sulc P, Backhaus S, Chertkov M (2014) Optimal distributed control of reactive power via the alternating direction method of multipliers. *IEEE Trans Energy Convers* 29. <https://arxiv.org/abs/1310.5748>
35. Todd D, Caufield M, Helms B, Starke M, Kirby B, Kueck J (2009) Providing reliability services through demand response: a preliminary evaluation of the demand response capabilities of Alcoa Inc. <https://www.ferc.gov/eventcalendar/Files/20100526085850-ALCOA%20Study.pdf>
36. Todorov E (2007) Linearly-solvable Markov decision problems. In: Schölkopf B, Platt J, Hoffman T (eds) *Advances in neural information processing systems*, vol 19. MIT Press, Cambridge, pp 1369–1376. <http://papers.nips.cc/paper/3002-linearly-solvable-markov-decision-problems.pdf>
37. Turitsyn K, Sulc P, Backhaus S, Chertkov M (2011) Options for control of reactive power by distributed photovoltaic generators. *Proc IEEE* 99(6):1063–1073
38. Zheng Y, Hill DJ, Zhang C, Meng K (2016) Non-interruptive thermostatically controlled load for primary frequency support. In: 2016 IEEE power and energy society general meeting (PESGM), pp 1–5. <https://doi.org/10.1109/PESGM.2016.7741469>

Distributed Control Design for Balancing the Grid Using Flexible Loads



Yue Chen, Md Umar Hashmi, Joel Mathias, Ana Bušić, and Sean Meyn

Abstract Inexpensive energy from the wind and the sun comes with unwanted volatility, such as ramps with the setting sun or a gust of wind. Controllable generators manage supply-demand balance of power today, but this is becoming increasingly costly with increasing penetration of renewable energy. It has been argued since the 1980s that consumers should be put in the loop: “demand response” will help to create needed supply-demand balance. However, consumers use power for a reason and expect that the quality of service (QoS) they receive will lie within reasonable bounds. Moreover, the behavior of some consumers is unpredictable, while the grid operator requires predictable controllable resources to maintain reliability. The goal of this chapter is to describe an emerging science for *demand dispatch* that will create *virtual energy storage* from flexible loads. By design, the grid-level services from flexible loads will be as controllable and predictable as a generator or fleet of batteries. Strict bounds on QoS will be maintained in all cases. The potential economic impact of these new resources is enormous. California plans to spend billions of dollars on batteries that will provide only a small fraction of the balancing services that can be obtained using demand dispatch. The potential impact on society is enormous: a sustainable energy future is possible with the right mix of infrastructure and control systems.

Y. Chen

Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, CO, USA

J. Mathias · S. Meyn

Department of ECE, University of Florida, Gainesville, FL, USA

M. U. Hashmi · A. Bušić (✉)

Inria, Paris, France

DI ENS, École normale supérieure, CNRS, PSL Research University, Paris, France

e-mail: ana.busic@inria.fr

1 Introduction

As more wind and solar energy come online, the system operators who run the power grid are faced with a problem: how do they compensate for the variable nature of the sun and wind?

Low-frequency variability from solar gives rise to the famous “duck curve” anticipated at CAISO [8]: it is predicted that ramps of 30% of the load over a few hours may be commonplace. In 2011, they introduced new market rules for pricing *flexible ramping products* to help combat this volatility. Figure 1 is taken from a 2014 presentation at the Southwest Power Pool (SPP) working group meeting.¹ The net load on March 8, 2014, shows the emergence of the CAISO duck curve. The price data illustrates that insufficient ramping resources can cause enormous spikes in wholesale power prices.

MISO (an ISO in central North America) recently alerted FERC to the need for new market rules to incentivize ramping products. They argue that the need for these services is increasing with the introduction of renewable energy: “*under its current market structure, short-term Net Load variations could create a situation where dispatchable resources have sufficient capacity, but there is a short-term scarcity event because MISO has inadequate ramp capability to respond to unexpected variations in Net Load . . . such ramp capability shortages could result in . . . dispatch intervals during which the price of energy can increase significantly due to scarcity pricing, even if the event does not present a significant reliability risk*”². MISO is not concerned about energy: they are lacking responsive resources, even though there is sufficient capacity.

The control systems diagram in Figure 2 provides a simple view of how the grid is operated today, in which wind and solar are viewed as sources of disturbances. In North America, the **GRID** block is in fact a subset of the grid called a *balancing*

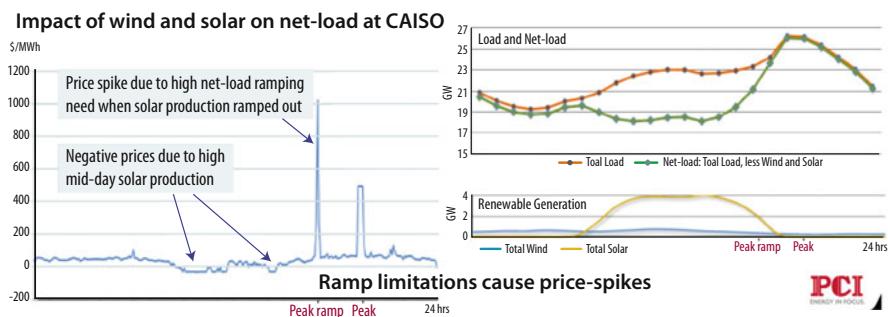


Fig. 1 Interaction between power and price dynamics at CAISO.

¹From Tony Delacluyse of PCI, with permission.

²<http://tinyurl.com/FERC-ER14-2156-000>

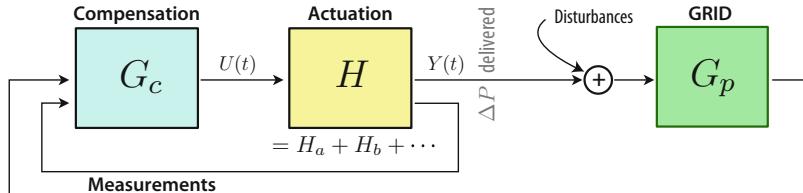


Fig. 2 Power grid control loop. A question addressed in this chapter: *where do we find H ?*.

region. The block denoted **Compensation** represents a *balancing authority* (BA). The grid-level measurements obtained by the BA are summarized as a scalar function of time called the area control error (ACE). It is a linear combination of two error signals: the deviation of local grid frequency from the nominal 60 Hz and the tie-line error – defined as the mismatch between scheduled and actual flow of power out of the balancing region. Command signals are broadcast to resources such as controllable generators so that the ACE signal is kept within desired bounds.

The compensator G_c is designed by the BA in a particular region. For example, PJM (an RTO in the eastern USA) creates their RegA and RegD signals by passing the ACE signal first through a PI compensator and then through a band-pass filter. In this case, the compensator G_c in Figure 2 is taken to be a PI controller, and the band-pass filters are embedded in the block denoted **actuation**. The decomposition “ $H = H_a + H_b + \dots$ ” represents many resources acting in parallel to provide actuation, which may include controllable generation and batteries.

It is anticipated that the basic architecture illustrated in Figure 2 will remain in place for many decades to come. The grid will become more adaptable to persistent disturbances or crisis through a combination of control techniques and hardware.

The term *ancillary service* refers to resources required to maintain supply-demand balance in the grid but do not necessarily supply energy. While controllable generation is the most common source of most ancillary services today, other technologies such as flywheels and batteries are increasingly popular because of their performance and because of new state and federal policies. The FERC report [37, pp. 23–24] contains a survey of experiments conducted by Beacon Power and Primus Power on the value of highly responsive resources for ancillary service. Primus claims approximately 76 percent more ACE correction compared with what can be obtained from generation sources. This is because a slower “... resource lags to the point of working against needed ACE correction.” In addition to their poorer performance, the use of generation for ancillary service may be costly in terms of fuel and emissions because the generators are not running at their ideal conditions and in fact may be online only to help regulate the grid.

The California Public Utilities Commission has encouraged utilities to invest billions of dollars on batteries, and many have responded. Southern California Edison (SCE) recently announced the “largest battery in the world.” It is capable of storing 32 MWhs (megawatt hours) of energy. It is not a single battery but a 6,300-square-foot facility that houses an enormous array of lithium-ion batteries.

1.1 Chapter Overview

Batteries may be a clean source of ancillary service, but currently they are an expensive solution. In addition to the large space required for large systems, batteries have finite life time and waste energy as they are charged and discharged to service the grid [17].

Distributed control architectures are described in this chapter to create *virtual energy storage* (VES) based on the inherent flexibility in power consumption from a majority of loads. The ancillary services that can be obtained include regulation (such as automatic generation control, or AGC), smooth peaks in load, address ramps from wind or solar generation, and help to recover gracefully from contingencies such as generation faults. It is believed that VES is a low-cost complement to batteries and power plants and may in the future provide the majority of required ancillary services.

The term *demand dispatch* is a convenient alternative to *demand response*; the latter is defined by policy makers and regulatory bodies (such as FERC) as load-shedding in exchange for some monetary reward. *Load-shedding is not the goal of the technology surveyed here.* In applications to both regulation and ramping services, the distributed control algorithms are designed so that power consumption is increased and decreased over time while keeping the total energy deviation over time at zero – *just like charging and discharging a battery.*

The control architectures described in this chapter are based on a series of papers on distributed control [11, 13, 35, 36]. The proposed frequency decomposition of VES resources was first introduced in [19, 20] in the context of commercial buildings and generalized in [1]. The key novel contribution in all of this work is the focus on “intelligence at the load,” manifested by local control loops. There are many benefits:

- (i) Communication infrastructure requirements are reduced, which hopefully leads to both improved security and higher consumer confidence regarding privacy.
- (ii) A simple control problem at the BA, since the single-input/single-output system is highly controllable.
- (iii) Strict bounds on quality of service (QoS) to the consumer are guaranteed.

This chapter does not consider market issues. It is assumed that consumer engagement will be achieved through contractual agreements and periodic credits, such as those offered by Florida Power & Light in their OnCall® program.

1.2 Some History and Further Reading

In the early 1980s, Schweppre wrote a series of influential articles on the value of new architectures for the grid [40, 41], with an emphasis on demand response based on either automation or prices. Tools for analysis were lacking at the time,

but many researchers came to fill the void. An influential example is the paper [29] that introduced ideas from statistical mechanics to model a large population of thermostatically controlled loads (TCLs).

There is substantial literature on indirect load control, where customers are encouraged to shift their electricity usage in response to real-time prices (several papers in this volume survey this literature, including the articles authored by Spence and by Moye). Dynamic prices can introduce uncertain dynamics, such as cyclical price fluctuations and increased sensitivity to exogenous disturbances, and present a risk to system stability [9, 15, 39].

Randomization is an essential element of the distributed control architecture described in this chapter. Its value has been widely recognized in academia as well as industry [42].

On the academic side, Matheiu's dissertation [32] and references [33, 34] were highly influential, motivating in part the research surveyed in this chapter and others [16, 26, 46, 48]. The control model in [32] is based on the mean-field setting of [29], with the introduction of a control signal from a central authority: at each time slot, a BA or aggregator broadcasts probability values $\{p_{\tau}^{\oplus}, p_{\tau}^{\ominus} : \tau \in \mathbb{R}\}$ where p_{τ}^{\oplus} (p_{τ}^{\ominus}) denotes the probability of turning the device on (off) when the temperature of the device is τ . The temperatures are binned to obtain a finite state-space aggregate model. This model is bilinear and partially observed, where the state x is the histogram of load temperature and power consumption. The bilinear control system is transformed to a linear model by defining products of probability and state as an input. The resulting linear state-space model has the same state, but the vector-valued input is now defined as products of the form $u_k = p_{\tau}^m x_j$ for some $\tau(k)$, $j(k)$, and $m(k) \in \{\oplus, \ominus\}$. Feedback control design is performed based on LQR. However, it is still necessary to recover the probability vector $\{p_{\tau}^m\}$. In this prior work, this is defined as the ratio of components of the input $u(t)$ and components of the *estimate* of the state at time t (see, e.g., eq. (11) of [34]). It is assumed that estimates are computed by the BA based on measurements of aggregate power consumption. A current challenge with this approach is the creation of sufficiently accurate state estimates for an inherently infinite-dimensional system. Challenges to state estimation are discussed in [13], where it is shown that the linearized mean-field model may not be observable. Robustness of this approach to bilinear control systems is another important area for future research.

The approach to distributed control surveyed in Sections 2 and 3 involves an entirely different approach to local control at each load. One example is the *individual perspective design* (IPD) described in Section 3.2. This can be regarded as an application of the MDP technique of Todorov [44], but only in one special case: the construction of [44] can be applied only if there is no exogenous stochastic disturbance in the load model. Contained in Section 3.2 are techniques to extend this design to a broader class of load models. These ideas were first applied to demand dispatch in [35] and have seen many extensions since. For more history the reader is referred to [14, 36], in addition to the papers surveyed in Section 3.2. While beyond the scope of this article, it is important to note that Todorov's "linearly solvable" MDP model [44] is similar to prior work such as [24], and the form of the solution could have been anticipated from well-known results in the theory of large

deviations for Markov chains [6]. It is pointed out in [45] that this approach has a long history in the context of controlled stochastic differential equations [18].

The remainder of the chapter consists of six sections organized as follows. Section 2 contains a high-level description of the control architecture, with details on distributed control contained in Section 3. The next three sections contain examples of distributed control of a large collection of resources in three settings: Section 4 concerns TCLs, Section 5 presents application to a population of residential pool pumps, and Section 6 describes application of similar methodology to a spatially distributed population of batteries. Conclusions and discussion of future research are contained in Section 7.

2 Distributed Control Architecture

This section contains a list of specific goals and general control design techniques that offer solutions.

2.1 Problem Statement

The grid operator requires resources to balance the grid at all times. The hypothesis of this chapter is that a large proportion of the needed resources can come in the form of virtual storage from flexible loads. Reliable grid services can be obtained from loads, but this requires a well-designed control architecture.

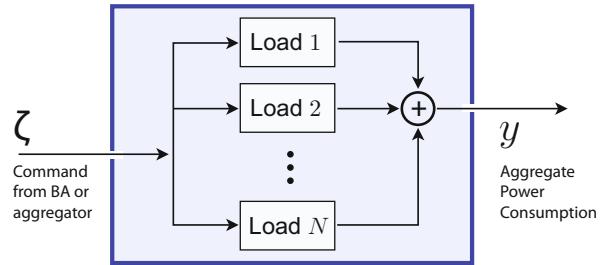
A particular hierarchical control architecture is developed in this chapter. One realization is illustrated by the feedback structure shown in Figure 2, in which the actuation block is composed of many resources acting in parallel, including generation, batteries, and virtual energy storage.

Assumptions regarding this control structure include

- (i) *Local control:* This will be based in part on randomized decision rules. This provides necessary degrees of freedom in shaping aggregate dynamics. Randomization also helps to prevent synchronization of the response from loads.
- (ii) *Information flow from loads:* Two-way information exchange between the BA and individual loads is *not* a component of this architecture. In [36] it is assumed that the BA measures aggregate power consumption from the loads under its authority. Alternatively, each load broadcasts its power state several times per day, and aggregate power consumption is estimated at the BA [13].

In [30, 31] it is argued that it is possible to create a reliable control system in which direct information flow from loads to BA is entirely absent. This requires more complex control at each load and hence is beyond the scope of this chapter.

Fig. 3 Control architecture: a common command signal is transmitted to each load in a particular class. The resulting input-output system from ζ to power consumption y is regarded as virtual energy storage.



- (iii) *Information flow from the BA:* A single regulation signal is broadcast to each collection of loads of the same class, as illustrated in Figure 3. This signal is designed based on grid-level measurements and a model of the aggregate behavior of the loads in each class.

The value of “local intelligence” at each load is vital for the envisioned architecture. Feedback loops at each load are used to ensure that QoS constraints are met and also so that the aggregate of loads will appear to the grid operator as a reliable resource – much like a battery system or a controllable generator.

Consumer choice will be an input to any VES system, and a monetary reward may be part of the arrangement. A contract for services can be established so that the consumer is rewarded for participation, without exposing him or her to the complexity and uncertainty of the grid. In this way, the BA or aggregator can design the system so that highly reliable grid services are obtained while respecting the QoS constraints of the consumer.

In the future it is possible that some loads will be grid friendly by design; the consumer will never know that their refrigerator is helping to regulate the grid.

2.2 Local Control

The lowest level of control in the proposed architecture is at an individual load, such as a water heater, refrigerator, agricultural water pump, or air-conditioner. The load is equipped with sensors. For example, the microprocessor in a water heater receives measurements of water temperature at one or more locations in the unit. It is also assumed that the load receives measurements from the grid. This could be purely local, such as the grid frequency measured locally [30, 31]. Theory is best developed in the setting where each load receives a signal from the BA.

The local control loop is designed to meet these potentially conflicting goals: 1. Ensure that the load is providing the desired services to the consumer, respecting strict bounds on QoS, and 2. Ensure that the *aggregate* of loads responds to a signal from the BA in a manner that is both predictable and beneficial to the grid.

There is one obvious challenge: the degrees of freedom are extremely limited for a typical load of interest. For example, a residential water heater or refrigerator can be in only one of a small number of power states. Contained in Section 3 are several

design techniques for local control that result in smooth aggregate behavior. This is possible without the introduction of complex scheduling rules or the solution of real-time optimization problems at the BA.

2.3 *Macro-control*

This high-level control layer may be a part of the traditional BA or through a load aggregator. The balancing challenges are of many different categories, on many different timescales:

- (i) Automatic generation control (AGC): timescales of seconds to 20 minutes.
- (ii) Balancing reserves: in the Bonneville Power Authority, the balancing reserves include both AGC and balancing on timescales of many hours.³
- (iii) Contingencies (e.g., a generator outage)

The final two challenges are observed in Figure 1:

- (iv) Peak shaving.
- (v) Smoothing ramps from solar or wind generation.

In this chapter it is assumed that these high-level control problems are addressed as they are today: the BA receives measurements of the grid and based on this information sends out signals to each resource in its domain. In many cases control loops are based on standard PI (proportional-integral) control design.

The difference here is that some resources are virtual, such as a collection of water heaters. A large collection of batteries distributed across the region might be regarded as a single resource – in this case, local control loops will be installed in each battery system so that the aggregate behaves as a single massive battery.

3 Mean-Field Control Design

Standard approaches for solving a stochastic control problem include stochastic dynamic programming and Markov decision processes (MDP) [38]. The future power grids will contain millions of smart components, which prohibit centralized decision-making using these techniques as they do not scale well with the number of different components in the system (both the state space and the control space of the model grow exponentially with the number of components). The extension of MDP models to the case of optimization problems involving many agents that are making decisions based on partial knowledge of the system is called DEC-POMDP (decentralized partially observable MDP). These problems are NEXP-hard for the

³Balancing on a slower timescale is achieved through real-time markets in some other regions of the USA and in every region under the jurisdiction of an RTO.

finite horizon optimization case [2], and undecidable in the infinite-horizon case [28].

In physics and probability theory, mean-field theory (MFT) approximates the behavior of a large number of small individual components which interact with each other. The effect of all the other individuals on any given individual is approximated by a single averaged effect, thus reducing a many-body problem to a one-body problem. The mean-field ideas first appeared in physics in the work of Pierre Curie and Pierre Weiss to describe phase transitions [23, 47]. Approaches inspired by these ideas have seen applications in epidemic models [3], computer network performance, and game theory [22, 27]. In power systems, they were first used to model the aggregate dynamic of the collection of water heaters in [29] and more recently in [25, 43, 46]. However, the global objective optimization under mean-field interactions remains very challenging, and an exact analysis is possible only under restrictive assumptions on the local dynamics and the cost structure. There is still a significant gap between the theoretical assumptions and the applications, and the results may be sensitive to the modeling errors.

The approach in [5, 11, 13] combines the mean-field theory with classical feedback control. The main idea consists in defining a parametrized family of randomized local decision rules that lead to an aggregate behavior with desirable control properties (e.g., passivity for the linearized aggregate input-output system).

This section provides an overview of key concepts and results of this approach.

3.1 Mean-Field Model

A nominal Markovian model for an individual load is created based on its typical operating behavior. This is described as a Markov chain with transition matrix denoted P_0 , with state space $\mathbf{X} = \{x^1, \dots, x^d\}$. For example, a water chiller turns on or off depending upon the temperature of the water. In this case, a state value x^i encodes water temperature as well as the power state (on or off).

A family of transition matrices $\{P_\zeta : \zeta \in \mathbb{R}\}$ is then constructed to define local decision-making. Each load evolves as a controlled Markov chain on \mathbf{X} , with common input $\zeta = (\zeta_0, \zeta_1, \dots)$. It is assumed that the scalar signal ζ is broadcast to each load. If a load is in state x at time t and the value ζ_t is broadcast, then the load transitions to the state x' with probability $P_{\zeta_t}(x, x')$. Letting X_t^i denote the state of the i th load at time t and assuming N loads, the empirical pmf (probability mass function) is defined as the average:

$$\mu_t^N(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{X_t^i = x\}, \quad x \in \mathbf{X}.$$

The mean-field model is the deterministic system defined by the evolution equations,

$$\mu_{t+1} = \mu_t P_{\zeta_t}, \quad t \geq 0, \quad (1)$$

in which μ_t is a row vector of dimension d . Under general conditions on the model and on μ_0 , it can be shown that μ_t^N is approximated by μ_t .

In [11, 13, 36] it is assumed that average power consumption is obtained through measurements or state estimation: Let $\mathcal{U}(x)$ denote power consumption when the load is in state x , where $\mathcal{U} : X \rightarrow \mathbb{R}_+$. The average power consumption is denoted

$$y_t^N = \frac{1}{N} \sum_{i=1}^N \mathcal{U}(X_t^i),$$

which is approximated using the mean-field model

$$y_t = \sum_x \mu_t(x) \mathcal{U}(x), \quad t \geq 0. \quad (2)$$

The next subsection describes the linearized dynamics. Sections 3.2–3.3 provide an overview of design techniques for the parametrized transition family $\{P_\zeta : \zeta \in \mathbb{R}\}$ to ensure that the linearized input-output model has desirable properties for control at the grid level.

3.1.1 Linearized Mean-Field Model

The mean-field model (1) is a state-space model that is linear in the state μ_t and nonlinear in the input ζ_t . The observation equation (2) is also linear as a function of the state. Assumptions imposed in [11, 13, 36] imply that the input is a continuous function of these values. The design of the feedback law $\zeta_t = \phi_t(y_0, \dots, y_t)$ is based on a linearization of this state-space model.

The linearized input-output model requires additional notation. The derivative of the transition matrix is also a $d \times d$ matrix, denoted

$$\mathcal{E}_\zeta = \frac{d}{d\zeta} P_\zeta \quad (3)$$

Denote $\tilde{\mathcal{U}}_\zeta = \mathcal{U} - \bar{\mathcal{U}}_\zeta$, with $\bar{\mathcal{U}}_\zeta = \pi_\zeta(\mathcal{U})$.

The invariant pmf π_ζ for P_ζ is regarded as the equilibrium state for the mean-field model (1), with respect to the constant input value $\zeta_t \equiv \zeta$. The linearization about this equilibrium is described in Proposition 1. The proof can be found in [36, Prop. 2.4].

Proposition 1 Consider a family of transition matrices $\{P_\zeta : \zeta \in \mathbb{R}\}$ that are continuously differentiable in ζ . Assume also that P_ζ is irreducible and aperiodic for each ζ . The unique invariant pmf π_ζ is an equilibrium for (1) when ζ takes on this constant value. The input-output model with state evolution (1), input ζ ,

and output (2) admits a linearization about this equilibrium. It is described as a d -dimensional state-space model with transfer function

$$G_\zeta(z) = C[Iz - A]^{-1}B \quad (4)$$

in which $A = P_\zeta^T$, $C_i = \tilde{\mathcal{U}}_\zeta(x^i)$ for each i , and

$$B_i = \sum_x \pi_\zeta(x) \mathcal{E}_\zeta(x, x^i), \quad 1 \leq i \leq d \quad (5)$$

3.2 Local Control Design

It is assumed throughout this chapter that the family of transition matrices used for distributed control is of the form

$$P_\zeta(x, x') := P_0(x, x') \exp(h_\zeta(x, x') - \Lambda_{h_\zeta}(x)) \quad (6)$$

in which h_ζ is continuously differentiable in ζ and Λ_{h_ζ} is the normalizing constant

$$\Lambda_{h_\zeta}(x) := \log\left(\sum_{x'} P_0(x, x') \exp(h_\zeta(x, x'))\right) \quad (7)$$

Each P_ζ is irreducible and aperiodic under the assumption that this is true for P_0 .

3.2.1 Myopic Design and the Exponential Family

A simple choice is the *myopic design*. This is obtained by setting $h_\zeta(x, x') = \zeta \mathcal{U}(x')$,

$$P_\zeta^{myop}(x, x') := P_0(x, x') \exp(\zeta \mathcal{U}(x') - \Lambda_\zeta(x)) \quad (8)$$

with the normalizing constant $\Lambda_\zeta(x) := \log\left(\sum_{x'} P_0(x, x') \exp(\zeta \mathcal{U}(x'))\right)$. This corresponds to a tilted probability transition matrix, favoring the transitions to states with lower power consumption when $\zeta < 0$ and to states with higher power consumption when $\zeta > 0$.

Advantages of this design include ease of implementation and the straightforward generalization to the continuous state-space case. This generalization will be illustrated in Sections 4 and 6.

It is possible to consider any other family of functions, linear with respect to ζ , leading to an exponential family for $\{P_\zeta : \zeta \in \mathbb{R}\}$,

$$h_\zeta(x, x') = \zeta H_0(x, x'). \quad (9)$$

The choice of H_0 will typically correspond to the linearization of a more advanced design around the value $\zeta = 0$ (or some other fixed value of ζ). One example is given in Section 3.3.

3.2.2 Individual Perspective Design

Consider a finite-time-horizon optimization problem: For a given terminal time T , let p_0 denote the pmf on strings of length T :

$$p_0(x_1, \dots, x_T) = \prod_{i=0}^{T-1} P_0(x_i, x_{i+1}),$$

where $x_0 \in X$ is assumed to be given. The scalar $\zeta \in \mathbb{R}$ is interpreted as a weighting parameter in the following definition of total welfare. For any pmf p , this is defined as the weighted difference,

$$\mathcal{W}_T(p) = \zeta \mathbf{E}_p \left[\sum_{t=1}^T \mathcal{U}(X_t) \right] - D(p \| p_0) \quad (10)$$

where the expectation is with respect to p , and D denotes relative entropy:

$$D(p \| p_0) := \sum_{x_1, \dots, x_T} \log \left(\frac{p(x_1, \dots, x_T)}{p_0(x_1, \dots, x_T)} \right) p(x_1, \dots, x_T)$$

It is easy to check that the myopic design is an optimizer for the horizon $T = 1$,

$$P_\zeta^{myop}(x_0, \cdot) \in \arg \max_p \mathcal{W}_1(p).$$

The infinite-horizon mean welfare is denoted,

$$\eta_\zeta^* = \lim_{T \rightarrow \infty} \frac{1}{T} \mathcal{W}_T(p_T^*) \quad (11)$$

The two terms in the welfare function (10) represent the two conflicting goals: To provide service to the grid and to reduce deviation of the load's behavior from the nominal. If the controlled probability p is chosen to be different from p_0 , it potentially reduces the QoS to the consumer, which is modeled by the term “ $-D(p \| p_0)$.”

Recall that $\mathcal{U}(X_t)$ is equal to the power consumption of the load at time t . If the grid operator desires lower power demand than the nominal value, this goal is modeled through the first term in (10) whenever the parameter ζ is negative.

A solution to the infinite-horizon problem is given by a time-homogeneous Markov chain whose transition matrix is obtained following the solution of an eigenvector problem, based on the $d \times d$ matrix

$$\widehat{P}(x, x') = \exp(\zeta \mathcal{U}(x)) P_0(x, x'), \quad x, x' \in X. \quad (12)$$

Let $\lambda > 0$ denote the Perron-Frobenius eigenvalue and v the eigenvector with nonnegative entries satisfying

$$\widehat{P}v = \lambda v \quad (13)$$

The proof of Proposition 2 is contained in [36, Section II], following [44].

Proposition 2 *If P_0 is irreducible, an optimizing p^* that achieves (11) is defined by a time-homogeneous Markov chain whose transition probability is defined by*

$$\check{P}_\zeta(x, x') = \frac{1}{\lambda} \frac{1}{v(x)} \widehat{P}(x, x') v(x'), \quad x, x' \in X. \quad (14)$$

3.3 Uncontrolled Dynamics

In many cases it is not possible to apply the IPD solution in the form (14) because a portion of the stochastic dynamics are not directly controllable. Consider a load model in which the full state space is the Cartesian product $X = X^u \times X^n$, where X^u are components of the state that can be directly manipulated through control.

In prior work [5, 6], the following conditional independence structure is assumed: for each state $x = (x_u, x_n)$ and each $\zeta \in \mathbb{R}$,

$$\begin{aligned} \check{P}_\zeta(x, x') &= R_\zeta(x, x'_u) Q_0(x, x'_n), \\ R_\zeta(x, x'_u) &= R_0(x, x') \exp(h_\zeta(x, x'_u) - \Lambda_{h_\zeta}(x)) \end{aligned} \quad (15)$$

where $\sum_{x'_u} R_\zeta(x, x'_u) = \sum_{x'_n} Q_0(x, x'_n) = 1$ for each x and ζ . The matrix Q_0 is out of our control – this models load dynamics and exogenous disturbances. For example, it may be used to model the impact of the weather on the climate of a building. The matrices $\{R_\zeta\}$ are a product of design.

It is reasonable to assume that \mathcal{U} is a function only of X^u ; the power state is directly controllable. In this case the myopic design (8) is unchanged, $h_\zeta(x, x'_u) = \zeta \mathcal{U}(x'_u)$.

The formulation of the IPD optimization problem is unchanged, but its solution is not in the form (14). A computational ODE approach is introduced in [5, 6]: for a

vector field \mathcal{V} whose domain and range are functions on $\mathbf{X} \times \mathbf{X}^u$,

$$\frac{d}{d\zeta} h_\zeta = \mathcal{V}(h_\zeta), \quad \zeta \in \mathbb{R}, \quad h_0 \equiv 1.$$

Besides its computational value, this approach provides a useful alternative to the myopic design. The function $H_0 = \mathcal{V}(h_0)$ can be used in the exponential family design (9). It is shown in [6] that this function is a solution to Poisson's equation for the nominal model: $P_0 H_0 = H_0 - \tilde{\mathcal{U}}_0$.

Motivation for the IPD design or its exponential family approximation is in part empirical. In nearly every numerical experiment conducted to date, it is found that the resulting input-output mean-field model appears nearly linear over a large range of ζ and also minimum phase. Moreover, in nearly all cases, the linearization (4) is *passive* when the delay is removed. That is, the transfer function $zC[Iz - A]^{-1}B$ is strictly positive real.

Passivity can be established mathematically for a restricted class of models [4] or using a different ODE called the system perspective design (SPD) [5].

3.4 Quality of Service and Opt Out

In analysis of QoS, it is convenient to consider a steady-state setting: the state process for each load is assumed to be a stationary process on the two-sided time interval. It is also useful to consider a functional form for QoS – the following conventions were introduced in [11].

Several QoS metrics may be considered simultaneously, but each is assumed to be of the following form. Assumed given is a function $\ell: \mathbf{X} \rightarrow \mathbb{R}$, defined so that $L_t^i := \ell(X_t^i)$ describes a “snapshot” indication of QoS for the i th load at time t . The function ℓ may represent the temperature of a TCL, cycling of an on/off load, or power consumption as a function of $x \in \mathbf{X}$.

Second is a stable transfer function denoted $H_{\mathcal{L}}$. The QoS of the i th load at time t is defined by passing L^i through the transfer function $H_{\mathcal{L}}$. Two classes of transfer functions $H_{\mathcal{L}}$ are considered in prior research and examples in this chapter:

- (i) Summation over a finite-time horizon T_f :

$$\mathcal{L}_t^i = \sum_{k=0}^{T_f} \ell(X_{t-k}^i). \quad (16)$$

- (ii) Discounted sum, with discount factor $\beta \in [0, 1]$:

$$\mathcal{L}_t^i = \sum_{k=0}^{\infty} \beta^k \ell(X_{t-k}^i). \quad (17)$$

When β is close to unity or T_f is very large, then these QoS metrics can be approximated by Gaussian random variable by appealing to the central limit theorem [11]. A Gaussian distribution indicates that QoS for some individuals in the population will sometimes take on unacceptable values.

QoS can be constrained by imposing an additional layer of control at each load. A simple mechanism is *opt-out control*.

The opt-out mechanism is based on predefined upper and lower limits, denoted b_+ and b_- . A load ignores a command from grid operator if it will result in $\mathcal{L}_{t+1}^i \notin [b_-, b_+]$ and takes an alternative action so that $\mathcal{L}_{t+1}^i \in [b_-, b_+]$. This ensures that the QoS metric of each load remains within the predefined interval for all time.

Numerical examples are presented in [11] for both residential pools and TCLs. Some of these results are surveyed in Section 5. Negative impact on tracking performance is observed in numerical experiments only when the QoS interval $[b_-, b_+]$ is small (e.g., b_+ is less than the mean plus one standard deviation of the distribution without opt-out control).

4 Example: Thermostatically Controlled Loads

This special case is dominant in much of the literature on demand dispatch. Examples of thermostatically controlled loads (TCLs) include refrigerators, water heaters, and air conditioning. Each of these loads is already equipped with primitive “local intelligence” based on a *deadband* (or *hysteresis interval*): there is a sensor that measures the temperature of the unit and turns the power on when the measured value reaches one end of this deadband.

The state process for a TCL at time t will be of the form

$$X(t) = (X_u(t), X_n(t)) = (m(t), \Theta(t)), \quad (18)$$

in which $m(t) \in \{0, 1\}$ denotes the power mode (the value “1” indicating the unit is on) and $\Theta(t)$ the inside temperature of the load. Exogenous disturbances that directly influence Θ include ambient temperature and usage: the inside temperature of a refrigerator is impacted by an open door, and the temperature of water in a water heater is influenced by the rate of flow of water out of the unit.

The remainder of this section is restricted to a residential water heater (WH). This will simplify discussion, and extensions to other TCLs are often straightforward.

4.1 Nominal Model

The standard ODE model of a water heater is the first-order linear system:

$$\frac{d}{dt}\Theta(t) = -\lambda[\Theta(t) - \Theta^a(t)] + \gamma m(t) - \alpha[\Theta(t) - \Theta^{in}(t)]f(t), \quad (19)$$

for constants $(\lambda, \gamma, \alpha)$, in which $\Theta(t)$ is the temperature of the water in the tank, $\Theta^a(t)$ is ambient temperature, $\Theta^{in}(t)$ is temperature of the cold water entering the tank (degrees Fahrenheit), $f(t)$ is flow rate of hot water from the WH (gallons/s), and $m(t)$ is the power mode of the WH (“on” indicated by $m(t) = 1$). The corresponding power consumed by a WH when $m(t) = 1$ is denoted P_{on} .

The upper and lower temperature limits that define the deadband are denoted Θ_- and Θ_+ , respectively. A standard residential water heater in the USA has the following typical behavior: at the moment that $\Theta(t)$ reaches the lower limit Θ_- , the unit turns on and remains on until the time t_+ at which $\Theta(t_+) = \Theta_+$. The unit then turns off and begins to cool. It may take 6 hours to once again reach the lower limit, while the time to heat the water is much shorter.

The nominal model used for local control design is based on an approximation of this typical behavior, in which with some probability the unit turns on before $\Theta(t)$ reaches Θ_- , and the unit may also turn off before reaching the maximum temperature Θ_+ . The definition of the nominal model is based on the specification of two cumulative distribution functions (CDFs) for the temperature at which the load turns on or turns off, denoted F^\oplus and F^\ominus . Random variables with these CDFs are denoted $\tilde{\Theta}^\oplus$ and $\tilde{\Theta}^\ominus$, so that

$$F^\oplus(\theta) = \mathbb{P}\{\tilde{\Theta}^\oplus \leq \theta\}, \quad F^\ominus(\theta) = \mathbb{P}\{\tilde{\Theta}^\ominus \leq \theta\}, \quad \theta \in \mathbb{R}.$$

It is always assumed that $\tilde{\Theta}^\oplus$ and $\tilde{\Theta}^\ominus$ take values in the interval $[\Theta_-, \Theta_+]$, which implies that $F^\oplus(\theta) = F^\ominus(\theta) = 1$ for $\theta \geq \Theta_+$ and $F^\oplus(\theta) = F^\ominus(\theta) = 0$ for $\theta < \Theta_-$.

A particular design for F^\ominus is obtained on fixing three parameters $\theta_0^\ominus \in [\Theta_-, \Theta_+]$ and constants $\varrho \in (0, 1)$ and $\kappa > 1$:

$$F^\ominus(\theta) = (1 - \varrho) \frac{[\theta - \theta_0^\ominus]_+^\kappa}{[\Theta_+ - \theta_0^\ominus]^\kappa}, \quad \theta \in [\Theta_-, \Theta_+],$$

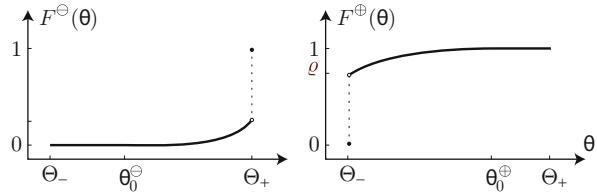
where $[x]_+ := \max(0, x)$ for $x \in \mathbb{R}$. In a symmetric model, the other CDF is defined by the transformation

$$F^\oplus(\theta) = 1 - \lim_{\theta' \downarrow \theta} F^\ominus(\Theta_+ + \Theta_- - \theta')$$

Figure 4 illustrates a particular special case of the symmetric model.

It is assumed that the local control operates in discrete time. By choice of time units, without loss of generality, it is assumed that the sampling interval is 1 unit. At time instance k , if the water heater is on (i.e., $m(k) = 1$), then it turns off at time $k + 1$ with probability

Fig. 4 Nominal model for a water heater: an instance of the symmetric model.



$$p^\ominus(k+1) = \frac{[F^\ominus(\Theta(k+1)) - F^\ominus(\Theta(k))]_+}{1 - F^\ominus(\Theta(k))}$$

If $\Theta(k+1) \leq \Theta(k)$, then this probability is zero. Similarly, if the load is off, then it turns on with probability

$$p^\oplus(k+1) = \frac{[F^\oplus(\Theta(k)) - F^\oplus(\Theta(k+1))]_+}{F^\oplus(\Theta(k))}$$

The nominal behavior of the power mode can be expressed

$$\begin{aligned} \mathbb{P}\{m(k) = 1 \mid \theta(k-1), \theta(k), m(k-1) = 0\} &= p^\oplus(k) \\ \mathbb{P}\{m(k) = 0 \mid \theta(k-1), \theta(k), m(k-1) = 1\} &= p^\ominus(k) \end{aligned} \quad (20)$$

The IPD and SPD designs were obtained in [5] based on a similar nominal model for a residential refrigerator.

The myopic design (15) is obtained through an exponential tilting:

$$p_\zeta^\oplus(k) := \frac{p^\oplus(k)e^\zeta}{p^\oplus(k)e^\zeta + 1 - p^\oplus(k)}, \quad p_\zeta^\ominus(k) := \frac{p^\ominus(k)}{p^\ominus(k) + (1 - p^\ominus(k))e^\zeta}$$

If $p^\oplus(k) > 0$, then the probability $p_\zeta^\oplus(k)$ is strictly increasing in ζ , approaching 1 as $\zeta \rightarrow \infty$; it approaches 0 as $\zeta \rightarrow -\infty$, provided $p^\oplus(k) < 1$.

4.2 System Identification

Power, temperature, and usage data from residential water heaters was obtained through our partners at ORNL.⁴ The constants $(\lambda, \gamma, \alpha)$ were estimated using least squares. The parameter values listed in Table 1 reflect the range of values observed in actual data.

⁴Water heater data provided by Ecotope, Inc., with funding from the Northwest Energy Efficiency Alliance (NEEA) and the Bonneville Power Administration (BPA).

Table 1 Parameters for nominal model for water heaters.

Temp. ranges	ODE pars.	Loc. control
$\Theta_+ \in [118, 122] \text{ F}$	$\lambda \in [8, 12.5] \times 10^{-6}$	$T_s = 15 \text{ sec}$
$\Theta_- \in [108, 112] \text{ F}$	$\gamma \in [2.6, 2.8] \times 10^{-2}$	$\kappa = 4$
$\Theta^a \in [68, 72] \text{ F}$	$\alpha \in [6.5, 6.7] \times 10^{-2}$	$\varrho = 0.8$
$\Theta^{in} \in [68, 72] \text{ F}$	$P_{on} = 4.5 \text{ kW}$	$\theta_0 = \Theta_-$

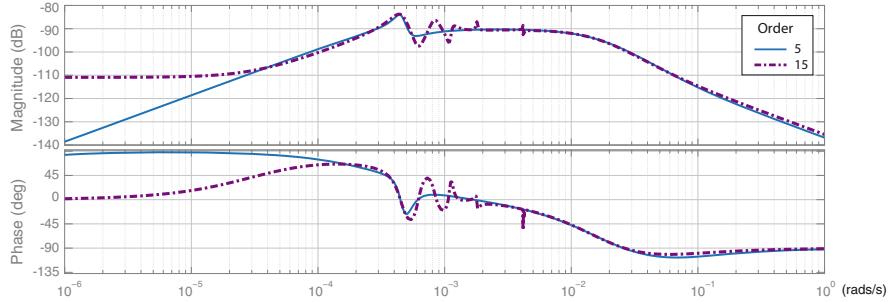


Fig. 5 Least-square estimates of the transfer function for water heaters.

A test-bed was created to simulate a collection of $N = 100,000$ water heaters with usage. Each evolves according to the ODE (19), but parameters were different for each of the N loads: parameters were chosen via uniform sampling of the values in Table 1. A simulation model for usage at each load was created, based on sampling from historical usage of actual water heaters.

The mean-field model is a nonlinear input-output system with input ζ and output equal to power deviation, y . An approximate linear model was obtained through least squares, in which the input ζ was taken to be the swept-sine: $\zeta(t) = 1.5 \sin(10^{-7}t^2)$ for $0 \leq t \leq 432 \times 10^5$ sec. (5 days). Figure 5 shows results from the estimation experiment for two different model orders. The Bode plots shown represent the approximate model in continuous time. The 5th-order model predicts that the gain of the linearization vanishes as the frequency tends to zero (DC). This is a physical reality for this example.

The linearization is minimum phase and stable. Its gain is approximately constant in the frequency range $[5 \times 10^{-4}, 10^{-2}] \text{ rad/s}$. It is expected that a collection of water heaters can accurately track signals in this frequency range.

4.3 Tracking

Design at the macrolevel is most easily performed for a model in continuous time. A PI controller $G_c(s) = K_P + K_I/s$ was designed based on the linearized mean-field model. The values $K_P = 10^5$ and $K_I = 500$ result in a crossover frequency

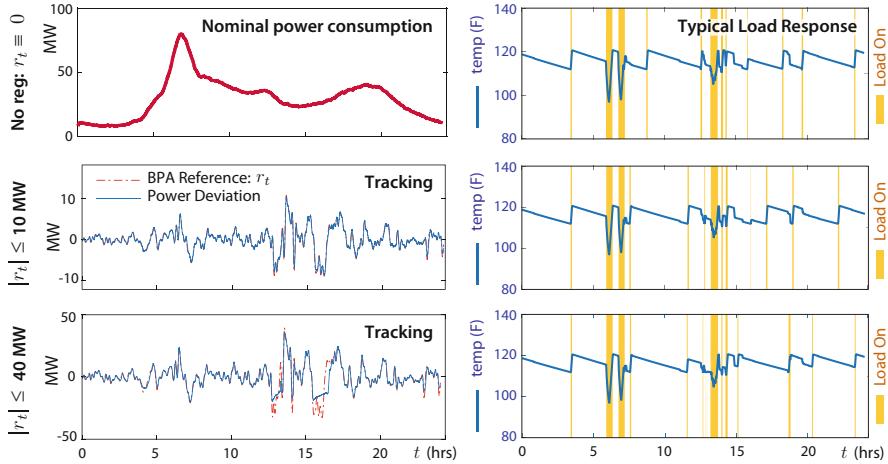


Fig. 6 Tracking results with 100,000 water heaters and the behavior of a single water heater in three cases, distinguished by the reference signal r . The morning peak in nominal power consumption is consistent with typical water usage included in the simulation experiments.

$\omega_c = 0.03$ rad/s (corresponding to a time period of approximately 3.5 minutes), with a 75° phase margin.

The balancing reserve signal from the Bonneville Power Administration (BPA) was used in the tracking experiments described in this section. A typical windy day, February 19, 2016, was chosen for the experiments described here. The signal was filtered using a second-order Butterworth high-pass filter with a cutoff frequency at 8×10^{-4} rad/s (corresponding to a sine wave with period of approximately 2 hours).

