Power System Modeling Under Uncertainty With Controllable Demand

by

Kári Hreinsson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2020 by the
Graduate Supervisory Committee:

Anna Scaglione, Chair
Kory Hedman
Junshan Zhang
Mahnoosh Alizadeh

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

With demand for increased efficiency and smaller carbon footprint, power system operators are striving to improve their modeling, down to the individual consumer device, paving the way for higher production and consumption efficiencies and increased renewable generation without sacrificing system reliability. This dissertation explores two lines of research. The first part looks at stochastic continuous-time power system scheduling, where the goal is to better capture system ramping characteristics to address increased variability and uncertainty. The second part of the dissertation starts by developing aggregate population models for residential Demand Response (DR), focusing on storage devices, Electric Vehicles (EVs), Deferrable Appliances (DAs) and Thermostatically Controlled Loads (TCLs). Further, the characteristics of such a population aggregate are explored, such as the resemblance to energy storage devices, and particular attentions is given to how such aggregate models can be considered approximately convex even if the individual resource model is not. Armed with an approximately convex aggregate model for DR, how to interface it with present day energy markets is explored, looking at directions the market could go towards to better accommodate such devices for the benefit of not only the prosumer itself but the system as a whole.

Til Ingu Láru og Katrínar Lóu

ACKNOWLEDGEMENTS

First, I am ever grateful to my advisor, Prof. Anna Scaglione, for the guidance and countless hours throughout these years. With her seemingly infinite knowledge of all things surrounding this research field, she would always point me in interesting directions and suggest novel solutions to seemingly impossible problems. I would like to thank my committee members, Prof. Mahnoosh Alizadeh, Prof. Kory Hedman and Prof. Junshan Zhang for all their feedback throughout the years, and in particular Prof. Hedman for all his classes, insights and practical considerations when it comes to power system research.

I would like to thank my SINE Lab colleagues, not only for all their help and advice, but also for all the lunch breaks, dinners, hikes and activities. I will always think fondly back to the times we spent together, and hope our paths will cross frequently in the future! I am ever grateful for getting the opportunity to study in Arizona, a place as opposite to Iceland as one can imagine, but amazing in so many surprising ways.

I would like to thank my parents, sisters, in-laws and friends for all the encouragement through the years. In particular I would like to thank my older sister who having both obtained a PhD in the USA and lived in Arizona, convinced me that this was not such a bad idea. Finally, I would not be here without my wife Inga Lára, who managed to keep up her morale through a long-distance relationship for all these years. I am looking forward to our future, and all the adventures it will entail.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ALGORITHMS

GLOSSARY

CAISO California ISO.

CLT Central Limit Theorem.

CT-MSUC Continuous-Time Multi-Stage Unit Commitment.

CTUC Continuous-Time Unit Commitment.

DA Deferrable Appliance.

DR Demand Response.

ED Economic Dispatch.

EV Electric Vehicle.

FERC Federal Energy Regulatory Commission.

IEEE Institute of Electrical and Electronics Engineers.

ISO Independent System Operator.

LED Light Emitting Diode.

MILP Mixed-Integer Linear Program.

MPC Model Predictive Control.

MSUC Multi-Stage Unit Commitment.

NOAA National Oceanic & Atmospheric Administration.

OPF Optimal Power Flow.

PCA Principle Component Analysis.

PH Progressive Hedging.

PTDF Power Transmission Distribution Factor.

PWL Piece-Wise Linear.

RMSE Root Mean Squared Error.

RTS IEEE Reliability Test System.

SCED Security-Constrained Economic Dispatch.

SCUC Security Constrained Unit Commitment.

SDDiP Stochastic Dual Dynamic Integer Programming.

SDDP Stochastic Dual Dynamic Programming.

SF Shapley-Folkman.

SUC Stochastic Unit Commitment.

SVD Singular Value Decomposition.

TCL Thermostatically Controlled Load.

TSO Transmission System Operator.

UC Unit Commitment.