Figure 6 shows results from several numerical experiments. The three rows are differentiated by the regulation signal: in the first row $r \equiv 0$; in the second, the absolute value of the regulation signal takes a maximum value of about 8 MW; and in the final row, the prior regulation signal was multiplied by 4. Exact tracking is not feasible over the entire period for the largest regulation signal (results shown in the bottom left plot), but the performance remains nearly perfect over time periods for which $|r_t|$ does not exceed about 90% of the nominal power consumption.

The second column shows evolution of temperature and the power mode for a typical load in the three cases. The seed for the random number generator was identical in each of the three experiments. It is amazing to see that the evolution of temperature and power mode is hardly impacted by local control.

These loads are equally valuable for contingency and ramping services. Figure 7 shows recent results that illustrate the potential. In these experiments the water flow was set to zero; in this case, the nominal power consumption for 100,000 loads is approximately 8 MW. Each plot is a particular sawtooth wave, scaled to reach the maximum lower limit of -8 MW.

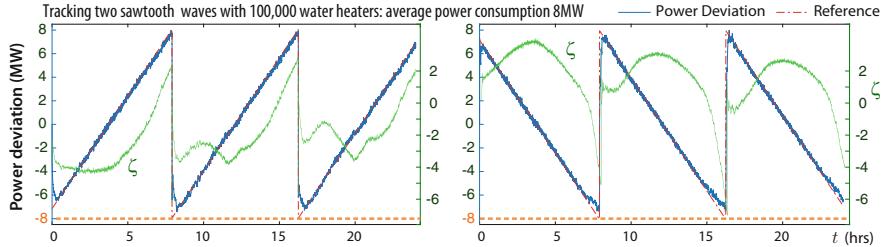


Fig. 7 Tracking a pair of sawtooth waves with 10^5 water heaters.

5 Example: Residential Pools

The paper [36] and its sequels [11–13] consider this system architecture in which the loads are a collection of pools. The motivation for considering pools is the inherent flexibility of pool cleaning and because the total load in a region can be very large. The maximum load is approximately 1 GW in California or Florida.

The state space for the discrete-time model is the finite set: $\mathbf{X} = \{(m, j) : m \in \{0, 1\}, j \in \{1, 2, \dots, J\}\}$. The first variable indicates the power mode, with $m = 1$ indicating the power is on. The second integer component is interpreted as follows: The load samples the grid signal periodically (the sampling increments are assumed to be deterministic or i.i.d. and distributed according to a geometric distribution). At the time of the t -th sample, if $X_t^i = (0, j)$, then the load has remained off for the past j sampling times and was turned off at sampling time $t - j$; the interpretation of $X_t^i = (1, j)$ is symmetrical.

A nominal model can be constructed in a manner similar to the case of TCLs. In this application, each CDF models the time at which the power mode changes. The IPD solution obtained using Proposition 2 is considered in [36] for a model without geometric sampling and in [13] for the present model. The linearized mean-field models obtained in [13, 36] are minimum phase and have a resonance at a frequency corresponding to a period of approximately 24 hours.

The numerical results that follow are based on a stochastic simulation of a large number of pools. Each pool consumes 1 kW when in operation. Both 12 and 8 hour nominal daily cleaning cycles are considered. Tracking results with a heterogeneous population of loads are described in [13, 36].

5.1 Tracking and Contingency Reserves

The first set of experiments concern tracking of the balancing reserves deployed over one week at BPA. The sampling time is taken to be every five minutes. The signal was filtered to remove the highest frequency components. Tracking the original signal is possible, but with reduced overall capacity [10]. In each example, the number of loads is equal to $N = 10^4$.

A theoretical limit on capacity is obtained by considering the fraction of pool pumps that are operating in nominal steady state:

$$\pi_0^\oplus = \sum_i \pi_0(1, i) \quad \text{where } \pi_0 \text{ is invariant for } P_0.$$

Upper and lower bounds on power deviation are defined as follows, in units of kWs:

$$\{\text{+Demand}^*, -\text{Supply}^*\} := \{(1 - \pi_0^\oplus) \times N, -\pi_0^\oplus \times N\}$$

This is approximately $\{+5, -5\}$ MW for 12 hr/day cycling, and $\{+6.6, -3.4\}$ MW for 8 hr/day. Results from simulation experiments shown in Figure 8 show that these limits are nearly attainable in each case.

The potential for virtual energy storage goes far beyond tracking a balancing reserve signal. Experiments were conducted in [10] to investigate the potential for providing contingency reserves in conjunction with balancing reserves. A reference signal was constructed based on the one used in the previous experiments, with two changes: during the period [40, 64] hours, the reference signal was *replaced* by a 4 MW power reduction, and during the period [100, 124] hours, the reference signal was *augmented* with a 3 MW power reduction. The PI control parameters were unchanged. Figure 9 shows again nearly perfect tracking.

In practice, the signal ζ should be transformed so that it is zero energy over the week – this will help to ensure that QoS constraints are not violated.

5.2 Quality of Service and Opt Out

The grid is receiving nearly perfect services – what about the service offered by each load to its owner?

In the experiments conducted to produce either of the plots in Figure 8, a histogram of total operation hours over the time horizon appears approximately Gaussian with mean value 78 hours (consistent with the 12hrs/day cleaning cycle for each pool.) The Gaussian approximation can be used to estimate the fraction of pools that are overcleaned or undercleaned over the week [12].

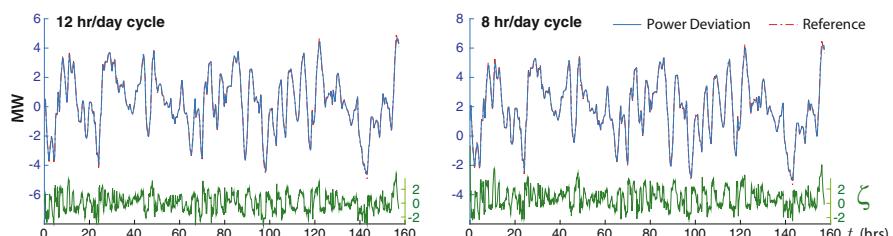


Fig. 8 Tracking is nearly perfect with reference scaled to the theoretical limit.

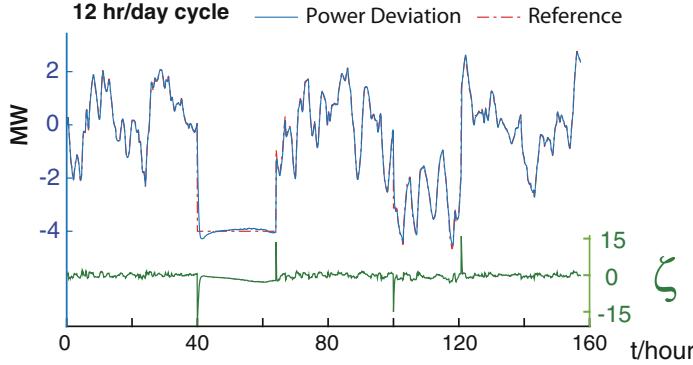


Fig. 9 Range of services provided by pools: contingency reserves and balancing can be supplied simultaneously.

To investigate the impact of opt-out control on QoS and capacity, consider a family of models parameterized by $0 \leq \varepsilon \leq 1$. The reference signal is obtained by scaling: $r_t^\varepsilon = \varepsilon r_t^1$, where r^1 is the reference signal considered in prior experiments.

Consider the following two QoS metrics suitable for this application:

- (i) Normalized power consumption of a load, $\ell(X_t^i) = \mathcal{U}(X_t^i) - \bar{y}^0$, where \bar{y}^0 is the nominal steady-state mean (under $\zeta = 0$).
- (ii) On/off cycling:

$$\ell^c(X_t^i, X_{t+1}^i) = \left| \sum_j \left(\mathbf{1}\{X_{t+1}^i = (1, j)\} - \mathbf{1}\{X_t^i = (1, j)\} \right) \right| \quad (21)$$

The discounted sum (17) was used to define \mathcal{L}_t^i in the experiments surveyed here. The discount factor $\beta = 2779/2880$ was chosen so that the discounted sum (17) is similar to the moving window QoS metric (16) with T_f corresponding to 10 days (recall the sampling period is five minutes).

Figure 10 illustrates an example of QoS improvement based on a 15% constraint on both QoS metrics, using the reference signal r^1 . The opt-out rate is very small in this case (much smaller than predicted by the corresponding tails of the histogram without opt out), and the tracking is nearly perfect.

Four QoS intervals were considered corresponding to constraints of, respectively, 5%, 10%, 15%, and 20%. For example, a cleaning QoS constraint of 5% corresponds to ± 3 cleaning hours – a very tight constraint over a 10-day time horizon. No lower bound was imposed on cycling QoS.

A normalized root-mean-square error (NRMSE) was adopted as the metric for grid-level tracking performance:

$$\text{NRMSE} = \frac{1}{\varepsilon} \frac{\text{RMS}(\mathbf{e}) - \text{RMS}(\mathbf{e}^0)}{\text{RMS}(\mathbf{r}^1)}, \quad (22)$$

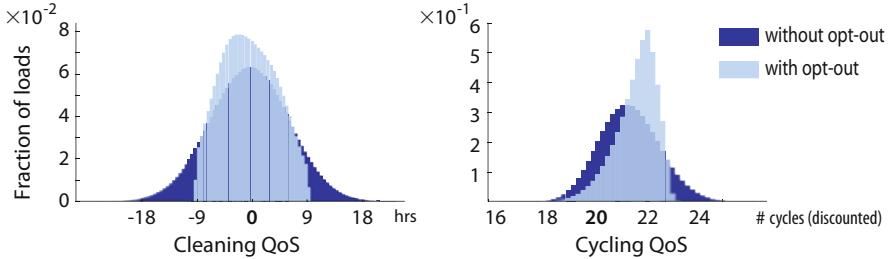


Fig. 10 Improvement of QoS with the introduction of local opt-out control.

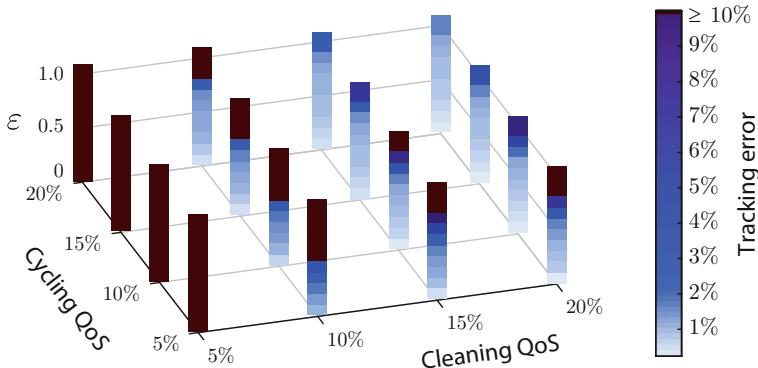


Fig. 11 Tracking performance with two QoS constraints – total cleaning hours and cycling. Opt-out control has little impact on tracking error over a large range of opt-out intervals.

where e and e^0 are error signals with and without the reference input and $\text{RMS}(f) = \sqrt{\frac{1}{T} \sum_{k=1}^T f_k^2}$ for any signal f over a time horizon T .

Tracking performance for a range of opt-out parameters is summarized in Figure 11 using 16 colored bars, distinguished by each pair of QoS constraints. Each bar represents tracking errors for different reference signal scaling factors, $0.1 \leq \varepsilon \leq 1$. The darkest color represents NRMSE (22) of 10% or greater, and lighter colors represent smaller values of NRMSE (indicated on the color bar label). This shows that opt-out control based on the two QoS metrics has little impact on tracking error over a large range of opt-out intervals.

6 Example: Battery Systems

Many believe that there will be a battery revolution over the next decade – small battery systems will be distributed across the grid at residential homes and in racks at gas stations where owners of electric vehicles can exchange their old battery.

Coordination of the population can be performed as described for residential loads, even though the physics and QoS constraints are very different. Longevity of a battery requires constraints on the state of charge (SoC), as well as ramping and temperature constraints.

A demand dispatch architecture is proposed in [7], in which the state space for an individual battery is again a Cartesian product: a particular state is denoted $x = (m, s)$, where $m \in \{\text{ch}, \text{dis}, \text{id}\}$ denotes charging mode and $s \in [0, 1]$ denotes the SoC. The power delivery at state x depends only on charging mode: $\mathcal{U}(\text{ch}, s) = \mathcal{U}_{\text{ch}} < 0$, $\mathcal{U}(\text{id}, s) = 0$, $\mathcal{U}(\text{dis}, s) = \mathcal{U}_{\text{dis}} > 0$.

The design of the family of transition matrices $\{P_\zeta\}$ on the state space $X = \{\text{ch}, \text{dis}, \text{id}\} \times [0, 1]$ is based on the myopic policy. The main difficulty compared to loads is that there is no obvious nominal model P_0 (for loads, this is taken as a stochastic perturbation of a deterministic model for $\zeta \equiv 0$.) The nominal model P_0 for batteries was chosen so that the invariant pmf π_0 would have most of its mass concentrated at SoC near 60%. The randomized decision rule is designed to encourage idle time for each battery and to avoid extreme SoC levels and frequent switching of modes.

Let $X_t^i = (M_t^i, S_t^i)$ denote the state of i th battery at time t . The SoC evolves as a controlled random walk: $S_{t+1}^i = S_t^i + h\delta_{\text{ch}}$, if $M_t^i = \text{ch}$, $S_{t+1}^i = S_t^i - h\delta_{\text{dis}}$, if $M_t^i = \text{dis}$, $S_{t+1}^i = S_t^i$, if $M_t^i = \text{id}$, where h is the time step length and δ_{ch} and δ_{dis} charging and discharging rates. The dynamics of the first component are governed by a “two-coin flip” randomized policy: In state (m, s) , a weighted coin is flipped to determine if the battery will stay in its current power mode. The design of the probability functions p_{ch} , p_{dis} , and $p_{\text{id}} : [0, 1] \rightarrow [0, 1]$ that model the probability to stay in the charging, discharging, or idle mode, respectively, is shown in Figure 12. If the outcome of the first coin flip is “mode change,” then a second coin flip is used to decide which of the remaining two modes the battery is going to switch to. This choice is done with the probabilities proportional to the values of the p -functions of the alternative power modes. For example, in state (ch, s) , the battery changes its mode to idle with probability $(1 - p_{\text{ch}}(s)) \times p_{\text{id}}(s) / (p_{\text{id}}(s) + p_{\text{dis}}(s))$.

The nominal design in Figure 12 was chosen by setting a *target SoC interval* to 40–80% SoC (to allow ramping capability while avoiding extreme SoC levels): If the battery is charging, it will remain charging with probability almost 1 until it reaches 40% SoC. The probability to keep charging then decreases and reaches almost 0 at 80% SoC. The design of p_{dis} is symmetrical. The function p_{id} has values almost 1 for 50–70% SoC values, and it is almost 0 outside the target interval.

Batteries are ideal for tracking signals of higher frequency – timescales of tens of seconds to many minutes. An example is the RegD signal used at PJM. It is found that tracking of this signal is nearly perfect using a combination of local control at the battery and a PI compensator at the BA (see [7] for details).

It might be assumed that the randomized control law would lead to excess cycling of batteries. In fact, the behavior of a typical battery behaved nearly deterministically. Typical behavior is illustrated in Figure 13.

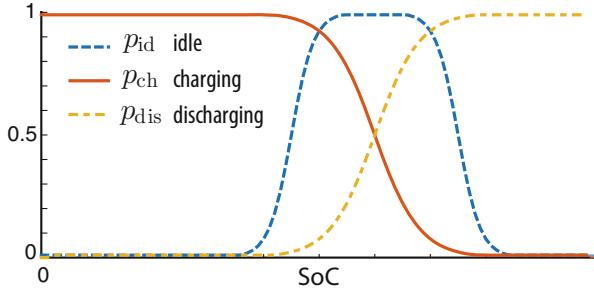


Fig. 12 Design of switching probability functions for the battery system. A weighted coin is flipped to determine if the battery will stay in its current power mode.

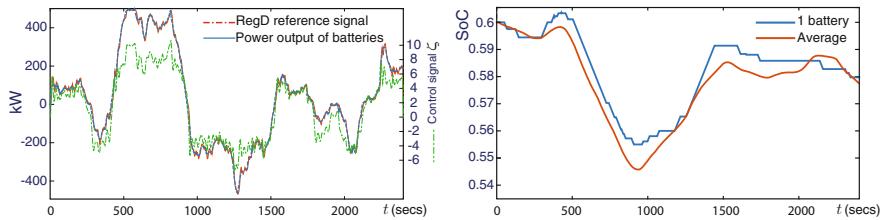


Fig. 13 Left: tracking the PJM RegD test signal with $N = 10^3$ batteries. Right: comparison of SoC of an individual and the average of the population.

A measure of battery activity is *mileage* [37], which is regarded here as an example of a QoS metric. For a time horizon T , mileage for battery k is denoted

$$\mathcal{L}_T^k = \sum_{t=1}^T |\mathcal{U}(X_t^k)|, \quad (23)$$

and \mathcal{L}_T denotes the average over $k = 1, \dots, N$. The *excess operation* is the normalized difference:

$$\mathcal{O}_T = \frac{\mathcal{L}_T - \mathcal{L}_T^*}{\mathcal{L}_T^*}, \quad \text{where } \mathcal{L}_T^* = \frac{1}{N} \sum_{t=1}^T |r_t|$$

It can be shown using Jensen's inequality that $\mathcal{O}_T \geq 0$ in the ideal case of perfect tracking. If *each* battery tracks the reference exactly, then $\mathcal{O}_T = 0$. In numerical experiments the value $\mathcal{O}_T \approx 3\%$ is typical.

7 Conclusions

With appropriate filtering and local control, loads can provide excellent grid services without two-way communication. While there is some cost to install hardware on appliances that can receive a signal from a balancing authority, in the long run this will be far less costly than batteries.

The numerical results presented in this paper, in particular the tracking results illustrated in Figures 6, 7, and 9, show that VES working in conjunction with traditional resources can provide balancing services, ramping services, and contingency reserves simultaneously. It is likely that water heaters, pools, and agriculture loads in California can provide the resources required to address their future grid service requirements.

Current research questions include:

- (i) The application of reinforcement learning may be valuable for learning the local control law, such as an extension of Z-learning [44] to the IPD approach.
- (ii) The numerical results presented here concern signals on timescales of tens of seconds and slower. Ancillary service on faster timescales corresponds to what is called primary reserves. Control design requires more care in this context because poor performance can induce grid instability [30, 31].
- (iii) Further research is required to better estimate capacity in terms of both energy and power [21].
- (iv) The impact of usage is not entirely understood. Numerical results presented in Section 4 suggest that this is not an obstacle in the case of water heaters. Air-conditioning is a greater challenge because variations in load are much greater.
- (v) A question posed in [31]: Does the load need to receive a signal from the BA? It is possible that some VES resources can provide valuable services using only local measurements. Frequency (as well as voltage) measurements can be obtained inexpensively at loads, and these measurements are similar to those used by the BA to construct analogs of our “ ζ ” today. The advantage of distributed control is reduced cost due to reduced communication between a BA and loads.

The BA will continue to regulate tie-line error, and they will continue to regulate frequency as well. It is hoped that the balancing resources required at the BA will be reduced through this extra layer of distributed control.

Acknowledgements Research was supported by the DOE Buildings Technologies Office as part of the Virtual Batteries project under the Buildings-to-grid Transactive Energy program (the authors would like to thank Mr. Joe Hagerman for his continued support), the National Science Foundation under grants CPS-0931416 and EPCN-1609131, the French National Research Agency grant ANR-16-CE05-0008, and the PGMO. We give our thanks to the anonymous reviewers for their suggestions and to Robert Moye at UF and Rainbow Energy for his encouragement and commentary.

References

1. Barooah P, Bušić A, Meyn S (2015) Spectral decomposition of demand-side flexibility for reliable ancillary services in a smart grid. In: Proceedings of the 48th annual Hawaii international conference on system sciences (HICSS), Kauai, pp 2700–2709
2. Bernstein DS, Givan R, Immerman N, Zilberstein S (2002) The complexity of decentralized control of Markov decision processes. *Math Oper Res* 27(4):819–840
3. Boudec JYL, McDonald D, Mundinger J (2007) A generic mean field convergence result for systems of interacting objects. In: Fourth international conference on the quantitative evaluation of systems (QEST 2007), September 2007, pp 3–18
4. Bušić A, Meyn S (2014) Passive dynamics in mean field control. In: Proceedings of the 53rd IEEE conference on decision and control, December 2014, pp 2716–2721
5. Bušić A, Meyn S (2016) Distributed randomized control for demand dispatch. In: IEEE conference on decision and control, December 2016, pp 6964–6971
6. Bušić A, Meyn S (2018) Ordinary differential equation methods for Markov decision processes and application to Kullback-Leibler control cost. *SIAM J Control and Optim.*, 56(1):343–366
7. Bušić A, Hashmi MU, Meyn S (2017) Distributed control of a fleet of batteries. In: American control conference, May 2017, pp 3406–3411
8. California ISO – Folsom, CA 95763-9014. ISO Today. Online www.caiso.com/Pages/TodaysOutlook.aspx
9. Callaway D, Hiskens I (2011) Achieving controllability of electric loads. *Proc IEEE* 99(1):184–199
10. Chen Y (2016) Markovian demand dispatch design for virtual energy storage to support renewable energy integration. PhD thesis, University of Florida, Gainesville, FL
11. Chen Y, Bušić A, Meyn S (2017) Estimation and control of quality of service in demand dispatch. *IEEE Trans Smart Grid* PP(99):1
12. Chen Y, Bušić A, Meyn S (2018) Ergodic theory for controlled Markov chains with stationary inputs. *Ann Appl Probab* 28(1):79–111
13. Chen Y, Bušić A, Meyn S (2017) State estimation for the individual and the population in mean field control with application to demand dispatch. *IEEE Trans Autom Control* 62(3):1138–1149
14. Chertkov M, Chernyak VY (2017) Ensemble control of cycling energy loads: Markov decision approach. In: IMA volume on the control of energy markets and grids. Springer, Berlin
15. Cho I-K, Meyn SP (2010) Efficiency and marginal cost pricing in dynamic competitive markets with friction. *Theor Econ* 5(2):215–239
16. Christakou K, Tomozei D-C, Le Boudec J-Y, Paolone M (2014) GECN: primary voltage control for active distribution networks via real-time demand-response. *IEEE Trans Smart Grid* 5(2):622–631
17. Fairley P (2015) Energy storage: power revolution. *Nature* 526:S102–S104
18. Fleming WH, Mitter SK (1982) Optimal control and nonlinear filtering for nondegenerate diffusion processes. *Stochastics* 8(1):63–77
19. Hao H, Middelkoop T, Barooah P, Meyn S (2012) How demand response from commercial buildings will provide the regulation needs of the grid. In: 50th Allerton conference on communication, control, and computing, pp 1908–1913
20. Hao H, Lin Y, Kowli A, Barooah P, Meyn S (2014) Ancillary service to the grid through control of fans in commercial building HVAC systems. *IEEE Trans Smart Grid* 5(4):2066–2074
21. Hao H, Sanandaji BM, Poolla K, Vincent TL (2015) Aggregate flexibility of thermostatically controlled loads. *IEEE Trans Power Syst* 30(1):189–198
22. Huang M, Malhamé RP, Caines PE (2006) Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun Inf Syst* 6(3):221–251
23. Kadanoff LP (2009) More is the same; phase transitions and mean field theories. *J Stat Phys* 137(5–6):777–797

24. Kárný M (1996) Towards fully probabilistic control design. *Automatica* 32(12):1719–1722
25. Kizilkale A, Malhame R (2013) Mean field based control of power system dispersed energy storage devices for peak load relief. In: 52nd IEEE conference on decision and control (CDC), pp 4971–4976
26. Kizilkale A, Malhame R (2014) A class of collective target tracking problems in energy systems: cooperative versus non-cooperative mean field control solutions. In: IEEE conference on decision and control, pp 3493–3498
27. Lasry J-M, Lions P-L (2007) Mean field games. *Jpn J Math* 2(1):229–260
28. Madani O, Hanks S, Condon A (2003) On the undecidability of probabilistic planning and related stochastic optimization problems. *Artif Intell* 147(1–2):5–34
29. Malhame R, Chong C-Y (1985) Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system. *IEEE Trans Automat Control* 30(9):854–860
30. Mathias J, Kaddah R, Bušić A, Meyn S (2016) Smart fridge/dumb grid? Demand Dispatch for the power grid of 2020. In: Proceedings of the 49th annual Hawaii international conference on system sciences (HICSS), pp 2498–2507
31. Mathias J, Bušić A, Meyn S (2017) Demand dispatch with heterogeneous intelligent loads. In: Proceedings of the 50th annual Hawaii international conference on system sciences (HICSS). arXiv 1610.00813
32. Mathieu J (2012) Modeling, analysis, and control of demand response resources. PhD thesis, University of California at Berkeley
33. Mathieu J, Callaway D (2012) State estimation and control of heterogeneous thermostatically controlled loads for load following. In: 45th international conference on system sciences. IEEE, Hawaii, pp 2002–2011
34. Mathieu J, Koch S, Callaway D (2013) State estimation and control of electric loads to manage real-time energy imbalance. *IEEE Trans Power Syst* 28(1):430–440
35. Meyn S, Barooah P, Bušić A, Ehren J (2013) Ancillary service to the grid from deferrable loads: the case for intelligent pool pumps in Florida. In: Proceedings of the 52nd IEEE conference on decision and control, pp 6946–6953
36. Meyn S, Barooah P, Bušić A, Chen Y, Ehren J (2015) Ancillary service to the grid using intelligent deferrable loads. *IEEE Trans Automat Control* 60(11):2847–2862
37. Pedroncelli R (2011) Frequency regulation compensation in the organized wholesale power markets – FERC 755. FERC Docket Nos. RM11-7-000 and AD10-11-000; Order No. 755 – Online, <http://tinyurl.com/FERC755>, 20 October 2011
38. Puterman ML (1994) Markov decision processes. Wiley, New York
39. Roozbehani M, Dahleh MA, Mitter SK (2012) Volatility of power grids under real-time pricing. *IEEE Trans Power Syst* 27(4):1926–1940
40. Scheppe FC (1978) Power systems ‘2000’: hierarchical control strategies. *IEEE Spectrum* 15:42–47
41. Scheppe F, Tabors R, Kirtley J, Outhred H, Pickel F, Cox A (1980) Homeostatic utility control. *IEEE Trans Power Apparatus Syst PAS-99(3):1151–1163*
42. Sharp J (2012) Electrical load disconnect device with electronic control, December 2012. US Patent 8328110
43. Tindemans SH, Trovato V, Strbac G (2015) Decentralized control of thermostatic loads for flexible demand response. *IEEE Trans Control Syst Technol* 23(5):1685–1700
44. Todorov E (2007) Linearly-solvable Markov decision problems. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in neural information processing systems, vol 19. MIT Press, Cambridge, pp 1369–1376
45. Todorov E (2009) Efficient computation of optimal actions. *Proc Natl Acad Sci* 106(28):11478–11483
46. Totu LC (2015) Large scale demand response of thermostatic loads. PhD thesis, Faculty of Engineering and Science, Aalborg University

47. Weiss P (1907) L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J Phys Theor Appl* 6(1):661–690
48. Ziras C, Vrettos E, Andersson G (2015) Primary frequency control with refrigerators under startup dynamics and lockout constraints. In: IEEE power and energy society general meeting. IEEE, Piscataway, pp 1–5

Disaggregating Load by Type from Distribution System Measurements in Real Time



Gregory S. Ledva, Zhe Du, Laura Balzano, and Johanna L. Mathieu

Abstract An electricity distribution network's efficiency and reliability can be improved using real-time knowledge of the total consumption/production of different load/generator types (e.g., air conditioning loads, lighting loads, photovoltaic generation) within the network. This information could be gathered from additional device-level sensors and communication infrastructure. Alternatively, this information can be inferred using existing network measurements and some knowledge of the underlying system. This work applies two online learning algorithms, dynamic mirror descent (DMD) and dynamic fixed share (DFS), to separate (or disaggregate), in real-time, feeder-level active demand measurements into two components: (1) the demand of a population of residential air conditioners and (2) the demand of the remaining loads served by the feeder. The online learning algorithms include models of the underlying load types, which are generated using historical building-level or device-level data. We develop methods to incorporate model prediction error statistics into the algorithms, develop connections between DMD and Kalman filtering, adapt the algorithms for the energy disaggregation application, and present case studies demonstrating that the algorithms perform disaggregation effectively.

1 Introduction

Power system entities such as utilities and third-party companies can improve an electricity distribution network's reliability and efficiency using real-time knowledge of the mix of load/generation within the distribution network. The mix of load/generation refers to the aggregate consumption/production of different types of load/generators, e.g., air conditioning loads, lighting loads, and photovoltaic generation. For example, a utility can better anticipate and plan for fault-induced

G. S. Ledva · Z. Du · L. Balzano · J. L. Mathieu (✉)

Department of Electrical Engineering and Computer Science, University of Michigan,
Ann Arbor, MI, USA

e-mail: gsledv@umich.edu; zhedu@umich.edu; girasole@umich.edu; jlmath@umich.edu

delayed voltage recovery (FIDVR) caused by motor stalling if it knows the real-time power consumption of small motor loads [2]. Companies that offer power system services via demand response are interested in knowing the time-varying, total electric load available for demand response. This knowledge can help them bid into ancillary services markets, or it could be used as a feedback signal within a load coordination algorithm [6, 12, 23, 26, 30, 32, 42, 43].

We define real-time feeder-level energy disaggregation as the problem of determining the mix of loads/generation connected to a distribution feeder as measurements arrive sequentially in time. This type of energy disaggregation can be accomplished by either computing the mix of loads/generation directly from device-level (i.e., submetering) data or by inferring the mix of loads/generation from distribution network and smart meter data. Acquiring real-time submetering data requires the installation of additional meters for tens of thousands of devices and also requires additional communication infrastructure to transmit the device-level data to a central location for real-time processing. Estimates of the per household costs associated with submetering are \$100 to over \$1,000 [1], which limits its practicality.

Alternatively, the mix of loads/generation can be inferred using existing infrastructure: a small number of distribution network measurements (e.g., the power demand served by each feeder) and historical data collected by smart meters. Smart meters capable of measuring household demand at frequent intervals¹ have been widely installed [29], but their communication limitations prevent their data from being available in real time [1]. However, historical data is available. Device-level demand could be estimated by disaggregating the household-level demand [1].

Inferring the mix of loads/generation can be achieved using online learning algorithms, a class of machine learning algorithms. In the single predictor setting, these algorithms use sequential data (or measurements) to update parameters (referred to here as states) within a predictor, which generates predictions about future data. In a setting with multiple predictors, called prediction with expert advice, algorithms use a defined set of predictors (referred to as experts), and they use the measurements to learn the best expert or best combination of experts, e.g., see [17, 22]. Much of the online learning literature assumes that the optimal state or the best expert does not vary in time, e.g., see [3, 7, 11, 22, 33]. Several papers (e.g., see [17–19, 45, 47]) provide performance bounds on these algorithms when the optimal state or best expert is allowed to be time-varying. However, in this case, the bounds are only meaningful (i.e., they scale sublinearly with respect to time) when the system (i.e., the state or best expert) varies relatively slowly in time.

Online convex programming is a method to solve online learning problems [39], and recent work [15, 38, 41] incorporates dynamic models into the online learning framework. Online convex programming uses a convex objective function to quan-

¹While most meters are currently configured to measure/record average power demand over 15 minute or hourly intervals, they generally have the ability to measure/record average power over much shorter intervals, for example, every minute.

tify the error between the predicted measurement computed by a predictor and the actual measurement. After each measurement is revealed, the predictor is updated as a function of the (possibly time-varying) convex objective function. Methods to solve convex optimization problems have been adapted to solve online convex programming, e.g., see [3, 11, 33]. Recently, [15, 38, 41] developed online convex optimization algorithms that handle highly time-varying systems by incorporating dynamic models of the systems. These algorithms establish performance bounds that depend on the accuracy of the underlying dynamic models rather than the variability of the state, allowing the algorithms to be effective in situations with highly time-varying states.

In this work, we apply two algorithms from [15], dynamic mirror descent (DMD) and dynamic fixed share (DFS) algorithms, to the feeder-level energy disaggregation problem. In our setting, we disaggregate a distribution feeder's demand measurements into two components: (1) the total power demand of a population of air conditioners and (2) the total power demand of all remaining loads served by the distribution feeder. The contributions of this work are as follows: (1) we summarize the DMD and DFS algorithms and provide simple examples of DMD implementations to provide intuition; (2) we develop methods to include model prediction error statistics into the DMD and DFS algorithms; (3) we establish connections between DMD and a discrete-time Kalman filter; and (4) we present simulations that show the effectiveness of the algorithms on the real-time feeder-level energy disaggregation problem and also show the influence of model prediction error statistics on the performance of the algorithms. We presented preliminary work in [27] and developed the data-driven case studies expanded upon in this chapter in [28].

The remainder of the chapter is organized as follows: Section 2 summarizes the problem framework that we consider for real-time feeder-level energy disaggregation and compares this problem framework to building-level energy disaggregation. Section 3 summarizes the DMD and DFS algorithms, discusses the inclusion of prediction error statistics into DMD, and makes comparisons between DMD and Kalman filtering; an appendix presents two simple example implementations of DMD. Section 4 describes the application of DFS to the feeder-level energy disaggregation problem, including a summary of the models used within DFS, algorithm implementation details, and discussion of a number of case studies. Finally, Section 5 presents the conclusions.

2 Framework for Real-Time Feeder-Level Energy Disaggregation

In our framework for real-time feeder-level energy disaggregation, we assume that a power system entity has access to real-time measurements of the active power demand served by a distribution feeder, real-time outdoor temperature

measurements, and historical feeder, temperature, and load data. We assume that the feeder serves both residential and commercial loads, and we also assume that the power system entity's objective is to determine the real-time demand of a population of residential air conditioners, referred to as the AC demand, from the feeder's total demand measurements. The other component of the feeder demand, consisting of the commercial and remaining residential demand, is referred to as the other load (OL) demand. The demand is measured at 1-minute intervals, and the measurements are the time-averaged active power demand over the interval. The real-time outdoor temperature measurements correspond to that of the physical area containing the underlying loads; note that we do not use weather data for individual loads. The temperature measurements are used within some models in real time where the models are parameterized with historical data.

We assume that the power system entity has access to four sources of historical data, which it uses to parameterize models that predict the two demand components during the real-time disaggregation. The historical data includes past feeder demand measurements, past outdoor temperature measurements, historical building-level demand data for both the residences and commercial buildings, and device-level demand data for the residential air conditioners constituting the AC demand. We assume that building- and device-level meters collect demand data at 1-minute intervals, but the data are not available in real time due to communication limitations [1].

The feeder-level energy disaggregation problem has similarities with building-level energy disaggregation [1, 8, 9, 24, 25, 36, 44], also known as non-intrusive load monitoring (NILM) [5, 10, 16, 46, 48], which separates the measured power demand of a building into the demand of individual loads or groups of loads within the building. Building-level energy disaggregation algorithms typically use an aggregate signal that is sampled at high frequencies (e.g., 10 KHz to over 1 MHz) and composed of 10–100 component loads. The algorithms generally leverage assumptions stemming from the relatively small number of underlying loads (e.g., a single device turns on or off per time-step [24] or that step changes can be seen in the aggregate signal [9]). Furthermore, disaggregation is generally performed offline. Building-level energy disaggregation algorithms include supervised and unsupervised approaches. In supervised approaches, historical disaggregated signals are available [4, 5, 8, 16, 25], and unsupervised approaches use only the aggregate signal [13, 21, 24, 40].

In contrast to building-level energy disaggregation, feeder-level energy disaggregation uses an aggregate signal composed of tens of thousands of underlying loads, but we are only interested in disaggregating load by type. We assume that the aggregate signal is sampled less frequently, i.e., on the order of seconds to minutes. The large number of loads and relatively slow measurement frequency renders many of the building-level approaches inadequate.

3 Summary and Discussion of the DMD and DFS Algorithms

An online learning algorithm predicts a system state $\widehat{\boldsymbol{\theta}}_t \in \Theta$ where the domain of the state $\Theta \subset \mathbb{R}^p$ is a bounded, closed, convex feasible set. After a prediction, denoted $\widehat{\boldsymbol{\theta}}_t \in \Theta$, is formed for time-step t , the system produces a measurement $\mathbf{y}_t \in \mathcal{Y} \subset \mathbb{R}^q$. A convex loss function $\ell_t : \Theta \rightarrow \mathbb{R}$ measures the accuracy of the prediction $\widehat{\boldsymbol{\theta}}_t$ with respect to \mathbf{y}_t , and by computing its gradient/subgradient, we can determine how to change the prediction to improve its accuracy with respect to the measurement. The loss function may also contain some known, possibly time-varying, function $h_t : \Theta \rightarrow \mathcal{Y}$ that computes a predicted measurement from the state prediction, i.e., $\widehat{\mathbf{y}}_t = h_t(\widehat{\boldsymbol{\theta}}_t)$. We assume that we can choose the form of the loss function, e.g., if our measurement function is $h_t(\boldsymbol{\theta}) = C\boldsymbol{\theta}$ then we can set $\ell_t(\widehat{\boldsymbol{\theta}}_t) = \|C\widehat{\boldsymbol{\theta}}_t - \mathbf{y}_t\|_2^2$. The goal of the online learning algorithm is to generate a sequence of predictions that result in low cumulative loss, which is the total achieved loss up to the current time-step.

In this section, we first summarize two online learning algorithms, DMD and DFS, which were originally developed in [15]. From a control systems perspective, DMD and DFS are similar to state estimation algorithms that consider a single model and multiple models, respectively, in forming the estimate of $\boldsymbol{\theta}_t$. Following the algorithm descriptions, we discuss the inclusion of prediction error statistics into DMD. We present a number of simple simulations in the appendix to provide intuition on several parameters and functions that can be chosen by the user within DMD and DFS.

3.1 The DMD Algorithm

The DMD algorithm uses a convex optimization formulation and a model to predict the system state at each time-step. The formulation has similarities to a discrete-time Kalman filter in that the Kalman filter and DMD iteratively use a model to advance the prediction to the next time-step and then adjust the prediction once the new measurement is available. Section 3.3 further develops connections between a Kalman filter and DMD.

The DMD algorithm formulation is

$$\widetilde{\boldsymbol{\theta}}_t = \arg \min_{\boldsymbol{\theta} \in \Theta} \eta^s \langle \nabla \ell_t(\widehat{\boldsymbol{\theta}}_t), \boldsymbol{\theta} \rangle + D(\boldsymbol{\theta} \|\widehat{\boldsymbol{\theta}}_t) \quad (1)$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \Phi(\widetilde{\boldsymbol{\theta}}_t) \quad (2)$$

where (1) adjusts the prediction using the new measurement and (2) is the model-based update that advances the adjusted prediction. In (1), $\widetilde{\boldsymbol{\theta}}_t$ is the adjusted

prediction, and we minimize the right-hand side over the variable $\boldsymbol{\theta}$. The parameter $\eta^s > 0$ is a step-size parameter, $\langle \cdot, \cdot \rangle$ is the standard dot product, $\nabla \ell_t(\widehat{\boldsymbol{\theta}}_t)$ is a subgradient of the convex loss function, and $D(\boldsymbol{\theta} \parallel \widehat{\boldsymbol{\theta}}_t)$ is a Bregman divergence, which penalizes the deviation between the optimizer $\boldsymbol{\theta}$ and the prediction $\widehat{\boldsymbol{\theta}}_t$. In (2), the model $\Phi(\cdot)$ advances the adjusted prediction $\widetilde{\boldsymbol{\theta}}_t$ to the next time-step. The step-size η^s influences how aggressively DMD adjusts $\widetilde{\boldsymbol{\theta}}_t$ to match the measurements versus trusting the model-based prediction; see the appendix for an illustrative example.

3.2 The DFS Algorithm

The DFS algorithm incorporates N^{mdl} experts where each expert contains one model from a set of N^{mdl} models $\mathcal{M}^{\text{mdl}} = \{1, \dots, N^{\text{mdl}}\}$. We use \mathcal{M}^{mdl} to denote both the set of experts and their corresponding models. DFS assumes that DMD produces each expert's prediction and then applies the Fixed Share algorithm [17] to form an overall prediction of the system state. The Fixed Share algorithm formulation is

$$w_{t+1}^m = \frac{\lambda}{N^{\text{mdl}}} + (1 - \lambda) \frac{w_t^m \exp(-\eta^r \ell_t(\widehat{\boldsymbol{\theta}}_t^m))}{\sum_{j=1}^{N^{\text{mdl}}} w_t^j \exp(-\eta^r \ell_t(\widehat{\boldsymbol{\theta}}_t^j))} \quad m \in \mathcal{M}^{\text{mdl}} \quad (3)$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \sum_{m \in \mathcal{M}^{\text{mdl}}} w_{t+1}^m \widehat{\boldsymbol{\theta}}_{t+1}^m \quad (4)$$

where (3) advances the weight of each expert and (4) forms the overall prediction as a weighed combination of the individual experts' estimates. In (3), $\widehat{\boldsymbol{\theta}}_t^m$ is the prediction of expert m at time-step t , w_t^m is the weight associated with expert m , $\lambda \in (0, 1)$ is a user-defined parameter that influences the weight that is shared among experts, and $\eta^r > 0$ is a user-defined parameter that influences how rapidly the algorithm can shift weight between the experts. DFS assumes that $\widehat{\boldsymbol{\theta}}_t^m$ is the value computed in (2) using model $\Phi^m(\cdot)$ for $m \in \mathcal{M}^{\text{mdl}}$. The weight w_t^m is based on each expert's total loss up until time t and the weight that is shared among all of the experts. Parameter λ controls the extent to which a single model can dominate the prediction: with λ near zero, a single model can dominate, and with λ near one, the overall prediction is forced to be close to the average of the individual predictions. For a given sequence of losses, the parameter η^r controls how rapidly the weights adjust, with larger values leading to faster weight changes. Setting η^r too high may lead to over-fitting, i.e., the weights may become erratic.

3.3 Including Model Prediction Error Statistics into DMD

In this section, we describe how the user can construct the DMD updates to include statistical information about the prediction errors. To support this claim, we design the DMD updates to match those of a Kalman filter, which specifically accounts for zero-mean and normally distributed errors in its model-based update and measurement predictions. Future work will develop methods to include error statistics corresponding to other probability distributions within DMD.

The loss function and divergence used within DMD must be convex, but their specific form can be chosen by the user. In choosing the convex loss and divergence functions, the user implicitly makes assumptions about statistics describing the model prediction accuracy within the measurement-based update of (1). For example, choosing the divergence function to be a squared ℓ_2 -norm, i.e., $D(\boldsymbol{\theta} \|\widehat{\boldsymbol{\theta}}_t) = \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\|_2^2$, treats all errors equally and weights larger errors more. This corresponds to the case where the error covariance matrix is equal to an identity matrix. However, using a Mahalanobis distance, i.e., a weighted squared ℓ_2 -norm, such as $(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)^T \widehat{P}_t^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)$ with some positive-definite matrix \widehat{P}_t , assumes that the errors in the model-based update have a covariance of \widehat{P}_t .

A Kalman filter's objective function is to minimize the mean-squared estimation error of the state under assumptions that the system is linear and that state and measurement prediction errors (often referred to as process and measurement noise) are independent in time, zero-mean, normally distributed, and independent from one another. The user must specify the error covariance matrices used within the closed-form update equations. Alternatively, within DMD, the user has the flexibility to select functions (rather than matrices). DMD can produce update equations identical to those of a Kalman filter by making the same assumptions required by a Kalman filter and then appropriately selecting the loss and divergence functions. In the remainder of this section, we summarize the discrete-time Kalman filter and show how to choose the model, divergence function, and loss function within DMD to produce update equations equal to those of a Kalman filter.

A discrete-time Kalman filter assumes an underlying system is [14, p. 190]

$$\boldsymbol{\theta}_{t+1} = A_t \boldsymbol{\theta}_t + \boldsymbol{w}_t \quad (5)$$

$$\mathbf{y}_t = C_t \boldsymbol{\theta}_t + \boldsymbol{v}_t \quad (6)$$

where \boldsymbol{w}_t is the process noise (which includes modeling error) and \boldsymbol{v}_t is the measurement noise. The formulation assumes that $\boldsymbol{w}_t \sim \mathcal{N}(\mathbf{0}, Q_t)$, $\boldsymbol{v}_t \sim \mathcal{N}(\mathbf{0}, R_t)$, and that A_t , C_t , Q_t , and R_t are known. The notation $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\alpha}, \beta)$ indicates that a random variable $\boldsymbol{\phi}$ is sampled from a normal distribution with mean $\boldsymbol{\alpha}$ and a symmetric, positive-definite covariance β .

A Kalman filter uses the assumptions on the underlying system and the known system parameters to estimate θ_t at each time-step while minimizing the mean-squared estimation error. The resulting update equations are [14, p. 190]

$$\tilde{\theta}_t = \hat{\theta}_t + \hat{P}_t C_t^T \left[C_t \hat{P}_t C_t^T + R_t \right]^{-1} (y_t - C_t \hat{\theta}_t) \quad (7)$$

$$\tilde{P}_t = \hat{P}_t - \hat{P}_t C_t^T \left[C_t \hat{P}_t C_t^T + R_t \right]^{-1} C_t \hat{P}_t \quad (8)$$

$$\hat{\theta}_{t+1} = A_t \tilde{\theta}_t \quad (9)$$

$$\hat{P}_{t+1} = A_t \tilde{P}_t A_t^T + Q_t \quad (10)$$

where $\hat{\theta}_t$ is the *a priori* state estimate, $\tilde{\theta}_t$ is an *a posteriori* state estimate, \hat{P}_t is the *a priori* estimation error covariance, and \tilde{P}_t is the *a posteriori* estimation error covariance. If the matrices within the system model are time-invariant, \hat{P}_t converges to a steady-state value, denoted \bar{P} . A steady-state Kalman filter uses \bar{P} in (7).

We choose the model, divergence function, and loss function within DMD to construct the DMD updates (1) and (2). We first set the DMD model to $\Phi(\hat{\theta}_t) = A_t \hat{\theta}_t$, which makes (2) the same as (9). Note that this corresponds to assuming the model is linear with a state-update matrix A_t , as in a Kalman filter. We then set the divergence and loss function to

$$D(\theta \|\hat{\theta}_t) = \frac{1}{2} \|(\hat{P}_t)^{-\frac{1}{2}} (\theta - \hat{\theta}_t)\|_2^2 \quad (11)$$

$$\ell_t(\hat{\theta}_t) = \frac{1}{2} \|(\hat{P}_t^y)^{-\frac{1}{2}} (C \hat{\theta}_t - y_t)\|_2^2 \quad (12)$$

where \hat{P}_t and \hat{P}_t^y are symmetric, positive-definite, covariance matrices corresponding to the model prediction errors and the measurement prediction errors, respectively. The quantity $G^{-\frac{1}{2}} = U (\Sigma^{-\frac{1}{2}}) U^T$ denotes a matrix square root of an arbitrary symmetric positive-definite matrix G , where U is orthonormal and Σ a diagonal matrix with positive entries on the diagonal. The square roots of \hat{P}_t and \hat{P}_t^y are also symmetric and positive definite [14]. Given the assumptions thus far, the matrices A_t , \hat{P}_t , and \hat{P}_t^y can be treated as parameters that are known at each time-step within DMD.

Given our choice of model, divergence function, and loss function, we can use (1) to derive a closed-form DMD update equation, which is the same as (7). We start from (1), substitute the divergence function, and then solve the convex program by finding the value of θ that sets the gradient of the convex objective function equal to 0. Following this, we substitute the gradient of the loss function. These steps result in

$$\tilde{\theta}_t = \arg \min_{\theta \in \Theta} \eta^s \langle \nabla \ell_t(\hat{\theta}_t), \theta \rangle + D(\theta \|\hat{\theta}_t) \quad (13)$$

$$= \hat{\theta}_t + \eta^s \hat{P}_t (-\nabla \ell_t(\hat{\theta}_t)) \quad (14)$$

$$= \hat{\theta}_t + \eta^s \hat{P}_t C_t^T (\hat{P}_t^y)^{-1} (y_t - C \hat{\theta}_t) \quad (15)$$

where $\hat{P}_t^y = (C_t \hat{P}_t C_t^T + R_t)$. Finally, setting $\eta^s = 1$ produces the same update as (7). Note that \hat{P}_t and \hat{P}_t^y are the same covariances as used in the Kalman filter. Their values and updates are assumed known, which is the same assumption we make when we use the Kalman filter.

4 Application of DFS to Real-Time Feeder-Level Energy Disaggregation

In this section, we apply the DFS algorithm to the problem framework described in Section 2. In the following, Section 4.1 details the construction of the underlying system (i.e., the plant), Section 4.2 describes the construction of the models used within the algorithms, Section 4.3 describes the implementation of the online learning algorithms and a Kalman filter, and Section 4.4 presents some case studies investigating the algorithm's performance. Note that we construct these case studies to investigate the effectiveness of DFS within the problem formulation in comparison to the effectiveness of a Kalman filter. Modifying the models developed and used within DFS as well as the user-defined loss and divergence functions within DMD may provide performance improvements.

4.1 Plant Construction

The plant, which is our representation of the underlying physical system, is composed of the active power demand of a set of commercial and residential loads connected to a distribution feeder. The time series consist of n^{steps} 1-minute time-steps over the course of 1 day, resulting in $n^{\text{steps}} = 1440$. We denote the measured total demand of the feeder as $y_t \in \mathbb{R}$, the AC demand as $y_t^{\text{AC}} \in \mathbb{R}$, the residential component of the OL demand as $y_t^{\text{OL, res}} \in \mathbb{R}$, and the commercial component of the OL demand as $y_t^{\text{OL, com}} \in \mathbb{R}$. The total demand is $y_t = y_t^{\text{AC}} + y_t^{\text{OL}}$ where $y_t^{\text{OL}} = y_t^{\text{OL, res}} + y_t^{\text{OL, com}}$.

We construct the y_t^{AC} , $y_t^{\text{OL, res}}$, and $y_t^{\text{OL, com}}$ time series using a feeder model, household demand data, air conditioner demand data, and commercial building demand data; additional details on the time series construction can be found in [28]. We assume the average daily active power demand of the commercial and residential loads is 5.8 MW and 2.1 MW, respectively, which is based on the feeder model R5-25.00-1 from GridLAB-D's feeder taxonomy [37]. The residential demand

consists of household demand data and air conditioner demand data from the Pecan Street, Inc. Dataport [35], where the aggregate residential demand corresponds to the summed daily demand of a set of 2,499 households. The AC demand y_t^{AC} corresponds to the summed demand of 2,269 primary air conditioner and blower units within those households. The residential OL demand $y_t^{\text{OL, res}}$ consists of the remainder of the aggregate household demand not included within the AC demand. The commercial OL demand $y_t^{\text{OL, com}}$ is the scaled sum of the whole-building demand from a big box retail store and a municipal building in the California Bay Area. We neglect losses in the power network, which, if included, would be part of the OL demand. We determine the set of houses, the air conditioner population, and the scaling factor for the commercial demand using data from August 3, as detailed in [28].

We also construct time series of the outdoor temperature for the physical area corresponding to the demand data, which is used in some models described in Section 4.2. The residential data corresponds to Austin, TX, and we use outdoor temperature data from [35]. The outdoor temperature for the commercial demand comes from the Concord, CA National Weather Service station [34].

4.2 Model Construction

In this section, we describe the models used within the DFS algorithm, where [28] provide more details. Section 4.2.1 details the linear regression models used to predict both the AC and OL demand, while Section 4.2.2 details the linear dynamic system models, specifically linear time-varying (LTV) system models used to predict the AC demand.

4.2.1 Linear Regression Models

The linear regression models of the AC and OL demand all have the same general form

$$\hat{y}_t^* = \boldsymbol{\alpha}^T \boldsymbol{\beta}_t$$

where \hat{y}_t^* is the prediction of the AC or OL demand, $\boldsymbol{\beta}_t$ is a vector of input features at time t , and $\boldsymbol{\alpha}$ is a vector of coefficients. The input features are the explanatory variables. The vector of coefficients forms a weighted combination of the input features, and their values are determined by applying least-squares error minimization to historical data including the input features and the demand signal. Examples of input features used within the models below include calendar variables such as time of week and weather variables such as outdoor temperature.

Below, we summarize the input features used in several regression models, including a simple regression model that forms a lookup table based on the time of day (TOD) and two multiple linear regression (MLR) models that use a vector of input features. The TOD regression models were generated using data from the week preceding August 3. We use residential data from June 24 to August 2, 2015, and commercial data from June 24 to August 2, 2009, to generate the MLR models, and we exclude anomalous data such as those corresponding to holidays.

TOD OL Demand Model The TOD OL demand model corresponds to a lookup table of OL demand predictions based on the time of day, generated by smoothing OL demand data from previous days. We construct TOD models for each weekday denoted, $\Phi^{\text{OL, Mon}}$, $\Phi^{\text{OL, Tues}}$, $\Phi^{\text{OL, Wed}}$, $\Phi^{\text{OL, Thu}}$, $\Phi^{\text{OL, Fri}}$, and their respective predictions are $\hat{y}_t^{\text{OL, Mon}}$, $\hat{y}_t^{\text{OL, Tues}}$, $\hat{y}_t^{\text{OL, Wed}}$, $\hat{y}_t^{\text{OL, Thu}}$, $\hat{y}_t^{\text{OL, Fri}}$.

MLR OL Demand Model The MLR OL demand model forms its predictions using two sets of input features and coefficient vectors, one for the residential OL demand and one for the commercial OL demand, where both sets of input features include calendar- and weather-based values. Two sets of input features are necessary because the OL demand data corresponds to two different physical areas. The input features for the residential OL demand are a time-of-week indicator vector, the outdoor temperature for Austin, TX, and the measured total demand at the last time-step, y_{t-1} . The commercial component of the model corresponds to “Baseline Method 1” from [31]. The input features for the commercial OL demand are a time-of-week indicator vector and the outdoor temperature of Concord, CA, at the given time of week. Whereas the residential component of the model has a single regression parameter for the outdoor temperature, the commercial component has separate temperature-based coefficients for each time of week. We denote the MLR OL demand model and its predictions as $\Phi^{\text{OL, MLR}}$ and $\hat{y}_t^{\text{OL, MLR}}$, respectively.

MLR AC Demand Model The input features of the MLR AC demand model $\Phi^{\text{AC, MLR}}$, with predictions $\hat{y}_t^{\text{AC, MLR}}$, are a time-of-week indicator vector and the lagged outdoor temperature for Austin, TX, raised to the first through fourth powers, i.e., the model includes a fourth-order polynomial in lagged temperature. The lag was chosen to maximize the cross correlation between the temperature and AC demand in the training data.

4.2.2 Linear Dynamic System Models

We also use two LTV dynamic system models to compute predictions of the AC demand. The on/off cycling of air conditioners varies with the outdoor temperature, and we generate LTV models from sets of linear time-invariant (LTI) models, originally developed in [20, 32], each corresponding to a different outdoor temperature. The first LTV model, denoted $\Phi^{\text{AC, LTV}^1}$, generates predictions, denoted $\hat{y}_t^{\text{AC, LTV}^1}$,

using a set of LTI models $\mathcal{M}^{\text{LTI1}}$ and the lagged, outdoor temperature. The second LTV model, denoted $\Phi^{\text{AC}, \text{LTV2}}$, generates predictions $\hat{y}_t^{\text{AC}, \text{LTV2}}$, using a separate set of LTI models $\mathcal{M}^{\text{LTI2}}$ and the time-averaged outdoor temperature over a window of previous minutes. The lag and the window are chosen to maximize the performance of the models on the training set. Both LTV models have the form

$$\begin{aligned}\hat{x}_{t+1}^* &= A_t^* \hat{x}_t^* \\ \hat{y}_t^{\text{AC}, *} &= C_t^* \hat{x}_t^*\end{aligned}$$

where superscript $*$ is replaced by LTV1 for $\Phi^{\text{AC}, \text{LTV1}}$ or LTV2 for $\Phi^{\text{AC}, \text{LTV2}}$. The first element of the vector $\hat{x}_t^* \in \mathbb{R}^2$ captures the portion of air conditioners that are drawing power, i.e., those that are on, and the second element captures the portion of air conditioners not drawing power, i.e., those that are off. The matrix A_t^* includes the probabilities that a given air conditioner (1) switches on during the time-step, (2) switches off during the time-step, (3) remains on, or (4) remains off. The matrix C_t^* scales the portion of air conditioners that are drawing power by an average power demand value to compute the prediction $\hat{y}_t^{\text{AC}, *}$. The LTI models in $\mathcal{M}^{\text{LTI1}}$ and $\mathcal{M}^{\text{LTI2}}$ have the same form as the LTV models, with time-invariant matrices identified from data corresponding to narrow ranges around specific outdoor temperatures. The time-varying matrices A_t^* and C_t^* are computed by linearly interpolating the elements of the two closest LTI models (where “closest” is measured in terms of lagged or average temperature). The LTI models are computed using air conditioner demand data from May 2 to August 2, 2015.

4.3 Algorithm Implementation Details

In this section, we describe the implementation of three algorithms used for the feeder-level energy disaggregation problem. First, we describe the implementation of a Kalman filter, which we use as the benchmark for the case studies presented in Section 4.4. We then describe an algorithm, referred to as P-DFS, that includes a modified version of DMD, referred to as P-DMD for pseudo-DMD. P-DMD includes measurement-based updates and model-based predictions but modifies the DMD equations (1) and (2) allowing us to include models of various forms, e.g., both LTV and MLR models, within the Fixed Share algorithm. Following this, we describe the DFS implementation. Within DFS, an expert applies DMD when the AC demand is modeled using an LTV model and P-DMD when the AC demand is modeled using an MLR model. Each of the methods detailed below incorporates model prediction error statistics explicitly. Several methods for constructing the covariances are detailed and investigated in Section 4.4. Note that in all implementations, we construct the convex program within DMD to

have a closed-form solution. Given this, the computational complexity of the DFS implementation is similar to that of a set of Kalman filters.

In all three algorithms, the model of the feeder consists of one AC demand model $\Phi^{AC}(\cdot)$ paired with one OL demand model $\Phi^{OL}(\cdot)$, i.e., the model is $\Phi(\cdot) = \{\Phi^{AC}(\cdot), \Phi^{OL}(\cdot)\}$. P-DFS and DFS use a set of models \mathcal{M}^{DFS} that consists of every pair of AC and OL demand models described in Section 4.2. The Kalman filter implementation applies to a set of models \mathcal{M}^{KF} that includes every possible pairing of an LTV AC demand model with an OL demand model.

4.3.1 Kalman Filter

The Kalman filter uses an LTV model to describe the underlying system, estimates the state of the AC demand model, i.e., $\theta_t = x_t^*$, and uses a pseudo-measurement of the AC demand $\tilde{y}_t^{AC} = y_t - \hat{y}_t^{OL}$ to adjust the model-based estimate where \hat{y}_t^{OL} is the predicted OL demand. A time-invariant process noise covariance Q is computed for each dynamic AC demand model using the historical AC demand measurements. In computing Q , the true state at each time-step is calculated using the measured AC demand and the AC demand model's matrices. The pseudo-measurement \tilde{y}_t^{AC} contains noise due to prediction errors in \hat{y}_t^{OL} , and a separate time-invariant covariance R is computed for each OL model. We compute Q and R using data for the week preceding August 3. We implement one Kalman filter for each model combination within \mathcal{M}^{KF} .

4.3.2 P-DFS Method

The models developed for this work have a variety of forms and different underlying parameters influencing the demand predictions. As a result, it is difficult to define a θ_t that is common across all models. To overcome this, we modify the DMD algorithm to decouple model-based updates and measurement-based updates, meaning the measurement-based updates do not influence the model-based updates. This allows the algorithm to be applied to the output of a given model, e.g., the demand predictions, rather than some underlying parameter while operating in the spirit of DMD. We first proposed this idea in [28].

We modify the model-based and measurement-based updates in DMD, i.e., (1) and (2), to formulate P-DMD, which is used within DFS to form the overall estimate. The P-DMD formulation is

$$\hat{\kappa}_{t+1}^m = \arg \min_{\theta \in \Theta} \eta^s \left\langle \nabla \ell_t(\hat{\theta}_t^m), \theta \right\rangle + D(\theta || \hat{\kappa}_t^m) \quad (16)$$

$$\check{\theta}_{t+1}^m = \Phi(\check{\theta}_t^m) \quad (17)$$

$$\hat{\theta}_{t+1}^m = \check{\theta}_{t+1}^m + \hat{\kappa}_{t+1}^m \quad (18)$$

for $m \in \mathcal{M}^{\text{DFS}}$. The value $\hat{\kappa}_t^m$ accumulates adjustments to the estimate $\hat{\theta}_t^m$ for model m based on the measurements, and we set $\hat{\theta}_0^m = \mathbf{0}$. The value $\check{\theta}_t^m$ is an open-loop state prediction, meaning that the measurements do not influence the prediction (in contrast with DMD), and (17) is the model-based update. Finally, (18) incorporates the accumulated measurement-based adjustment $\hat{\kappa}_t^m$ into the model-based prediction. The AC and OL demand models generate their predictions independently from one another, and so (18) can be rewritten as

$$\hat{\theta}_{t+1}^m = \Phi(\check{\theta}_t^m) + \hat{\kappa}_{t+1}^m \quad (19)$$

$$= \begin{bmatrix} \Phi^{\text{AC}}(\check{\theta}_t^m) \\ \Phi^{\text{OL}}(\check{\theta}_t^m) \end{bmatrix} + \hat{\kappa}_{t+1}^m. \quad (20)$$

The Fixed Share equations (3) and (4) are then applied to the predictions.

In P-DFS, θ_t is the AC and OL demand, i.e., $\theta_t = [y_t^{\text{AC}} \ y_t^{\text{OL}}]^T$. We choose the loss function to be (12) and divergence function to be (11) with $\hat{\kappa}_t^m$ as the second argument rather than $\hat{\theta}_t^m$. The resulting closed-form update (16) is

$$\hat{\kappa}_{t+1}^m = \hat{\kappa}_t^m + \eta^s \hat{P}_t C^T (\hat{P}_t^y)^{-1} (y_t - C\hat{\theta}_t) \quad (21)$$

where $C = [1 \ 1]$. The estimation error covariance $Q_t \in \mathbb{R}^{2 \times 2}$ and the measurement noise covariance $R_t \in \mathbb{R}^1$ are used to compute \hat{P}_t and \hat{P}_t^y . We set $Q_t = \text{diag}(R_t^{\text{AC}}, R_t^{\text{OL}})$, where $\text{diag}(\cdot)$ forms a diagonal matrix from the scalar arguments. The values $R_t^{\text{AC}} \in \mathbb{R}$ and $R_t^{\text{OL}} \in \mathbb{R}$ correspond to the variances of the AC and OL demand models' prediction errors. We detail several sets of assumptions and methods for computing the parameters R_t , R_t^{AC} , R_t^{OL} , \hat{P}_t , and \hat{P}_t^y in Section 4.4.

4.3.3 DFS Method

This method also uses the set of models \mathcal{M}^{DFS} . The formulation applies DMD to the LTV AC demand models and P-DMD to all other models, including the OL demand models. The individual model-based estimates are then used as expert predictions within the Fixed Share algorithm. We set $\theta_t = [(x_t^*)^T \ y_t^{\text{OL}}]^T \in \mathbb{R}^3$, where $*$ is LTV1 or LTV2, allowing inclusion of the LTV model dynamics within (1). The model-based update is

$$\hat{\theta}_{t+1}^m = \begin{bmatrix} \Phi^{\text{AC}}(\tilde{\theta}_t^m) \\ \Phi^{\text{OL}}(\check{\theta}_t^m) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \hat{\kappa}_{t+1}^m$$

where we update the AC demand predictions using DMD and the OL demand predictions using P-DMD. The closed-form measurement-based update of the AC

demand component is (15). The estimation error covariance $Q_t \in \mathbb{R}^{3 \times 3}$ and the measurement noise covariance $R_t \in \mathbb{R}$ are used to compute \hat{P}_t and \hat{P}_t^y . The process noise matrix is $Q_t = \text{blkdiag}(Q_t^{\text{AC}}, R_t^{\text{OL}})$ where $\text{blkdiag}(\cdot)$ constructs a block diagonal matrix from the arguments. The matrix $Q_t^{\text{AC}} \in \mathbb{R}^{2 \times 2}$ corresponds to the process noise of the AC demand model. We use $A_t = \text{blkdiag}(A_t^\star, 0) \in \mathbb{R}^3$ to update \hat{P}_t and \hat{P}_t^y , which assumes that errors in the AC demand model are decoupled from errors in the OL demand model and that the errors of the OL demand model are independent at each time-step. We detail several methods for constructing Q_t^{AC} , R_t^{OL} , \hat{P}_t , and \hat{P}_t^y in Section 4.4.

4.4 Case Studies

In this section, we describe the setup and summarize the results for a set of case studies investigating the performance of DFS and P-DFS on the feeder-level energy disaggregation problem. We simulate a set of days using data from August 3–5, 10–14, 17, and 18 where we excluded weekends and days on which demand response actions were taken by the commercial buildings. The Kalman filters are used as benchmarks. Specifically, we denote BKF as the “best” Kalman filter achieving the lowest *ex post* root mean square estimation error (RMSEE), and we denote AKF as the average RMSEE of the set of Kalman filters. For an arbitrary time series ψ_t and its estimate $\hat{\psi}_t$ over n^{steps} time-steps, the RMSEE is defined as

$$\varepsilon^{\text{RMSE}} = \sqrt{\sum_{t=1}^{n^{\text{steps}}} (\psi - \hat{\psi}_t)^2 / n^{\text{steps}}}. \quad (22)$$

Table 1 lists the values of the parameters η^s and η^r used in each scenario; we set $\lambda = 1.0 \times 10^{-5}$ in all scenarios. We tuned η^r using the simulation for August 3, where the goal was to achieve fast weight transitions without over-fitting, i.e., without erratic weights. Qualitative tuning is appropriate as an optimal value for a given simulated day is not necessarily the optimal value for other simulated days. Parameter η^s was tuned similarly. In the next subsection, we detail three methods for constructing the covariances, referred to as “identity,” “historical,” and “real time” in Table 1.

Table 1 Parameters η^s and η^r used in DFS and P-DFS

Method	P-DFS	P-DFS	P-DFS	DFS	DFS	DFS
Covariance	Identity	Historical	Real time	Identity	Historical	Real time
η^s	0.4	0.5	0.5	0.013	0.5	1.0
η^r	1.0×10^{-5}	10	1.0×10^{-3}	1.0×10^{-5}	10	1.0×10^{-3}

4.4.1 Covariances for DFS and P-DFS

In this section, we detail three methods for generating the covariance matrices used within the DFS and P-DFS algorithms. The first method does not explicitly include any model prediction error statistics into the measurement-based updates of DMD and P-DMD. The second method uses historical data from the week preceding August 3 to compute covariance matrices. The third method uses an unrealistic assumption, i.e., that the total, AC, and OL demand are measured at each time-step and used to compute the exact covariance at each time-step. The details of each method are as follows.

1. Identity: we assume that \hat{P}_t and \hat{P}_t^y are appropriately sized identity matrices for both DFS and P-DFS.
2. Historical: DFS and P-DFS assume that the process noise covariance is time-invariant, i.e., $Q_t = Q$ and that the measurement noise covariance is $R_t \approx 0$ as the total demand measurements are accurate. The covariances Q^{AC} , R^{AC} , and R^{OL} used within the two variations of Q are computed using historical estimation errors, and Q^{AC} is used within the Kalman filter. DFS updates \hat{P}_t according to (8) and (10), and P-DFS sets $\hat{P}_t = Q$. Both methods set $\hat{P}_t^y = (C\hat{P}_t C^T + R_t)$.
3. Real time: DFS and P-DFS assume that $\hat{P}_t = Q_t$ where the covariances are computed at each time-step using measurements of the AC and OL demand. Variance R_t is computed at each time-step using measurements of the total demand. Both algorithms set $\hat{P}_t^y = (C\hat{P}_t C^T + R_t)$.

4.4.2 Results

We next summarize the results for each scenario described above. Figure 1 presents time series of the total demand, OL demand, AC demand, their respective estimates, and the model weights from the August 4 simulation while running P-DFS with covariances generated from historical data. In Figure 1d, Φ^{Other} is used to denote the combined weight of all model combinations not explicitly specified. Table 2 summarizes the mean, minimum, and maximum RMSEE for each demand component across the simulated days and scenarios. Figure 2 presents time series of the AC demand and various estimates across several scenarios from the August 11 simulations.

From Figure 1c, it is clear that, in this case, the P-DFS algorithm effectively estimates the AC demand in real time. In this scenario, BKF achieves an RMSEE of 148.4 kW for the AC demand, and the P-DFS algorithm performs similarly, achieving an RMSEE of 155.0 kW for the AC demand. It should be noted that the P-DFS algorithm is determining the model of the underlying system in real time, as can be seen in Figure 1d. Alternatively, the BKF algorithm selects the most accurate model after the simulation, which is not feasible in practice. For comparison, AKF achieves an RMSEE of 173.1 kW. The weights within P-DFS are initially dominated by Φ^{Other} , which makes sense as the weight of each model

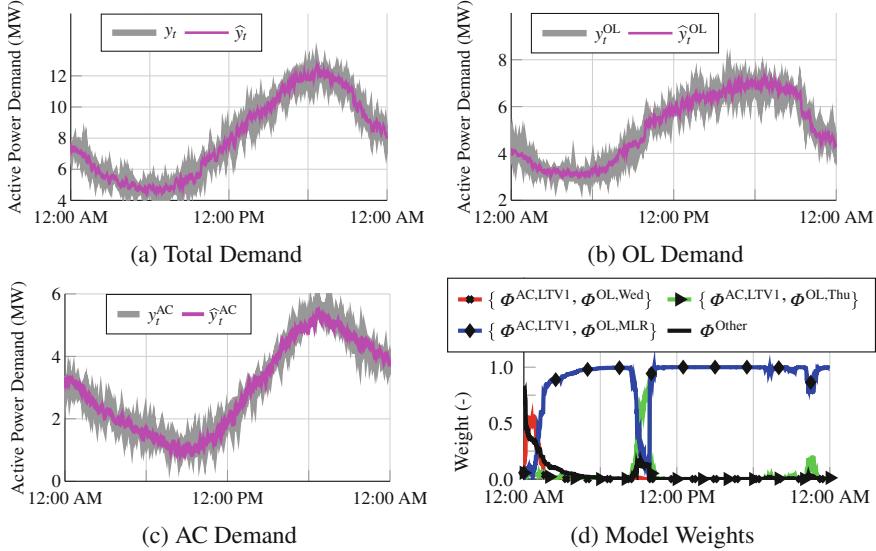


Fig. 1 Time series of the total, OL, and AC demands versus their estimates as well as times series of the weights from the August 4 simulation while running P-DFS with historical covariances

Table 2 Mean, minimum (Min), and maximum (Max) RMSEE in kW over 10 simulated days for each algorithm and covariance computation method

Method	Covariance	Total Demand			AC Demand			OL Demand		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
P-DFS	Identity	88.9	100.0	110.5	151.0	220.6	325.8	150.8	222.3	327.2
P-DFS	Historical	98.4	114.8	123.2	155.0	252.2	371.5	150.2	250.1	372.5
P-DFS	Real time	146.6	154.3	168.4	120.2	125.3	131.8	104.8	114.5	130.5
DFS	Identity	175.4	199.1	224.8	194.2	230.9	314.5	145.0	216.2	312.7
DFS	Historical	100.5	119.5	126.1	192.0	259.8	311.5	190.6	265.5	320.2
DFS	Real time	120.8	125.2	129.1	104.0	116.5	140.1	96.6	109.4	131.9
BKF	Historical	—	—	—	148.4	195.3	318.9	—	—	—
AKF	Historical	—	—	—	173.1	259.4	357.5	—	—	—

combination is initialized to the same value. As the simulation progresses, the weight shifts to $\{\Phi^{\text{AC,LTV1}}, \Phi^{\text{OL,MLR}}\}$, which is the most accurate model. At points of the simulation, it loses accuracy, and the weight shifts to other model combinations during those times. The total demand is estimated closely, which can be achieved based on the parameter settings as discussed in Section ‘‘Varying the Gradient Descent Step Size η^s ’’ in Appendix. Finally, it should be noted that while P-DFS did not achieve lower RMSEE than BKF in this case, in some cases it does outperform BKF.

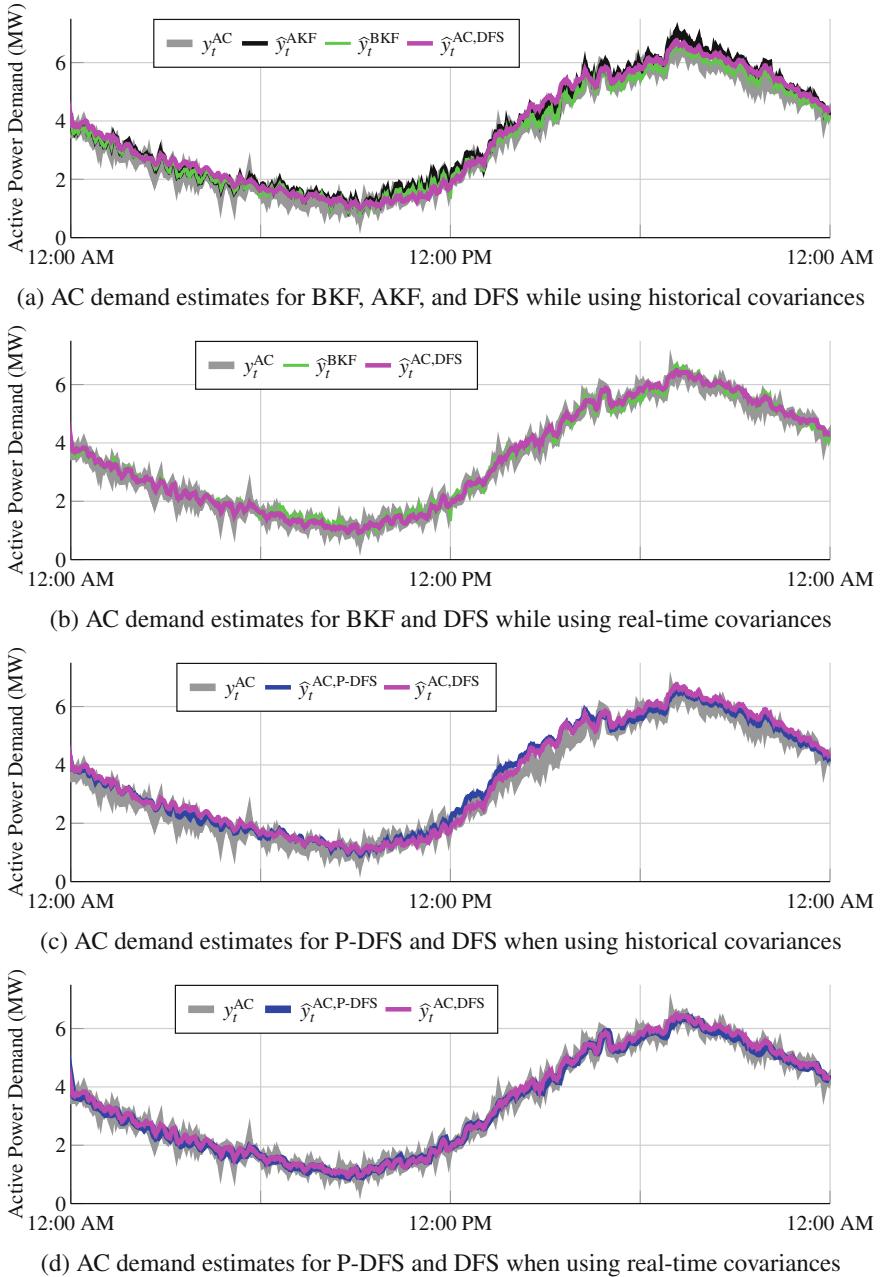


Fig. 2 Time series of the AC demand and various estimates from the August 11 simulations

As Table 2 shows, P-DFS achieves AC demand RMSEEs that are worse than BKF but generally better than the AKF when using realistic (i.e., historical) covariance data. DFS achieves AC demand RMSEE that is comparable to the AKF. Figure 2a shows time series for AKF, BKF, and DFS using historical covariances. When using unrealistic (i.e., real-time) covariance data, both DFS and P-DFS outperform BKF, which is still using historical data to compute the covariance matrices. An example of this is shown in Figure 2b.

Figure 2c provides example time series of the AC demand estimates for P-DFS and DFS when using historical covariances, and Figure 2d provides similar example time series when using real-time covariance data. As can be seen in Figure 2c, the P-DFS algorithm generally achieves better RMSEE for the AC demand than the DMD algorithm. However, as can be seen in Figure 2d, DFS achieves lower RMSEE than P-DFS when real-time errors are used to generate the covariance matrices. Part of the reasoning for this is that the LTV AC demand models only include two states, and for a given outdoor temperature, the models rapidly converge to a steady-state value. When running DFS, this means that the measurement-based adjustment at a given time-step may not have an effect on the model's predictions after several time-steps. Alternatively, the P-DFS formulation continually adjusts the model predictions based on its accuracy, and by separating these adjustments from the model, these adjustments persist.

Also, our method of computing the covariances with historical data degrades performance. This implies that our assumptions regarding the errors are overly coarse. However, with the inclusion of unrealistically accurate covariance information, which is done when using real-time covariance data, the performance of the DFS and P-DFS algorithms improves dramatically.

5 Conclusions

In this chapter, we summarized the real-time feeder-level energy disaggregation problem and an online learning algorithm, dynamic fixed share (DFS), that we adapted and applied to the problem. It was shown that the dynamic mirror descent (DMD), which is used within the DFS algorithm, can be constructed to be equivalent to a discrete-time Kalman filter through proper choice of user-defined functions and parameters. In addition, simple examples were constructed to illustrate aspects of the DMD algorithm. Two implementations of DFS-based algorithms were described. The first modifies the DFS algorithm to incorporate combinations of models with different model structures resulting in estimates of output, rather than the state. The second implemented the original DFS algorithm. Finally, case studies were presented that indicate the online learning algorithms are capable of performing real-time feeder-level energy disaggregation and that model prediction error statistics can be effectively incorporated into these algorithms.

Future work will further explore connections between Kalman filtering and online learning methods, enabling application of results across both well-studied fields. Another topic of future work is addressing the simultaneous problems of active manipulation and online estimation of the AC demand, e.g., in a demand response program.

Acknowledgements This research was funded by NSF Grant #ECCS-1508943. We also thank the Pacific Gas and Electric Company for the commercial building electric load data.

Appendix

In this appendix, we present two examples that demonstrate the influence of several of the user-defined functions and parameters within DMD. The first example shows how the choice of η^s impacts the estimate in the presence of measurement noise. The second example illustrates how the choice of divergence and loss functions impact the estimates generated by (1). These examples are constructed to isolate impact of the component of interest. In reality, the various parameters and function choices influence each other in nontrivial ways, which generally cannot be known *a priori*.

In the examples below, the plant model, whose state we are trying to estimate, consists of (5) and (6), where $C = [0 \ 1]$ and

$$A = \begin{bmatrix} \cos(\pi/500) - \sin(\pi/500) \\ \sin(\pi/500) \cos(\pi/500) \end{bmatrix}. \quad (23)$$

The state is $\theta_t \in \mathbb{R}^2$, its initial value is $\theta_0 = [0 \ 1]^T$, $w_t \in \mathbb{R}^2$, and $v_t \in \mathbb{R}$. We assume that w_t and v_t satisfy the assumptions of a Kalman filter, and their covariances are $Q \in \mathbb{R}^{2 \times 2}$ and $R \in \mathbb{R}$, respectively, where we detail their values in each example.

Varying the Gradient Descent Step-Size η^s

The parameter η^s influences how closely DMD adjusts the state estimate $\widehat{\theta}_t$ to match the (possibly noisy) measurement versus trusting the predictions of the system model $\Phi(\cdot)$. In this example, we assume the plant model contains no process noise, i.e., $Q = 0$, and the measurement noise covariance is $R = 1$. The DMD model $\Phi(\widehat{\theta}_t)$ is set to the plant model (5) and (6) excluding w_t and v_t . The divergence is set to $D(\theta \|\widehat{\theta}_t) = \frac{1}{2} \|\theta - \widehat{\theta}_t\|_2^2$, and the loss function is set to $\ell_t(\widehat{\theta}_t) = \frac{1}{2} \|C\widehat{\theta}_t - y_t\|_2^2$. The resulting closed-form measurement-based update (1) is

$$\widetilde{\theta}_t = \widehat{\theta}_t + \eta^s C^T (y_t - C\widehat{\theta}_t). \quad (24)$$

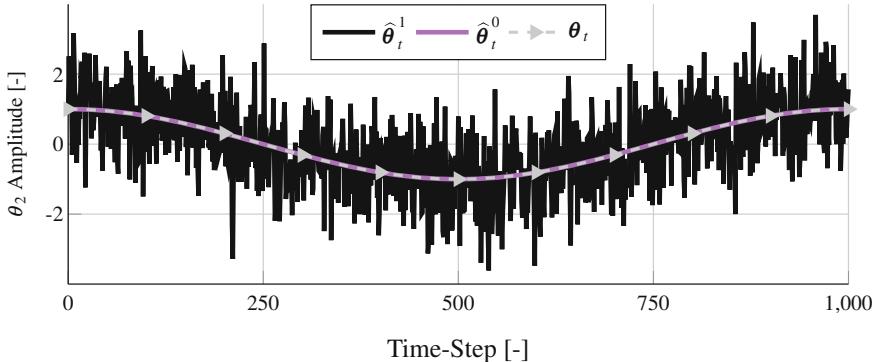


Fig. 3 Time series of the second element of θ_t , $\hat{\theta}_t^0$, and $\hat{\theta}_t^1$

We apply DMD for two different values of η^s (i.e., $\eta_t^0 = 0.0$ and $\eta_t^1 = 1.0$, where the resulting estimates are denoted $\hat{\theta}_t^0$ and $\hat{\theta}_t^1$, respectively) and compare the results. All estimates are initialized at the true state.

Figure 3 presents the resulting time series of the second elements of θ_t , $\hat{\theta}_t^0$, and $\hat{\theta}_t^1$; we exclude time series of the first elements as they exhibit similar characteristics. The second term of (24) is 0 for $\hat{\theta}_t^0$, and so there is no adjustment to the state estimate based on the measurement. As a result, $\hat{\theta}_t^0$ matches θ_t exactly because the model within DMD exactly matches the plant model. Alternatively, for $\hat{\theta}_t^1$, the convex program adjusts the state estimate to match the noisy measurements rather than trusting DMD's model, resulting in significant estimation error.

Varying the Choice of Divergence and Loss Functions

The choice of the divergence and loss functions within DMD influences the algorithm's measurement-based adjustments. In this example, we vary DMD's measurement-based update by using two choices for the divergence and loss functions – one that includes covariance matrices explicitly and one that does not. We also simulate a Kalman filter to empirically show that the DMD estimates match those of a Kalman filter when the divergence and loss functions are constructed as described in Section 3.3.

In this example, we assume the measurement noise covariance is $R = 2$, and the process noise covariance is

$$Q = \begin{bmatrix} 0.25 & 0.1 \\ 0.1 & 0.25 \end{bmatrix}.$$

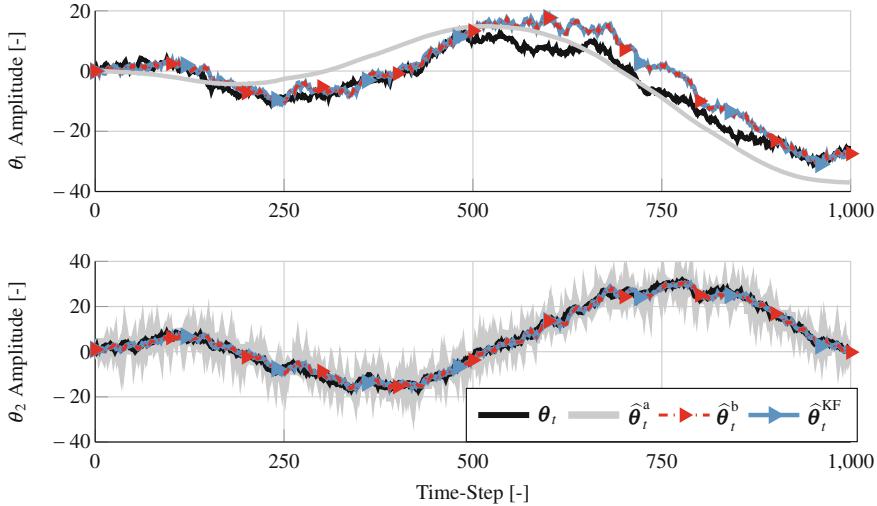


Fig. 4 Time series of θ_t , $\hat{\theta}_t^a$, $\hat{\theta}_t^b$, and $\hat{\theta}_t^{KF}$

We construct a steady-state discrete-time Kalman filter, whose estimates are denoted $\hat{\theta}_t^{KF}$, using the underling system model and covariances. Both DMD formulations use $\eta^s = 1$. The first DMD formulation, whose estimates are denoted $\hat{\theta}_t^a$, uses the same loss function, divergence function, and resulting measurement-based update equation as in the previous example. The second DMD formulation, whose estimates are denoted $\hat{\theta}_t^b$, uses the divergence and loss functions needed to produce measurement-based updates that are equivalent to those of the Kalman filter, i.e., (15), and we set $\hat{P}_t = \bar{P}$ and $\hat{P}_t^y = [\bar{C}\bar{P}C^T + R]$. Note that the second formulation explicitly includes accurate model prediction error statistics via the covariances, whereas the first estimate implicitly assumes the covariances are identity matrices.

Figure 4 presents the time series of θ_t , $\hat{\theta}_t^a$, $\hat{\theta}_t^b$ and $\hat{\theta}_t^{KF}$. Note that the estimates $\hat{\theta}_t^b$ and $\hat{\theta}_t^{KF}$ coincide exactly, empirically supporting our claim that we can choose the DMD model, divergence function, and loss function to achieve a measurement-based update equivalent to that of Kalman filter. In estimating the second element of θ_t , we first note that both $\hat{\theta}_t^a$ and $\hat{\theta}_t^b$ follow the general trajectory of the true state, but $\hat{\theta}_t^b$ is noticeably smoother than $\hat{\theta}_t^a$. By including the covariance matrices into the measurement-based update, $\hat{\theta}_t^b$ is better able to account for the measurement noise resulting in a less erratic estimate and reduced estimation error versus $\hat{\theta}_t^a$. In estimating the first element of θ_t , both methods have significant deviations from the true state value; however, the root mean square estimation error in $\hat{\theta}_t^b$ is smaller than that of $\hat{\theta}_t^a$, indicating that $\hat{\theta}_t^b$ is more accurate over the duration of the simulation.

Again, the inclusion of accurate statistical information into the measurement-based update has led to a more accurate estimate.

It should be noted that this example was constructed such that the Kalman filter is the optimal estimator. In reality, the assumptions of the Kalman filter rarely hold, as in the case studies presented in Section 4.4. A Kalman filter can still be applied with varying degrees of success, but it may not be the optimal estimator. The DMD algorithm relaxes some of the underlying assumptions, which allows greater flexibility in designing the updates, but the theoretical guarantees of the Kalman filter do not apply. Additionally, it should be noted that in this example we assume we have a perfect estimate of the covariance matrices. In Section 4.4, we show that including inaccurate model prediction error statistics into the DMD algorithms within DFS can degrade the performance of DFS, while including accurate statistics (which can be hard to obtain in practice) can substantially improve the estimation accuracy of DFS.