CHAPTER 1

INTRODUCTION

Under the threat of global warming, the energy sector must move to decrease emissions through reduced consumption of fossil fuels. In this vein, many avenues are being pursued, the most prominent being increased efficiency and the growing share of renewable (emission free) sources of energy, of which solar and wind energy are growing rapidly. In this context, efficiency does not only mean a more efficient way to consume energy (such as switching out an incandescent light bulb for Light Emitting Diode (LED)), but also a more efficient way to generate energy, which is primarily achieved through improved modeling of generation and transmission to decrease losses as well as operating generators closer to their peak efficiency and turning-off unnecessary high emission generation. Conventional power plants, such thermal plants based on gas, coal or nuclear fission, along with reservoir-based hydro, tend to have a stable supply and/or a large storage of fuel (or water) nearby, meaning that in the short term (days or weeks) plant operators can increase or decrease their output power as they see fit. Here, renewable generation, primarily solar and wind but also run-of-river hydro, are fundamentally different in that the source of energy is exogenous to the control of the operator in its generation, it can only be curtailed. A related issue seen by some power system operators with high penetration of solar power is the significant ramps required from conventional generation in the morning and evening hours, often leading to negative prices of electricity and ramping shortages. All of this highlights that a much larger portfolio for storage and fast ramping resources is needed, well beyond the hydro-power resources that are used today. In the future this portfolio is likely to include a significant amount of demand response programs and distributed storage. Modeling these resources and capturing their value in balancing the grid is of paramount importance for their future integration.

The responsibility of managing the overall power system lies with the Independent System Operator (ISO) (or Transmission System Operator (TSO), but ISO will be used throughout this dissertation) that centrally coordinates generation (and to some extent consumption) through energy and ancillary service markets using various power system models such as Unit Commitment (UC) or Economic Dispatch (ED), with the primary objective of maintaining a reliable and economic system. Power system models such as UC/ED approximate the true physical characteristics and constraints of the power system, in order to make the models solvable in a matter of seconds or minutes, even for large scale systems.

This dissertation explores two threads of addressing the aforementioned issues, introduced in the following sections.

1.1 Handling Variability and Uncertainty

Presently, uncertainty in power systems is managed by scheduling reserve capacity in advance to compensate for errors in (net-)load forecasts. The vast literature dealing with Stochastic Unit Commitment (SUC) and of the Security Constrained Unit Commitment (SCUC) problems suggests an alternative approach that captures the exogenous uncertainties directly into the decision process. In fact, SUC ensures that a feasible solution exists for all considered scenarios, and that the expected associated costs are minimized. The most common SUC formulations are two-stage SUC problems, pioneered by [Wiebking(1977)] (see e.g. [Zheng *et al.*(2015)] for an overview of the considerable literature in this area).

Compared to the two-stage SUC, the *Multi-Stage* Stochastic Unit Commitment (MSUC) takes a sequence of decisions, helping to achieve smoother boundary conditions on the commitment variables, with the downside that the complexity is often prohibitive. First in [Takriti *et al.*(1996)], and then in a series of follow up work (see e.g. [Carpentier *et al.*(1996), Nowak and Römisch(2000), Shiina and Birge(2004), Papavasiliou *et al.*(2011), Analui and Scaglione(2017)] and the references therein), many authors worked on curbing the Multi-

Stage Unit Commitment (MSUC) computational complexity. It is natural to use the more accurate representation of the uncertainties in SUC to optimally oversee a more economic commitment of reserves. While the literature above shows that improved handling of uncertainty leads to greater reliability, it ignores the fact that load and generation mostly changes continually and smoothly, and instead models it through step-wise discrete-time functions. The coarse representation of inter-hourly ramping events also affects the system reliability and urges for the perusal of a continuous-time approach.

In this dissertation, more specifically Chapter 2 and published papers [Hreinsson *et al.*(2018), Hreinsson *et al.*(2019)], leveraging existing work [Parvania and Scaglione(2016)] focused on the deterministic Continuous-Time Unit Commitment (CTUC), the Continuous-Time Multi-Stage Unit Commitment (CT-MSUC) formulation is introduced, in which an underlying load scenario tree is used to decide the baseline day-ahead dispatch, commitment and reserve capacity, considering continuous-time generation trajectories as part of the decision variables. In a nutshell, the continuous-time representation increases the number of variables per branch, capturing the trajectory as a polynomial spline. This change allows to schedule the reserve capacity and power, accounting for the future expected real-time cost in dispatching them, providing a more accurate representation of the future inter-hourly ramping needs and capturing large scale-storage bids as well. Compared to [Parvania and Scaglione(2016)], both the diurnal variability as well as the stochastic nature of the net-load are captured and the simulations show that this does help scheduling the right set of units to perform reliably in real time. In a separate line of research, a number studies have been carried out that address sub-hourly scheduling problem as a solution to increase system flexibility and more efficient reserve allocation in presence of net-load variability. In [Lopez *et al.*(2018)] the authors have assessed the conditions that drive sub-hourly scheduling in UC models. In addition, [Deane *et al.*(2014)] examined modeling of power system with significant levels of wind generation at varying temporal resolutions and captured the asso-