References

1. Armel K, Gupta A, Shrimali G, Albert A (2013) Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy* 52:213–234
2. Baone C, Xu Y, Kueck J (2010) Local voltage support from distributed energy resources to prevent air conditioner motor stalling. In: Innovative smart grid technologies (ISGT), Gothenburg
3. Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper Res Lett* 31(3):167–175
4. Berges M, Goldman E, Matthews H, Soibelman L (2009) Learning systems for electric consumption of buildings. In: International workshop on computing in civil engineering, Austin
5. Berges M, Goldman E, Matthews H, Soibelman L (2010) Enhancing electricity audits in residential buildings with nonintrusive load monitoring. *J Ind Econ* 14(5):844–858
6. Can Kara E, Kolter Z, Berges M, Krogh B, Hug G, Yuksel T (2013) A moving horizon state estimator in the control of thermostatically controlled loads for demand response. In: Proceedings of smartGridComm, Vancouver
7. Cesa-Bianchi N, Long PM, Warmuth MK (1996) Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Trans Neural Netw* 7(3):604–619
8. Dong R, Ratliff L, Ohlsson H, Sastry S (2013) A dynamical systems approach to energy disaggregation. In: Proceedings of the IEEE conference on decision and control, Firenze
9. Dong R, Ratliff L, Ohlsson H, Sastry S (2013) Energy disaggregation via adaptive filtering. In: Proceedings of the Allerton conference, Monticello
10. Dong R, Ratliff L, Ohlsson H, Sastry S (2014) Fundamental limits of nonintrusive load monitoring. In: Proceedings of the HiCoNS, Berlin
11. Duchi JC, Shalev-Shwartz S, Singer Y, Tewari A (2010) Composite objective mirror descent. In: 23rd annual conference on learning theory (COLT), Haifa, pp 14–26
12. Esmaeil Zadeh Soudjani S, Abate A (2015) Aggregation and control of populations of thermostatically controlled loads by formal abstractions. *IEEE Trans Control Syst Technol* 23(3):975–990
13. Gonçalves H, Ocneanu A, Bergés M, Fan R (2011) Unsupervised disaggregation of appliances using aggregated consumption data. In: Proceedings of the 1st KDD workshop on data mining applications in sustainability (SustKDD), San Diego

14. Grewal MS, Andrews AP (2015) Kalman filtering. Wiley, New York
15. Hall EC, Willett RM (2015) Online convex optimization in dynamic environments. *IEEE J Sel Top Sign Proces* 9(4):647–662
16. Hart G (2010) Nonintrusive appliance load monitoring. *Proc IEEE* 80(12):1870–1891
17. Herbster M, Warmuth MK (1998) Tracking the best expert. *Mach Learn* 32(2):151–178
18. Herbster M, Warmuth MK (2001) Tracking the best linear predictor. *J Mach Learn Res* 1(Sep):281–309
19. Jadbabae A, Rakhlin A, Shahrampour S, Sridharan K (2015) Online optimization: competing with dynamic comparators. In: Proceedings of the eighteenth international conference on artificial intelligence and statistics, pp 398–406
20. Kalsi K, Elizondo M, Fuller J, Lu S, Chassin D (2012) Development and validation of aggregated models for thermostatic controlled loads with demand response. In: Proceedings of Hawaii international conference on systems science, Wailea
21. Kim H, Marwah M, Arlitt MF, Lyon G, Han J (2011) Unsupervised disaggregation of low frequency power measurements. In: Proceedings of the 2011 SIAM international conference on data mining, vol 11, pp 747–758
22. Kivinen J, Warmuth MK (1999) Averaging expert predictions. In: European conference on computational learning theory. Springer, Berlin, pp 153–167
23. Koch S, Mathieu J, Callaway D (2011) Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services. In: Proceedings of the power systems computation conference, Stockholm
24. Kolter J, Jaakkola T (2012) Approximate inference in additive factorial HMMs with application to energy disaggregation. In: Proceedings of the international conference on artificial intelligence and statistics, La Palma
25. Kolter J, Batra S, Ng A (2010) Energy disaggregation via discriminative sparse coding. In: Proceedings of the advances in neural information processing systems (NIPS), Vancouver
26. Ledva G, Vrettos E, Mastellone S, Andersson G, Mathieu J (2015) Applying networked estimation and control algorithms to address communication bandwidth limitations and latencies in demand response. In: Hawaii international conference on systems science (HICSS), Grand Hyatt, Kauai
27. Ledva GS, Balzano L, Mathieu JL (2015) Inferring the behavior of distributed energy resources with online learning. In: Proceedings of the Allerton conference, Monticello, pp 187–194
28. Ledva GS, Balzano L, Mathieu JL (2018) Real-time energy disaggregation of a distribution feeder's demand using online learning. *IEEE Trans Power Syst* (in press). DOI: [10.1109/TPWRS.2018.2800535](https://doi.org/10.1109/TPWRS.2018.2800535)
29. Lee MP, Aslam O, Foster B, Hou S, Kathan D, Pechman C, Young C (2015) Assessment of demand response and advanced metering. Staff report, Federal energy regulatory commission. <https://www.ferc.gov/legal/staff-reports/2015/demand-response.pdf>
30. Mathieu J, Callaway D (2012) State estimation and control of heterogeneous thermostatically controlled loads for load following. In: Proceedings of the Hawaii international conference on systems science, Wailea, pp 2002–2011
31. Mathieu J, Gadgil A, Callaway D, Price P, Kiliccote S (2010) Characterizing the response of commercial and industrial facilities to dynamic pricing signals from the utility. In: Proceedings of ASME 2010 4th international conference on energy sustainability, Phoenix
32. Mathieu J, Koch S, Callaway D (2013) State estimation and control of electric loads to manage real-time energy imbalance. *IEEE Trans Power Syst* 28(1):430–440
33. Nemirovsky AS, Yudin DB (1983) Problem complexity and method efficiency in optimization. In: Wiley-Interscience series in discrete mathematics. Wiley, New York
34. NOAA (2009) NNDC climatic data online, Satellite and Information Service National Climate Data Center. <http://www7.ncdc.noaa.gov/CDO/dataproduct>
35. Pecan Street Inc (2016) Dataport. <https://dataport.pecanstreet.org/>
36. Powers J, Margossian B, Smith B (1991) Using a rule-based algorithm to disaggregate end-use load profiles from premise-level data. *IEEE Comput Appl Power* 4(2):42–47

37. Schneider KP, Chen Y, Chassin DP, Pratt RG, Engel DW, Thompson SE (2008) Modern grid initiative distribution taxonomy final report. Technical report, Pacific Northwest National Laboratory (PNNL), Richland
38. Shahrampour S, Jadbabaie A (2018) Distributed online optimization in dynamic environments using mirror descent. *IEEE Trans Autom Control*, 63(3), 714–725.
39. Shalev-Shwartz S (2011) Online learning and online convex optimization. *Found Trends Mach Learn* 4(2):107–194
40. Shao H, Marwah M, Ramakrishnan N (2012) A temporal motif mining approach to unsupervised energy disaggregation. In: Proceedings of the 1st international workshop on non-intrusive load monitoring, Pittsburgh, vol 7
41. Simonetto A, Mokhtari A, Koppel A, Leus G, Ribeiro A (2016) A class of prediction-correction methods for time-varying convex optimization. *IEEE Trans Signal Process* 64(17):4576–4591
42. Vrettos E, Mathieu J, Andersson G (2014) Control of thermostatic loads using moving horizon estimation of individual load states. In: Proceedings of the power systems computation conference (PSCC), Wroclaw
43. Vrettos E, Mathieu J, Andersson G (2014) Demand response with moving horizon estimation of individual thermostatic load states from aggregate power measurements. In: Proceedings of the American control conference (ACC), Portland
44. Wytock M, Kolter J (2014) Contextually supervised source separation with application to energy disaggregation. In: Proceedings of the conference on artificial intelligence (AAAI), Québec City, pp 486–492
45. Yang T, Zhang L, Jin R, Yi J (2016) Tracking slowly moving clairvoyant: optimal dynamic regret of online learning with true and noisy gradient. In: Proceedings of the 33rd international conference on machine learning, New York
46. Zeifman M, Roth K (2011) Nonintrusive appliance load monitoring: review and outlook. *IEEE Trans Consum Electron* 57(1):76–84
47. Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the international conference on machine learning (ICML), Washington
48. Zoha A, Gluhak A, Imran M, Rajasegarar S (2012) Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey. *Sensors* 12:16838–16866

Risk-Aware Demand Management of Aggregators Participating in Energy Programs with Utilities



William D. Heavlin, Ana Radovanović, Varun Gupta, and Seungil You

Abstract Electric utilities typically offer demand-side management (DSM) programs in order to reduce peak demand and to shift supply risks. These same programs engender a new business model, that of the energy aggregators. Energy aggregators seek to harvest the DSM incentives by strategically deferring the loads under their control. Examples of deferrable loads are electric vehicles (EVs) and heating, ventilation, and air-conditioning (HVAC) systems. To choose appropriately from a utility's menu of programs, the aggregator must forecast both temperature and load and should also estimate the uncertainties associated with these forecasts. Further, the aggregator can work to mitigate these uncertainties by managing flexible loads under their control.

We propose a formulation that unifies the various kinds of deferrable loads and explicitly balances the trade-off between user discomfort and monetary costs. Our main contribution comes from incorporating the uncertainty of temperature and load forecasts into the optimal choice of DSM program selection.

1 Introduction

Large flexible loads, such as air-conditioning, domestic heating, and electric vehicles, induce daily demand peaks and troughs in electric power, to which utilities often respond by increasing supply from environmentally unfavorable and

W. D. Heavlin · A. Radovanović (✉)
Google, Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA

S. You
Kakao Mobility, 231, Pangyoek-ro, Bundang-gu, Seongnam-si, Gyeonggi-do,
Republic of Korea, 13494
e-mail: sean.you@kakaomobility.com

V. Gupta
University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637, USA
e-mail: guptav@uchicago.edu

inefficient power generators [27]. Renewable supply sources, e.g., wind and solar, are relevant for meeting daily demand peaks but are highly variable. Furthermore, in the USA at least, current distribution networks limit the role of renewables in the grid. At any rate, the associated uncertainty of such renewables increases the need for dynamic response on the part of electric power suppliers [42]. In direct response to irregular demand and, partly, in indirect response to the irregular supply from renewables, the smart grid [3] seeks to incorporate information-based technologies, through smart meters [14], demand-side management (DSM) [11, 33], and power line-based communication [24].

The present work focuses on the DSM programs most commonly offered by utilities. These efforts aim at shaping consumption patterns both by decreasing peak demand and by shifting such demand to off-peak periods. DSM achieves this by a combination of direct curtailment requests and consumer price signals. While economic theory strongly favors price-based systems, in the case of electric power, we need an economic agent sufficiently sensitive to time-of-day price changes to modify his/her electricity consumption. This has led to the emergence of a new business model, the energy aggregator [25, 38], which acts on behalf of a group of energy consumers to mitigate energy costs as a trusted agent and works with the utility to reduce demand as circumstances warrant. In this chapter we focus on the distribution-level aggregators, we focus on distribution-level aggregators acting as intermediaries acting as intermediaries between consumers and utility.

Even though it can have a significant impact on the expected cost, uncertainty in temperature forecasts, load parameters, and environmental conditions has been largely neglected in demand planning. Small changes in the demand curve can induce large changes in cost, first because of the capacity-based demand charges set by utilities and second because of the payment rules of demand response (DR) programs. Typically, an aggregator can select a tariff offered by the utility. Such tariffs define both time-of-use prices of marginal energy and aggregate-level surcharges that are commonly a function of monthly peak power consumption [5]. In addition, demand response (DR) programs allow an aggregator to nominate the demand levels that it is capable of curtailing, both in response to high wholesale prices and/or to lowered electricity system reliability. Such nominations are typically submitted before the beginning of each month yet obligate the aggregator to respond to event notifications that it receives only one day ahead of time or even on the same day.

In this chapter, we present an algorithm by which such an agent, the aggregator, optimally balances its consumers' needs with those of the utility. Our analysis focuses on two canonical devices with deferrable loads: (1) heating, ventilation, and air-conditioning (HVAC) systems and (2) electric vehicles (EVs). The algorithm incorporates both the time-dependent consumer benefit from receiving electric power and the time-sensitive marginal costs of producing power. We concentrate mainly on day-ahead, cost-effective demand management, but the framework extends naturally to real time. Central to our analysis is the treatment of the uncertainty in the environmental conditions and load parameters. Such estimates of uncertainty in turn enable us to estimate the impact on the resulting cost function. These calculations enable aggregators to cooperate strategically with power suppliers to implement demand reduction [17, 40] while managing risk.

Since our focus is on distribution-level aggregators (which is the most common and regulation-approved form of participation within the USA), physical constraints such as line capacities which are relevant in an electric grid as line capacities, which are relevant in an electric grid, do not apply here. A key takeaway of the presented work is demonstrating the impact of uncertainty on the expected cost function. An aggregation assumption allows us to estimate the impact of noise efficiently, bypassing the commonly used, computationally slow, Monte Carlo simulation approach [10].

This chapter is organized as follows. Section 2 is a short survey of the past work in areas of energy consumption and generation modeling, including the most common approaches in cost-effective management of aggregated energy assets. Section 3 includes our mathematical model and the framework by which we capture uncertainties of flexible devices. In order to motivate the selected cost functions, we also describe typical utility programs. In Section 4, we formulate the power scheduling optimization and the approximate distributions of the resulting optimal schedules. The risk-aware program selection strategy is discussed in Section 5. In Section 6, we present a numerical example, and in Section 7, we summarize our conclusions.

2 Literature Review

Given the vast amount of literature on energy markets and demand response, it will not be possible for us to cite all the relevant papers, even restricting to energy aggregators (also called load-serving entity, LSE). Instead, we give a brief classification of the literature based on a few important dimensions that most papers differ along: (i) market design and coordination mechanism, (ii) optimization metric, (iii) computational approaches, and (iv) demand models.

Markets and Coordination Mechanism: As in our formulation, one of the most common assumptions is that of an aggregator that can directly control the individual devices which it serves. Therefore, the decision variables in the optimization problem are the day-ahead bids, as well as the real-time dispatch/unit commitment decisions (e.g., [26, 32, 41]). A second assumption is where the aggregator publishes prices to the loads it serves, and the loads “self-dispatch” by optimizing their individual utility functions (e.g., [7, 21]). We note that this is a difference only in spirit, since the final solutions, for example, the day-ahead bids obtained under either model, would usually be the same. A third much rarer coordination mechanism is found in some papers on aggregation of HVACs, where the aggregator cannot observe the state of each individual HVAC. In such cases, the aggregator publishes a common probabilistic policy (e.g., [28, 44]). Another dimension under the broad theme of market design is on two-stage (day-ahead and real-time) vs multistage markets. All the papers mentioned above are for the two-stage markets. In papers on the multistage markets, the focus is not on detailed modeling of loads but on the impact of information structure and contract design on optimal procurement decisions (e.g., [36]).

Optimization Metric: When demand is considered elastic, the usual optimization metric is the expected monetary cost and user discomfort cost/utility, as in the present chapter (e.g., [7, 29, 31, 36]). When the load is considered inelastic, the usual optimization metric is the expected profit/payment of the aggregator (e.g., [21, 41]). Some authors also consider optimizing various measures of risk such as conditional value at risk (cVaR) (e.g., [32]), asymmetric cost functions (e.g., [26]), and risk-constrained optimization (e.g., [32]).

Computational Approaches: For two-stage stochastic problems (e.g., when load constraints or utilities are separable across time, in which case the two stages are day ahead and real time), a common approach is Monte Carlo optimization using scenario generation. One recent example is [32] where, under the assumption of independent normally distributed forecast uncertainties, the authors use scenario generation followed by scenario reduction to convert a two-stage stochastic optimization problem to a mixed-integer linear program. Also [26] uses Monte Carlo scenario generation, and [41] formulates a stochastic linear program. For multistage problems (e.g., aggregate constraints across time as in EVs), a common approach uses rolling horizon look-ahead stochastic dynamic programming (e.g., [6, 15]). The authors in [12, 36] obtain a closed-form expression for the optimal procurement contracts for a multistage procurement problem by solving the dynamic programming problem analytically. Other approaches for optimization under uncertainty are robust optimization (e.g., [8, 45]), and chance constrained optimization (e.g., [43]). Chance-constrained and robust optimization have been recently applied to solving Optimal Power Flow (OPF) problems, and to coordinate control of energy assets on the transmission network. The latter requires physical constraints such as line flows limits [4, 37].

Models for Demand/Loads: A common but less accurate approach for modeling loads assumes that they are given as exogenous forecasts with uncertainty (e.g., [32]). Among higher-fidelity models of electric loads and thermal loads (HVACs), the popular models are first-order (e.g., [7, 28]), second-order (e.g., [44]), and more detailed energy-mass balance-based models (e.g., [2, 30, 31, 39]). [20] proposes an alternate model associating a thermal device with a nominal control curve and a set of feasible perturbation curves. For EVs, the popular models for demand are aggregate charge in exogenously defined windows (e.g., [7]) and detailed models specifying individual trips and durations (e.g., [31]). Evolution of state of charge is specified by first-order models (e.g., [7, 41]). For models on EV usage patterns, see [18] and references therein.

3 Mathematical Model

We consider a large collection (an aggregate) of deferrable loads: electric vehicles (EVs), and building HVAC systems, on a large commercial campus. Our goal is to manage daily power demand in a cost-efficient way. Demand-side management

works on two time scales: (i) On a monthly basis, the load aggregator selects one of a set of programs designated by the utilities and in doing so chooses a cost-vs-demand curve. (ii) One day ahead, the aggregator seeks to allocate power consistent with the current cost-demand curve while minimizing costs. Both (i) and (ii) make use of temperature and power demand forecasts, which are inherently noisy and imperfect.

In this paper, we propose a computationally efficient and practical way to adjust power allocations in order to reduce expected costs. This method can take advantage of updates to actual temperatures, EV driver availability, and building occupancies. Furthermore, we show how the selection of different program parameters affects overall costs and cost uncertainty and how the choice of such parameters can be adjusted cost-effectively.

To model the behavior of deferrable devices, HVACs and EVs, we use a familiar discrete-time linear system (see e.g., [16, 23]). At the time index $k + 1$, the state of an HVAC device is

$$X_i[k + 1] - X_i[k] = \theta_i(Z[k] - X_i[k]) - \eta_i u_i[k] \quad (1)$$

where $X_i[k]$, $Z[k]$, and $u_i[k]$ are the indoor and outdoor temperatures and power consumption at time k and θ_i and η_i are the thermal transmission and cooling coefficients of the i th load.

Analogously, in the case of an EV, we use the relation

$$X_i[k + 1] - X_i[k] = \eta_i A_i[k] u_i[k], \quad (2)$$

where $X_i[k]$ and $u_i[k]$ are the charged energy and power consumption at time k and η_i the charging efficiency of the i th load. Here, $A_i[k] \in \{0, 1\}$ denotes an indicator of availability of the i th load at time k . In other words, if EV i is attached to the charging station at time k , then $A_i[k] = 1$; otherwise, $A_i[k] = 0$.

Notice that (1) and (2) have quite similar forms. We unify these two state equations as

$$X_i[k + 1] - X_i[k] = \theta_i(Z[k] - X_i[k]) + \eta_i A_i[k] u_i[k]. \quad (3)$$

Here, if the i th load is an HVAC, then $A_i[k] = -1$ for all k . Otherwise, if the i th load is an EV, then $\theta_i = 0$ for all k .

Note that, if the outside temperature $Z[k]$, and the EV's availability $A_i[k]$ are known, for any given power consumptions $u_i(k)$, the above defined changes in state are completely deterministic. However, since power allocation decisions are made in advance, there are some uncertainty in $Z[k]$ and $A_i[k]$ and, therefore, a randomness in the induced state, $X_i[k]$.

In addition to the considered uncertainties in the outside temperature $Z[\cdot]$ and drivers' availability $A_i[\cdot]$, (3) could include an additive, modeling "noise," characterized in the estimation of parameters θ_i and η_i . The methodology presented in this chapter is directly applicable to this modeling extension. Without loss of generality, to preserve simplicity in the exposition, we omit it.

3.1 Cost Objectives

In demand management, in addition to the tariff-prescribed energy and demand payments to the utility, there is the countervailing customer discomfort induced by the aggregator exploiting their loads' flexibility.

In the case of EVs, the regret is the penalty for not fully charging in the time available. In this case, a user cares only about whether her/his EV is charged by the time he/she wants to drive it, and, therefore, the related discomfort is a function of state of charge when he/she disconnects the EV. To that end, we assume that the corresponding cost at time k is proportional to the probability of a user disconnecting in interval $(k, k + 1)$, i.e., $P_i[k] \stackrel{\text{def}}{=} \mathbb{P}[A_i[k - 1] = 1, A_i[k] = 0]$.

In case of HVACs, the user discomfort term represents the penalty for deviating from the target temperature set point. Further, this discomfort at time k is proportional to the occupancy $O_i[k]$ of the building i .

In this paper, we represent the user discomfort by the quadratic function

$$\frac{1}{2} R_i[k](X_i[k] - \bar{x}_i[k])^2,$$

where

$$R_i[k] = \begin{cases} P_i[k] & \text{for EV} \\ O_i[k] & \text{for HVAC} \end{cases}.$$

For HVAC cooling, $\bar{x}_i[k]$ is the desired indoor temperature of the i th load at time k and for EV charging $\bar{x}_i[k]$ is the full-charge battery capacity and physical upper bound on $X_i[k]$.

The energy cost term is the dollar-denominated *economic* cost that any aggregator is charged and is therefore dependent on the utility tariffs and selected programs. Each tariff incorporates time-of-use pricing, as well as demand charges (see [5]). Demand charges depend on the monthly peak demand and severely penalize consumers when their monthly peak demand exceeds demand levels set by the selected tariff. In addition, in other programs, such as demand response (DR) (see [1]), the aggregator can earn revenue (and thereby reduce total costs) by shedding demand at a proposed price (bid) and can be penalized if it fails to do so. In these programs, aggregators nominate their demand reduction capacities before each month. Depending on the program, utilities notify aggregators of DR events either one day prior to or on the same day. The performance of each aggregator is calculated as the difference between the actual demand and a utility-curated baseline load, during the event hours. Obviously, such evaluations are complicated by utility-specific definitions of baseline loads.

One example of a day-ahead DR program is Capacity Bidding Program (CBP, [34]); Figure 1 illustrates the expected payment as a function of the nominated

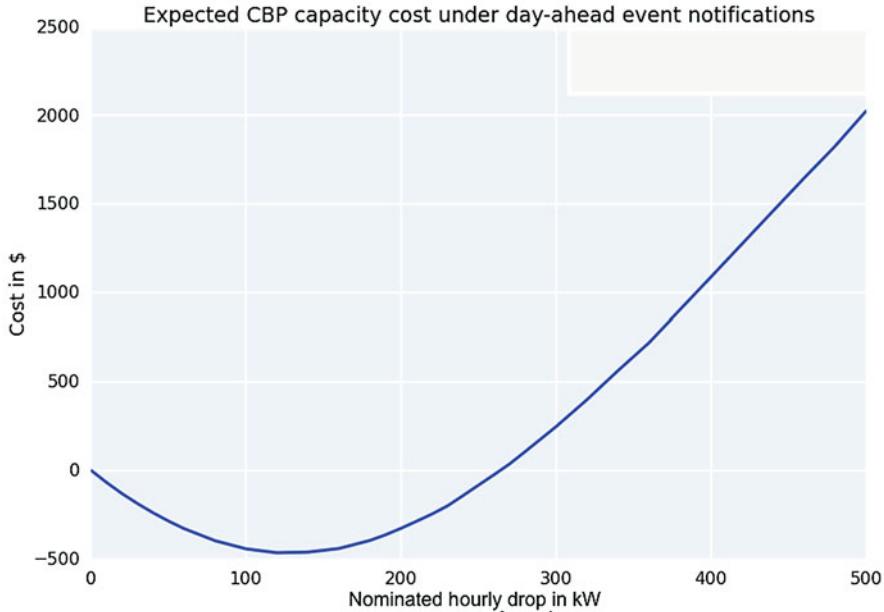


Fig. 1 Expected CBP capacity cost as the function of the nominated capacity.

capacity. Demand in commercial buildings can be reduced by changing temperature set points. In this case, empirically, the loads that are shed are rather variable and depend on diverse factors such as outside temperature and building occupancy. A deeper treatment of the components of this variability is beyond the scope of this paper. The interested reader is referred to [35] for more.

Motivated by the nature of the above discussion, we represent these costs by strictly convex functions:

$$c(u_{\text{total}}[k], k, \mathcal{P}_j)$$

where $u_{\text{total}}[k] \stackrel{\text{def}}{=} \sum_i u_i[k]$ denotes the aggregated power consumption corresponding to the time index k , while \mathcal{P}_j represents parameters from the j th program. The cost function $c(\cdot, \cdot, \cdot)$ incorporates the time-of-use pricing, so it depends on k .

Utility programs usually have multiple options $\{\mathcal{P}_j\}_{j=1}^{n_c}$ (including both tariffs and DR), with different energy cost functions:

$$\{c(u_{\text{total}}[k], k, \mathcal{P}_1), \dots, c(u_{\text{total}}[k], k, \mathcal{P}_{n_c})\},$$

where n_c is the number of offered programs. Our goal is to determine power schedule u that jointly minimizes the expected energy and discomfort cost, while controlling the uncertainty in the resulting charges, here represented by risk-aware

constraints. Furthermore, we assume that the set of programs discussed above potentially includes grid regulation services (demand response, frequency regulation, etc.) provided through the local utility.

A common approach in addressing the above-described cost minimization of the aggregated load uses model predictive control (MPC) [9, 19, 22]. By replacing stochastic processes $Z[k]$, $A[k]$, and $R[k]$ with their expectations, we can find an optimal schedule $u[k]$, $\forall k$, which minimizes a certain cost function. However, ignoring the intrinsic uncertainty of these processes prohibits offering any guarantees for the variability of the resulting energy cost, which, in general, results in large discrepancies between the minimum cost as computed and the actual cost paid to the utility.

The standard way of incorporating diverse variabilities is to apply Monte Carlo simulation. By generating many sample paths of $Z[k]$, $A[k]$, and $R[k]$, we can estimate the distribution of the optimal schedule U , and the resulting energy cost $\sum_k c(\sum_i U_i[k], k, \mathcal{P}_j)$, for a given program \mathcal{P}_j . However, note that this process involves solving the cost minimization problem by computing optimal $u_i[k]$ s for each sample path and each utility program \mathcal{P}_j , resulting in a computationally expensive and, potentially, very slow approach. In addition, often, the Monte Carlo approach does not provide clear insights on the impact of parameter values, and their interactions on the optimal cost, which is probably the most significant drawback of this method.

In this work, we propose a computationally tractable way of approximating the distribution of the optimal schedule $U[k]$, $\forall k$, and use this approximation to perform risk-aware minimization of the expected cost. By simulation, we show that this approximation and the resulting optimal schedule are quite close to that obtained using the Monte Carlo simulation, with computational burden equivalent to a single optimization per utility program.

4 Approximation of Optimal Power Schedules

In this section, we describe an optimization approach that is sensitive to the distribution of the optimal power schedule. For notational convenience, let

$$\begin{aligned}\mathbf{Z} &:= (Z[0], \dots, Z[K-1]) \\ \mathbf{R} &:= (R_1[0], \dots, R_1[K-1], \dots, R_N[K-1]) \\ \mathbf{A} &:= (A_1[0], \dots, A_1[K-1], \dots, A_N[K-1]).\end{aligned}$$

A convenient assumption asserts that random vectors, \mathbf{Z} , \mathbf{R} , and \mathbf{A} , are mutually uncorrelated. This assumption can be easily justified from the construction of the discomfort cost terms in Section 3. Note that all the randomness of processes $R_i[\cdot]$ stems from the uncertainty in building occupancy in case of the HVAC load,

while the uncertainty of $A_i[\cdot]$ depends on the charging availability to EV drivers, which, in most contexts, does not depend on the outside temperature. However, this independence assumption is not essential for the proposed approach and is used only to improve the clarity of exposition.

Let $\mathbf{z} = (z[0], \dots, z[K-1])$, $\mathbf{r} = (r_1[0], \dots, r_N[K-1])$, and $\mathbf{a} = (a_1[0], \dots, a_N[K-1])$ denote realizations (sample paths) of \mathbf{Z} , \mathbf{R} , and \mathbf{A} , respectively. For this sample path, we define the following optimization that computes the optimal schedule $\mathbf{u} = (u_1[0], \dots, u_1[K-1], \dots, u_N[K-1])$:

$$\begin{aligned} & \underset{\mathbf{u}}{\text{minimize}} \quad \sum_{k=0}^{K-1} \sum_{i=1}^N \frac{r_i[k]}{2} (x_i[k] - \bar{x}_i[k])^2 + \lambda \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i[k], k, \mathcal{P}_j \right) \\ & \text{subject to} \quad x_i[k+1] = (1 - \theta_i)x_i[k] + \eta_i a_i[k] u_i[k] + \theta_i z[k], \quad \{x_i[0]\} \text{ is given} \\ & \quad 0 \leq u_i[k] \leq \bar{u}_i[k], \end{aligned} \quad (4)$$

where the first term of the objective function is the discomfort cost, the second term is the j th program's energy cost, and $\bar{u}_i[k]$ is the upper bound of the power consumption of the i th device at time k . The two cost terms do not have the same units. While energy cost is in dollars, discomfort cost has different units depending on the type of load. It is therefore necessary to calibrate one to another. This is the role of parameter λ , and its tuning allows for exploring the trade-offs between the discomfort and energy costs. Here, we do not consider tuning λ .

The current state $x_i[k]$ satisfies

$$x_i[k+1] = (1 - \theta_i)^{k+1} x_i[0] + \theta_i \sum_{l=0}^k (1 - \theta_i)^{k-l} z[l] + \eta_i \sum_{l=0}^k (1 - \theta_i)^{k-l} a_i[l] u_i[l]. \quad (5)$$

By substituting this within the discomfort function, we define the cost function

$$\mathcal{R}(\mathbf{u}; \mathbf{z}, \mathbf{r}, \mathbf{a}) = \sum_{k=0}^{K-1} \sum_{i=1}^N \frac{r_i[k]}{2} (x_i[k] - \bar{x}_i[k])^2,$$

where $x_i[k]$ is given by (5). In addition, if we define the feasible set $\mathcal{U} = \{\mathbf{u} : 0 \leq u_i[k] \leq \bar{u}_i[k]\}$, we end up with the more compact representation of (4):

$$\underset{\mathbf{u} \in \mathcal{U}}{\text{minimize}} \quad \mathcal{R}(\mathbf{u}; \mathbf{z}, \mathbf{r}, \mathbf{a}) + \lambda \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i[k], k, \mathcal{P}_j \right). \quad (6)$$

Note that the first term depends on the realizations of random vectors, \mathbf{Z} , \mathbf{R} , and \mathbf{A} , so the optimal solution of (6) is also a random vector. Denote by $\mathbf{U}(\mathcal{P}_j)$ the optimal solution of (6), when the j th program is chosen. Since for any realizations \mathbf{z} ,

\mathbf{r} , and \mathbf{a} , we can determine the optimal solution $\mathbf{U}(\mathcal{P}_j)$. $\mathbf{U}(\mathcal{P}_j)$ is measurable with respect to \mathbf{Z} , \mathbf{R} , and \mathbf{A} . In this section, we propose approximating the distribution of the optimal $\mathbf{U}(\mathcal{P}_j)$.

4.1 Expected Cost Minimization

Consider the following expected cost minimization:

$$\underset{\mathbf{u} \in \mathcal{U}}{\text{minimize}} \quad \mathbb{E}[\mathcal{R}(\mathbf{u}; \mathbf{Z}, \mathbf{R}, \mathbf{A})] + \lambda \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i[k], k, \mathcal{P}_j \right), \quad (7)$$

where \mathcal{R} is a convex function with respect to \mathbf{u} , implying that the above optimization is also convex.

First, we compute the expected cost $\mathbb{E}[\mathcal{R}(\mathbf{u}; \mathbf{Z}, \mathbf{R}, \mathbf{A})]$. Let $\Delta_z[k] = Z[k] - \mathbb{E}[Z[k]]$, $\Delta_{r_i}[k] = R_i[k] - \mathbb{E}[R_i[k]]$, $\Delta_{a_i}[k] = A_i[k] - \mathbb{E}[A_i[k]]$, and $\Delta_{x_i}[k] = X_i[k] - \mathbb{E}[X_i[k]]$. Then, from (5), we have

$$\Delta_{x_i}[k] = \sum_{l=0}^{k-1} (1 - \theta_i)^{k-l} (\theta_i \Delta_z[k] + \eta_i \Delta_{a_i}[l] u_i[l]).$$

The discomfort induced by the i th device at time k is given by

$$\frac{R_i[k]}{2} (X_i[k] - \bar{x}_i[k])^2 = \frac{R_i[k]}{2} \left((x_i[k] - \bar{x}_i[k])^2 + 2\Delta_{x_i}[k](x_i[k] - \bar{x}_i[k]) + \Delta_{x_i}[k]^2 \right).$$

From the zero-correlation assumption discussed at the beginning of this section, the second term has expectation zero, resulting in

$$\mathbb{E} \left[\frac{R_i[k]}{2} (X_i[k] - \bar{x}_i[k])^2 \right] = \frac{\mathbb{E}[R_i[k]]}{2} (x_i[k] - \bar{x}_i[k])^2 + \frac{\mathbb{E}[R_i[k]]}{2} \mathbb{E} \left[\Delta_{x_i}[k]^2 \right]. \quad (8)$$

By replacing all random quantities by their means, the first term can also be obtained via a deterministic convex optimization problem.

To derive the second term in (8), we use

$$\begin{aligned} \Delta_{x_i}[k]^2 &= \sum_{l=0}^{k-1} \sum_{l'=0}^{k-l} \left((1 - \theta_i)^{k-l} \theta_i \Delta_z[l] + \eta_i \Delta_{a_i}[l] u_i[l] \right) \\ &\quad \times \left((1 - \theta_i)^{k-l'} \theta_i \Delta_z[l'] + \eta_i \Delta_{a_i}[l'] u_i[l'] \right), \end{aligned}$$

which, after taking expectations, results in

$$\mathbb{E} \left[\Delta_{x_i}[k]^2 \right] = \underbrace{\sum_{l=0}^{k-1} \sum_{l'=0}^{k-1} (1 - \theta_i)^{2k-l-l'} \eta_i^2 \text{Cov}(A_i[l], A_i[l']) u_i[l] u_i[l']}_{Q_i(u_i)[k]} + \dots,$$

where the remaining terms do not depend on u_i . Therefore, the expected cost minimization, (7), is equivalent to the following convex optimization:

$$\begin{aligned} \underset{\mathbf{u} \in \mathcal{U}}{\text{minimize}} \quad & \mathcal{R}(\mathbf{u}; \mathbb{E}[\mathbf{Z}], \mathbb{E}[\mathbf{R}], \mathbb{E}[\mathbf{A}]) + \lambda \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i[k], k, \mathcal{P}_j \right) \\ & + \frac{1}{2} \sum_{i=1}^N \sum_{k=0}^{K-1} r_i[k] Q_i(u_i)[k]. \end{aligned} \quad (9)$$

Note that the third term in (9) weights the uncertainty $Q_i(u_i)[k]$ by the discomfort cost parameter $r_i[k]$.

Denote the optimal solution of (9) by $\mathbf{u}_c(\mathcal{P}_j)$. In the following subsection, we make a Taylor series approximation around this solution $\mathbf{u}_c(\mathcal{P}_j)$.

4.2 Variance Estimation

To calculate the variance of the total cost, we examine the impact of random quantities, \mathbf{Z} , \mathbf{R} , and \mathbf{A} , on the optimal solution \mathbf{U} . Consider the following Lagrangian:

$$\begin{aligned} \mathcal{L}(\mathbf{u}; \mathbf{Z}, \mathbf{R}, \mathbf{A}, \mu^+, \mu^-) = & \mathcal{R}(\mathbf{u}; \mathbf{Z}, \mathbf{R}, \mathbf{A}) + \lambda \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i[k], k, \mathcal{P}_j \right) \\ & + \sum_{i,k=0}^{K-1} \mu_{i,k}^+ (u_i[k] - \bar{u}_i) + \sum_{i,k=0}^{K-1} \mu_{i,k}^- (-u_i[k]), \end{aligned}$$

where each $\mu_{i,k}^+$ and $\mu_{i,k}^-$ are dual variables corresponding to the box constraints in the feasible set \mathcal{U} in optimization (9). For notational convenience, denote $\mathbf{B} = (\mathbf{Z}, \mathbf{R}, \mathbf{A})$, the concatenation of all random quantities, and $\mathbf{b} = \mathbb{E}[\mathbf{B}]$. Since \mathbf{U} is the optimal solution of (6), $\mathbf{U} = \operatorname{argmin}_{\mathbf{u}} \mathcal{L}(\mathbf{u}; \mathbf{B})$. From the optimality condition, we have

$$\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{U}; \mathbf{B}, \mu^+, \mu^-) = 0.$$

By applying Taylor's expansion around the center point, $(\mathbf{u}_c, \mathbf{b}, \mu_c^+, \mu_c^-)$, where μ_c^+ and μ_c^- are the optimal dual variables corresponding to \mathbf{u}_c , we have

$$\begin{aligned} \nabla_{\mathbf{u}} \mathcal{L}(\mathbf{U}; \mathbf{B}, \mu) &= 0 \\ &\approx \nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}_c; \mathbf{b}, \mu_c^+, \mu_c^-) + H_{uu}(\mathbf{U} - \mathbf{u}_c) + H_{ub}(\mathbf{B} - \mathbf{b}). \end{aligned} \quad (10)$$

This chapter does not include a formal justification for using the approximation in (10). Proving the approximation formally would strongly rely on smoothness in the convex cost objective and "small" uncertainty in \mathbf{B} , causing a nonsignificant change in the number of tight box constraints.

The Hessian matrices, H_{uu} and H_{ub} , are given by

$$\begin{aligned} [H_{uu}]_{ij} &:= \frac{\partial^2 \mathcal{L}}{\partial [\mathbf{u}]_i \partial [\mathbf{u}]_j} \\ [H_{ub}]_{ij} &:= \frac{\partial^2 \mathcal{L}}{\partial [\mathbf{u}]_i \partial [\mathbf{b}]_j}, \end{aligned}$$

where $[H_{uu}]_{ij}$ and $[H_{ub}]_{ij}$ are the (i, j) th component of H_{uu} and H_{ub} and $[\mathbf{u}]_i$ and $[\mathbf{b}]_i$ are the i th component of \mathbf{u} and \mathbf{b} , respectively.

From (10), we can conclude that

$$H_{uu} \left(\mathbf{U} - \left(\mathbf{u}_c - H_{uu}^{-1} \nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}_c; \mathbf{b}, \mu_c^+, \mu_c^-) \right) \right) \approx -H_{ub}(\mathbf{B} - \mathbf{b}),$$

which implies the linear approximation of optimal solution \mathbf{U} :

$$\mathbf{U} \approx \mathbf{u}_c - H_{uu}^{-1} \nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}_c; \mathbf{b}, \mu_c^+, \mu_c^-) - H_{uu}^{-1} H_{ub}(\mathbf{B} - \mathbf{b}). \quad (11)$$

Observe that expression (11) has the form of a linear control scheme. Indeed, (11) can also be derived using the quadratic approximation of the Lagrangian (approximating $\mathcal{L}(\mathbf{U}; \mathbf{B})$ as quadratic in control \mathbf{U} and linear in input \mathbf{B} in the neighborhood of $(\mathbf{u}_c; \mathbf{b})$).

Therefore, we obtain the following mean and covariance estimate of \mathbf{U} :

$$\mathbb{E}[\mathbf{U}] \approx \underbrace{\mathbf{u}_c - H_{uu}^{-1} \nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}_c; \mathbf{b}, \mu_c^+, \mu_c^-)}_{\mathbf{u}^*} \quad (12)$$

$$\text{Cov}[\mathbf{U}] \approx \underbrace{H_{uu}^{-1} H_{ub} \text{Cov}[\mathbf{B}] H_{ub}^T H_{uu}^{-1}}_{\mathbf{V}^*}. \quad (13)$$

In principle, \mathbf{u}^* and \mathbf{V}^* may be overly crude approximations of the mean and covariance of \mathbf{U} . However, note that the energy cost of the aggregator only depends on the aggregate power consumption. We utilize this aggregation property (large number of loads N) to estimate aggregation statistics below.

Specifically, consider the following aggregate power consumption:

$$\begin{aligned}\mathbf{U}_{\text{total}} &:= \left[\sum_i U_i[0], \dots, \sum_i U_i[K-1] \right]^T \\ &= \underbrace{\left[I_K \dots I_K \right]}_{S \in \mathbb{R}^{K \times NK}} \mathbf{U},\end{aligned}$$

where I_K is the identity matrix in $\mathbb{R}^{K \times K}$. Then (12)–(13) imply the following:

$$\begin{aligned}\mathbb{E}[\mathbf{U}_{\text{total}}] &\approx \mathbf{u}_{\text{total}}^* := S\mathbf{u}^*, \\ \text{Cov}[\mathbf{U}_{\text{total}}] &\approx \mathbf{V}_{\text{total}}^* := S\mathbf{V}^*S^T.\end{aligned}\tag{14}$$

In some cases, the method for selecting a cost-effective utility program might depend on more than just the first two moments of the total loads' power consumption. Using the empirical studies in [13, 35], which utilize multivariate normal random vectors to model day-ahead outside temperature and building occupancy, we apply the property that the linear transformation of multivariate normal random vectors is a multivariate normal vector to obtain

$$\mathbf{U}_{\text{total}} \sim \mathcal{N}(\mathbf{u}_{\text{total}}^*, \mathbf{V}_{\text{total}}^*),\tag{15}$$

where $\mathcal{N}(\cdot, \cdot)$ denotes a multivariate normal distribution. The previous approximation incorporates our data-driven observations that drivers' availability patterns match very well the corresponding building's daily occupancy profiles. The approximation in (15) is used in Section 5 to evaluate risk proneness for each of the n_c programs. In the next section, we illustrate how to incorporate the characterization of \mathbf{V} into the selection of the optimal program.

5 Risk-Aware Program Selection

Using the derived approximation in (15), we can estimate the economic cost, while incorporating aggregator's tolerance to risk when selecting the optimal program \mathcal{P}_j . In particular, we express the aggregator's sensitivity to risk by its tolerance to the uncertainty in the dollar-denominated energy cost term.

One possible program selection rule consists of selecting \mathcal{P}_j that solves

$$\begin{aligned}&\underset{\mathcal{P}_1, \dots, \mathcal{P}_{n_c}}{\text{minimize}} \quad \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i^*[k], k, \mathcal{P}_j \right) \\ &\text{subjectto} \quad \sqrt{\text{Var} \left[\sum_{k=0}^{K-1} c \left(\sum_{i=1}^N U_i[k], k, \mathcal{P}_j \right) \right]} \leq \xi \sum_{k=0}^{K-1} c \left(\sum_{i=1}^N u_i^*[k], k, \mathcal{P}_j \right),\end{aligned}\tag{16}$$

where ξ represents the aggregator's tolerance of risk; smaller values imply a greater sensitivity to uncertainty of future energy charges.

Applying the approximation in (15) leads to a closed-form expression on the left-hand side of the constraint in (16). For example, consider the following linear energy cost function $c(\cdot, \cdot, \cdot)$ as

$$c(U_{\text{total}}[k], k, \mathcal{P}) := \sum_{k=0}^{K-1} \gamma[k] U_{\text{total}}[k],$$

where the time-of-use pricing $\gamma[k]$ depends also on the choice of program P_j . Then

$$\text{Var}\left[\sum_{k=0}^{K-1} \gamma[k] U_{\text{total}}[k]\right] \approx \sum_{k=0}^{K-1} \sum_{k'=0}^{K-1} \gamma[k] \gamma[k] [\mathbf{V}_{\text{total}}^*]_{kk'},$$

where $[\mathbf{V}_{\text{total}}^*]_{ij}$ is the (i, j) component of $\mathbf{V}_{\text{total}}^*$. Similarly, using (15), we can apply the same procedure for any polynomial $c(\cdot, \cdot, \cdot)$, to derive a similar closed-form approximation. By analogy, one can use the approximation in (15) to compute an appropriate variance term for a general function $c(\cdot, \cdot, \cdot)$, e.g., by numerical integration.

In this way, the left-hand side of (16) allows us to identify the most appropriate risk-aware programs. Choosing that program \mathcal{P}_j that minimizes the expected energy charge $\sum_{k=0}^{K-1} c(u_{\text{total}}^*[k], k, \mathcal{P}_j)$ finalizes the decision process.

6 Simulation Results

To illustrate the key ideas of the proposed methodology, in this section, we report some simulations.

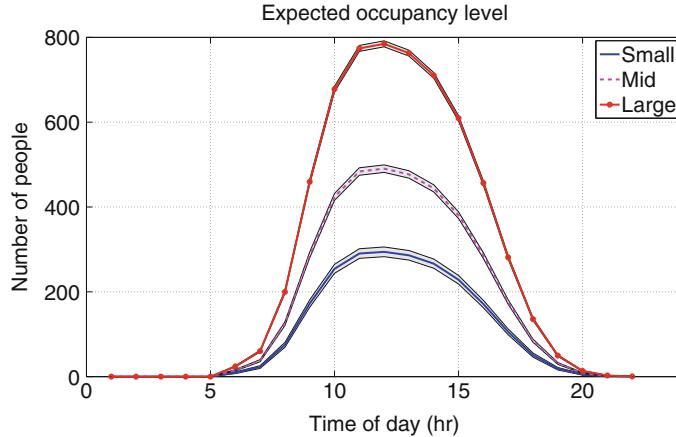
6.1 Simulation Model

For simplicity, our simulation concentrates on HVAC systems. We assume that the load aggregator has 10 small-sized, 10 medium-sized, and 10 large-sized buildings (30 buildings in total). Table 1 summarizes the parameters of these three building types.

To model the occupancy level, \mathbf{O} , we draw independent Gaussian random variables, $\mathbf{O}_{\text{small}}$, \mathbf{O}_{med} , and $\mathbf{O}_{\text{large}}$, corresponding to the occupancy levels of the small-, medium-, and large-sized buildings, respectively. Figure 2 shows mean values of $\mathbf{O}_{\text{small}}$, \mathbf{O}_{med} , and $\mathbf{O}_{\text{large}}$; the exact covariance matrix is available upon request.

Table 1 System parameters.

Building type	Capacity	Efficiency, η ($^{\circ}\text{C}/\text{kW}$)	θ	U (kW)
Small	300	0.2	0.25	30
Mid	500	0.15	0.3	30
Large	800	0.1	0.35	30

**Fig. 2** $E[\mathcal{O}]$ of each building type. Gray area shows $\pm 2\sigma$, where σ is the standard deviation of each \mathcal{O} .

To model the error between actual and forecasted temperatures, Δ_z , we use multivariate Gaussian distribution with the first-order autocorrelation between time periods. The targeted temperature $\bar{x}_i(k)$ is given by 18°C from 6:00 am to 6:00 pm, and 21°C , otherwise. Figure 3 shows the forecasted outdoor and targeted indoor temperatures over the scheduling day.

Finally, we assume that the energy cost function for j th program is given by

$$c(u, k, \mathcal{P}_j) = c_{\text{tou}}(k)u + \rho_j u^2.$$

The first term represents the time-of-use marginal energy costs. The second term, which penalizes large values of allocated power, represents the costs associated with incurring demand charges. ρ_j is a parameter that depends on the utility program \mathcal{P}_j . Table 2 captures $c_{\text{tou}}(k)$.

The discomfort cost trade-off parameter λ is set to $\lambda = 10$, as there are ten of each building type.

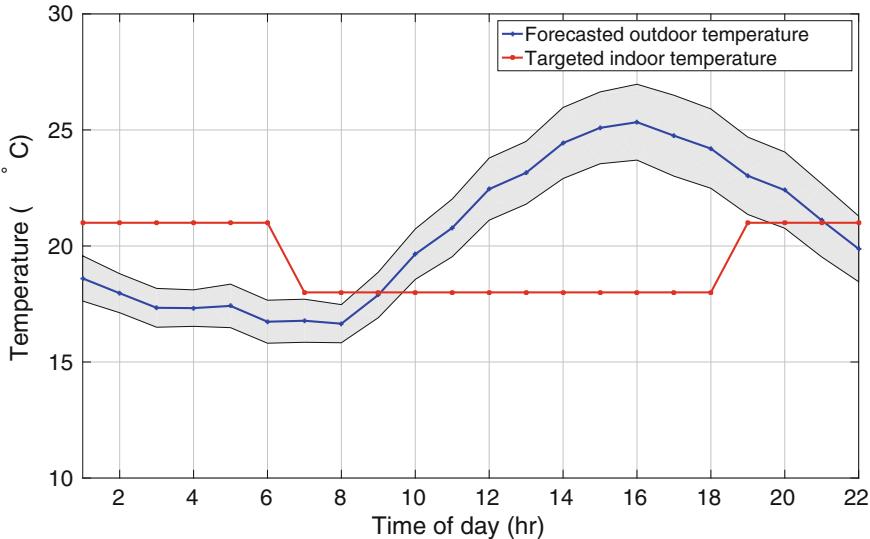


Fig. 3 The forecasted outdoor and the targeted indoor temperatures. The gray area indicates one standard deviation of the forecasting error.

Table 2 Time of usage cost.

From	To	TOU cost (\$/kW)
8:00 pm	10:00 am	0.1038
10:00 am	1:00 pm	0.1805
1:00 pm	6:00 pm	0.2958
6:00 pm	8:00 pm	0.1805

6.2 Simulation Results

First, we compute the day-ahead optimal schedule, $\mathbf{u}_{\text{total}}^*$, by solving (9). Figure 4 shows the optimal power demand for different program parameters ρ_j aggregated across all buildings.

To validate the approximation (15), we estimate the true probability distribution of optimal schedules \mathbf{U}_i using Monte Carlo simulation. To this end, we first generate sample paths for occupancy levels, as described in the previous section. We likewise generate sample paths for actual temperatures. Then, for each sample path, we solve (6) to obtain the optimal solution corresponding to that problem instance. Our empirical studies show that 10^5 sample paths compute mean and variance of $\mathbf{U}_{\text{total}}$ with high accuracy. Note the significant increase in power consumption between 10:00am and 12:00pm due to the cost-effective pre-cooling.

Next, we compute $\mathbf{u}_{\text{total}}^*$, $\mathbf{V}_{\text{total}}^*$ using the formulas (12)–(13). In order to apply these formulas, we solve (9), the result of which incorporates duals, $\mu_{i,k}^+$ and $\mu_{i,k}^-$, and allows us to compute Hessians needed for the approximation. Figure 5 shows $\mathbf{u}_{\text{total}}^*$ and the square root of the diagonal of $\mathbf{V}_{\text{total}}^*$ (estimated standard deviation).

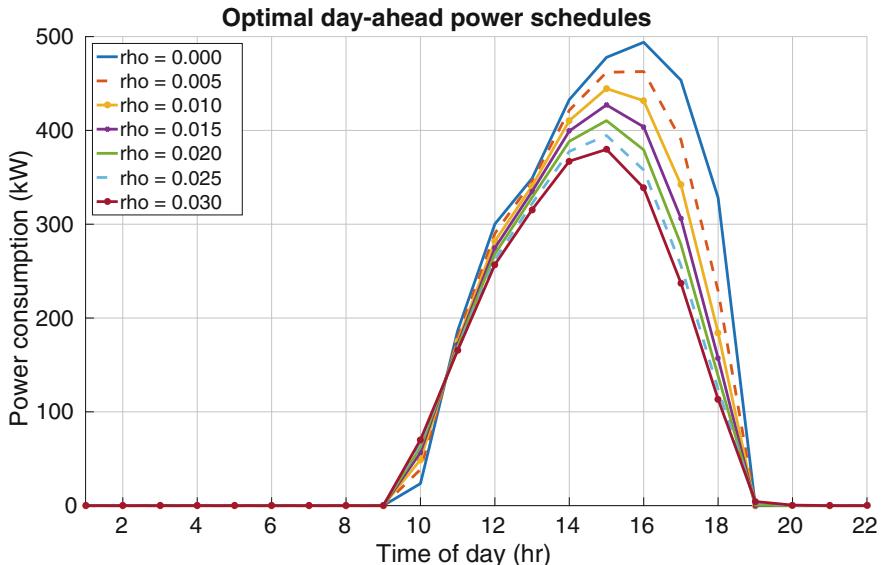


Fig. 4 Aggregated day-ahead optimal schedule, u_{total}^* , for different parameters ρ_j . Larger ρ_j yields smaller peak demands.

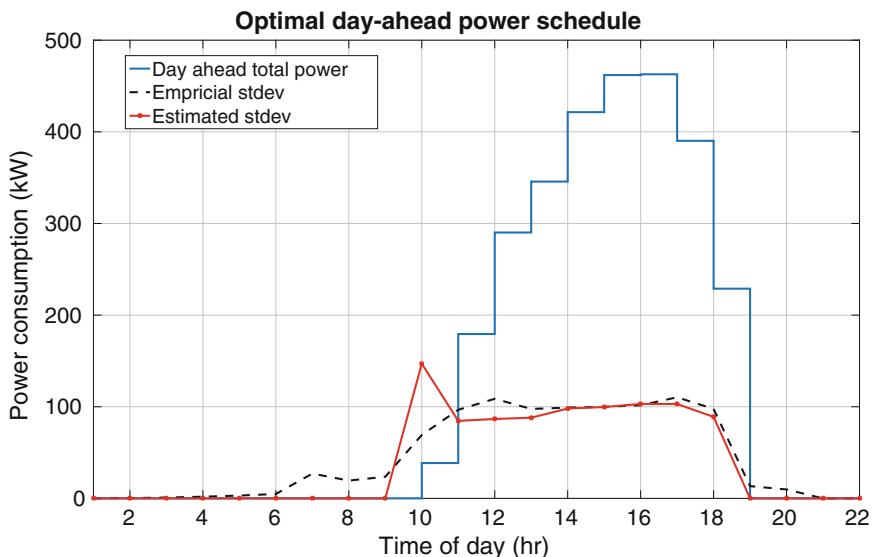


Fig. 5 Approximation of the total consumption variance when $\rho = 0.005$.

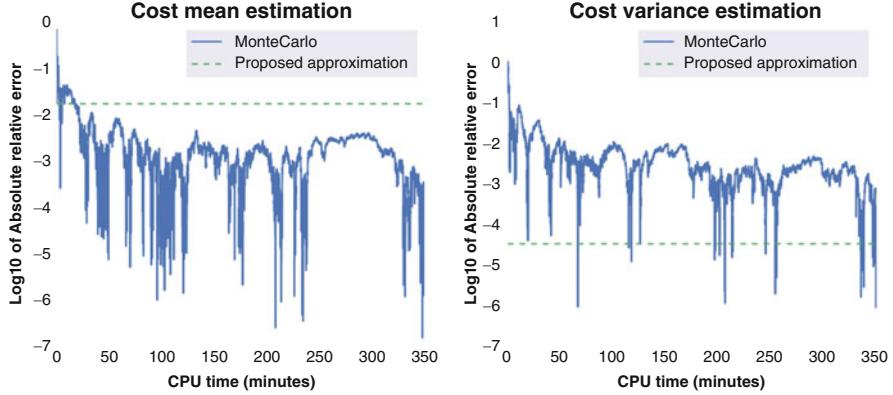


Fig. 6 Timing results of the Monte Carlo simulation. The x-axis represents the CPU time in minutes, and the y-axis gives the ratio of the absolute error to the exact values of the mean and variance. The left plot shows the CPU timing and the corresponding accuracy of the optimal energy cost; the right plot gives the required CPU timing to evaluate the standard error.

Although the proposed approximation V_{total}^* performs less accurately from 6:00am to 10:00am, the impact of this on the overall energy cost $\sum_k c(u_{\text{total}}[k], k, \mathcal{P}_j)$ is negligible. Most of the energy usage happens between 12:00p.m. and 19:00p.m., and the approximation performs well over this period.

Figure 6 shows the CPU time required by the Monte Carlo simulation to characterize the optimal cost. On 3.5 GHz machine, the proposed method takes 14.91 seconds in CPU time, due mostly to solving the optimization (9). Using the same machine, Monte Carlo simulation requires 10 minutes ($\approx 40 \times$ longer) to evaluate the expected value of the optimal cost and 6 hours ($\approx 1440 \times$ longer) to evaluate its standard error.

In Figure 7, we plot the coefficient of variation $\frac{\sqrt{\text{Var}\left[\sum_{k=0}^{K-1} c\left(\sum_{i=1}^N U_i[k], k, \mathcal{P}_j\right)\right]}}{\sum_{k=0}^{K-1} c\left(\sum_{i=1}^N u_i^*[k], k, \mathcal{P}_j\right)}$

and the expected value on log-10 scale. Observe that for larger ρ , the energy cost is higher, but the corresponding coefficient of variation is smaller. For example, if the aggregator chooses $\xi = 10^{-5}$, then the aggregator should also choose the utility program with $\rho = 0.005$.

7 Conclusions

The business model of load aggregators exploits two advantages of load aggregation. First, the larger scale that aggregation affords makes it cost-effective to manage loads wisely. Load aggregators can therefore act as rational economic agents in a way not economically justified for smaller, building-scale loads. Second, the risk

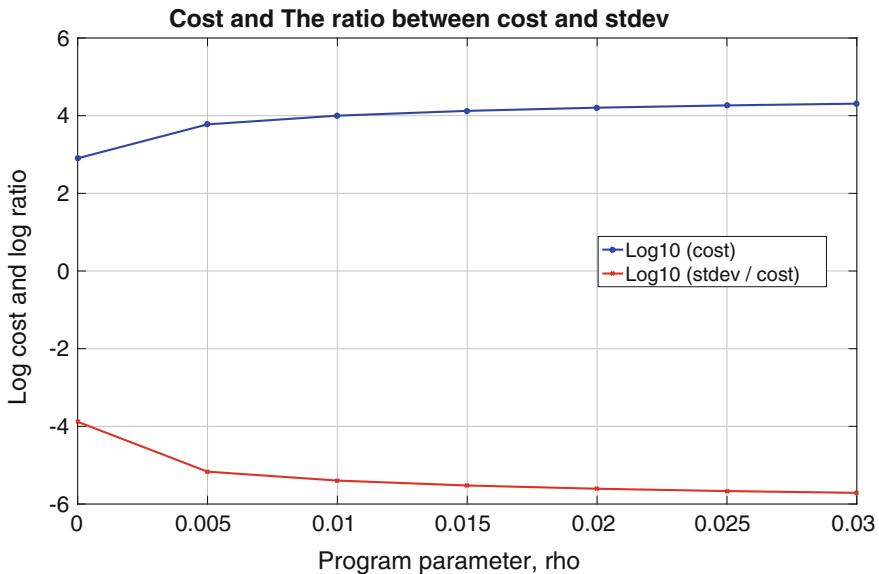


Fig. 7 Expected energy costs and the coefficients of variation as a function of ρ .

mitigation that aggregation offers also enables a market-based transfer of risk from utilities to aggregators, which is the essence of demand response (DR) programs. Aggregation induces a certain kind of central limit effect. Some components of demand uncertainty are relatively independent (e.g., variability in building occupancy). In contrast, some dimensions of demand uncertainty are common to all components, such as the uncertainty due to imperfect temperature forecasts. The structure of total uncertainty, and the balance between the multiple independent and the commonly shared components, continues as a topic for further research.

This paper considers a cost-optimal demand management of an aggregator with a diverse set of flexible loads. We pay special attention to studying the impact on demand planning of the uncertainty from predicting driver availability, forecasting temperatures, and projecting building occupancy. Motivated by the range of objectives that apply when managing Google campus, we assume that utilities offer an aggregator a range of options for tariffs and demand response programs. These programs can significantly impact monthly energy payments. Our optimization objective explicitly incorporates the treatment of user discomfort, making the final decision criteria both more realistic and more adaptable.

Our key contributions derive from incorporating uncertainty in the decision-making framework and from developing algorithms that are computationally efficient. The resulting framework allows us to identify favorable demand reduction schemes in a manner that is properly adapted to the associated risks. Finally, note that our scheme enables a load aggregator to adapt appropriately, both to

a wider range of demand, induced, for example, by high temperatures, and to a greater inventory of supplies, such as those offered by weather-dependent renewable sources.

References

1. Albadri M, El-Saadany E (2008) A summary of demand response in electricity markets. *Electr Power Syst Res* 78(11):1989–1996
2. Bacher P, Madsen H (2011) Identifying suitable models for the heat dynamics of buildings. *Energy Build* 43(7):1511–1522
3. Balijepalli VM, Pradhan V, Khaparde S, Shereef R (2011) Review of demand response under smart grid paradigm. In: Innovative Smart Grid Technologies-India (ISGT India), 2011 IEEE PES. IEEE, Piscataway, pp 236–243
4. Bienstock D, Chertkov M, Harnett S (2014) Chance constrained optimal power flow: risk-aware network control under uncertainty. *Proc Natl Acad Sci* 56:461–495
5. California Public Utilities Commission: Utility tariff information. <http://www.cpuc.ca.gov/PUC/energy/Electric+Rates/utiltariffs/>
6. Caramanis M, Foster JM (2009) Management of electric vehicle charging to mitigate renewable generation intermittency and distribution network congestion. In: Decision and control, 2009 held jointly with the 2009 28th Chinese control conference. Proceedings of the 48th IEEE conference on CDC/CCC 2009. IEEE, Piscataway, pp 4717–4722
7. Chen L, Li N, Jiang L, Low SH (2012) Optimal demand response: problem formulation and deterministic case. In: Control and optimization methods for electric smart grids. Springer, New York, pp 63–85
8. Chen Z, Wu L, Fu Y (2012) Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization. *IEEE Trans Smart Grid* 3(4):1822–1831
9. Chen C, Wang J, Heo Y, Kishore S (2013) MPC-based appliance scheduling for residential building energy management controller. *IEEE Trans Smart Grid* 4(3):1401–1410
10. Chen X, Wei T, Hu S (2013) Uncertainty-aware household appliance scheduling considering dynamic electricity pricing in smart home. *IEEE Trans Smart Grid* 4(2):932–941
11. Chiu WY, Sun H, Poor HV (2013) Energy imbalance management using a robust pricing scheme. *IEEE Trans Smart Grid* 4(2):896–904
12. Cho IK, Meyn, SP (2010) Efficiency and marginal cost pricing in dynamic competitive markets with friction. *Theor Econ* 5(2):215–239
13. Erickson V, Lin Y, Kamthe A, Bramhe R, Surana A, Cerpa E, Sohn D, Narayanan S (2009) Energy efficient building environment control strategies using real-time occupancy measurements. In: Proceedings of the first ACM workshop on embedded sensing systems for energy-efficiency in buildings (BuildSys '09). ACM, New York, pp 19–24
14. FERC Staff Report (2006) Assessment of demand response and advanced metering. Technical Report, Docket AD06-2-00, Federal Energy Regulatory Commission
15. Foster JM, Caramanis MC (2013) Optimal power market participation of plug-in electric vehicles pooled by distribution feeder. *IEEE Trans Power Syst* 28(3):2065–2076
16. Gan L, Topcu U, Low SH (2013) Optimal decentralized protocol for electric vehicle charging. *IEEE Trans Power Syst* 28(2):940–951
17. Gonzalez Vaya M, Andersson G (2013) Optimal bidding strategy of a plug-in electric vehicle aggregator in day-ahead electricity markets. In: 2013 10th international conference on the European energy market (EEM). IEEE, Piscataway, pp 1–6

18. Grahn P, Munkhammar J, Widén J, Alvehag K, Söder L (2013) Phev home-charging model based on residential activity patterns. *IEEE Trans Power Syst* 28(3):2507–2515
19. Halvgaard R, Poulsen NK, Madsen H, Jørgensen JB (2012) Economic model predictive control for building climate control in a smart grid. In: Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES. IEEE, Piscataway, pp 1–6
20. Hao H, Sanandaji BM, Poola K, Vincent TL (2015) Aggregate flexibility of thermostatically controlled loads. *IEEE Trans Power Syst* 30(1):189–198
21. Ilic MD, Xie L, Joo JY (2011) Efficient coordination of wind power and price-responsive demand part I: theoretical foundations. *IEEE Trans Power Syst* 26(4):1875–1884
22. Kennel F, Gorges D, Liu S (2013) Energy management for smart grids with electric vehicles based on hierarchical MPC. *IEEE Trans Ind Inf* 9(3):1528–1537
23. Kraning M, Chu E, Lavaei J, Boyd SP (2014) Dynamic network energy management via proximal message passing. *Found Trends Optim* 1(2):73–126
24. National Energy Technology Laboratory (2007) NETL Modern Grid Initiative – Powering Our 21st-Century Economy
25. Lambert Q (2012) Business Models for an Aggregator-Is an aggregator economically sustainable on Gotland?. MsC thesis, XR – EE – ICS 2012:003, Stockholm, Sweden
26. Li T, Shahidehpour M, Li Z (2007) Risk-constrained bidding strategy with stochastic unit commitment. *IEEE Trans Power Syst* 22(1):449–458
27. Masters G (2013) Renewable and efficient power systems. Wiley, Chichester
28. Mathieu JL, Kamgarpour M, Lygeros J, Andersson G, Callaway DS (2015) Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads. *IEEE Trans Power Syst* 30(2):763–772
29. Mohsenian-Rad AH, Leon-Garcia A (2010) Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Trans Smart Grid* 1(2):120–133
30. Molina-Garcia A, Kessler M, Fuentes JA, Gomez-Lazaro E (2011) Probabilistic characterization of thermostatically controlled loads to model the impact of demand response programs. *IEEE Trans Power Syst* 26(1):241–251
31. Nguyen DT, Le LB (2014) Joint optimization of electric vehicle and home energy scheduling considering user comfort preference. *IEEE Trans Smart Grid* 5(1):188–199
32. Nguyen DT, Le LB (2015) Risk-constrained profit maximization for microgrid aggregators with demand response. *IEEE Trans Smart Grid* 6(1):135–146
33. Office of Energy, G.o.W.A. (2010) U.S. Energy Information Administration. Electric Utility Demand Side Management. <https://www.eia.gov/electricity/data/eia861/dsm/>
34. Pacific Gas and Electric Company: ELECTRIC SCHEDULE E-CBP. http://www.pge.com/tariffs/tm2/pdf/ELEC_SCHEDULE_E-CBP.pdf
35. Radovanović A, Heavlin D, Kiliccote S (2016) Optimized risk-aware nomination strategy in demand response markets. In: The 3rd ACM international conference on systems for energy-efficient built environments (BuildSys ‘16). ACM, New York
36. Rajagopal R, Bitar E, Wu F, Varaiya P (2012) Risk limiting dispatch of wind power. In: 2012 American control conference (ACC). IEEE, Piscataway, pp 4417–4422
37. Roald L, Misra S, Chertkov M, Andersson G (2015) Optimal power flow with weighted chance constraints and general policies for generation control. In: Proceedings of 54th IEEE conference on decision and control. IEEE, Piscataway
38. St. John J (2013) Europe’s new models for demand response. Greentech Media
39. Tashrouch B, Molhim M, Al-Rousan M (2005) Dynamic model of an HVAC system for control analysis. *Energy* 30(10):1729–1745
40. Tsikalakis AG, Hatziargyriou ND (2011) Centralized control for optimizing microgrids operation. In: 2011 IEEE power and energy society general meeting. IEEE, Piscataway, pp 1–8
41. Vagropoulos SI, Bakirtzis AG (2013) Optimal bidding strategy for electric vehicle aggregators in electricity markets. *IEEE Trans Power Syst* 28(4):4031–4041
42. Whitaker C, Newmiller J, Ropp M, Norris B (2008) Distributed photovoltaic system design and technology requirements. Technical Report, SAND2008-0946, Sandia National Laboratories

43. Wu Z, Gu W, Wang R, Yuan X, Liu W (2011) Economic optimal schedule of CHP microgrid system using chance constrained programming and particle swarm optimization. In: 2011 IEEE Power and Energy Society general meeting. IEEE, Piscataway, pp 1–11
44. Zhang W, Lian J, Chang CY, Kalsi K (2013) Aggregated modeling and control of air conditioning loads for demand response. *IEEE Trans Power Syst* 28(4):4655–4664
45. Zhang Y, Gatsis N, Giannakis GB (2013) Robust energy management for microgrids with high-penetration renewables. *IEEE Trans Sustain Energy* 4(4):944–953

Toward Resilience-Aware Resource Allocation and Dispatch in Electricity Distribution Networks



Devendra Shelar, Saurabh Amin, and Ian Hiskens

Abstract This contribution presents an approach to improve the resilience of electricity distribution networks (DNs) to a class of cyber-physical failures by way of optimal allocation of distributed energy resources (DERs). The approach is motivated by the need to adapt the well-known security-constrained optimal power flow problem to DNs with remotely controllable (and, hence, vulnerable) distributed generation sources or loads. To this end, we model the interaction between the system operator (SO) and an external adversary as a three-stage sequential game. In this game, the SO allocates the available resources (Stage 0) and also responds to the adversary's action by optimally dispatching them (Stage 2). The adversary, on the other hand, compromises a subset of vulnerable components with the objective of inducing operating bound violations (Stage 1). We consider qualitatively different allocation strategies in Stage 0 and develop a scalable greedy heuristic to solve Stages 1–2 (i.e., bilevel optimization problem). We utilize this greedy heuristic to obtain structural insights about optimal adversarial compromises and desirable allocation strategies of the SO.

D. Shelar · S. Amin (✉)

Department of Civil and Environmental Engineering, Massachusetts Institute of Technology,
Cambridge, MA 02139 USA

e-mail: shelard@mit.edu; amins@mit.edu

I. Hiskens

Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109, USA
e-mail: hiskens@umich.edu

Table of Notations

Radial DN parameters

\mathcal{N}	set of nodes in DN
\mathcal{E}	set of edges in DN
\mathcal{G}	tree topology of the radial DN $\mathcal{G} = (\mathcal{N}, \mathcal{E})$
r_j	resistance of line $(i, j) \in \mathcal{E}$
x_j	reactance of line $(i, j) \in \mathcal{E}$
z_j	impedance $z_j = r_j + jx_j$ of line $(i, j) \in \mathcal{E}$
\mathcal{P}_i	path from the root node to node i
Z_{ij}	$Z_{ij} := \sum_{k \in \mathcal{P}_i \cap \mathcal{P}_j} z_k$ common path impedances between nodes i and j
j	complex square root of -1, $j = \sqrt{-1}$

Nodal quantities of node $i \in \mathcal{N}$ in state $\eta \in \{o, c\}$

pc_i^η, qc_i^η	active and reactive power consumed at node i
pg_i^η, qg_i^η	active and reactive power generated at node i
v_i^η	complex voltage at node i
$\underline{V}_i, \bar{V}_i$	lower and upper bounds on square of voltage magnitude at node i
f_i^η	frequency of energy resource at node i

Edge quantities of edge $(i, j) \in \mathcal{E}$ in state $\eta \in \{o, c\}$

p_j^η, q_j^η	active and reactive power flowing on line (i, j)
----------------------	--

Attacker model

δ_i	$\delta_i = 1$ if EV i is compromised; 0 otherwise.
------------	---

SO model

$\bar{\gamma}_i$	maximum allowed fraction of load control
γ_i	fraction of load control at load i
sg^c	SO response set-point of DER i
ϕ	$\phi := (sg^c, \gamma^c)$ SO response strategy

1 Introduction

In this chapter we introduce the *resilience-aware optimal power flow* (RAOPF) problem and discuss its relevance to optimal allocation and dispatch of contingency resources in the face of cyber-physical failures in electricity distribution networks. Our contribution is motivated by the need to adapt (and extend) the classical security constrained optimal power flow (SCOPF) problem [1, 14] to the contingencies resulting from targeted compromise (attack) of remotely accessible nodes in

distribution networks (DNs), e.g., security attacks to DERs or electric vehicle (EV) charging facilities. We model DN as a radial network with bulk generator (BG) at the substation node as well as spatially distributed DERs. We assume that the BG has a finite ramp rate; thus, regulation of system frequency becomes relevant in our formulation (in addition to voltage regulation).¹ The RAOPF problem provides optimal dispatch of DERs and optimal shedding of controllable loads to limit the cost of maintaining regulation objectives during attack-induced contingencies.

The underlying challenge that motivates for our work is optimal resource allocation to improve resilience of DNs to simultaneous component failures that can lead to contingency events. We view DERs and controllable loads as *resources* that can be used (dispatched) after the contingency events. For a given attack (or a compromised set of components), we say that a resource allocation is *more resilient* than another if an appropriately defined post-contingency cost (weighted sum of network costs and the cost of load control) is less than the cost in the latter case. Furthermore, we say that a resource allocation is *optimal* if it minimizes the sum of cost of resource allocation and the “worst-case” post-contingency cost under a set of failure scenarios. To capture these properties, we formulate a three-stage optimization problem with network and resource constraints to evaluate the total cost for a range of resource allocation strategies under security attacks to the DN nodes. We call this formulation as RAOPF to emphasize the resiliency improving aspect of the resulting allocation. Our solution illustrates important trade-offs in allocating spatially distributed resources by accounting for the nature of their contribution (active *vs.* reactive power) and their spatial location (upstream *vs.* downstream).

The RAOPF problem is constrained by the power flow equations which are physical laws and, therefore, must be satisfied. The other constraints include technological specifications of BG (droop characteristics), DERs (apparent power capability, active and reactive power set points), EV facilities (charging rate), and loads (controllable versus noncontrollable parts). Finally, the operating constraints, which model the frequency and voltage regulation as well as line capacities, are imposed in the nominal mode. However, one or more of these operating constraints may be violated as a result of an adversarial action of the attacker; in our formulation, such violations result in a contingency. Thus, we view a *contingency* as a sudden, unplanned incident caused due to failure of one or more components that has a direct effect on the operating constraints of the DN [5].

To prevent or limit the impact of contingencies, we allow DERs to be allocated at the nodes of the DN, in addition to the supply by the BG; see Figure 1. Any point on the supply-demand balance line is a resource allocation that determines the amount of power supplied by the BG and the amount of power supplied by the DERs. If the controllable loads are also curtailed, then the supply-demand line shifts inward due to reduction in aggregate demand. In our formulation, the capacity of an

¹Thus, our formulation is especially relevant to resiliency issues in isolated microgrids.

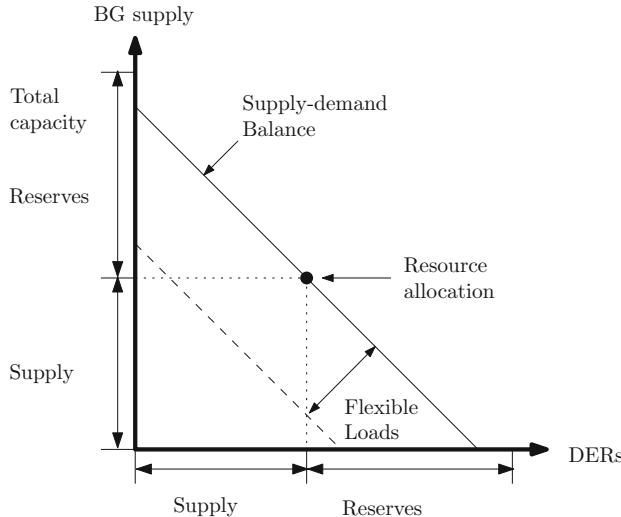


Fig. 1 An illustration of power allocation through a BG and spatially distributed DERs.

energy resource (BG or the DERs) in *excess* of the power supplied by the resource determines the *reserves* provided by that energy resource.

In the post-contingency situation, violations of operational constraint(s) must be contained by the system operator (SO). If such violations are not resolved in a timely manner, additional components may fail, which can result in new contingencies. For example, significant loss of DER supply in highly loaded DNs may result in a drop in node voltages below a critical threshold causing other supply sources to trip, potentially resulting in a network effect (or cascade) [17]. Thus, planning for sufficient resources is essential so that the SO is able to meet regulation objectives in contingency situations. Typically, these objectives include voltage regulation (VR), frequency regulation (FR), and capacity management (CM) [7]. In particular, lack of adequate active power resources can cause loss of frequency regulation, and the scarcity of reactive power resources can lead to voltage fluctuations. In addition, in many situations, the capacity of one or more lines limits the reallocation of power that is needed to serve demand during contingencies [5, 16]. These factors have been identified as crucial for resilience of electricity grids [3, 5, 23] and are poised to become significant even for DNs.

In recent years, thanks to technological improvements and reduced cost of deployment, DERs have emerged as a promising solution for provision of reserves; in particular, by means of active and reactive power control [7, 21]. These functionalities are enabled by the appropriate power electronics and allow the DERs to respond to a range of fluctuations in a fast manner (order of milliseconds) as opposed to the slower response via traditional means, which is typically in the order of few seconds to few minutes. Thus, allocation of DERs as reserves to facilitate

fast response for meeting regulation objectives is an important problem in its own right; in this work, we instantiate this problem in the context of DN resilience.

Our work is also motivated by the SCOPF formulation which is used for contingency planning in transmission networks. In many transmission systems, the SOs solve some form of the SCOPF problem for the operational planning and dispatch by considering a given (a priori known) set of reliability failures [8]. By solving SCOPF, the SO is able to compute a resource allocation strategy which allows for timely response to any contingency resulting from these reliability failures [1, 4, 14]. The idea behind our formulation is similar to SCOPF; the main distinction is that we capture the contingency situations resulting from the action of a strategic attacker to DN components.

We argue that the RAOPF problem can be used by the SO to compute the optimal resource allocation as well as response for DNs under strategic disruptions of supply-demand nodes. The problem is challenging because the individual objectives VR, FR, and CM are not exactly aligned with each other. As a result, there are trade-offs in the optimal resource allocation, which our modeling framework captures. Admittedly, the focus of RAOPF is limited to adversarial compromise of supply-demand DN nodes, and its extension to *all* possible N-k contingencies is an open question.² Still, RAOPF provides important insights regarding the structure of the optimal attack and the SO's strategies (both allocation and dispatch of reserves).

We formulate the RAOPF problem as a Stackelberg game consisting of three levels (stages). The upper-level (Stage 0) problem represents the SO's problem of resource allocation for optimal power flow and planning of reserves in anticipation of an attack. The middle-level (Stage 1) problem represents a contingency model that captures the impact of attacker-induced failures on the aggregate supply-demand balance. In the lower-level (Stage 2) problem, the SO controls available reserves to utilize the existing reserves and, if required, also impose load shedding. In the last two stages of the game, the objective of the attacker (resp., SO) is to maximize (resp., minimize) the post-contingency cost (i.e., weighted sum of cost incurred due to violation in VR, FR, and CM) and cost of load shedding subject to constraints due to power flow and DER/load models. In Stage 0, the SO's objective is to minimize the sum of cost of resource allocation and the maximin post-contingency cost.

The decisions in each of the three stages can be summarized as follows:

- *Stage 2*: Given a fixed reserve allocation and a fixed contingency, what is the optimal SO response in terms of dispatch of available resources (and load shedding)?
- *Stage 1*: Given a fixed reserve allocation and the assumed attacker model, what is the optimal attack that maximizes the post-contingency cost, assuming the SO will respond optimally?

²If we explicitly enumerated all N-k contingencies, the number of constraints will increase exponentially with N and k. For example, with N = 100 and k = 10, the number of constraints will be of order 10^{13} , which makes the problem computationally hard.

- *Stage 0:* What should be the optimal allocation of supply resources across the BG and DERs, assuming the optimal strategies of the attacker and the SO in Stages 1 and 2, respectively?

In Section 2, we introduce our DN model and operating constraints. Then, in Section 3, we formulate the last two stages of the RAOPF problem as a bilevel optimization problem. Next, in Section 4, we present our computational approach to the bilevel problem and evaluate its performance with the help of a case study. In Section 5, we append Stage 0 to the bilevel problem and present the complete formulation of the RAOPF problem. Developing a computationally tractable solution approach to the RAOPF formulation is part of our ongoing work (and thus, it is beyond the scope of this contribution); however, we present a few insights on the optimal attacker strategy and also discuss the main trade-offs faced by the SO in minimizing the post-contingency cost. These trade-offs directly influence the qualitative structure of SO’s resource allocation strategy. While the allocation strategies that we consider are not necessarily optimal (in the sense of our RAOPF problem), we argue that their qualitative structure is relevant for construction of optimal allocation strategy. Finally, we draw some concluding remarks in Section 6.

2 Network Model

In this section, we first introduce the basic notations in our network model and define the state variables. Then, we describe the operating constraints, namely, the power flow equations and operating limits. These constraints also include approximate models that relate the deviations in the frequency and nodal voltages in pre- and post-contingency modes (i.e., before and after an adversarial compromise).

2.1 Radial Distribution Network Model

We build on the classical model for radial DNs [9, 19, 22] ; see Section 6 for notations. Consider a tree network of nodes and distribution lines $\mathcal{G} = (\mathcal{N} \cup \{0\}, \mathcal{E})$, where \mathcal{N} denotes the set of all DN nodes. The substation node is labeled as 0. Let $N := |\mathcal{N}|$. A distribution line connecting node j to its parent node i in the tree network is denoted $(i, j) \in \mathcal{E}$. Each distribution line $(i, j) \in \mathcal{E}$ has a complex impedance $z_j = r_j + jx_j$, where $r_j > 0$ and $x_j > 0$ denote the resistance and inductance of the line (i, j) , respectively, and $j = \sqrt{-1}$.

We distinguish between two *modes*, denoted $\eta \in \{o, c\}$, where o and c denote the pre- and post- contingency modes, respectively. The state vector in mode η , denoted $x^\eta \in \mathbb{R}^{4N}$, is defined as

$$x^\eta := [pc^\eta, qc^\eta, pg^\eta, qg^\eta],$$

where pc_j^η and pg_j^η (resp., qc_j^η and qg_j^η) denote the active (resp., reactive) power consumption and generation at node j . For a given mode η , let p_j^η and q_j^η denote the active and reactive power flowing from node i to node j on the line $(i, j) \in \mathcal{E}$ and v_i^η denote the voltage magnitude of node i ; see power flow equations in (1) below. Throughout this chapter, x^η , v^η , pc^η , qc^η , pg^η , qg^η , p^η , and q^η are row vectors of appropriate dimensions.

In our model, the BG is connected to the substation node 0, and any other node $i \in \mathcal{N} \setminus \{0\}$ may or may not have a DER connected to it. Let f_0^η denote the frequency of the BG and f_i^η denote the frequency of DER at node i . Throughout, we will assume that the frequencies of individual DERs are synchronized with that of the BG, i.e., $f_i^\eta = f_0^\eta$. Thus, we refer the BG frequency as the *system frequency* and drop the subscript 0 in f_0^η . We will nominally assume that $f^o = 60$ Hz and $v_0^o = 1$ pu in the pre-contingency mode.

2.2 Constraints

The constraints in our network model comprise of the power flow equations, voltage/frequency deviation models, operating limits in the pre-contingency mode, and models of generators (BG and DERs) and loads (EV and non-EV components).

Linear power flows (LPF): For a state x^η , the standard LPF model can be written as [9, 11]

$$p_j^\eta = \sum_{k:(j,k) \in \mathcal{E}} p_k^\eta + pc_j^\eta - pg_j^\eta \quad \forall j \in \mathcal{N}, \eta \in \{o, c\} \quad (1a)$$

$$q_j^\eta = \sum_{k:(j,k) \in \mathcal{E}} q_k^\eta + qc_j^\eta - qg_j^\eta \quad \forall j \in \mathcal{N}, \eta \in \{o, c\} \quad (1b)$$

$$v_j^\eta = v_i^\eta - r_j p_j^\eta - x_j q_j^\eta \quad \forall (i, j) \in \mathcal{E}, \eta \in \{o, c\} \quad (1c)$$

Here, (1a) (resp., (1b)) is the active (resp., reactive) power conservation equations; (1c) relates the voltage drop and the power flows. We will use the notation \mathcal{X}^o and \mathcal{X}^c to denote the sets of states that satisfy (1) for $\eta = o$ and $\eta = c$, respectively.³

Frequency and voltage deviation models: In our model, the ramp rate of BG is a limiting factor and impacts the deviation in system frequency as well as the deviation in nodal voltages between pre- and post-contingency modes. Following [2, 10], the change in frequency and substation voltage from the pre-contingency state x^o to post-contingency state x^c are related as follows:

³Note that, in this contribution, we used the LPF model for the sake of simplicity and computational tractability. However, our main ideas are also relevant to DN with nonlinear power flows.

$$f^o - f^c = -\mathbf{f}^{reg} (p_0^o - p_0^c) \quad (2a)$$

$$v_0^o - v_0^c = -\mathbf{v}^{reg} (q_0^o - q_0^c), \quad (2b)$$

where \mathbf{f}^{reg} is the frequency regulation (or droop) constant of the BG that captures the change in frequency (in Hz) per unit additional active power supplied into the substation node and \mathbf{v}^{reg} is the voltage regulation constant of the BG that captures the per unit change in voltage per unit additional reactive power supplied into the substation node.

Operating limits: Let f_i^{min} and f_i^{max} denote the (given) allowable lower and upper bounds within which the system frequency should operate for the DER at node i , and define $\underline{\mathbf{f}} := \max_{i \in \mathcal{N}} f_i^{min}$ and $\overline{\mathbf{f}} := \min_{i \in \mathcal{N}} f_i^{max}$. Similarly, let $\underline{\mathbf{v}}_i$ and $\overline{\mathbf{v}}_i$ denote the lower and upper bounds within which the voltage at node i should be maintained. Finally, let \bar{s}_j denote the maximum power carrying capacity of line (i, j) .

Now, we can state the operating limits for the pre-contingency state x^o :

$$\underline{\mathbf{f}} \leq f^o \leq \overline{\mathbf{f}} \quad (3a)$$

$$\underline{\mathbf{v}}_i \leq v_i^o \leq \overline{\mathbf{v}}_i \quad \forall i \in \mathcal{N} \quad (3b)$$

$$(p_j^o)^2 + (q_j^o)^2 \leq \bar{s}_j^2 \quad \forall j \in \mathcal{N} \text{ s.t. } (i, j) \in \mathcal{E} \quad (3c)$$

where (3a) and (3b) specify the lower and upper bounds for the system frequency and nodal voltages, and (3c) models the capacity of the distribution lines.

In principle, similar regulation requirements can also be stated for the post-contingency state x^c . However, in our framework, the post-contingency state is a result of attacker-SO interaction and thus cannot be expressed explicitly. Thus, we choose to model the worst-case contingency (see Section 3) and consider violations in operating limits in the post-contingency mode as costs (as opposed to constraints).

Bulk generator and DER model: Let $sg_i := pg_i + jqg_i$ denote the complex power supplied by the generator at the node i , where pg_i and qg_i denote the active and reactive power components. The generator output is constrained as follows:

$$sg_i \in \mathcal{S}_i,$$

where \mathcal{S}_i is assumed to be a convex set [6, 22]. We consider the following convex sets as candidates for \mathcal{S}_i :

$$\mathcal{S}_i^{circ} := \{(p, q) \mid 0 \leq p \leq \overline{pg}_i, \underline{qg}_i \leq q \leq \overline{qg}_i, p^2 + q^2 \leq \bar{s}_i^2\}, \text{ or} \quad (4)$$

$$\mathcal{S}_i^{poly} := \{(p, q) \mid 0 \leq p \leq \overline{pg}_i, \underline{qg}_i \leq q \leq \overline{qg}_i, A_i^p p + A_i^q q \leq b_i\}, \quad (5)$$

where \overline{pg}_i denotes the maximum active power bound for the DER output and \underline{qg}_i and \overline{qg}_i denote the minimum and maximum reactive power bounds. Note that if

node i has no DER, we can conveniently choose $\bar{s}g_i = 0$. Finally, we denote the set of feasible set points for all the generators (i.e., BG and DERs) by $\mathcal{S} := \prod_{i \in \mathcal{N}} \mathcal{S}_i$.

Load models: For the sake of illustration, we consider that electric vehicles (EVs) connected to the DN are the only nodes vulnerable to compromise by the attacker. Without loss of generality, we assume that each node has an EV load and a non-EV load. For the mode η , let se_i^η and sn_i^η denote power consumed by the EV and non-EV load at node i . Then, the total power consumed, sc_i^η , can be written as

$$sc_i^\eta = se_i^\eta + sn_i^\eta. \quad (6)$$

Next, we introduce non-EV and EV load models.

Non-EV load model: We assume that non-EV loads are constant power loads.⁴ Let \bar{sn}_i denote the nominal demand of non-EV load at node i . However, to maintain the operating limits of the DN in the post-contingency mode, we allow the SO to shed a part of nominal load. This flexibility is modeled by introducing a parameter $\gamma_i^\eta \in [0, \bar{\gamma}_i]$, where $\bar{\gamma}_i \in [0, 1]$ denotes the maximum load control capability at the node i . As an example, $\bar{\gamma}_i = 0.1$ would mean that a maximum of 10% of the non-EV load at node i can be shed. Thus, the actual power consumed by the non-EV load can be expressed as follows:

$$sn_i^\eta = (1 - \gamma_i^\eta)\bar{sn}_i. \quad (7)$$

For simplicity, we also assume that the SO fulfills all non-EV demand in the pre-contingency mode, i.e. $\gamma_i^o = 0 \forall i \in \mathcal{N}$.

EV load model: Typically, EV loads are modeled as constant power loads. For simplicity, we only allow two charging rates for each EV, viz., slow and fast. Let $\mathcal{S}_i^e = \{\underline{se}_i, \bar{se}_i\}$ denote the set of charging rates of EV at node i , where \underline{se}_i (resp., \bar{se}_i) is the slow (resp., fast) charging rate of EV at node i . Thus, the power consumed by the EV load is given by

$$se_i^\eta = \delta_i^\eta \bar{se}_i + (1 - \delta_i^\eta) \underline{se}_i, \quad (8)$$

where the binary variable $\delta_i^\eta = 0$ (resp., $\delta_i^\eta = 1$) indicates the slow (resp., fast) charging rate.

Henceforth, we will limit our attention to attacker-induced compromise of EVs, i.e., we focus on a scenario in which a subset of EVs can be simultaneously compromised by an external adversary to induce the contingency mode. Before moving further, we want to emphasize that we selected the specific scenario of attack to EVs for the sake of concreteness. Indeed, our approach can be adopted

⁴More generally, non-EV loads can be modeled using the constant impedance (Z), constant current (I), constant power (P), or a general ZIP model. The non-EV power consumption can also change due to frequency deviations. Our network model can be extended to include these general load models.

to other scenarios that require resource allocation and dispatch on part of the SO to resolve the supply-demand imbalance created as a result of cyber-physical failures (attack).

3 Attacker-SO Interaction (Bilevel Problem)

In this section, we describe the attacker-SO interactions during the contingency caused by compromise of vulnerable EV loads. Specifically, we consider the contingency caused by a simultaneous compromise of EV loads from low to high charging rates which results in a sudden increase in the aggregate demand [12]. The attacker selects the EVs in a targeted manner to induce violations in one or more operating limits, which can result in an increased cost of regulation for the SO in the post-contingency mode. To limit this cost, the SO responds by dispatching the DERs as contingency reserves and, if necessary, by exercising load control. Thus, the attacker's (resp., SO's) objective is to maximize (resp., minimize) the post-contingency cost (sum of attacker-induced network operating costs and forced/load shedding).

We model the attacker-SO interaction as a sequential game in which the attacker moves first and the SO responds next. We now describe these stages in detail.