ciated costs that are not accounted for in hourly models. Evidently, there are trade-offs such as problem size, data procurement and computational times in higher resolution scheduling. In past work [Parvania and Scaglione(2016)], preserving the same number of continuous variables as the proposed methods via third order splines, the authors already showed that the sub-hourly UC performed better than the hourly UC but still worse than the continuous-time formulation. Another tangential line of follow up work leveraging splines in operational decisions are [Parvania and Khatami(2017), Scaglione(2016)] which address continuous time pricing for deterministic economic dispatch (without commitment decisions) problems.

The proposed method further improves on existing literature with the following extensions:

(a) With the inclusion of models and bids for energy storage,

(b) By allowing quadratic cost curves for ramping, power and energy,

(c) By modeling a multi-bus system with continuous-time flows across transmission lines, and

(d) With a more generic formulation not only limited to third order polynomials.

The inclusion of storage in conventional UC models has been explored in e.g. [Khatami *et al.*(2017), Lorca and Sun(2017), Bakirtzis *et al.*(2018), Taylor(2015)] and the references within, but none combines the continuous time and multi-stage stochastic formulations presented here. Storage devices are modeled by constraining the range of the integral of their power, as such they are approximated to have zero losses, and dispatch and cost curves are modeled in a continuous range from negative to positive power dispatch.

1.2 Demand Response Modeling

The consumption side of power systems has traditionally been considered inelastic or incontrollable, and to maintain a balance between production and consumption, ISO have instead focused on controlling generation. With significant improvements in communication infrastructure, computational power and the electrification of transportation, there are now large residential and commercial loads that can communicate cheaply in real-time, opening up new opportunities.

Controllable DR is often composed of large numbers of small participants whose behavior is described through non-convex models, making it intractable for direct inclusion in ISO models (e.g. UC/ED). It is broadly accepted that separate entities, called *aggregators*, would provide an intermediary interface. As there are no practical bidding formats to include the true description of load flexibility in the market, aggregators often present themselves as price-sensitive demand or virtual generators, inventing offer/bid curves in traditional market formats to capture their flexibility as close as possible.

1.2.1 Exploring Demand Response Flexibility

Focusing on direct load control, the integration of small individual loads or an aggregates of such loads into present days scheduling and market models is not clear. There are little obstacles to integrating loads that can be translated into conventional energy bids/offers, possibly with constraints resembling a generator. However, most loads that fall under the residential/commercial DR classification are unlike generators or industrial loads, in that they consume more or less a constant amount of energy, but are flexible in the precise consumption pattern. Several papers formulate optimization programs for system operators or aggregators [Parvania *et al.*(2014)], some consider individual EV constraints (see e.g. [Sortomme and El-Sharkawi(2012), Sanchez-Martin *et al.*(2012), Sojoudi and Low(2011),

Yao *et al.*(2013)]), while others focus solely on an aggregate of TCLs [Callaway(2009), Koch *et al.*(2011), Hao *et al.*(2013), Alizadeh and Scaglione(2013), Kalsi *et al.*(2012), Mathieu *et al.*(2013b), Ramanathan and Vittal(2008)].

In [Subramanian *et al.*(2012), Nayyar *et al.*(2013)] the authors consider the scheduling of an aggregate of simple flexible loads, the former explores various scheduling strategies while the latter looks at how such an aggregate can be considered as storage offering flexibility to absorb stochastic variations of renewable power and reduce need of or offer reserves. In [Barot and Taylor(2014)], [Trangbæk *et al.*(2011)] the authors take a more general approach and explore how the feasible region for an aggregate load model can be described by convex polytopes, by computing the Minkowski sums of their individual feasible regions. Specifically, [Barot and Taylor(2014)] finds an outer approximation of the feasible region while [Trangbæk *et al.*(2011)] calculates the exact sum for simple loads. These approaches are strictly limited to devices described by convex constraints, with the different options presenting a tradeoff between computational complexity and the precision of the represented feasible region. Chapters 3 and published paper [Hreinsson *et al.*(2016)] explore the approximation to the Minkowski sums approach in [Foster and Caramanis(2013)] obtained by clustering EVs and DAs [Alizadeh *et al.*(2015)]. The goal is to aid the interpretation of this mathematical constructs in the context of energy reserves. The novel contribution lies in the characterization of the equivalent aggregate DR resource in terms of:

(1) potential load curtailment, stored negative energy and,

(2) flexibility and in the formulation of a statistical model for the aggregate resource that allows to predict future equivalent storage potential within a certain confidence level.

Monetary objectives are ignored, such as cost of switching, utility or inconvenience, leaving that for future work.

1.2.2 Thermostatically Controlled Load Modeling

TCLs are among the most promising candidate appliances for DR programs [Callaway and Hiskens(2011)], as they can offer considerable flexibility and have virtually no ramping limits unlike conventional spinning generation. The first simplified dynamical model for a TCL population was introduced in [Chong and Malhamé(1984)]. Its aim was to capture the rebound peak observed after TCLs were interrupted during an emergency. While some of the recent work is concerned with the response of TCLs to real-time pricing (e.g. [Lu and Chassin(2004), Zhong *et al.*(2013), Zhao *et al.*(2013), Yoon *et al.*(2014), Behboodi *et al.*(2018)]), here the focus is on direct load control models that are decoupled from the economic signals that entice consumer participation. Instead of considering load curtailment, most of the recent work on DR of TCLs is based on changing thermostat set-points in order to adjust the load profile [Kundu *et al.*(2011), Callaway(2009), Bashash and Fathy(2011), Mathieu *et al.*(2012), Mahdavi *et al.*(2017)] and realize a certain collective response for load following, regulation and frequency control. Most population dynamics models in the literature are reminiscent of the model in [Chong and Malhamé(1984)]. In such models, TCLs are assumed to have a certain dead-band; the control consists of switching fractions of the TCLs population between the ON and OFF states prematurely, relative to the time they hit the corresponding temperature threshold, in order to create the intended deviation from the otherwise-uncontrolled load profile. Work refining this basic idea is in e.g. [Callaway(2009), Bashash and Fathy(2011), Koch *et al.*(2011), Alizadeh *et al.*(2015), Zhang *et al.*(2013)]. Another relevant line of work has used battery models to capture the TCL population. Variants of this idea have been equivalent thermal battery models that are controlled to follow the least costly control trajectory [Mathieu *et al.*(2013a)], and generalized battery model [Hao *et al.*(2015)] with lower and upper bounds for electric power consumption. For the latter, the control is achieved by aggregating and looping through individual TCLs in a priority-

stack, switching units *early* until the desired load profile is obtained. Similar reference-signal control techniques are described by [Meyn *et al.*(2013), Totu *et al.*(2016), Tindemans *et al.*(2015)]. The authors of [Elghitani and Zhuang(2017)] consider generic DR *blocks* of certain power, duration and slack, prioritizing and optimizing the service of blocks to minimize costs. In [Ming *et al.*(2017)] the authors describe a robust economic dispatch market clearing process, subject to a generic DR uncertainty bounded by a certain confidence through the scenario approach [Campi *et al.*(2009)].

The literature cited above assumes ambient temperatures to be constant, however, knowing that temperature can both change quickly and randomly, a new method proposed in Chapter 3 and associated papers [Hreinsson and Scaglione(2017), Hreinsson *et al.*(2020a)] attempt to capture the relationship between temperature and the range of load attainable by controlling a large number of TCLs. Most of the literature on TCL control focuses on scalable control solutions, but because of the constant ambient temperature limitation, does not offer a direct method to include TCL-DR in (stochastic) power system planning models such as ED. Only recently work has emerged that allows for variations in temperature [Mathieu *et al.*(2015), Mahdavi *et al.*(2016), Mahdavi *et al.*(2017), Vrakopoulou *et al.*(2017), Li *et al.*(2017)]. The authors of [Mahdavi *et al.*(2016)] introduce a model that tracks electric load of TCLs subject to changing ambient temperature, and later in [Mahdavi *et al.*(2017)] apply that model to implement a Model Predictive Control (MPC) model that reacts to a reference load signal by changing thermostat reference temperatures. A recent contribution in a similar vein to the formulation in Chapter 3 is [Vrakopoulou *et al.*(2017), Li *et al.*(2017)], where the authors use a time-varying equivalent battery model for an aggregate of space heaters from [Mathieu *et al.*(2015)], and solve a robust Optimal Power Flow (OPF) given the uncertainty of ambient temperature and of transmission-level wind infeed. Uncertainty is managed by ensuring feasibility against a certain number of scenarios [Vrakopoulou *et al.*(2017)] or by making the assumption that the uncertainty can be