Attack stage: Let $\Psi_k := \{\delta \in \{0, 1\}^N \mid \sum_{i \in N} \delta_i \leq k\}$ denote the set of feasible strategies of a resource-constrained attacker. In our model, the attacker chooses a subset of EVs to compromise and sets their rate of charging to $\delta^a \in \Psi_k$. Here, $\delta_i^a = 1$ means that EV at node i is compromised and starts charging at the faster rate; $\delta_i^a = 0$ implies otherwise. The attacker's action is constrained as follows:

$$\sum_{i \in N} \delta_i^a \leq k, \quad (9)$$

where (9) states that at most k EV nodes are compromised. Recalling (8), we know that the attacker's action determines the effective charging rates in the post-contingency mode:

$$\delta^c = \delta^a. \quad (10)$$

The resource constraint (9) on attacker's action captures his limited capability in compromising spatially distributed EVs. We justify this constraint in the following way: First, the EV nodes are likely to be heterogeneous in their design and manufacturer type. The attacker may not have specific attack paths for each EV type. Second, the process of EV integration with DNs is gradual in nature, and there aren't any security regulations that the EV facilities must implement. Some of them may install intrusion prevention/detection tools to safeguard the software controlling the charging rate and/or preventing the EVs from overcharging; however, the remaining

facilities will remain vulnerable. Third, certain electric cars may have a buggy control software that is vulnerable to a virus, which can compromise certain types of EV facilities [13]. Hence, the number of facilities that could be compromised simultaneously may be proportional to the number of electric cars with the buggy control software.

Without much loss of generality, we assume that the EVs when fully charged do not remain connected to the DN and, hence, are not vulnerable to attack; i.e., the attacker only targets the EVs that are not fully charged. As a consequence, we do not include the state-of-charge constraints of the EVs in our model. Furthermore, to induce the maximal impact in the post-contingency mode, the attacker will only target EVs that were charging at the slow rate in the pre-contingency mode. Hence, for simplicity, we can assume that for all EV nodes, $\delta_i^o = 0$ in (8), i.e., $se_i^o = \underline{se}_i$.

Note that attacks to other components (e.g., DERs, non-EV loads) can be modeled in a similar manner. For example, in our previous work [19, 20], we considered attacks that manipulated DER set points. Thus, while the specific channel of attack might vary across different scenarios, the net effect is change in network state between pre- and post-contingency modes (to see this, notice how (6)–(8) and (10) affect (1) and (2)). Also note that, although issues such as reverse power flows and overvoltages do not arise in our model, they may become relevant in other scenarios, e.g., when the attacker introduces sudden disconnection of loads or simultaneously turns a large number of EVs to slow charging rate. We expect that even in such scenarios, the basic nature of attacker-SO interaction will be similar to our model.

SO response stage: Let $\Phi := \mathcal{S} \times \Gamma$, where $\Gamma := \prod_{i \in \mathcal{N}} [0, \bar{\gamma}_i]$. In our model, the SO responds to attacker actions by choosing the set points of the non-compromised DERs and, if needed, imposes load curtailment at one or more nodes according to a strategy $[sg^c, \gamma^c] =: \phi \in \Phi$. Essentially, the SO chooses new set points sg^c of non-compromised DERs and load control parameters γ^c to reduce the post-contingency cost. These choice variables are captured by strategy vector ϕ .

We make the standard assumption that the SO knows the nominal non-EV (\overline{sn}) and EV demand (\underline{se}). Additionally, we assume that the SO has full observability of network state; this can be achieved by continuously monitoring nodal voltages. Under this assumption, the SO can determine the identity of compromised EVs and use this knowledge to compute the optimal response to attack. Relaxing this assumption would entail designing SO response with imperfect state information. While this issue is of practical relevance, we do not pursue it here.

For a fixed resource allocation in the pre-contingency mode (i.e., for given x^o), we can now represent attack and SO response stages in the following maximin (or bilevel) formulation as follows:

$$\begin{aligned} [\text{Maxmin}] \quad \mathcal{L}(x^o) := & \max_{\delta^a \in \Psi_k} \min_{\phi \in \Phi} C_{\text{loss}}(x^o, x^c(\delta^a, \phi)) \\ & \text{s.t. (1), (2), (6)–(8), (10)} \end{aligned} \tag{11}$$

Here, we model the post-contingency cost as a sum of the cost due to the voltage bound violation (C_{VR}), the cost due to frequency bound violation (C_{FR}), and the cost due to load control:⁵

$$C_{\text{loss}} := C_{\text{VR}} + C_{\text{FR}} + C_{\text{LC}} \quad (12a)$$

$$C_{\text{VR}}(x^o, x^c) := W_{\text{VR}} \max_{i \in \mathcal{N}} \max(v_i - v_i^c, v_i^c - \bar{v}_i, 0) \quad (12b)$$

$$C_{\text{FR}}(x^o, x^c) := W_{\text{FR}} \max(f - f^c, f^c - \bar{f}, 0) \quad (12c)$$

$$C_{\text{LC}}(x^o, x^c) := W_{\text{LC}} \cdot \gamma^c, \quad (12d)$$

where W_{VR} and W_{FR} denote the coefficients assigned to the voltage and frequency regulation objectives and the vector $W_{\text{LC}} \in \mathbb{R}_{+}^{\mathcal{N}}$ represents the cost of unit load shedding after the contingency. Note that, in (12b), the cost of voltage regulation is defined as the maximum voltage bound violation over all nodes.

Although the [Maxmin] problem does not consider nonlinear power flow, it turns out that optimal value of this problem is a lower bound of the maximin loss in the post-contingency mode under nonlinear power flows [20]. Furthermore, under certain additional assumptions, we can also use solution to the [Maxmin] problem for an appropriately modified LPF model to upper bound the maximin loss. For more details on establishing these bounds, we refer the reader to [20].

4 Greedy Heuristic Approach for [Maxmin] Problem

We now focus on solving the [Maxmin] problem which is a bilevel mixed integer linear program with the inner problem being a linear program. A standard approach to solving such problems is the KKT-based reformulation approach which gives a single level mixed-integer linear program (MILP) [15, 25, 26]. In principle, the MILP reformulation can be used to solve the [Maxmin] problem for small-sized networks. However, scaling this approach to larger networks is not straightforward and entails finding reasonable upper bounds on the Lagrange multipliers in the KKT conditions. In our previous work [19, 20], we have investigated an alternative approach which exploits the properties of linear power flows on radial networks to develop a greedy heuristic that is scalable to large-sized networks. We apply this heuristic to the [Maxmin] problem. With the help of a case study, we also compare the results obtained from this heuristic with those obtained by the KKT-based MILP reformulation approach and brute force (when possible).

Before proceeding further, we need to introduce some additional notation. For any given node $i \in \mathcal{N}$, let \mathcal{P}_i be the path from the root node to node i . Thus, \mathcal{P}_i is

⁵For simplicity, we only focus on voltage and frequency regulation, and do not consider congestion management (CM) as a regulation objective. That is, we assume that constraints (3c) will not be active.

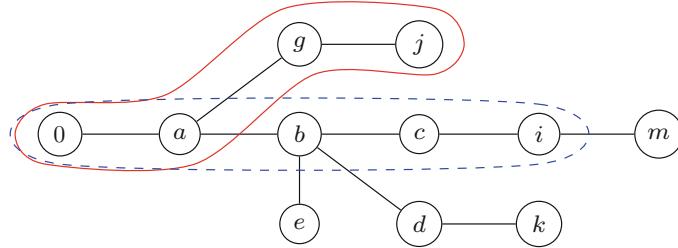


Fig. 2 Precedence description of the nodes for a tree network. Here, $j \prec_i k$, $e =_i k$, $b \prec k$, $\mathcal{P}_j = \{a, g, j\}$, $\mathcal{P}_i \cap \mathcal{P}_j = \{a\}$.

an ordered set of nodes starting from the root node and ending at node i , excluding the root node; see Figure 2. We say that node j is an *ancestor* of node k ($j \prec k$) or, equivalently, k is a successor of j iff $\mathcal{P}_j \subset \mathcal{P}_k$. We define the *relative ordering* \preceq_i , with respect to a “pivot” node i as follows:

- j precedes k ($j \preceq_i k$) iff $\mathcal{P}_i \cap \mathcal{P}_j \subseteq \mathcal{P}_i \cap \mathcal{P}_k$.
- j strictly precedes k ($j \prec_i k$) iff $\mathcal{P}_i \cap \mathcal{P}_j \subset \mathcal{P}_i \cap \mathcal{P}_k$.
- j is at the same precedence level as k ($j =_i k$) iff $\mathcal{P}_i \cap \mathcal{P}_j = \mathcal{P}_i \cap \mathcal{P}_k$.

We define the common path impedance between any two nodes $i, j \in \mathcal{N}$ as the sum of impedances of the lines in the intersection of paths \mathcal{P}_i and \mathcal{P}_j , i.e., $Z_{ij} := \sum_{k \in \mathcal{P}_i \cap \mathcal{P}_j} z_k$, and denote the resistive (real) and inductive (imaginary) components of Z_{ij} by R_{ij} and X_{ij} , respectively. Z , R , and X denote the corresponding matrices of appropriate sizes.

We can use the abovementioned notion of node precedence to describe the structure of an optimal attack given a *fixed* SO action (or response).

4.1 Optimal Attack for Fixed SO Response

Following the standard approach [18, 24], we define the master problem [Maxmin – a] (respectively, subproblem [Maxmin – d]) for fixed SO action $\phi \in \Phi$ (respectively, fixed attacker action $\delta \in \Psi_k$) as follows:

$$\begin{aligned}
 [\text{Maxmin} - \text{a}] \quad & \delta^*(\phi) \in \underset{\delta \in \Psi_k}{\operatorname{argmax}} C_{\text{loss}}(x^o, x^c(\delta, \phi)) \\
 & \text{s.t. (1), (2), (6)–(8), (10)} \\
 [\text{Maxmin} - \text{d}] \quad & \phi^*(\delta) \in \underset{\phi \in \Phi}{\operatorname{argmin}} C_{\text{loss}}(x^o, x^c(\delta, \phi)) \\
 & \text{s.t. (1), (2), (6)–(8), (10)}
 \end{aligned}$$

Recall that the inner problem [Maxmin – d] is a linear program, whereas the outer problem [Maxmin – a] is a mixed-integer program. We now focus on understanding the properties of the master problem which will help in developing a greedy heuristic for the [Maxmin] problem.

For a fixed SO response, the cost of load control becomes constant. Hence, the post-contingency cost, C_{loss} , only comprises of C_{VR} and C_{FR} terms. We make three claims which provide insights about the attacker's optimal attack strategy. We refer the reader to [19, 20] to gain intuition about formal proofs of these claims.

Let $\Delta_j(f)$ denote the change in the system frequency due to an individual disruption of EV at node j . Then, thanks to LPF model, if two EVs are identical, then the change in system frequency due to individual disruption of the EVs will also be identical regardless of the location of the EVs in the network:

$$\mathbf{Claim \ 1} \quad \mathcal{S}_j^e = \mathcal{S}_k^e \implies \Delta_j(f_0) = \Delta_k(f_0).$$

Claim 1 implies that if the attacker focuses only on maximizing FR, then the attacker has no preference between attacking one of the two identical EVs regardless of their location in the network.

Now, with a slight abuse of notation, let $\Delta_j(v_i)$ denote the change in the voltage at node i due to an individual disruption of EV at node j . Our second claim is as follows: if the EVs at node j and k are identical, and node j is upstream of node k relative to node i ($j \prec_i k$), then the impact on v_i due to individual EV disruption at node j will be smaller than the impact due to individual EV disruption at node k , that is:

$$\mathbf{Claim \ 2} \quad \mathcal{S}_j^e = \mathcal{S}_k^e \quad \text{and} \quad j \prec_i k \implies \Delta_j(v_i) < \Delta_k(v_i).$$

Finally, let $\Delta_J(v_i)$ (resp., $\Delta_J(f)$) denote the change in the voltage at node i (resp., system frequency) due to disruption of EVs at nodes $j \in J$. Then, our third claim directly follows from the linearity of LPF model:

$$\mathbf{Claim \ 3} \quad \Delta_J(v_i) = \sum_{j \in J} \Delta_j(v_i) \quad \text{and} \quad \Delta_J(f) = \sum_{j \in J} \Delta_j(f).$$

In summary, while voltage regulation is affected by both spatial structure and extent of compromise, the frequency regulation is only affected by the latter factor.

4.2 Greedy Heuristic

Based on our claims in Section 4.1, we propose our the following greedy heuristic. (This heuristic was first presented in [18].) But first, we need to introduce Algorithm 1 which computes an optimal attack for a given (fixed) SO response, i.e., it solves [Maxmin – a].

Consider an arbitrary “pivot” EV as a candidate node targeted by the attacker, who aims to maximize the weighted sum of losses due to voltage and frequency bound violations. (Again, since we are considering SO action as fixed, the cost of load control can be ignored.) Thus, the attacker's objective is maximize the affine

Algorithm 1 Pivot node algorithm

```

1: Calculate  $v^o$  (pre-contingency voltage profile).
2: for  $i \in \mathcal{N}$  do
3:   for  $j \in \mathcal{N}$  do
4:     Compute  $\Delta_j(v_i, f)$ 
5:     Sort  $j$  s in decreasing order of  $\Delta_j(v_i, f) \rightarrow (\pi_1, \dots, \pi_N)$       // (Claims 1 and 2)
6:     Set  $J_i^* = (\pi_1, \dots, \pi_k)$  by choosing first  $k$  nodes.
7:     Calculate  $\Delta_{J_i^*}(v_i, f) = \sum_{j \in J_i^*} \Delta_j(v_i, f)$       // (Claim 3)
8:   end for
9: end for
10: Find  $i^* = \operatorname{argmax}_{k \in \mathcal{N}} (L_k + \Delta_{J_k^*}(v_k, f))$ 
11: return  $J_{i^*}^*$ .

```

function $L_i = W_{VR}(\underline{v}_i - v_i^c) + W_{FR}(\underline{f} - f^c)$. In fact, for compromise of any pivot EV node, the resulting effect on (or contribution to) L_i can be computed very easily, thanks to the linear power flow assumption. Let this effect be denoted by $\Delta_j(v_i, f)$. Now, sort the EV nodes in decreasing order of the effects on L_i due to their individual disruptions $\Delta_j(v_i, f)$, and pick the top k nodes.⁶ Assuming that the attacker will target these k EV nodes, compute the optimal SO response and the post-contingency loss.

Then, repeat the same procedure with a different node as a pivot node. If the post-contingency loss with the new node as the pivot node is higher, update the values for the current best attacker strategy and the current best post-contingency cost. Iterate over the remaining nodes and repeat the procedure until all the nodes are exhausted.

Now, we can propose our greedy heuristic (GH), which iterates between solving the master problem (with fixed SO actions) and the subproblem (with fixed attacker actions), with successively increasing maximin values of post-contingency losses. In the first iteration, fix the SO response to the pre-contingency values, i.e., $sg^c = sg^o$ and $\gamma^c = \gamma^o$, and compute the optimal attacker strategy as the solution of [Maxmin – a] by implementing the pivot node algorithm. Then, consider this attacker strategy as fixed, and compute an optimal SO response as well as the post-contingency cost by solving [Maxmin – d]. In the next iteration, consider the new SO response as fixed, and again compute the optimal attacker strategy. Then, fixing the new attacker strategy, compute the optimal post-contingency cost. If this cost is smaller than the previously computed post-contingency cost, we terminate the heuristic. Otherwise, we continue to iterate between the master and the subproblems until we get some attacker strategy twice. Since, the number of optimal attacker strategies is finite, the greedy heuristic is bound to terminate. However, we observe that the heuristic converges to optimality in few iterations. Indeed, we observed that our heuristic provides optimal solutions in less than five iterations for medium-sized networks of size 37.

⁶A similar pivot node algorithm is presented in [12].

4.3 Evaluation of the Greedy Heuristic

We describe a set of computational experiments to evaluate the performance of the greedy heuristic (GH) in solving the two-stage subgame. Specifically, we compare the GH solution against the solutions obtained by the KKT approach mentioned at the beginning of this section and also brute force (BF). We also evaluate the effect of weights on post-contingency costs for a range of k values; see attacker's resource constraint (9).

Network Setup. Our simulation setup is as follows: We consider a modified IEEE 37 node network as shown in Figure 3. Each line has an identical impedance of $z_j = 0.01 + 0.02j$, and each node has one DER and one non-EV load. The set of feasible DER set points is given by

$$\mathcal{S}_i^{poly} = \{p + jq \mid p \geq 0, -a \leq q \leq a, 4p + 3q \leq 5a, 4p - 3q \leq 5a\},$$

where $a = 0.04$ is a parameter; see (5). In the slow-charging mode, each EV load is $se_i^o = 2(1 + 0.33j)a$. In the fast-charging mode, each EV draws twice the power drawn in slow-charging mode: $se_i^a = 4(1 + 0.33j)a$. The non-EV demand at each node is $\bar{s}_i = 0.03 + 0.01j$, and the maximum load control parameter is $\bar{\gamma}_i = 0.5$, i.e., 50% of the non-EV load can be shed at each node. For the sake of simplicity, we assume that all DERs, non-EV loads, and EVs are homogeneous. Furthermore,

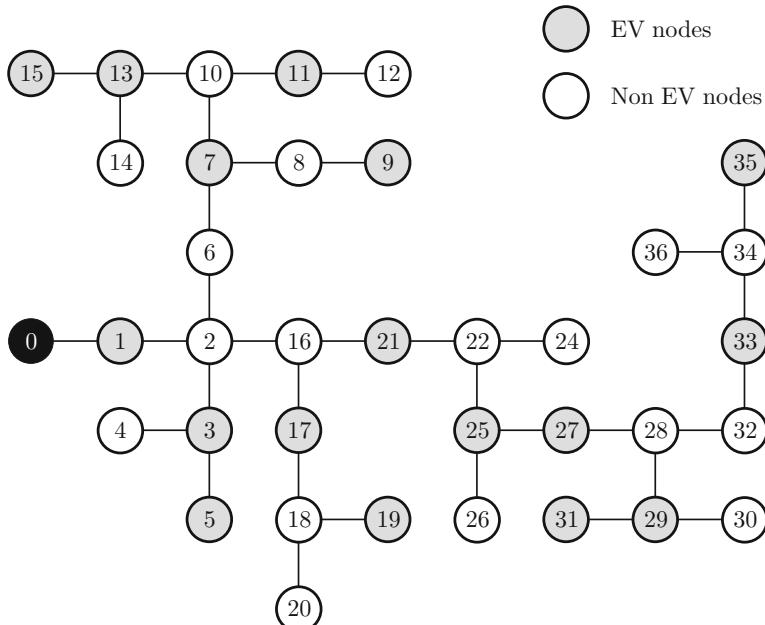
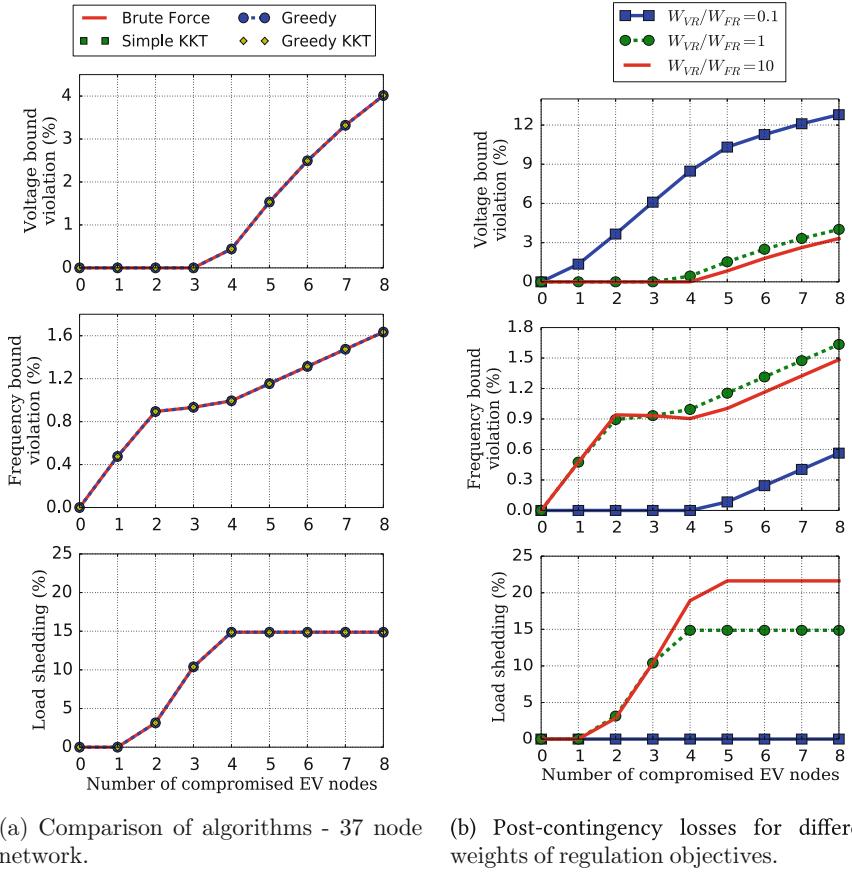


Fig. 3 Modified IEEE 37 node network.



(a) Comparison of algorithms - 37 node network.

(b) Post-contingency losses for different weights of regulation objectives.

Fig. 4 Evaluation of the greedy heuristic.

the black node in Figure 3 is the substation node, the gray-colored nodes are the nodes with EVs, and the remaining nodes do not have EVs.

We assume the following cost parameters: the cost of per unit load shedding, per unit voltage bound violation, and per unit frequency-bound violation is chosen to be $W_{LC} = 1$, $W_{VR} = 250$, and $W_{FR} = 250$, respectively; see (12). The voltage and frequency regulation constants in (2) are chosen as $v^{reg} = 0.01$ and $f^{reg} = 0.02$.

GH vs. KKT vs. BF. Figure 4a shows percentage voltage bound violation $\left(100 \max_i \max \left(\frac{v_i - v_i^c}{v_0}, \frac{v_i^c - \bar{v}_i}{v_0}, 0\right)\right)$, percentage frequency bound violation $\left(100 \max \left(\frac{f - f^c}{f_0}, \frac{f^c - \bar{f}}{f_0}, 0\right)\right)$, and percentage load shedding $\left(\frac{100}{N} \sum_i \gamma_i^c\right)$ as the number of EV nodes compromised increases. The pre-contingency set points are chosen to be $sg_i^o = (0.9 + 0.33j)a$. Note that the GH provides an optimal solution in this setting. Also note that, for the chosen weight parameters and $k = 1$,

C_{VR} and C_{LC} are both zero, but C_{FR} is positive, which implies that SO tolerates some frequency bound violation to maintain voltage regulation and full demand satisfaction. This shows that SO tolerates some frequency bound violation at the expense of no load control. For slightly higher intensity attacks ($k = 2, 3$), the SO starts imposing load control. However, as k increases further, the load control saturates at 15% for $k \geq 4$, although the total load control capability is 50%. This observation has been detailed in our previous work [19]. Intuitively, initial shedding of downstream loads reduces the post-contingency cost because the active and reactive power reduction contributes to reduction in both C_{FR} and C_{VR} . Indeed, when the SO exhausts the load control capability of the downstream nodes, control of nodes that are upstream is not as beneficial in reducing C_{VR} . Hence, the saturation in cost of load control.

In Figure 4b, we fixed the W_{LC} and varied the W_{VR}/W_{FR} ratio. The different W_{VR}/W_{FR} ratios correspond to different weights given to voltage and frequency regulation objectives. Note that for $W_{VR}/W_{FR} = 0.1$, the SO exerts no load control, but for higher W_{VR}/W_{FR} , there is load control. This indicates that the load control is more effective in reducing C_{VR} than in reducing C_{FR} . Indeed, a reduction in the load reduces both active and reactive power demand. However, the C_{FR} is affected only by active power reduction (see (2a)), whereas the C_{VR} is affected by both active and reactive power reduction (see (1c)). Hence, load control directly reduces C_{VR} and also indirectly reduces C_{FR} . Again, the C_{LC} reaches a saturation level after the downstream nodes' capability of load control is exhausted. Additionally, as the W_{VR}/W_{FR} ratio increases, the saturation level is reached for a higher intensity attack and also attains a higher saturation value.

5 Resource Allocation via Resilience-Aware OPF

In this section, we extend the [Maxmin] bilevel formulation to a trilevel framework with the outermost level denoting the resource allocation stage. We call this extended formulation the RAOPF problem:

$$\begin{aligned} [\text{RAOPF}] \quad \mathcal{L} := \min_{x^o \in \mathcal{X}^o} & C_{\text{alloc}}(x^o) + C_{\text{loss}}(x^o(u), x^c(\delta^{oa}, \phi^*)) \\ \text{s.t. (3), (6)} - (8) \\ (\delta^{oa}, \phi^*) \in \arg \max_{\delta^a \in \Psi_k} \min_{\phi \in \Phi} & C_{\text{loss}}(x^o(u), x^c(\delta^a, \phi)) \\ \text{s.t. } x^c \in \mathcal{X}^c, (2), (6) - (8), (10) \end{aligned} \tag{13}$$

As mentioned in Section 1, the SO's objective in Stage 0 is to determine the resource allocation (i.e., output of the generators sg^o) that minimizes the total cost of resource allocation (C_{alloc}) and the maximin post-contingency loss incurred in the last two stages of the game.

The RAOPF problem (13) belongs to a class of mixed-integer non-convex trilevel problems which are typically computationally hard to solve. However, after the MILP reformulation of the last two stages ([Maxmin]), the overall RAOPF can be shown to be a mixed-integer bilevel nonlinear program (MIBNLP). Although MIBNLP are NP-hard problems, few computational approaches have been proposed in the literature for solving of MIBNLP problems based on branch and bound techniques [26]. We do not focus on implementing these techniques here but instead focus on simple examples which provide us interesting and practically relevant insights on the SO's allocation/dispatch and attacker's strategy.

By way of simple examples, we first illustrate the key trade-offs faced by the SO in maintaining regulation objectives (Section 5.1). Next, we describe the structure of optimal attack in two cases: with and without adequate resources (Section 5.2). Finally, we present some insights about resource allocation strategies (Section 5.3) and compare two qualitatively different resource allocation strategies (Section 5.4).

5.1 Insights on Optimal SO Response

The fact that the regulation objectives VR, FR, and CM are not aligned with each other can be seen by considering a simple two-node network in Figure 5. It has a BG with $f^{reg} = 0.1$ and $v^{reg} = 0.1$. Node 1 has a load with $pc_1^o = 0.4$ pu and $qc_1^o = 0.2$ pu. Node 1 also has a DER which can be modeled as in Figure 6a with apparent power capability of $\bar{s}\bar{g} = 0.4$ pu. The pre-contingency output of the DER is set to $pg_1^o = 0.2$ pu and $qg_1^o = 0.2$ pu. The line parameters are $r_1 = 0.2$ pu and $x_1 = 0.4$ pu.

Now, consider the contingency created by a sudden change of load to twice its pre-contingency value, i.e., $pc_1^c + jqc_1^c = 0.8 + 0.4j$. This trade-off in maintaining the regulation objectives (FR, VR, and CM) is apparent from the difference in optimal DER outputs needed to address each of these objectives individually. Indeed, the DERs alone may not be able to completely resolve the contingency; under our assumptions, the remaining supply-demand imbalance is eventually covered by the BG.

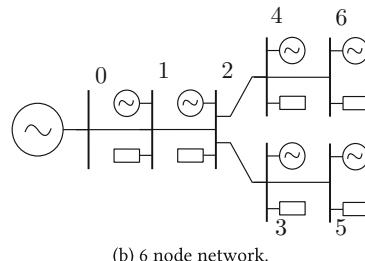
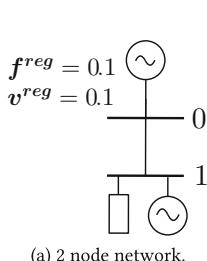
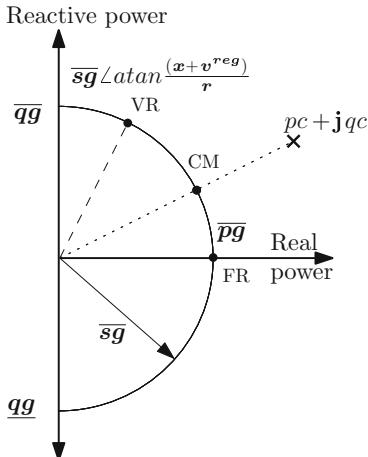
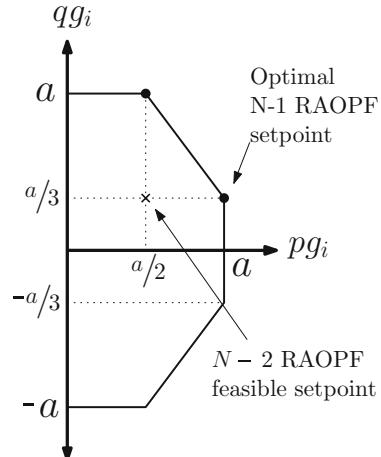


Fig. 5 DN topologies.



(a) SO response for each of the three regulation objectives in 2 node example network.



(b) Resource allocation strategies considered for the 6 node network.

Fig. 6 Trade-offs in maintaining regulation objectives and DER set points for reserve allocation.

Let $\Delta p := (pc_1^c - pg_1^c) - (pc_1^o - pg_1^o)$ be the net change in active power consumed at node 1. Similarly, let $\Delta q := (qc_1^c - qg_1^c) - (qc_1^o - qg_1^o)$ be the net change in reactive power consumed at node 1. Now, consider the following cases which correspond to the SO addressing each regulation objective individually (again, see Figure 6a for the corresponding DER set points):

- Using (2a), the drop in system frequency can be approximated as $f^{reg} \Delta p$. Thus, to achieve maximum FR, the SO should minimize $f^{reg} \Delta p$.
- Using (1) and (2b), the voltage drop at node 1 can be approximated as $r \Delta p + (x + v^{reg}) \Delta q$. Thus, to best maintain VR, the SO should minimize this quantity.
- Finally, the power flow on line (0,1) can be expressed as $(pc_1^c - pg_1^c) + j(qc_1^c - qg_1^c)$. Thus, for CM, the SO should minimize $r((pc_1^c - pg_1^c)^2 + (qc_1^c - qg_1^c)^2)$.

The optimal DER set point for each of the abovementioned cases can be expressed in closed form and are given below (these set points are illustrated in Figure 6a):

$$pg_1^{c\star} + j qg_1^{c\star} = \begin{cases} \overline{sg} \angle \arctan 0 & \text{for maximum FR} \\ \overline{sg} \angle \arctan \left(\frac{(x+v^{reg})}{r} \right) & \text{for maximum VR} \\ \overline{sg} \angle \arctan \left(\frac{qc_1^c}{pc_1^c} \right) & \text{for maximum CM} \end{cases}$$

Clearly, the optimal response DER set points for FR, VR, and CM are distinct. Thus, the optimal DER set points depend on the weight of each regulation objective

Table 1 Trade-offs between FR, VR, and CM.

Network objectives	Post-contingency x_1 qg_1/pg_1	Objective Values		
		Δf	Δv_1	$\mathbf{r}_1(p_1^2 + q_1^2)$
FR	0	0.01	0.22	0.05
VR	$(x + v^{reg})/r$	0.041	0.051	0.076
CM	qc^c/pc^c	0.015	0.119	0.031

in the post-contingency cost; see (12). For the chosen parameters in the two-node network, the DER set points and their corresponding impact on regulation objectives are summarized in Table 1.

5.2 Insights on Optimal Attacker Strategy

Now let us study the structure of optimal attack under no DER response by considering a six-node example network as shown in Figure 5b. We will define the load and DER parameters in terms of a constant scalar $a = 0.1$ pu. Let $b = a/3$. Assume that each line has identical impedance $z = \mathbf{r} + j\mathbf{x}$, where $\mathbf{r} = 0.03$ pu and $\mathbf{x} = 0.06$ pu. At each node $i \in \mathcal{N}$, we assume the non-EV load $sn_i^o = a + jb$. Consider the EV load $se_i^o = a + jb$ for $i \in \{3, 4, 5, 6\}$, and the EV load as $se_i^o = 1.4(a + jb)$ for $i \in \{1, 2\}$. We assume that if the EVs are compromised, then their load becomes twice of that of their pre-contingency demand, i.e., $\bar{s}\mathbf{e}_i = 2s\mathbf{e}_i$. Let us consider the pre-contingency DER set points to be $sg_i^o = a + jb$. The frequency regulation constant f^{reg} is 1 Hz/pu, i.e., the frequency drops by 1 Hz if the supply-demand deficit suddenly increases by 1 pu and the voltage regulation constant v^{reg} is 0 pu. The frequency bounds are $\underline{f} = 59.8$ and $\overline{f} = 60.2$ Hz. The voltage bounds are $\underline{v}_i = 0.9$ and $\overline{v}_i = 1.1$.

Recursively using the voltage drop equation (1c), we can compute the voltage profile as follows:

$$\forall i \in \mathcal{N}, \eta \in \{o, c\}, \quad v_i^\eta = v_0^\eta \mathbf{1} - \sum_{j \in \mathcal{N}} \mathbf{R}_{ij}(pc_j^\eta - pg_j^\eta) - \mathbf{X}_{ij}(qc_j^\eta - qg_j^\eta). \quad (14)$$

Using (14), we can compute the pre-contingency voltage profile:

$$v^o = [0.965 \ 0.938 \ 0.928 \ 0.928 \ 0.923 \ 0.923].$$

We can check that this six-node DN satisfies regulation objectives under any single EV node attack. For example, when the EV at node 1 or 2 is compromised, the net active power demand increases by $1.5a$ pu. Hence, the frequency only drops to 59.85 Hz, which is above frequency lower bound. Similarly, if node 5 or node 6 is compromised, then the minimum voltage in the DN is 0.907, which is above the

voltage lower bound. In case of compromise of an EV at an intermediate node 3 or 4, we can similarly ensure that the regulation objectives are fulfilled, as these nodes are smaller in size than nodes 1 or 2 and are located upstream to the nodes 5 and 6. Consequently, the impact of EV compromise at node 3 or 4 will be smaller than nodes 1 or 2 (resp., nodes 5 or 6) in terms of frequency (resp., voltage) drop. Thus, in the terminology of classical SCOPF problem, this network is resilient to the N-1 contingencies, each concerning the compromise of a single EV node.

Now, we consider the case when the attacker compromises $k = 2$ EV nodes. Let's consider three different subcases.

(a) $\mathbf{W}_{\text{VR}} = \mathbf{0}, \mathbf{W}_{\text{FR}} > \mathbf{0}$: In this case, the attacker's goal is to maximize C_{FR} . Then, by Claim 1, the attacker's optimal strategy will be to compromise EVs at nodes 1 and 2 because nodes 1 and 2 have the largest EVs. In this case, the location of EVs in the DN does not matter from the attacker's perspective.

(b) $\mathbf{W}_{\text{VR}} > \mathbf{0}, \mathbf{W}_{\text{FR}} = \mathbf{0}$: Now, the attacker's goal is to maximize C_{VR} . Following Claim 2, the attacker's optimal strategy is to compromise EVs at nodes 4 and 6. Since the net demand at each node is positive, power only flows from the substation to the downstream nodes. As a result, node 6 has the lowest voltage in the DN. Voltages at all nodes will reduce if EVs are compromised, but the voltage at node 6 will reduce the most if nodes 4 and 6 are compromised (by Claims 2 and 3). Therefore, we observe that the attacker chooses to compromise downstream EVs. Note that due to symmetric nature of the DN, compromising EVs at nodes 3 and 5 is also an optimal attack strategy for this case.

(c) $\mathbf{W}_{\text{VR}} > \mathbf{0}, \mathbf{W}_{\text{FR}} > \mathbf{0}$: In this case, the attacker's goal is to maximize weighted sum of C_{FR} and C_{VR} . We observe that for a certain range of values for $\frac{W_{\text{VR}}}{W_{\text{FR}}}$ ratio, the optimal attack strategy is to compromise nodes 2 and 6. The attacker compromises an upstream node 2 instead of a downstream node 4 to increase the loss of FR even though the loss in VR may reduce. Additionally, we see that although nodes 1 and 2 have identical EVs, attacker will choose to compromise node 2 because of his preference for downstream EV nodes maximizes loss of VR.

Thus, we observe that when the $\frac{W_{\text{VR}}}{W_{\text{FR}}}$ ratio is small, the attacker chooses to compromise large EV nodes which may or may not be spatially co-located. However, as the $\frac{W_{\text{VR}}}{W_{\text{FR}}}$ ratio increases the optimal attack starts to target downstream nodes in a clustered manner.

5.3 Insights on Resource Allocation

Next, among the optimal attacker strategies determined in Section 5.2, we consider the following attack scenarios each involving simultaneous compromise of $k = 2$ EV nodes: (a) nodes 1 and 2 are compromised (i.e., $\delta = [1, 1, 0, 0, 0, 0]$), and (b) nodes 4 and 6 are compromised ($\delta = [0, 0, 0, 1, 0, 1]$). For each of these two scenarios, we evaluate the costs due to loss in VR and FR components of the total post-contingency cost when DER reserves are not present and compare these costs with the case when DER reserves are available.

(i) Network with No DER Resources

Assume that all the DERs are operating at $sg_i^o = a + b\mathbf{j}$ pu. At this initial set point, there is no available active or reactive power reserve from the DERs.

Attack Scenario (a)

The net increase in active power load is $3a = 0.3$ pu. This change results in $f^c = 59.7$ Hz. Hence, some amount of load shedding will be required to bring the frequency back to the acceptable range.

Attack Scenario (b)

Under this attack, if the SO does not respond, then the post-contingency voltages will be

$$\mathbf{v}^c = [0.952 \ 0.912 \ 0.902 \ 0.898 \ 0.898 \ 0.888].$$

Clearly, the voltage bounds will be violated at nodes 4 and 6, and some load shedding is required to bring voltages back to acceptable range. Note that the voltages at nodes 4 and 6 are smaller than the voltages at nodes 3 and 5. This is due to the proximity of load compromises to nodes 4 and 6.

(ii) Network with DER Reserves

Now assume that the pre-contingency DER set points are $sg_i^o = 0.5a + b\mathbf{j}$ pu. This gives us active and reactive power reserves of $0.5a + 2b\mathbf{j}$; see Figure 6b. Note that this is an overestimate of actually available reserves, because if active power reserves are fully used, then reactive power reserves cannot be used at all and vice versa. We chose this DER set point only for the ease of calculation; it is certainly not an optimal reserve allocation in the face of two-sized EV attacks. Under this resource allocation, the pre-contingency voltage profile will be

$$\mathbf{v}^o = [0.956 \ 0.921 \ 0.908 \ 0.908 \ 0.902 \ 0.902],$$

which also satisfies the voltage bounds.

Attack Scenario (a)

Again, the total load suddenly increases as a result of EV attacks to nodes 1 and 2. Now, each DER can rapidly respond to the contingency, and if the SO

increases their generation from the initial set point $sg_i^o = 0.5a + bj$ to final set point $sg_i^c = a + bj$, then the additional active power injected from the DERs is $6(a - a/2) = 3a$ pu. Hence, the net change in active power between pre- and post-contingency situation is 0. As a result, there is no change in frequency despite two EVs being compromised. Although there is a drop in voltage because of a net increase in reactive power demand, the voltage bounds are also satisfied. Hence, load shedding is not required.

Attack Scenario (b)

Due to the compromise of downstream EV nodes, the minimum voltage in DN will violate the bounds in the absence of a DER response. Fortunately, this situation can be avoided if the reactive power supply is increased and the set points of all DERs are changed to $0.5a + aj$. The resulting post-contingency voltage profile will be as follows:

$$v^c = [0.97 \ 0.945 \ 0.94 \ 0.93 \ 0.938 \ 0.922].$$

Thus, all voltage bounds are met with this DER reserve.

Using this illustrative example, we have tried to argue that with sufficient reserves as well as appropriate SO response, the DN can withstand contingencies resulting from compromise of multiple ($k = 2$) EV nodes. In this example, we see that both frequency and voltage regulation objectives can be maintained without any load control because the DER reserves were sufficient to provide the active and reactive power supply needed to avoid the frequency and voltage bound violations.

5.4 Further Insights on Resource Allocation Stage (Stage 0)

Finally, we study two possible SO strategies for optimal resource allocation in Stage 0. We retain the same network setup as in Section 4. First, we focus on “uniform” resource allocation, i.e., all DERs have identical pre-contingency set points. For this resource allocation, we use the greedy heuristic to compute optimal attacker strategy and the SO response. Second, based on our observations regarding the SO response, we suggest a feasible “heterogeneous” resource allocation, i.e., DERs having different pre-contingency set points while keeping the total DER output identical to that of the former case. Finally, we compare the worst-case post-contingency losses for the two resource allocation strategies.

Trade-off between active and reactive power allocation. First, we show that there exists a trade-off between active and reactive power resource allocation to meet the objectives of FR and VR. We assume no load control, i.e., $\bar{y} = 0$ and vary the initial DER resource allocation as shown in Figure 7a. Two different values of

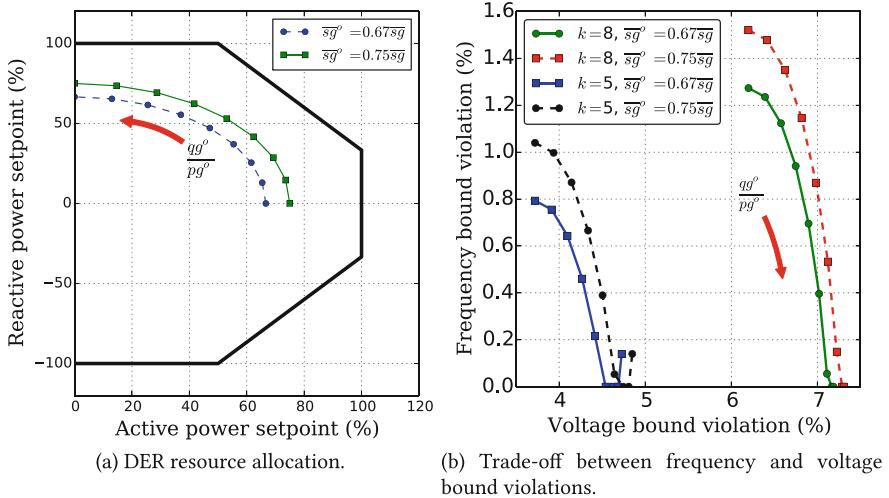


Fig. 7 Post-contingency losses for different weights of regulation objectives.

\overline{sg}^o are chosen, and the resource allocation is varied in the increasing order of $\frac{qg_i^o}{pg_i^o}$ (see Figure 7a). For each combination of \overline{sg}^o and $\frac{qg_i^o}{pg_i^o}$ ratio, the optimum maximin post-contingency losses are computed for two attack intensities; see Figure 7(b).

We can draw some useful observations from this figure: as the intensity of the attack k (i.e., number of compromised EV nodes) increases or the apparent reserves allocated decrease (i.e., \overline{sg}^o increases), the post-contingency voltage bound violation increases. Note that for both $k = 5$ and $k = 8$, as $\frac{qg_i^o}{pg_i^o}$ ratio increases, the voltage bound violation increases since the reactive power reserves are reduced. The frequency bound violation decreases initially for higher allocation of active power reserves. Interestingly enough, for $k = 5$ and for large enough $\frac{qg_i^o}{pg_i^o}$ ratio, we can see the frequency bound violation increases again. This can be explained as follows: for large enough $\frac{qg_i^o}{pg_i^o}$ ratio, the reactive power reserve reduces. Hence, to do VR, the SO increases both active and reactive power output of the DERs. However, since the attack intensity is small, the net change in active power after the attack becomes positive and large enough to cause violation of upper frequency bound.

Diversification in SO response. Second, we show that the optimal SO response admits a diversification strategy, where some DERs supply more active power than reactive power (i.e., their contribution to FR is more than that to VR), while other DERs supply more reactive power than active power (i.e., their contribution to VR is more than that to FR).

Consider the 13-node network as shown in Figure 8a. For $k = 4$, the optimal attacker strategy is to compromise EV nodes $\{5, 6, 7, 8\}$ or $\{9, 10, 11, 12\}$. Due to symmetry, assume that the latter EV node set is compromised. These four nodes

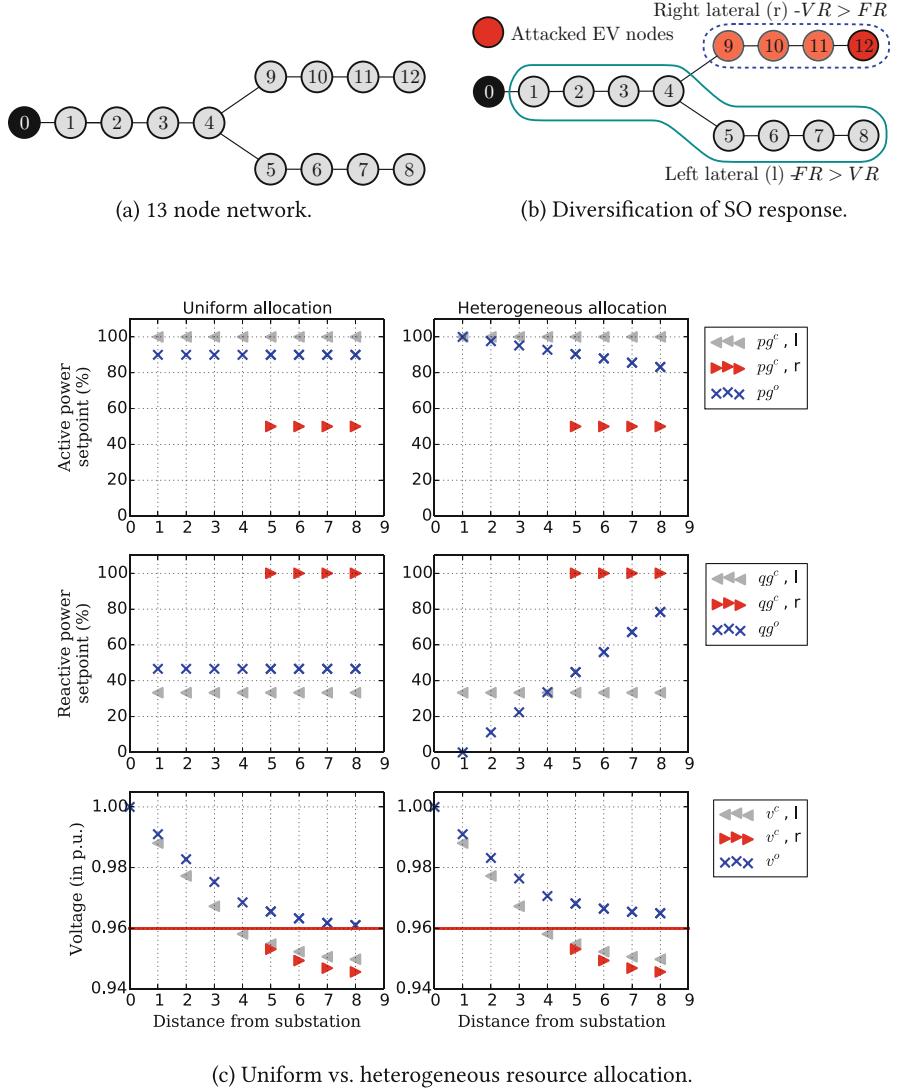


Fig. 8 Diversification of nodes for voltage vs. frequency regulation.

form the right lateral, denoted by (r), and the remaining nodes form the left lateral, denoted by (l). Consider uniform resource allocation, as shown in Figure 8c. The pre-contingency output of the DERs is 90% and 47% of the maximum active and the maximum reactive power output, i.e., $sg_i^o = 0.9\bar{p}g_i + 0.47\bar{q}g_i$. Before the attack, the voltages of the nodes in the left lateral are equal to the corresponding nodes in the right lateral. After the attack, the voltages in the right lateral fall below that of the left lateral. Hence, the DERs in the right lateral start contributing to VR,

by generating $sg_i^c = (0.5 + j)\bar{sg}_i$. However, the rest of the DERs contribute more to the FR by generating $sg_i^c = (1 + 0.33j)\bar{sg}_i$. This shows that the DERs diversify in their roles to contribute to different objectives.

Diversification in DER resource allocation. Finally, we evaluate the pre-contingency state vector and post-contingency cost for a heterogeneous resource allocation strategy and compare with the uniform allocation strategy. Recall from our experiment above that the downstream DERs are likely to contribute more to VR than to FR. Therefore, we may choose the initial DER set points as shown in Figure 8c, such that downstream DERs contribute more reactive power as compared to upstream DERs. Now, consider the following heterogeneous allocation strategy: as the distance of the node from the substation increases, let us choose a higher reactive power set point and lower active power set point. Note that we keep the sum total of active and reactive power output of the DERs to be the same as in the case of uniform allocation. Interestingly, we observe that the post-contingency losses are identical for both uniform and heterogeneous resource allocation. However, the pre-contingency voltage profile is better for the heterogeneous resource allocation as opposed to uniform resource allocation. We expect that a better voltage profile will allow the SO to incur lesser regulation cost in the pre-contingency state.

6 Concluding Remarks

In this chapter, we presented a new problem, resilience-aware optimal power flow (RAOPF), to evaluate the resilience of DNs to a class of contingencies resulting from adversarial compromise of EV nodes. We posed RAOPF as a three-stage sequential game. We primarily focused on the last two stages and considered the first-stage resource allocation as fixed. For the sake of simplicity, we only considered linear power flows. We studied the structure of the problem when the attacker compromises a subset of EV nodes to suddenly increase the net demand in the DN, and the SO responds by activating (or dispatching) available DER reserves and imposing load control, if needed. This problem can be solved by reformulating it to a MILP via using KKT conditions for the innermost (third stage) problem. While this approach is a classical one, it does not exploit the structure of the problem (i.e., power flows on tree networks). Based on our earlier work [19, 20], we proposed a greedy heuristic that enables much faster computation of attack strategy to maximinimize the SO's post-contingency loss.

Our computational results show that the greedy heuristic computes near-optimal strategies. Importantly, the optimal SO strategy shows a diversification of resources where the downstream DERs contribute to voltage regulation more than they contribute to frequency regulation. Based on this insight, we proposed a heterogeneous resource allocation strategy, where we keep the total active and reactive power reserves constant but allow the downstream DERs to contribute more reactive power. We observe that the post-contingency losses are identical for both heterogeneous and uniform resource allocation, but the pre-contingency voltage profile is better for the former case.

In summary, our main contributions are (a) an approach to speed up the computation of attacker-SO strategies (relative to classical MILP approach) by utilizing properties of power flow on radial DNs and (b) insights into optimal resource allocation and DER dispatch when the SO faces trade-offs in maintaining regulation objectives during contingencies that result from simultaneous node compromises.

References

1. Alsac O, Stott B (1974) Optimal load flow with steady-state security. *IEEE Trans Power Apparatus Syst* PAS-93(3):745–751. ISSN 0018-9510. <https://doi.org/10.1109/TPAS.1974.293972>
2. Andersson G (2012) Dynamics and control of electric power systems. Lecture Notes in Electrical Engineering, Power Systems Laboratory, ETH, Zurich, pp 57–90
3. Andersson G, Donalek P, Farmer R, Hatzigaryiou N, Kamwa I, Kundur P, Martins N, Paserba J, Pourbeik P, Sanchez-Gasca J, Schulz R, Stankovic A, Taylor C, Vittal V (2005) Causes of the 2003 major grid blackouts in north america and europe, and recommended means to improve system dynamic performance. *IEEE Trans Power Syst* 20(4):1922–1928. ISSN 0885-8950. <https://doi.org/10.1109/TPWRS.2005.857942>
4. Bhaskar MM, Srinivas M, Sydulu M (2010) Security constraint optimal power flow (SCOPF)-a comprehensive survey. *Int J Comput Appl* 11(6):42–52
5. Bienstock D (2015) Electrical transmission system cascades and vulnerability: an operations research viewpoint. SIAM, Philadelphia
6. Braun M, Strauss P (2008) A review on aggregation approaches of controllable distributed energy units in electrical power systems. *Int J Distrib Energy Resour* 40(4):297–319
7. Camacho EF, Samad T, Garcia-Sanz M, Hiskens I (2011) Control for renewable energy and smart grids. In: The impact of control technology. Control Systems Society, New York, pp 69–88
8. Capitanescu F, Martinez Ramos JL, Panciatici P, Kirschen D, Marcolini AM, Platbrood L, Wehenkel L (2011) State-of-the-art, challenges, and future trends in constrained optimal power flow. *Electr Power Syst Res* 81(8):1731–1741
9. Chiang HD, Baran ME (1990) On the existence and uniqueness of load flow solution for radial distribution power networks. *IEEE Trans Circuits Syst* 37(3):410–416. ISSN 0098-4094. <https://doi.org/10.1109/31.52734>
10. De Brabandere K, Bolsens B, Van den Keybus J, Woyte A, Driesen J, Belmans R (2007) A voltage and frequency droop control method for parallel inverters. *IEEE Trans Power Electron* 22(4):1107–1115. ISSN 0885-8993. <https://doi.org/10.1109/TPEL.2007.900456>
11. Farivar M, Neal R, Clarke C, Low S (2012) Optimal inverter VAR control in distribution systems with high PV penetration. In: 2012 IEEE Power and Energy Society general meeting, pp 1–7. <https://doi.org/10.1109/PESGM.2012.6345736>
12. Kundu S, Hiskens IA (2014) Overvoltages due to synchronous tripping of plug-in electric-vehicle chargers following voltage dips. *IEEE Trans Power Deliv* 29(3):1147–1156. ISSN 0885-8977. <https://doi.org/10.1109/TPWRD.2014.2311112>
13. Lee A (2014) Electric sector failure scenarios and impact analyses. Technical report, National Electric Sector Cybersecurity Organization Resource (NESCOR), Electric Power Research Institute (EPRI), Palo Alto. <http://www.smartgrid.epri.com/doc/NESCOR%20failure%20scenarios%2006-30-14a.pdf>
14. Monticelli A, Pereira MVF, Granville S (1987) Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Trans Power Syst* 2(1):175–180. ISSN 0885-8950. <https://doi.org/10.1109/TPWRS.1987.4335095>

15. Moore JT, Bard JF, The mixed integer linear bilevel programming problem. *Oper Res* 38(5):911–921
16. Report on the grid disturbance on 30th and 31st July 2012. http://www.cercind.gov.in/2012/orders/Final_Report_Grid_Disturbance.pdf
17. Salam AA, Mohamed A, Hannan MA (2008) Technical challenges on microgrids. *ARPN J Eng Appl Sci* 3(6):64–69
18. Salmeron J, Wood K, Baldick R (2004) Analysis of electric grid security under terrorist threat. *IEEE Trans Power Syst* 19(2):905–912. ISSN 0885-8950. <https://doi.org/10.1109/TPWRS.2004.825888>
19. Shelar D, Amin S (2017) Security assessment of electricity distribution networks under der node compromises. *IEEE Trans Control Netw Syst* 4(1):23–36
20. Shelar D, Giraldo J, Amin S (2015) A distributed strategy for electricity distribution network control in the face of der compromises. In: 2015 54th IEEE conference on decision and control (CDC), pp 6934–6941. <https://doi.org/10.1109/CDC.2015.7403312>
21. Tonkoski R, Lopes LAC, El-Fouly THM (2011) Coordinated active power curtailment of grid connected pv inverters for overvoltage prevention. *IEEE Trans Sustain Energy* 2(2):139–147
22. Turitsyn K, Sulc P, Backhaus S, Chertkov M. Options for control of reactive power by distributed photovoltaic generators. *Proc IEEE* 99(6):1063–1073. ISSN 0018-9219. <https://doi.org/10.1109/JPROC.2011.2116750>
23. Van Cutsem T, Vournas C (1998) Voltage stability of electric power systems, vol 441. Springer Science & Business Media, Berlin
24. Wood RK (2011) Bilevel network interdiction models: formulations and solutions. In: Wiley encyclopedia of operations research and management science. Wiley, Hoboken
25. Yao Y, Edmunds T, Papageorgiou D, Alvarez R (2007) Trilevel optimization in power network defense. *IEEE Trans Syst Man Cybern Part C Appl Rev* 37(4):712–718. ISSN 1094-6977. <https://doi.org/10.1109/TSMCC.2007.897487>
26. Zeng B, An Y (2014) Solving bilevel mixed integer program by reformulations and decomposition. *Optimization Online* 1–34

A Cautionary Tale: On the Effectiveness of Inertia-Emulating Load as a Cyber-Physical Attack Path



Hilary E. Brown and Christopher L. DeMarco

Abstract Recent research has explored the potential for distributed, consumer-based equipment to participate in control action seeking to improve grid dynamic performance. Renewable resources are displacing synchronous generators, reducing the electrically coupled rotating inertia supplied to the system as a percentage of generation. However, this loss may be mitigated by feedback control emulating the dynamics of rotating inertia and so-called “emulated inertia” control may be implemented in distributed, consumer-based resources. The case study presented illustrates that emulated inertia feedback is also extremely well-suited to subversion by a cyberattacker. In particular, local inertia-emulating feedback can create wide-area instabilities with only slight modification of feedback parameters. The amount of affected load can be relatively modest and the attacker can “target” particular generators, producing oscillations that would likely trip rate-of-change-of-frequency protective relays within one minute. The authors believe this scenario is particularly troubling, because it is likely that distributed consumer-based control systems will lack the strong cybersecurity protection afforded large generation resources.

1 Background and Motivation

A long literature has explored the potentially severe consequences of cyber-physical attacks on the power system, in particular the subversion of feedback control in the power grid. In the late 1990s, one of the authors explored the potential for subversion of governor control loops [12–14, 18]. In the context of a quickly deregulating power industry, this type of malicious control was explored as a form of economic gamesmanship. Fortunately, there have been no documented cases where such subversion caused generator outages. Since 2007, when the US Department

H. E. Brown · C. L. DeMarco (✉)

Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, USA
e-mail: hilary.brown@ieee.org; cdeMarco@wisc.edu

of Homeland Security demonstrated the “Aurora” cyber-physical attack, causing the physical destruction of the targeted generator [35], the vulnerability of electric grids to cyber-physical attack has been explored by researchers. Three years later, the Stuxnet worm was found in industrial control systems and the specter of state-sponsored cyberattacks loomed large [22]. Although many different aspects of cyber-physical vulnerabilities have been explored, grid electromechanical dynamics suggest the presence of a unique vulnerability: lightly damped, wide-area oscillatory modes require relatively little energy to be destabilized. This opens up the possibility for subverted control systems to induce wide-area instabilities with a relatively small control effort. The original malicious control in the late 1990s began to explore this area, but the topic of targeted feedback control subversion deserves more investigation as power system cyberattacks have moved from theory to practice, as illustrated by the December 2015 attack on the Ukrainian power grid [3, 4].

Traditionally, the architecture of grid control has limited the number and types of equipment that can serve as actuators, usually large, central station generators, supplemented by a modest number of transmission-based controls. The philosophy of the responsive grid may be viewed as a democratization of grid control, with a vast expansion of the number of grid elements that may participate as control actuators. It will be a large challenge to coordinate innumerable actuators spread across thousands of square kilometers, which interact through a coupled dynamical system. Traditional grid control design offers one solution to the coordination problem: let frequency serve as geographically distributed signal, which carries (nearly) global information on system performance but can be measured locally.

Recognizing the multitude of threats, the National Electric Reliability Corporation (NERC) has adopted strong cybersecurity standards to protect generators and utility control centers [11]. However, it is unclear how such standards would apply to distributed resources, which individually may not pose a threat, but could cause destabilization when their action is coordinated through the use of only local frequency measurements. This work is motivated by the potential of distributed, consumer-based equipment to improve dynamic performance and efficiency. The case study presented here explores the double-edged sword inherent to such distributed control, particularly considering “emulated inertia” controls. The same distributed control that can contribute to improving grid dynamic performance is also well-suited to subversion by a cyberattacker. In an era in which grid operations may plausibly be targeted by entities with significant resources and little concern for consequences in an attack’s aftermath [38], the sophisticated analysis necessary for an offline attack design is not guaranteed to be a deterrent.

Recently, a number of other authors have examined wide-area instability as a possible mechanism of cyberattacks on grid protection and control. References [24] and [23] consider the potential for remotely controlled system breakers to initiate instability. Like the work presented here, [10] uses continuously acting feedback control to cause instability via remote changes to control gains on voltage control devices. References [28–30] extend the control theory underlying such designs, but their case studies were limited to test systems using only the linear, dc approximation of network power flow. Perhaps closest in spirit to this work is

that of [2], which studies “dynamic load altering attacks,” wherein feedback control altering load is employed to cause system-wide frequency instability. That work’s design used simple proportional-integral controllers and was illustrated using a 6-bus test system with classical generator models.

The work here extends the characterization of power system cyber-physical attack vulnerabilities in several ways: 1) the case study uses more accurate, detailed generator models and examines behavior in large 179-bus test system; 2) the control design uses an observer, locally estimating a low-dimensional projection of the system state from a measurement of local frequency, so that compromised locations reinforce one another without communication; and 3) this work examines a likely future scenario in which load aggregators offer inertial emulation services to the grid and this inertia-emulating control is “turned against” the system. In addition, the attack is shown to be robust to errors in the attackers’ model, exploring in particular uncertainties in generator parameters, system loading, and topology.

To begin, Section 2 will describe the theory behind feedback loop subversion. Then, the application of that control theory to the power system will be discussed in Section 3, with special attention paid to the differential equations that represent the dynamic behavior of the system generators and controlled load points. Finally, Sections 3.3 and 4 will describe the test system and the results of the case study, respectively.

2 Malicious Control

The subversion of emulated inertia feedback control is really a simple exercise in eigenvector/eigenvalue placement, but the information requirements for the attacker to *design* the attack are high. The attacker must start with knowledge of the linearization of system dynamics, the state matrix \mathbf{A} , and the resulting eigenvalues and eigenvectors ($\lambda_1, \underline{v}_1, \lambda_2, \underline{v}_2, \lambda_3, \underline{v}_3, \dots$).

One pair of complex eigenvalues (preferably a lightly damped electromechanical swing mode) is chosen to be destabilized, say $\lambda_1^0 = \lambda_2^{0*}$, modified to new unstable eigenvalues $\lambda_1 = \lambda_2^*$. A set of n_C buses are assumed to have subverted control, determining an input matrix \mathbf{B} , with the physical inputs being changes to the power commanded at these buses. The attacker likewise chooses a set of n_T buses that will be targeted to experience largest magnitudes of unstable oscillations. Our premise is that an attacker seeking to maximize grid disruption would target generating stations equipped with rate-of-change-of-frequency protective relays, which would disconnect those generators upon large $|d\omega/dt|$.

If the system studied is linearized about the operating point, the state space representation is

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx}\end{aligned}\tag{1}$$

where x is a vector of states, y is a vector of outputs, and \mathbf{C} is the output matrix. Any pre-existing feedback control is incorporated into \mathbf{A} . For the system studied here, \mathbf{B} has non-zero entries only at the locations of compromised emulated inertia control, and \mathbf{C} has non-zero entries only at the locations corresponding to the measured frequency states at those same bus locations.

The attacker constructs a full state feedback matrix \mathbf{F} such that

$$(\mathbf{A} - \mathbf{BF})\underline{v}_1 = \lambda_1 \underline{v}_1 \quad (2)$$

$$(\mathbf{A} - \mathbf{BF})\underline{v}_k = \lambda_k \underline{v}_k \quad (3)$$

where (2) represents the destabilized mode with v_1 as the new unstable eigenvector and (3) is valid for the remaining unmodified eigenvectors, $k = 3, 4, \dots, n$. The unstable eigenvector is constructed to maximize the component magnitudes at the targeted machines, compared to all other components. That such a feedback matrix exists is discussed in [25], where the conditions required are satisfied in the linearized representation of the power system.

The feedback design above yields a full, centralized state feedback, but the goal of malicious control dictates that each local controller should operate independently, with no communication between compromised nodes. This suggests a design in which the appropriate row of \mathbf{F} above is replicated at each controller location and each controller is supplemented by a local observer to estimate the system state on which \mathbf{F} acts.

The following important observations can be made about this construction. First, a bit of algebra on (2) and (3) reveals that any \mathbf{F} must have rows made of linear combinations of real and imaginary parts of the left eigenvector for the original λ_1^0 . Moreover, if the attacker wishes to estimate $\mathbf{F}x$ from measurements, the attacker does not need the whole state x , but only a projection of state onto the two dimensional invariant subspace associated with mode represented by λ_1, λ_2 . From a control implementation perspective, each local observer/controller is simply a second-order linear filter, which can be implemented using (or by subverting the coefficients of) one biquad filter block!

The ability of the observer system to have the same eigenvalues as the original feedback system is addressed in [9]. This classical result, known as the “separation property” or “separation principle,” states that the eigenvalues achieved by state feedback remain the same when that feedback is applied to a state estimated by a Luenberger observer, assuming the state is directly measurable. In other words, the design of the state feedback may be “separated” from the estimation of the state, as done here. In contrast to the approach presented here using multiple local observers, the classical result assumed a single “system-wide” estimator. Although the proof of the separation property as presented in [9] is not immediately translatable, it provides strong qualitative evidence for the validity of the approach and numeric case studies demonstrate that the eigenvalues do indeed move as desired. Ideally,

only the targeted eigenvalues are moved. In practice, however, the remaining eigenvalues are slightly perturbed due to small inaccuracies introduced by the local observer.

3 Application to the Power System

To represent the dynamic behavior of the generators and controlled load points, the differential-algebraic equations were linearized about the operating point. This section details the dynamic equations, as well as the assumptions made about the rate-of-change-of-frequency (ROCOF) protective relays that are assumed to be present at the generator buses.

3.1 Dynamic Equations of Generators

This work advances the existing work on cyberattack vulnerabilities because it represents generators using more complicated models, the one-axis model and Type I exciter. Following the conventions in [34], the nonlinear state equations representing the dynamic behavior of generator i are:

$$\begin{aligned}
 \dot{\delta}_i &= (\omega_i - \omega_o) \\
 \dot{\omega}_i &= \frac{T_{Mi}}{M_i} - \frac{E'_{qi} I_{qi}}{M_i} - \frac{(x_{qi} - x'_{di})}{M_i} I_{di} I_{qi} - \frac{D_i}{M_i} (\omega_i - \omega_o) \\
 \dot{E}'_{qi} &= -\frac{E'_{qi}}{T'_{doi}} - \frac{(x_{di} - x'_{di})}{T'_{doi}} I_{di} + \frac{E_{fdi}}{T'_{doi}} \\
 \dot{E}_{fdi} &= -\frac{(K_{Ei} + S_{Ei}(E_{fdi}))}{T_{Ei}} E_{fdi} + \frac{V_{Ri}}{T_{Ei}} \\
 \dot{R}_{fi} &= -\frac{R_{fi}}{T_{fi}} + \frac{K_{fi}}{T_{fi}^2} E_{fdi} \\
 \dot{V}_{Ri} &= -\frac{V_{Ri}}{T_{Ai}} + \frac{K_{Ai} R_{fi}}{T_{Ai}} - \frac{K_{fi} K_{Ai}}{T_{fi} T_{Ai}} E_{fdi} + \frac{K_{Ai}}{T_{Ai}} (V_{ref,i} - V_i)
 \end{aligned} \tag{4}$$

where the dots indicate the derivative with respect to time. For generator i , the rotor angle is δ_i and the frequency is ω_i . The base frequency is ω_o and is assumed to be 60 Hz. The constant M_i is given by $M_i = 2H_i/\omega_o$, where H_i is the normalized generator inertia in seconds. The damping coefficient is denoted D_i , the mechanical torque is T_{Mi} , and x_{di} , x_{qi} , and x'_{di} are the machine reactances and transient reactances in the d - and q -axes. The d - and q -axes currents are I_{di} and I_{qi} . The scaled flux

linkage is E'_{qi} , with a time constant of T'_{doi} . The field excitation is E_{fdi} , the rate feedback is R_{fi} , and V_{Ri} is the amplifier output. The time constants associated with the excitation system are T_{Ei} , T_{fi} , and T_{Ai} , and the excitation gains are K_{Ei} , K_{Ai} , and K_{fi} . The voltage reference setting is $V_{ref,i}$ and the terminal voltage is V_i . All generators are assumed to have the same exciter saturation function, $S_{Ei}(E_{fdi}) = 0.0039e^{1.555E_{fdi}}$. Although this formulation may not fully represent the details of the most modern excitation systems, it should closely approximate the behavior of many installed systems [16]. The results of simulations with expanded generator dynamic models are presented in [8].

The generator stator algebraic equations are [34]:

$$\begin{aligned} 0 &= V_i \sin(\delta_i - \theta_i) - x_{qi} I_{qi} \\ 0 &= E'_{qi} - V_i \cos(\delta_i - \theta_i) - x'_{di} I_{di} \end{aligned} \quad (5)$$

where θ_i is the bus angle. Together, the generator stator equations and power balance equations at each node comprise the algebraic equations in a differential-algebraic model of the power system.

3.2 Rate of Change of Frequency Protection

As noted previously, this work postulates an attacker targeting ROCOF protection on generators. A draft of NERC's standard PRC-024-1 stated that the ROCOF must be greater than or equal to 2.5 Hz/s to trip in the no-trip frequency zone [36], but this language was eventually removed to allow trips for "documented limitations" regardless of ROCOF [37]. Additionally, a PPA Energy report notes that most generator turbine types cannot maintain transient stability with a ROCOF of 2 Hz/s, but can do so at 1 Hz/s if not operating at leading power factor [33]. The work here characterizes a malicious attack as "successful" when a ROCOF exceeding 2 Hz/s is induced at a targeted generator. At this value, it is assumed that the protection system would disconnect the generator from the grid. Explorations into the best techniques to measure ROCOF remain an active area of ongoing discussion [27, 32]. Given the severity induced by the following examples, we believe there is a high probability of tripping, regardless of measurement algorithm. Malicious control causing two generators to trip offline would be an N-2 contingency which may not have been considered by the planning coordinator in stability studies.

3.2.1 Emulated Inertia Overview

As the name suggests, emulated inertia control seeks to emulate the dynamics of electrically coupled rotating inertia, as traditionally provided by synchronous generators. As synchronous generators are displaced by power electronically cou-

pled renewable resources, the amount of system inertia decreases. If one accepts the premise that significant system inertia is a desirable property, the decrease in rotating inertia may be a problem. The potential of emulated inertia controls at distributed points as a means to compensate for reduced system inertia has been explored by many authors under the label of “virtual synchronous generators” [1, 6, 15, 17, 20, 39], including incorporation in photovoltaic systems [26] and high-voltage dc terminals [41]. In particular, this work assumes a scenario in which load points have storage locally available and can adjust their demand in order to provide inertia-emulating behavior; a representative example is that of electric vehicle chargers [40].

From a feedback control perspective, emulated inertia is extremely simple. It is a particular choice of local linear feedback control that produces an actuation signal varying the active power injection or withdrawal at a bus, in response to the bus frequency (i.e., $\Delta P = K(s)\Delta\omega$). For idealized emulated inertia, $K(s) = Ms$, which represents a pure differentiator with a gain analogous to an inertia, M . A graphical representation of how the emulated inertia control interacts with the system is shown in Figure 1. While $K(s) = Ms$ is ideal, any practical derivative control has additional filtering. In practice, therefore, $K(s)$ is likely implemented as one or more biquad filters (i.e., a cascade of second-order transfer functions with programmable coefficients).

System nodes with local storage become controllable and we propose the following state equations to model behavior at each bus exercising such control

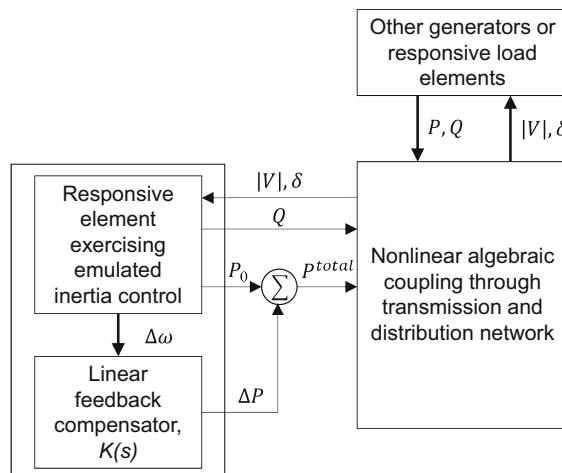


Fig. 1 Relationships between the responsive grid elements with emulated inertia control and the power system network

$$\begin{aligned}\dot{\theta}_i &= \omega_{Li} - \omega_o \\ \dot{\omega}_{Li} &= -\frac{1}{M_{Li}}(P_{mismatch} + D_{Li}(\omega_{Li} - \omega_o))\end{aligned}\tag{6}$$

where θ_i is the angle at bus i , ω_{Li} is the bus frequency, and M_{Li} and D_{Li} are the equivalent load inertia and damping constants, respectively. $P_{mismatch}$ is the sum of the power supplied by the network, the power demanded by the load, and the additional controlled power input. The above formulation closely approximates those used in [1] and [26].

To the authors knowledge, distributed emulated inertia control is not yet widely adopted in utility practice. Hence, typical implementation information is not available and we have made the assumptions detailed here and in Section 3.3 out of necessity.

3.3 Test System

As noted in Section 1, cyber-physical grid attacks through the subversion of protection and control systems have received considerable attention in the last decade. However, the literature reveals few studies with more than 50 buses, and none in large systems that include detailed models of synchronous generators and their control systems. Indeed, most studies have been limited to classical generator models, dc power flow models of network, and include an infinite bus. This work's case study of a large test system with detailed generator models is an attempt to remedy these shortcomings.

The Western Systems Coordinating Council (WSCC), now the Western Electricity Coordinating Council, oversees the electrical system of the Western Interconnection of the United States. The test system used here is a reduced representation of the WSCC system circa 1985 and was developed in the early 1990s during a joint project between the Brazilian Electric Energy Research Center (CEPEL-Brazil), the Electric Power Research Institute (EPRI), and the University of Wisconsin to explore control system interactions. This was before the release of critical energy infrastructure data was restricted. The system has 179 buses, 263 branches including transformers, and 29 generators. A one-line diagram is shown in Figures 2 and 3. The complete system information for the 179-bus representation has been presented in several publicly available sources [21, 31], and additional generator dynamic parameters are available in [19], which uses a network that has been further reduced. The dynamic parameters of the excitation system were chosen consistent with a stable linearization at the chosen operating point, based on typical values found in [34]. With Table 1, references [19, 21, 31], and the operating point information in the Appendix, the interested reader should be able to validate the simulations herein. The per unit base for this work is 125 MVA.

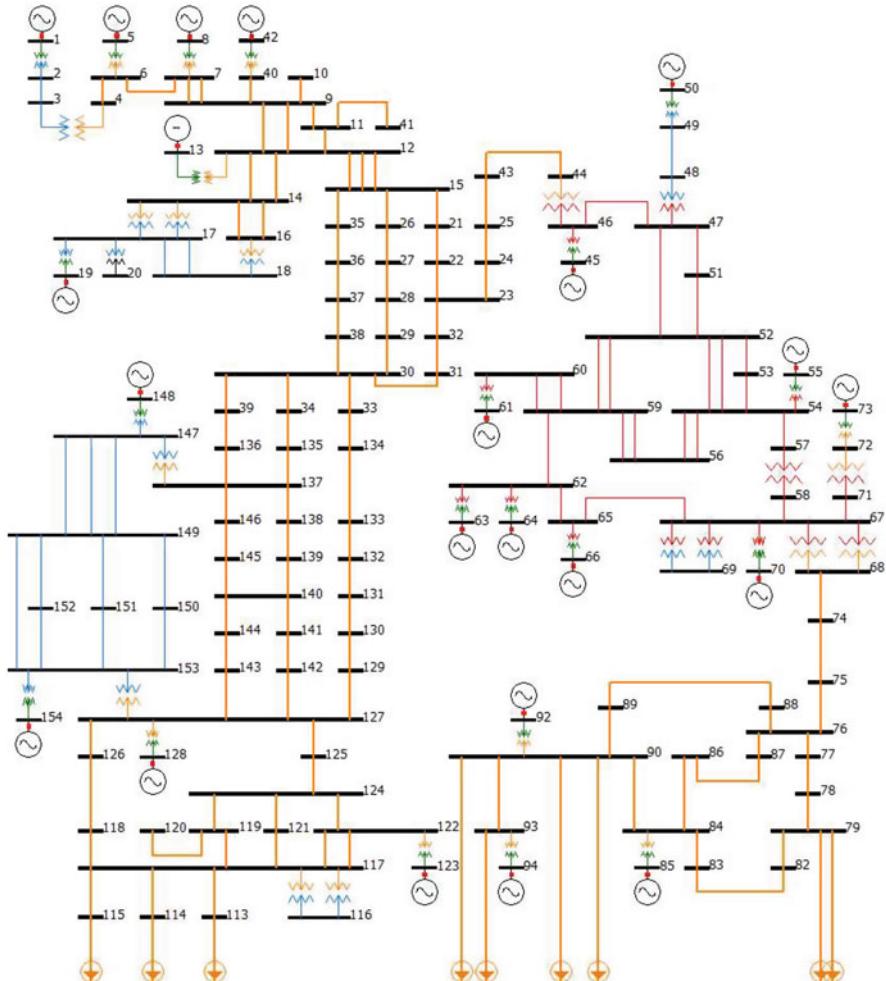


Fig. 2 Partial one-line diagram of the WSCC 179-bus test system, the system continues at the arrows in Figure 3

In this work, all 29 generators are represented with the one-axis model with dynamics of an IEEE Type I excitation control system and primary (governor) frequency control (no infinite bus). The network behavior is assumed to be positive sequence with a balanced phasor representation and is simulated using the nonlinear algebraic power flow equations. Off-nominal transformers are included but are not actively controlled during the time horizon studied. The loads are assumed to have fixed P-Q demand, supplemented at select locations with the dynamics of emulated inertia control. Five locations with emulated inertia control are assumed to be subverted in any particular study, and the choice varies with the mode to

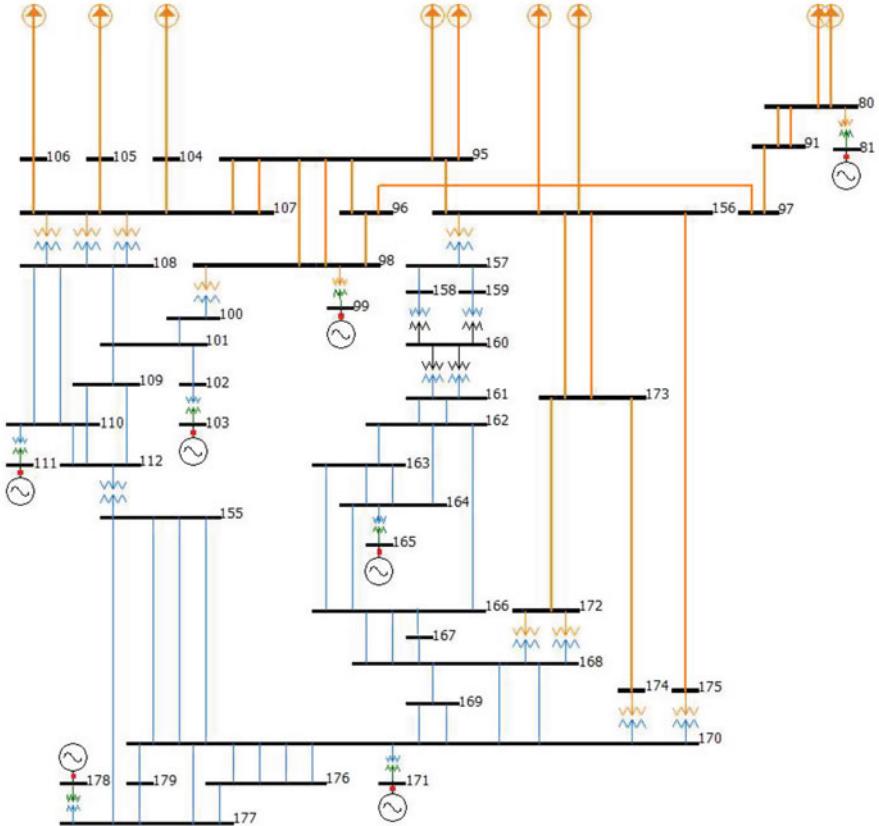


Fig. 3 Partial one-line diagram of the WSCC 179-bus test system, the system continues from the arrows in Figure 2

be destabilized and the generators that are targeted. The candidate locations for emulated inertia control are shown in Table 2, along with the assumed values of the parameters. A discussion of the selection of M and D values for emulated inertia control in a smaller test system is found in [7].

4 Demonstration of Vulnerability

The attacker must answer some key design questions: which mode should be destabilized, which generator buses should be targeted, and, finally, which load buses should be subverted? Taken together, exact, rigorous answers to this set of question are computationally challenging. Here, we examine the use of Hautus

Table 1 Generator dynamic constants in the WSCC 179-bus model

Generator	Bus	D_i (pu)	T'_{doi} (s)	T_{Ai} (s)	K_{Ai} (pu)	T_{Fi} (s)	K_{Fi} (pu)	T_{Ei} (s)	K_{Ei} (pu)
1	1	0.05	5.874	0.1978	8.60	0.3361	0.0605	0.3016	0.9604
2	5	0.05	6.216	0.2061	9.10	0.3636	0.0655	0.3262	1.0389
3	8	0.05	6.060	0.2050	8.84	0.3472	0.0625	0.3115	0.9919
4	19	0.05	5.976	0.1949	20.63	0.3390	0.0610	0.3041	0.9684
5	42	0.05	6.165	0.1919	20.13	0.3493	0.0629	0.3134	0.9980
6	45	0.05	5.871	0.1951	19.17	0.3524	0.0634	0.3161	1.0068
7	50	0.05	5.790	0.1902	19.79	0.3515	0.0633	0.3154	1.0043
8	55	0.05	5.712	0.1958	19.31	0.3428	0.0617	0.3075	0.9793
9	61	0.05	5.991	0.2073	20.25	0.3498	0.0630	0.3139	0.9995
10	63	0.05	5.850	0.1982	20.28	0.3364	0.0606	0.3018	0.9612
11	64	0.05	6.066	0.2076	8.95	0.3406	0.0613	0.3055	0.9730
12	66	0.05	6.129	0.1915	19.52	0.3657	0.0658	0.3281	1.0449
13	70	0.05	6.174	0.2076	8.69	0.3501	0.0630	0.3141	1.0002
14	73	0.05	5.979	0.1957	19.82	0.3605	0.0649	0.3235	1.0301
15	81	0.05	6.000	0.2007	3.91	0.3632	0.0654	0.3258	1.0376
16	85	0.05	5.748	0.1916	9.11	0.3503	0.0631	0.3143	1.0009
17	92	0.05	5.727	0.1948	9.05	0.3435	0.0618	0.3082	0.9815
18	94	0.05	5.841	0.1942	19.88	0.3442	0.0620	0.3088	0.9834
19	99	0.05	6.180	0.2027	19.92	0.3390	0.0610	0.3041	0.9685
20	103	0.05	6.024	0.2038	20.53	0.3605	0.0649	0.3234	1.0301
21	111	0.05	5.865	0.2022	19.00	0.3362	0.0605	0.3016	0.9606
22	123	0.05	6.135	0.1906	19.26	0.3402	0.0612	0.3052	0.9721
23	128	0.05	6.057	0.1988	8.59	0.3555	0.0640	0.3189	1.0156
24	148	0.05	5.757	0.2020	20.03	0.3460	0.0623	0.3104	0.9886
25	154	0.05	5.709	0.2003	20.44	0.3328	0.0599	0.2986	0.9509
26	165	0.05	6.243	0.1984	19.73	0.3461	0.0623	0.3105	0.9889
27	171	0.05	6.237	0.2068	19.13	0.3501	0.0630	0.3141	1.0002
28	178	0.05	5.904	0.1921	20.92	0.3369	0.0606	0.3022	0.9625
29	13	0.05	6.000	0.2000	10.00	0.3500	0.0630	0.3140	1.0000

matrix singular-value-based measures of controllability and observability, described in [5], instead of network graph heuristics. The control Hautus matrix is defined as

$$H_C = [\lambda I - \mathbf{A} | \mathbf{B}] \quad (7)$$

where I is the identity matrix with correct dimension. The minimum SV of the control Hautus matrix is denoted as σ_k for a particular mode k . The mode of interest to the attacker is associated with the complex pair of eigenvalues selected to be destabilized. Then, five potential locations are chosen from a list of twenty and, for each possible selection yielding a different \mathbf{B} input matrix, σ_k was calculated. The group with the maximum σ_k was chosen as the SV-based placement.

Table 2 Dynamic constants of possible emulated inertia control at different network locations

Bus Number	$M_{Li} (s^2/rad)$	$D_{Li} (\text{pu})$	$P (\text{pu})$
4	0.0494	0.1483	4.000
43	0.0502	0.1478	8.000
46	0.0522	0.1571	4.880
52	0.0501	0.1469	3.662
56	0.0517	0.1518	3.032
65	0.0492	0.1535	6.720
79	0.0517	0.1469	8.000
80	0.0501	0.1455	9.333
90	0.0491	0.1426	7.218
91	0.0509	0.1533	6.848
96	0.0503	0.1481	9.840
97	0.0476	0.1478	3.248
101	0.0510	0.1530	3.019
108	0.0503	0.1466	8.528
112	0.0493	0.1466	3.209
116	0.0475	0.1568	6.221
153	0.0515	0.1548	7.072
163	0.0512	0.1535	2.560
168	0.0481	0.1559	6.462
176	0.0487	0.1517	7.102

Then, there is a question about what constitutes a “successful” design from the attacker’s point of view. If we assume that the attacker wants the attack to be subtle (which was not the case during the attack on the Ukrainian power grid, where it was soon clear that it was a cyberattack [3]), then the following questions may be considered when deciding on the “success” of an attack:

1. Are the largest magnitude frequency deviations found at the targeted generator buses?
2. Does the power change commanded at the subverted control points stay within $\pm 100\%$ of the nominal load?
3. From a system perspective, is the total power change commanded at any moment not “noticeable,” e.g., does it stay within $\pm 2\%$ of the total system load?

Again, the subverted control design is local and purely linear; however, its impact is evaluated (empirically) by examining the state trajectories in the full nonlinear model.

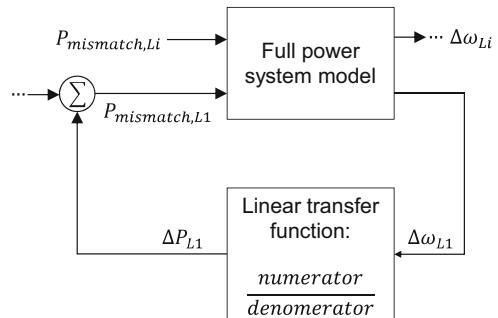
For this demonstration, two lightly damped electromechanical modes were selected to be destabilized. Table 3 shows the modes, original eigenvalues (λ^0), destabilized eigenvalues (λ), targeted generators, and the chosen locations for the subverted control. For each of these modes, a base simulation was completed, where it was assumed that the parameters, topology, and system loading were known exactly. The transfer functions between $\Delta\omega_{Li}$ and ΔP_{Li} are listed in Table 4, according to the description in Figure 4. We note that the form of the transfer

Table 3 Modes, eigenvalues, and subverted control locations for the case study

Mode	λ^0	λ	Targeted Generators	Subverted Buses
I	$-0.2428 \pm j 6.6246$	$1.25 \pm j 6.6246$	7	43, 52, 108, 116, 176
II	$-0.1571 \pm j 10.1452$	$1.5 \pm j 10.1452$	12, 13	43, 65, 108, 116, 153

Table 4 Transfer functions at local load control buses

Mode	Load Location	Denominator	Numerator
I	43		$-0.50 \times 10^{-4}s - 14.2 \times 10^{-4}$
	52		$-38.2 \times 10^{-4}s + 12.5 \times 10^{-4}$
	108	$s^2 - 0.29s + 44.11$	$4.96 \times 10^{-6}s - 34.3 \times 10^{-6}$
	116		$0.27 \times 10^{-6}s + 5.57 \times 10^{-6}$
	176		$-0.29 \times 10^{-6}s + 2.37 \times 10^{-6}$
II	43		$-0.13 \times 10^{-7}s + 1.21 \times 10^{-7}$
	52		$-36.7 \times 10^{-4}s + 0.33 \times 10^{-4}$
	108	$s^2 - 0.39s + 102.3$	$-0.10 \times 10^{-6}s + 7.73 \times 10^{-6}$
	116		$0.16 \times 10^{-7}s - 13.3 \times 10^{-7}$
	176		$0.51 \times 10^{-9}s - 5.28 \times 10^{-8}$

Fig. 4 The transfer function at local control point L_i 

functions indicate that the controllers are unstable, indicating that limits on local control parameters could be hardwired to prevent parameters that could lead to an unstable local controller from being selected.

Several robustness simulations were completed to evaluate the performance of the designed control in the presence of incorrect parameter values in the attacker's model of the underlying system. In other words, the same control was applied to time domain simulations with different parameters and topology. In the case of uncertain loading, the original control was designed for a system that was relatively heavily loaded, as measured by its nearness to the edge of stability during linearization. In order to evaluate loading uncertainty, the control was designed for a lower loading point, approximately 85% of the values provided in the Appendix, and it was simulated for $\pm 15\%$.

As described in Section 3.2, it is assumed that a ROCOF value of 2 Hz/s would cause the protection systems to disconnect a generator and it is that point which

Table 5 Time and control effort to reach the first ROCOF value above 2 Hz/s

Case	Variation	Mode I Time (s)	Mode I Max $ \Delta\text{Load} $ (pu)	Mode II Time (s)	Mode II Max $ \Delta\text{Load} $ (pu)
Base	—	101.16	0.4205	91.00	0.2168
Gen. Parameters	$\pm 2\%$	101.17	0.4493	91.33	0.2180
	$\pm 6\%$	102.11	0.4835	91.33	0.2052
	$\pm 11\%$	101.20	0.4234	90.42	0.2091
Topology	Line 1 Open	102.15	0.4406	91.52	0.2065
	Line 2 Open	102.10	0.4193	91.17	0.2148
	Line 3 Open	101.10	0.4295	91.17	0.2148
Load Scaling	-15%	103.20	0.5784	94.94	0.2622
	Designed	101.80	0.5256	94.67	0.2357
	+15%	101.00	0.4595	90.98	0.2239

is used to determine the “success” of the attack. The time domain simulations are shown for the time in which the maximum load change commanded remains within ± 1 pu (or $\pm 100\%$ the nominal value). The control effort that it takes to reach a ROCOF of 2 Hz/s and the time that threshold is first reached are summarized for all simulations in Table 5. The reader may observe from this table that the control strategy is only slightly degraded, with a slightly longer time interval and a slight increase in control effort to achieve the objective. The detailed results of each mode will be described in Sections 4.1 and 4.2.

4.1 Result for Mode I

The generator parameters were varied from their base values by an increasing window of $\pm 2\%$, $\pm 6\%$, and $\pm 11\%$, with an assumed uniform distribution about the designed values. The frequency of the targeted generator for Mode I is shown in Figure 5, while the ROCOF measured at the targeted generator is shown in Figure 6. In these simulations, the curves for $\pm 2\%$ and $\pm 11\%$ were nearly indistinguishable from the base on the scale of the shown plots. For these simulations, the differing levels of uncertainty had a negligible effect on the maximum frequency deviation (approximately ± 0.3 Hz). It is seen that uncertainty in the generator parameters shifts the behavior in time, but the targeted generator still exhibited ROCOF excursions in excess of 2 Hz/s. The control effort at the load point with the maximum change is shown in Figure 7 and the net change in load at each moment in time is shown in Figure 8. Although the curves follow one another fairly closely, it is noted that it took an additional cycle for the case with the largest uncertainty in generator parameters to reach the ROCOF threshold. The load change command stays within ± 0.6 pu to achieve the frequency and ROCOF values described above, while the total net power change command is less than 0.5% of the total system load.

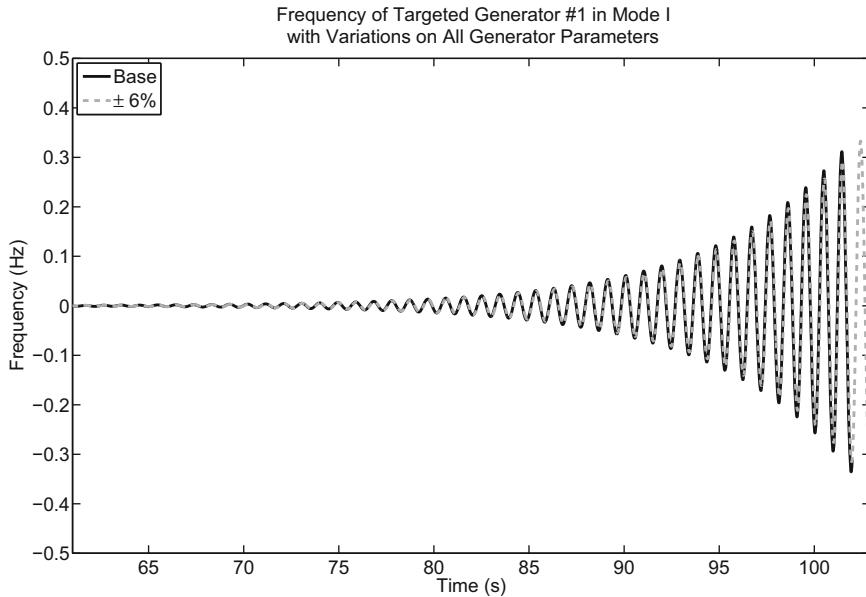


Fig. 5 The frequency measured at the targeted generator 7 for Mode I for uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 11\%$ were indistinguishable from the base

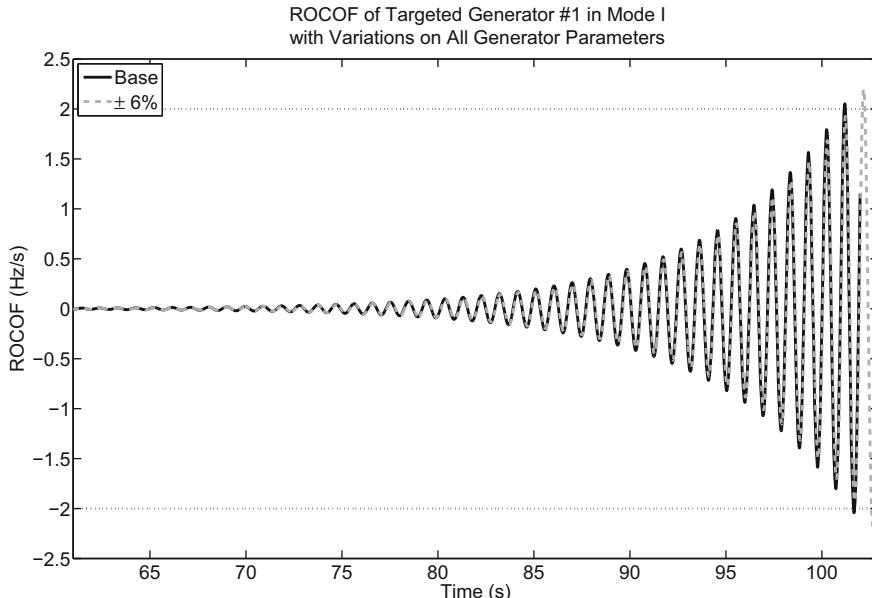


Fig. 6 The ROCOF measured at the targeted generator 7 for Mode I for uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 11\%$ were indistinguishable from the base

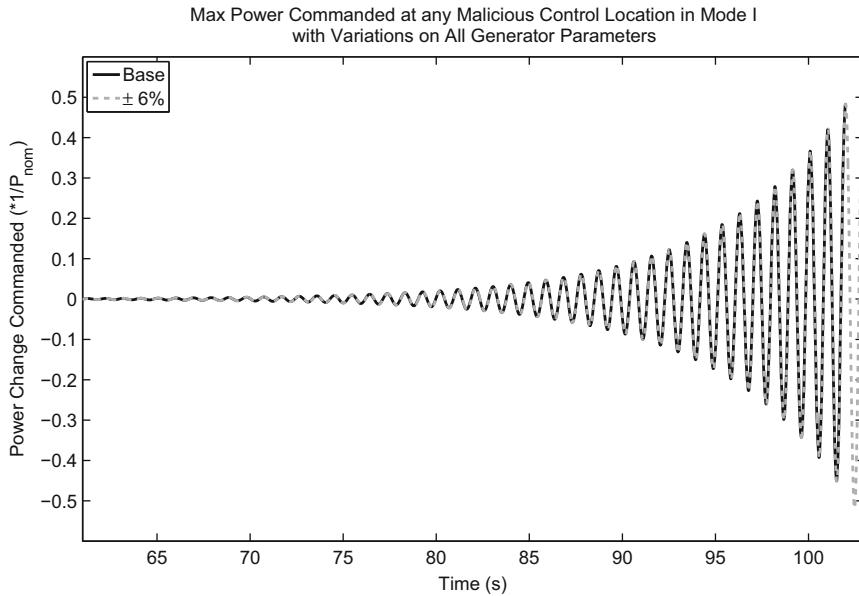


Fig. 7 The change in load the control point with the maximum absolute change for Mode I, considering uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 11\%$ were indistinguishable from the base

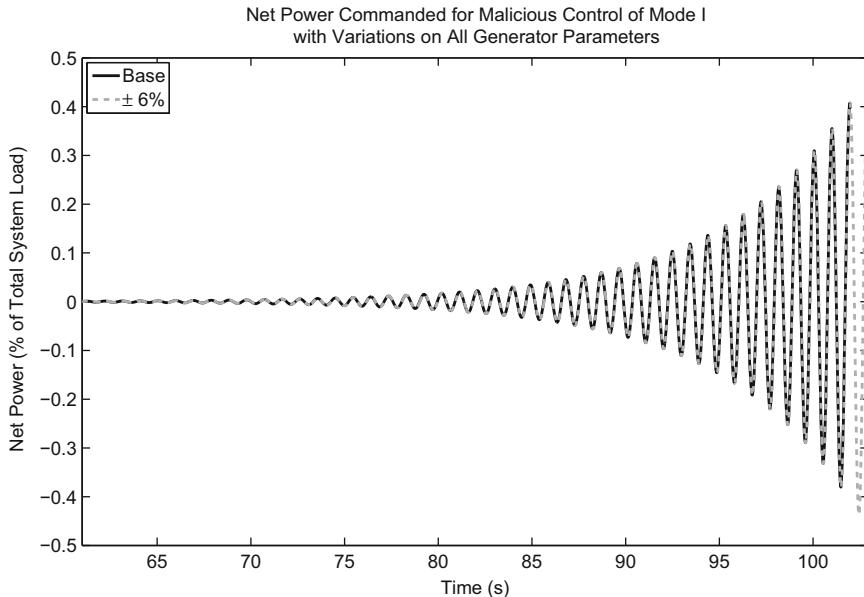


Fig. 8 The total net change in load for Mode I, considering uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 11\%$ were indistinguishable from the base

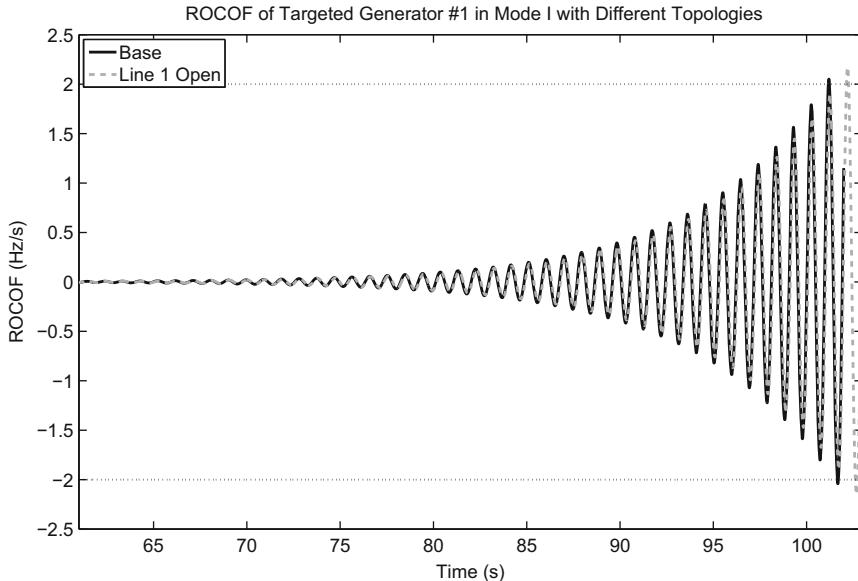


Fig. 9 The ROCOF measured at the targeted generator 7 for Mode I for uncertainty in the system topology. The curves for lines 2 and 3 out of service were indistinguishable from the base

To test robustness under topology changes, the base control design was applied to a network with one of three lines taken out of service. This is a reasonable test, because the power system often has lines taken out of service for maintenance or upgrades. Line 1 is the line between buses 32 and 33; lines 2 and 3 are the lines between buses 88 and 89 and 118 and 126, respectively. For Mode I, the base case and the two cases with line 2 or 3 out of service were indistinguishable. The ROCOF values for the base case and line 1 out of service are shown in Figure 9. The frequencies, maximum load commands, and net load values are very similar to those shown in previous figures. The difference in topology only delays the time when the ROCOF limit is reached by approximately 1 s, with a very small increase in control effort (0.0201 pu).

To represent a scenario in which an attacker might not have accurate information regarding the system's real time operating point, the malicious control was redesigned for a lower loading level. Then, the time domain simulations were performed when the system loading was increased or decreased. The plot of the ROCOF values is shown in Figure 10, where the increased loading curve was indistinguishable from the moderate loading curve. Changing the system loading only shifts the time at which the ROCOF limit is reached, not whether it is reached. The frequency, control effort, and net load change were similar to those shown previously. It can be noted in Table 5 that the case with heavier loading requires slightly less control effort to reach the ROCOF threshold, while lighter loading requires more effort.

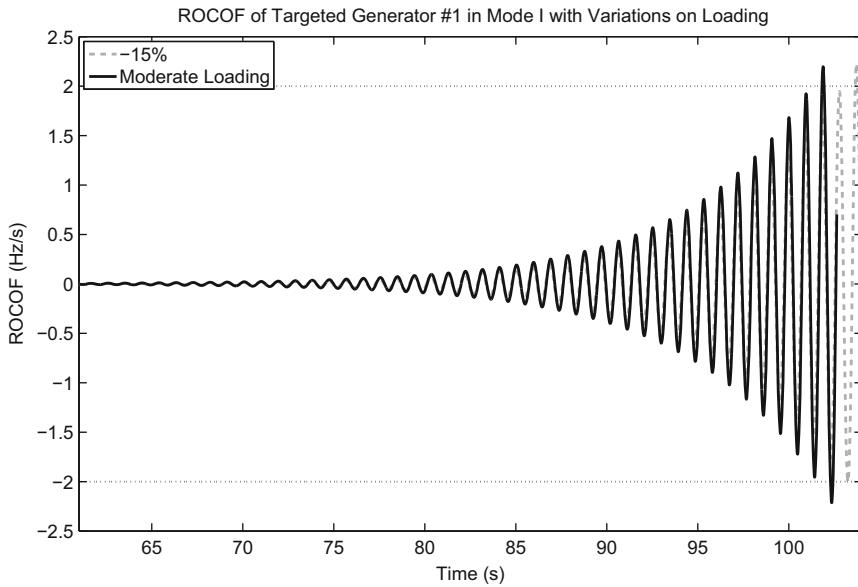


Fig. 10 The ROCOF measured at the targeted generator 7 for Mode I for uncertainty in the system loading. The increased loading curve was indistinguishable from the moderate loading curve

4.2 Results for Mode II

For Mode II, the results for variations in the generator parameters are shown in Figures 11, 12, 13, and 14. Figure 11 shows the difference in frequency for variations of uncertainty in generator parameters. The curves for the simulations with $\pm 2\%$ and $\pm 6\%$ were indistinguishable from the base. The ROCOF values are shown in Figure 12, and uncertainty in parameters merely shifts the moment in time when the ROCOF value reaches the threshold. The control effort, or the maximum change in load commanded, is shown in Figure 13, while the net load change is shown in Figure 14. From these figures, the control effort is seen to stay within ± 0.5 pu, and the net load change is less than 0.5%. Of note, in this set of robustness tests, greater uncertainty in the generator parameters resulted in slightly less control effort required to reach the ROCOF threshold, as seen in Table 5.

For uncertainty in topology, the curves of the ROCOF values in Mode II are shown in Figure 15 and control effort to reach a ROCOF of 2 Hz/s are shown in Table 5. The case with line 2 open was indistinguishable from the case with line 1 open, while the case with line 3 open was indistinguishable from the base. The ROCOF values for uncertainty in loading are shown in Figure 16. The curve for

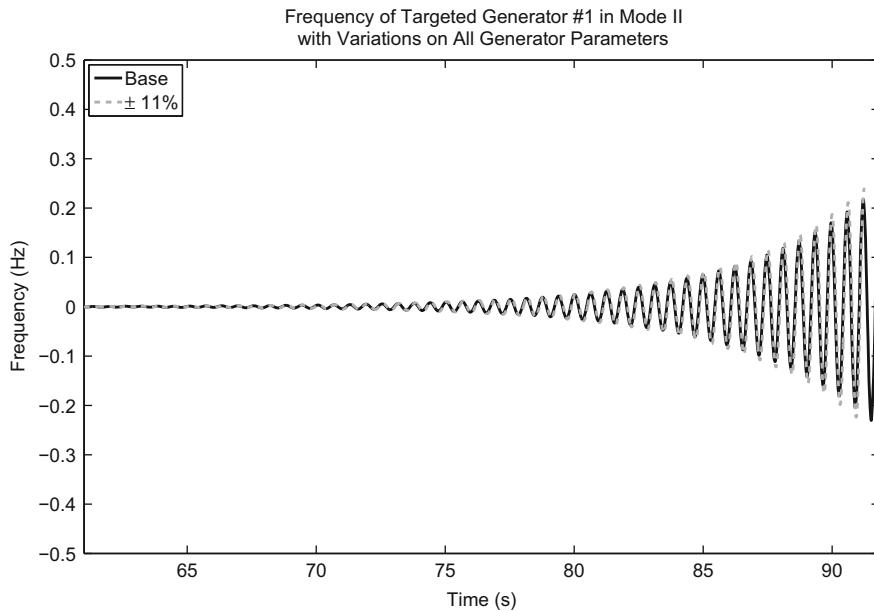


Fig. 11 The frequency measured at the targeted generator 12 for Mode II for uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 6\%$ were indistinguishable from the base

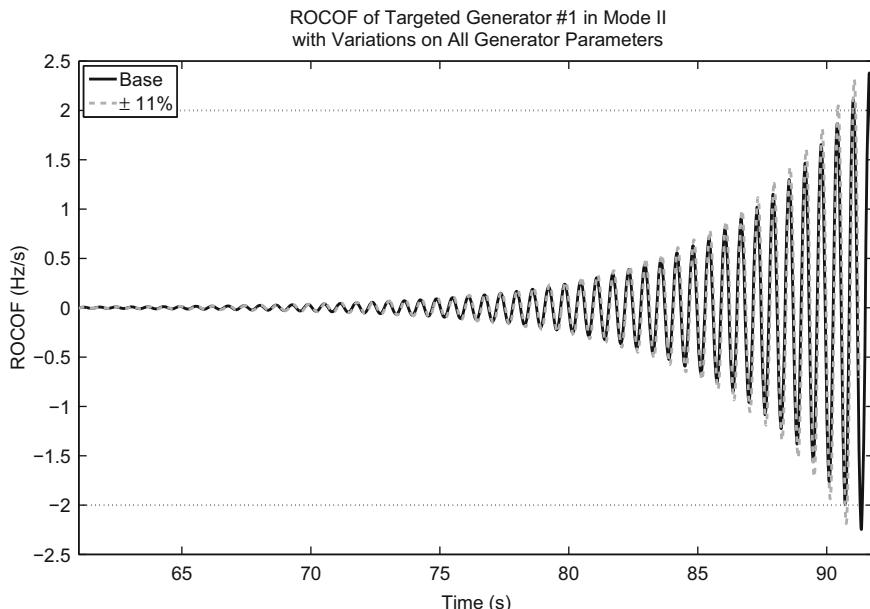


Fig. 12 The ROCOF measured at the targeted generator 12 for Mode II for uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 6\%$ were indistinguishable from the base

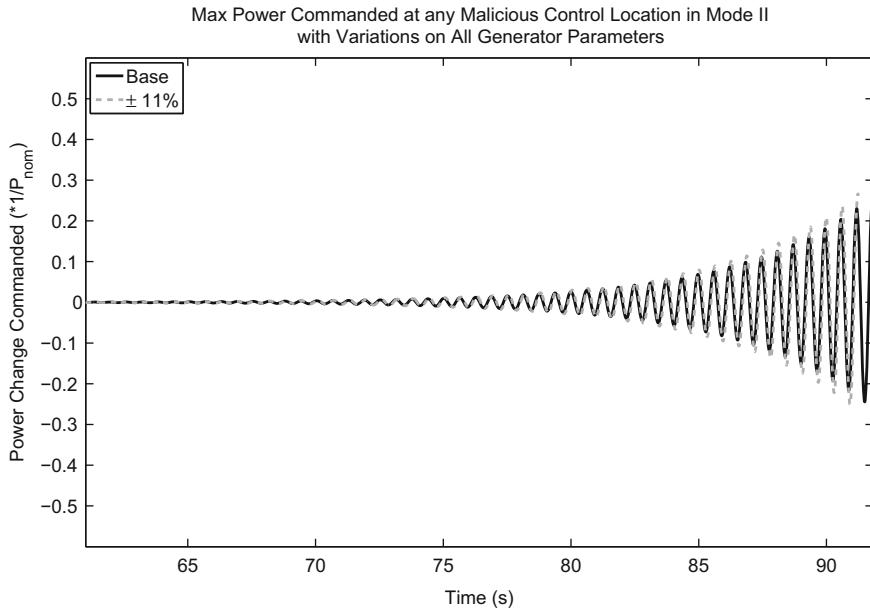


Fig. 13 The change in load the control point with the maximum absolute change for Mode II, considering uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 6\%$ were indistinguishable from the base

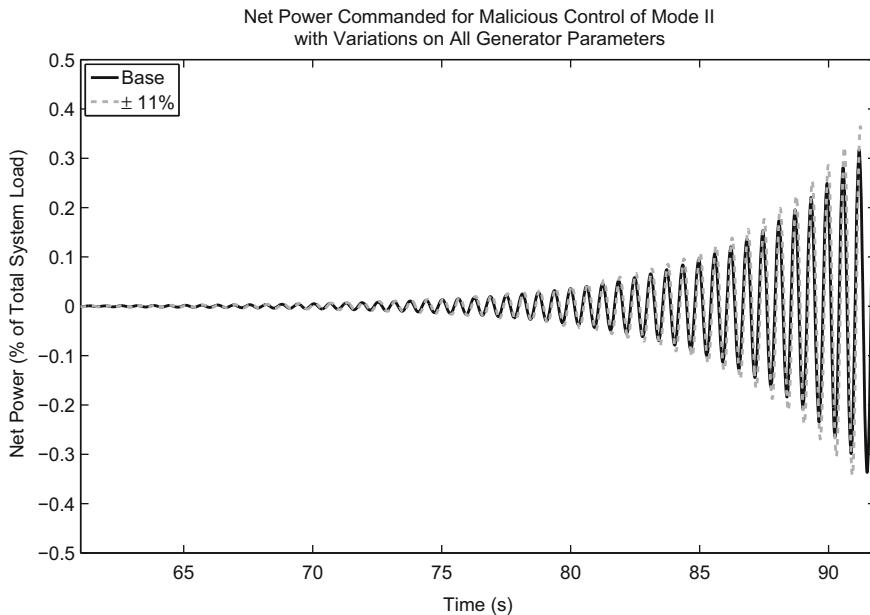


Fig. 14 The total net change in load for Mode II, considering uncertainty in the generator parameter values. The curves for $\pm 2\%$ and $\pm 6\%$ were indistinguishable from the base

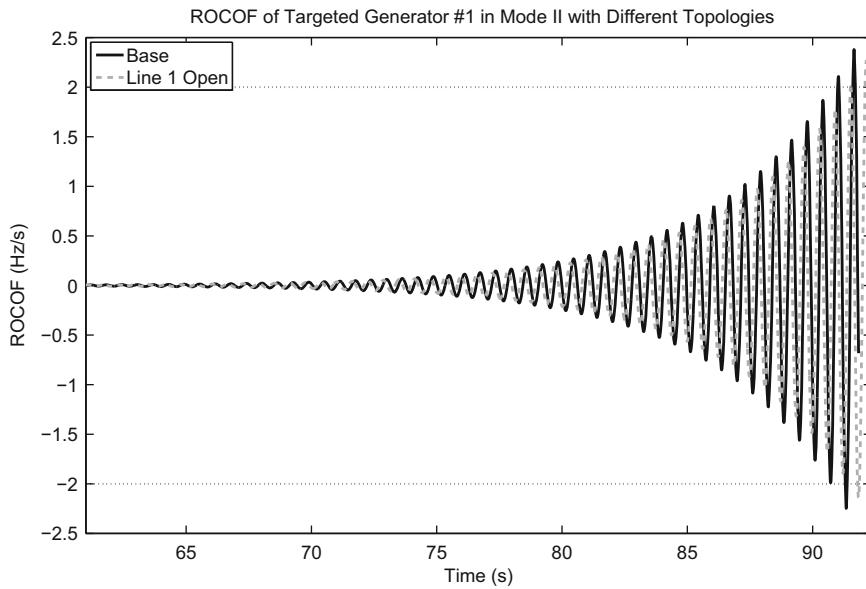


Fig. 15 The ROCOF measured at the targeted generator 12 for Mode II for uncertainty in the system topology. The curve for line 2 open was indistinguishable from the curve for line 1 open; the curve for line 3 open was indistinguishable from the base

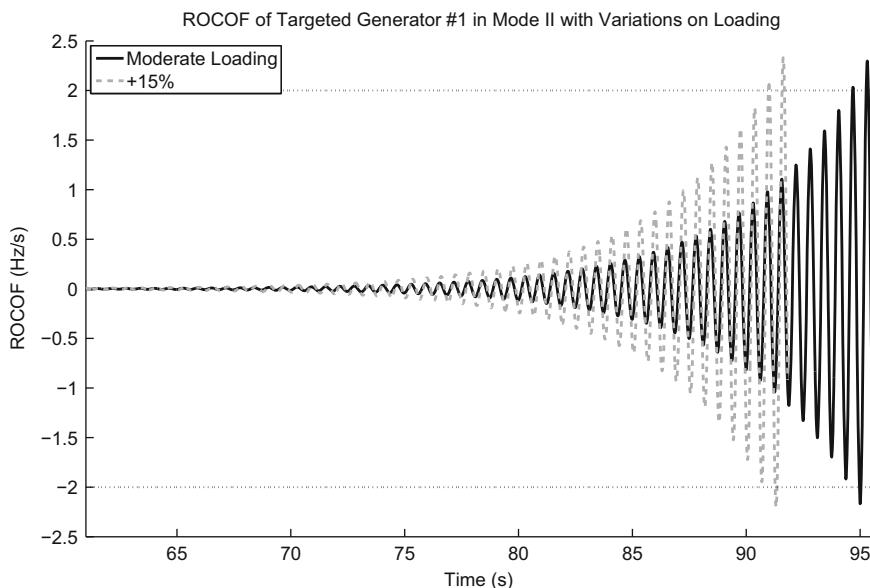


Fig. 16 The ROCOF measured at the targeted generator 12 for Mode II for uncertainty in the system loading. The curve for reduced system loading was indistinguishable from moderate loading

reduced system loading was indistinguishable from moderate loading. Again, it is seen that changes in the underlying system only shift the point in time when the ROCOF threshold is reached. As in Mode I, heavier system loading reduces the control effort, while decreased system loading increases the required control effort.

5 Conclusion

In this paper, simulations demonstrating the power grid's vulnerability to cyber-physical attack through compromised feedback loops were presented. The malicious control, which originated from undermined synthetic inertia stabilizers, caused conditions which would likely trip generators offline. Furthermore, this result persisted under robustness testing, where the only difference in performance was a question of “when” the tripping conditions were reached, not “if.”

The work presented in this paper is intended as a caution against complacency regarding the need for cybersecurity in distributed grid control, even at the load level. Demonstrating the vulnerability is fairly straightforward, but it raises more interesting questions. The field of control systems has long examined the trade-offs between robustness and performance, while quantitative finance similarly examines reward versus risk. In the control architecture of a responsive grid, it will be valuable to “design in” a consideration of performance versus (cyber-) risk. Based on the experience so far, intrusion avoidance motivated by the consumer and post-intrusion protective responses originating with utility operations are likely inadequate strategies. The results in this paper support the idea that control parameter limits could be used to preclude local instability. However, it does not follow that those limits would completely address this vulnerability. Future work may explore the possibility of more surreptitious control designs, where the local controller transfer functions are stable, but the instability arises through the coupling to the system. Other avenues may consider further limitations on the amount of load at each controlled node which may be subverted by an attack.

Acknowledgments This research was supported by the US Department of Energy Lawrence Berkeley National Laboratory, through the project “Algorithmic/Computational Research within Power System Vulnerability,” under a prime contract No. DE-AC02-05CH11231.

Appendix

The operating point used for the simulations detailed in this chapter is included in Table 6, since it is slightly different from that used in [21].

Table 6 Bus data for the WSCC 179-Bus Test System, including operating point around which the system is linearized

	Bus	V (pu)	$\angle V$ (°)	P (MW)	Q (MVAr)	Bus	V (pu)	$\angle V$ (°)	P (MW)	Q (MVAr)	Bus	V (pu)	$\angle V$ (°)	Type (MW)	P (MW)	Q (MVAr)	
1	1.020	41.91	2	4480	1150.20	61	1.050	9.87	2	1780	534.60	121	1.039	-27.12	3	0	0
2	1.002	36.88	3	-3600.1	-699.9	62	0.975	15.92	3	-3000	127.1	122	1.045	-25.02	3	-50	-24.9
3	0.994	26.74	3	0	0	63	0.950	23.97	2	1048	-132.90	123	0.980	-20.61	2	765	-4206.30
4	1.106	21.78	3	-500	0	64	1.000	34.30	2	2050	464.80	124	1.039	-27.72	3	-305	-192
5	1.000	13.05	2	4450	1011.10	65	1.040	16.81	3	-840	831.5	125	1.000	-25.92	3	0	0
6	1.041	9.19	3	-4400	-1000.2	66	1.000	20.13	2	962	148.80	126	1.006	-26.25	3	0	0.1
7	1.067	4.08	3	-5000	2245.9	67	1.014	16.65	3	-239	-259.8	127	1.020	-26.46	3	-5661	-502.4
8	1.000	18.47	2	9950	1854.00	68	1.075	13.97	3	0	-257.7	128	1.000	-18.18	2	3467	1654.40
9	1.090	-6.82	3	-3500	711.6	69	1.013	16.11	3	-139.7	-23.8	129	0.969	-31.58	3	0	0.1
10	1.112	-6.88	3	0	0	70	1.000	23.87	2	2160	-30.50	130	1.071	-21.37	3	0	0
11	1.088	-15.84	3	-2584	-394	71	0.987	5.73	3	0	0	131	1.033	-25.98	3	0	0
12	1.107	-11.53	3	-3200	1290.1	72	0.985	-2.47	3	-1750	56	132	1.102	-17.30	3	0	0
13	1.000	0.00	1	5373	326.03	73	1.040	4.94	2	800	123.00	133	1.077	-21.63	3	189	-61.4
14	1.109	-13.52	3	44.4	-21.7	74	1.072	15.74	3	0	-0.2	134	1.051	-25.22	3	0	0
15	1.112	-14.34	3	66.6	-1438.8	75	1.066	-2.77	3	0	0	135	1.054	-21.47	3	0	0
16	1.109	-13.62	3	-0.3	1097.2	76	1.068	-0.12	3	0	-890.9	136	1.053	-21.45	3	0	0
17	1.082	-14.64	3	67.3	1150.9	77	1.057	1.20	3	0	0	137	1.076	-19.52	3	0	-194.8
18	1.078	-15.15	3	-3136.7	103.6	78	1.039	-5.74	3	0	0	138	1.087	-15.78	3	0	0
19	1.055	-7.56	2	1301	431.50	79	1.054	-4.39	3	-617	-883.2	139	1.038	-24.71	3	0	0
20	1.085	-16.94	3	-160	-31.2	80	1.046	-4.33	3	-793.3	-527.9	140	1.052	-22.06	3	0.7	-305.5
21	1.117	-15.05	3	0	0	81	0.960	3.93	2	2640	378.10	141	1.065	-19.05	3	0	0
22	1.113	-13.86	3	0	0	82	1.024	-6.20	3	0	0	142	1.011	-28.08	3	0	0.2
23	1.109	-14.84	3	0.1	0	83	1.047	2.84	3	0	0	143	0.993	-30.82	3	0	0.1

(continued)

Table 6 (continued)

Bus	V (pu)	$\angle V$ (°)	Type	P (MW)	Q (MVAr)	V (pu)	$\angle V$ (°)	Type	P (MW)	Q (MVAr)	V (pu)	$\angle V$ (°)	Type	P (MW)	Q (MVAr)	
24	1.042	-9.07	3	0	-426.3	84	1.072	1.09	3	-90	-506.7	144	1.063	-18.06	3	0
25	1.107	-15.37	3	0	-487.8	85	1.000	7.60	2	1690	195.60	145	1.038	-24.71	3	0
26	1.126	-16.58	3	0	0	86	1.068	2.50	3	0	0	146	1.087	-15.78	3	0
27	1.117	-14.40	3	0	0	87	1.064	-1.54	3	0	0	147	1.140	-15.32	3	-148
28	1.116	-16.89	3	0	0	88	1.057	8.91	3	0	0	148	1.020	-4.35	2	-323.4
29	1.122	-14.70	3	0	0	89	1.037	-17.37	3	0	0	149	1.141	-19.80	3	25.70
30	1.103	-17.29	3	339	-125.6	90	1.048	-8.16	3	-902.3	-693.6	150	1.144	-22.98	3	-210.4
31	1.103	-12.88	3	0	-0.1	91	1.029	-18.76	3	-856	-19.6	151	1.144	-22.81	3	0
32	1.111	-15.59	3	-0.1	0.1	92	1.020	-0.20	2	982.7	-128.80	152	1.142	-22.99	3	-10
33	1.107	-13.95	3	0	0	93	1.066	-4.16	3	0	-448.9	153	1.140	-26.82	3	-27.5
34	1.118	-15.12	3	0	0	94	1.050	4.73	2	1680	446.60	154	1.050	-22.12	2	-136.2
35	1.125	-16.71	3	0	0	95	1.048	-22.18	3	-204.4	27.8	155	1.035	-21.93	3	-884
36	1.117	-14.40	3	0	0	96	1.036	-25.93	3	-1230	-73	156	1.053	-17.99	3	192.30
37	1.115	-17.03	3	0	0	97	1.031	-23.31	3	-406	-41	157	1.047	-19.71	3	-271.3
38	1.122	-14.72	3	0	0	98	1.036	-25.55	3	-3097.9	-322	158	1.034	-25.45	3	0
39	1.116	-15.49	3	0	0	99	1.050	-21.03	2	1690	593.80	159	1.034	-25.45	3	-0.1
40	1.078	-10.06	3	-3000	-300	100	1.034	-25.94	3	0	0	160	1.030	-26.54	3	-237.2
41	1.131	-6.05	3	1525	50	101	1.004	-30.27	3	-377.4	-64.5	161	1.024	-26.65	3	63.4
42	1.000	-1.60	2	2910	953.30	102	1.011	-30.28	3	-3191	-630	162	1.022	-26.35	3	-138
43	0.955	-0.88	3	-1000	-250	103	1.020	-23.92	2	3195	1032.50	163	1.021	-26.29	3	-28.2
44	1.098	-23.01	3	0	-496.2	104	1.045	-25.74	3	0	-0.1	164	1.029	-24.64	3	-117
45	1.009	-12.94	2	1640	285.70	105	1.046	-25.74	3	0	0	165	1.000	-19.85	2	-65.3
46	1.007	-17.20	3	-610	-1320.6	106	1.048	-25.78	3	0	-0.1	166	1.021	-25.73	3	0.1
															-121	-24.9

47	1.026	0.25	3	-33.9	-11.9	107	1.051	-26.69	3	0	0.1	167	1.020	-25.55	3	-135	-27.1
48	1.027	0.62	3	-148	7.9	108	0.986	-28.79	3	-1066	-365.3	168	1.022	-24.78	3	-807.8	-132.1
49	1.040	10.39	3	-255	-100	109	1.007	-27.83	3	-175	-17.9	169	1.026	-23.76	3	-205.2	-17.4
50	1.000	14.02	2	445	91.70	110	1.002	-26.87	3	-3117.9	-78.1	170	1.030	-22.57	3	-121	-25.4
51	1.024	2.76	3	-185	-78.5	111	1.010	-14.15	2	2200	393.70	171	1.020	-15.45	2	110	29.10
52	1.031	4.97	3	-457.7	-212.1	112	1.017	-23.11	3	-401.1	-80.2	172	1.039	-22.61	3	0	0.1
53	1.024	6.94	3	-141.2	-71.4	113	1.032	-28.45	3	0	0	173	1.055	-17.61	3	1862	1078.3
54	1.033	14.43	3	-116.1	-511.8	114	1.031	-28.47	3	0	0	174	1.077	-21.17	3	0	0.1
55	1.050	19.46	2	1665	-31.40	115	1.031	-28.52	3	0	0	175	1.068	-20.56	3	0	-184
56	1.047	8.53	3	-379	-67.6	116	1.120	-30.46	3	-777.6	-750	176	1.028	-23.63	3	-887.7	6.4
57	1.047	14.96	3	-31.6	-50.5	117	1.045	-27.49	3	-55.6	-405	177	1.029	-21.51	3	0	0
58	1.041	15.41	3	0	0	118	1.035	-27.83	3	0	0	178	1.020	-18.90	2	200	-52.20
59	1.051	6.68	3	250	-12.7	119	1.037	-28.93	3	-265	-13.9	179	1.031	-22.30	3	72.8	17
60	1.051	4.94	3	-2053	45.6	120	1.034	-29.21	3	-40	-21.5						

References

1. Alipoor J, Miura Y, Ise T (2015) Power system stabilization using virtual synchronous generator with alternating moment of inertia. *IEEE J Emerg Sel Top Power Electron* 3(2):451–458
2. Amini S, Mohsenian-Rad H, Pasqualetti F (2015) Dynamic load altering attacks in smart grid. *IEEE power energy society innovative smart grid technologies conference*. <https://doi.org/10.1109/ISGT.2015.7131791>
3. Assante M (2016) Confirmation of a coordinated attack on the Ukrainian power grid. Available via SANS Industrial Control Systems Security Blog. <https://ics.sans.org/blog/2016/01/09/confirmation-of-a-coordinated-attack-on-the-ukrainian-power-grid>. Cited 15 Jan 2016
4. Baker S, Filipiak N, Timlin K (2011) In the dark: crucial industries confront cyberattacks. Center for Strategic and International Studies, Washington. Available via McAfee. <http://www.mcafee.com/us/resources/reports/rp-critical-infrastructure-protection.pdf>. Cited 30 Apr 2015
5. Baone C (2012) Coordinated control of wind generation and energy storage for power system frequency regulation. Doctoral Thesis, University of Wisconsin-Madison, Madison
6. Bömer J, Burges K, Nabe C, Pöller M (2010) All island TSO facilitation of renewables studies: Final Report for Work Package 3. Ecofys, Cologne. Available via Ecofys. <http://www.ecofys.com/en/publication/all-island-tso-facilitation-of-renewables-studies>. Cited 22 Oct 2015
7. Brown HE, DeMarco CL (2016) Synthetic inertia and small signal stability. *N Am Power Symp*. <https://doi.org/10.1109/NAPS.2016.7747848>
8. Brown HE, DeMarco CL (2017) Risk of cyber-physical attack via load with emulated inertia control. *IEEE Trans Smart Grid*. <https://doi.org/10.1109/TSG.2017.2697823>
9. Chen C-T (2009) Linear system theory and design. Oxford series in electrical and computer engineering. International 3rd ed. Oxford University Press, Oxford
10. Chen B, Mashayekh S, Butler-Purry KL, Kundur D (2013) Impact of cyber attacks on transient stability of smart grids with voltage support devices. In: IEEE Power Energy Society general meeting. <https://doi.org/10.1109/PESMG.2013.6672740>
11. CIP-002-4 Cyber security and critical cyber asset identification: rationale and implementation reference document (2010). North American Electric Reliability Corporation, Atlanta. Available via NERC. http://www.nerc.com/docs/standards/sar/Project_2008-06_CIP-002-4_Guidance_clean_20101220.pdf. Cited 12 Feb 2016
12. DeMarco CL (1998) Design of predatory generation control in electric power systems. In: Proceedings of the Hawaii international conference on system sciences. IEEE. <https://doi.org/10.1109/HICSS.1998.656009>
13. DeMarco CL (2000) Eigenvector assignment in power system controller design: illustration through predatory control. In: IEEE Power and Energy Society general meeting. <https://doi.org/10.1109/PES.2000.867461>
14. DeMarco CL, Sariashkar JV, Alvarado F (1996) The potential for malicious control in a competitive power systems environment. In: Proceedings of the IEEE international conference on control applications. IEEE. <https://doi.org/10.1109/CCA.1996.558870>
15. Driesen J, Visscher K (2008) Virtual synchronous generators. In: IEEE Power and Energy Society general meeting. <https://doi.org/10.1109/PES.2008.4596800>
16. Eremia M, Bulac C (2013) Description and modeling of the excitation systems. In: Eremia M, Shahidehpour M (eds) *Handbook of electrical power system dynamics: modeling, stability, and control*. IEEE Press series on power engineering. Wiley, Hoboken
17. Gonzalez-Longatt FM (2012) Effects of the synthetic inertia from wind power on the total system inertia: simulation study. In: International symposium on environment-friendly energies and applications. <https://doi.org/10.1109/EFEA.2012.6294049>
18. Gorski T, DeMarco C (1998) Application of dynamic generation control for predatory competitive advantage in electric power markets. In: Ilic M, Galiana F, Fink L (eds) *Power systems restructuring: engineering and economics*. Kluwer Academic, Boston
19. Huang L (2003) Electromechanical wave propagation in large electric power systems. Doctoral Thesis, Virginia Polytechnic Institute and State University, Blacksburg

20. Karapanos V, Kotsopoulos P, Hatzigyriou N (2015) Performance of the linear and binary algorithm of virtual synchronous generators for the emulation of rotational inertia. *Electr Power Syst Res* 123:119–127
21. Khatib A-RA (2002) Internet-based wide area measurement applications in deregulated power systems. Doctoral Thesis, Virginia Polytechnic and State University, Blacksburg
22. Kushner D (2013) The real story of Stuxnet: How Kaspersky LAb tracked down the malware that stymied Iran's nuclear-fuel enrichment program. *IEEE Spectrum*. <https://doi.org/10.1109/MSPEC.2013.6471059>
23. Liu S, Mashayekh S, Kundur D, Zourntos T, Butler-Purry KL (2012) A smart grid vulnerability analysis framework for coordinated variable structure switching attacks. In: IEEE Power and Energy Society general meeting. <https://doi.org/10.1109/PESGM.2012.6344617>
24. Liu S, Chen B, Zourntos T, Kundur D, Butler-Purry K (2014) A coordinated multi-switch attack for cascading failures in smart grid. *IEEE Trans Smart Grid* 5(3):1183–1195
25. Moore BC (1976) On the flexibility offered by state feedback in multivariable systems beyond closed loop eigenvalue assignment. *IEEE Trans Automat Control* 21(5):689–692
26. Nanou SI, Papakonstantinou AG, Papathanassiou SA (2015) A generic model of two-stage grid-connected PV systems with primary frequency response and inertia emulation. *Electr Power Syst Res* 127:186–196
27. Palermo, J (2016) International review of frequency control adaptation. DGA Consulting, Hunters Hill. Available via the Australian Energy Market Operator. https://www.aemo.com.au/-/media/Files/Electricity/NEM/Security_and_Reliability/Reports/FPSS-International-Review-of-Frequency-Control.pdf. Cited 22 Mar 2017
28. Pasqualetti F, Dörfler F, Bullo F (2012) Cyber-physical security via geometric control: distributed monitoring and malicious attacks. In: IEEE conference on decision and control. <https://doi.org/10.1109/CDC.2012.6426257>
29. Pasqualetti F, Dörfler F, Bullo F (2013) Attack detection and identification in cyber-physical systems. *IEEE Trans Automat Control* 58(11):2715–2729
30. Pasqualetti F, Dörfler F, Bullo F (2015) Control-theoretic methods for cyberphysical security: geometric principles for optimal cross-layer resilient control systems. *IEEE Control Syst Mag* <https://doi.org/10.1109/MCS.2014.2364725>
31. Phadke AG (1993) Real-time phasor measurement for improved monitoring and control. Electric Power Research Institute, Palo Alto. Available via EPRI. <http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=TR-103640>. Cited 28 Nov 2016
32. Phadke AG, Thorp JS, Adamiak MG (1983) A new measurement technique for tracking voltage phasors, local system frequency, and rate of change of frequency. *IEEE Trans Power Apparatus Syst PAS-102(5)*:1025–1038
33. Rate of Change of Frequency (ROCOF): Review of TSO and generator submissions Final Report (2013). PPA Energy, TNEI Services Ltd, Guildford. Available via the Commission for Energy Regulation. <http://www.cer.ie/docs/000260/cer13143-%28a%29-ppa-tnei-rocof-final-report.pdf>. Cited 22 Feb 2016
34. Sauer PW, Pai MA (1997) Power system dynamics and stability. Stipes Publishing, LLC, Champaign
35. Staged Cyber Attack Reveals Vulnerability in Power Grid (2007). Available via YouTube. <https://www.youtube.com/watch?v=fJyWngDco3g>. Cited 3 May 2015
36. Standard PRC-024-1: Generator Performance During Frequency and Voltage Excursions, Standard Development Roadmap (2012). North American Electric Reliability Corporation, Atlanta. Available via NERC. http://www.nerc.com/pa/Stand/Project%20200709%20%20Generator%20Verification%20%20PRC0241/PRC-024-1_redline_to_initial_ballot_2012Feb22.pdf. Cited 29 Feb 2016
37. Standard PRC-024-1: Generator Frequency and Voltage Protective Relay Settings (2014). North American Electric Reliability Corporation, Atlanta. Available via NERC. www.nerc.com. Cited 29 Feb 2016
38. Terrorism and the Electric Power Delivery System (2012). National Research Council of the National Academies, Washington. <https://doi.org/10.17226/12050>

39. van Wesenbeeck MPN, de Haan SWH, Varela P, Visscher K (2009) Grid tied converter with virtual kinetic storage. IEEE PowerTech. <https://doi.org/10.1109/PTC.2009.5282048>
40. Vaya MG, Andersson G (2013) Combined smart-charging and frequency regulation for fleets of plug-in electric vehicles. In: IEEE Power and Energy Society general meeting. <https://doi.org/10.1109/PESMG.2013.6672852>
41. Zhu J, Booth CD, Adam GP, Roscoe AJ, Bright CG (2013) Inertia emulation control strategy for VSC-HVDC transmission systems. Trans Power Syst 28(2):1277–1287