described by a jointly-Gaussian distribution [Li *et al.*(2017)]. In contrast to [Vrakopoulou *et al.*(2017), Li *et al.*(2017)], the proposed method does not formulate a robust problem, but a stochastic optimization, where the joint uncertainty of net-load and temperature is captured through a scenario tree, and unlike the thermal battery model of [Mathieu *et al.*(2015)], it is based around a state-space population model which allows to more accurately capture the complex inter-temporal relationships between temperature and constraints on power and energy. The novelty of the proposed work can be summarized as follows:

(a) It derives equivalent operational decisions/constraints of DR aggregates directly as a function of varying ambient temperature,

(b) it shows how to utilize this mapping in a decision problem that incorporates temperature forecasts or temperature scenario trees, capturing the uncertainty on the TCL model.

Numerical results showcase the accuracy of the proposed representation and the benefits of using this optimization framework for DR.

1.2.3 Non-convex Loads and the Shapley-Folkman Lemma

Chapter 4 (and paper submission [Hreinsson *et al.*(2020b)]) looks at non-convex loads, how the Shapley-Folkman (SF) lemma can be mobalized to achieve a deeper understanding on how DR aggregators can come closer to presenting their actual constraints and costs to the ISO.

Vast research has been devoted to the interface between aggregators and conventional energy markets, trying to maximize the aggregators profit (see e.g. [Parvania *et al.*(2013), Di Somma *et al.*(2018), Kohansal and Mohsenian-Rad(2015), Chen *et al.*(2016a), Henríquez *et al.*(2017), Kowli and Meyn(2011), Samadi *et al.*(2015), Ottesen *et al.*(2016)]), some with focus on specific DR resources [Ruiz *et al.*(2009), Lu(2012), Mathieu *et al.*(2013c), Coffman *et al.*(2019), Contreras-Ocana *et al.*(2017)], or on their synergy with renewables [Call-

away(2009), Ortega-Vazquez *et al.*(2013), Subramanian *et al.*(2013)], while others discuss more generic bidding strategies [Li *et al.*(2011a), Pandžić *et al.*(2013)]. Several authors proposed an interface between aggregators and individual DR resources based on distributed pricing [Chen *et al.*(2010), Li *et al.*(2011b), Li *et al.*(2015b), Li *et al.*(2015a), Jiang and Low(2011), Chang *et al.*(2013)], usually employing distributed primal dual decompositions. A similar approach is employed between aggregators and ISOs in [Gatsis and Giannakis(2013)], describing an iterative distributed economic dispatch formulation. Closer to this dissertation, is the body of research on low order models of DR aggregates, i.e. how to package DR aggregates to be handed off to the ISO, where [Ruiz *et al.*(2009)] considers this as a virtual generator, while [Hao *et al.*(2014), Nayyar *et al.*(2013)] consider those as virtual storage devices, or a polytope in a joint decision and load space, e.g. [Alizadeh *et al.*(2014a)]. In [Barot and Taylor(2017)] and earlier work, the authors aggregate a collection of individuals by finding the outer approximation of the Minkowski sum from the individual constraints, while [Zhao *et al.*(2017)] uses a geometric approach to find the largest inner and smallest outer approximations of the actual feasible load set as a homothetic transformation of a prototype polytope, allowing for efficient bounding of the Minkowski sum of individual resources. Later in [Nazir *et al.*(2018)] an algorithm is proposed that improves on [Zhao *et al.*(2017)] by decomposing the feasible set into sub-sets before fitting the largest homothet inside the sub-sets. In [Müller *et al.*(2017)] and derived work the authors describe aggregation and disaggregation of flexible resources based on zonotopes and utilize those to maximize profits for an aggregator that is a price-taker.

This dissertation revisits the problem of designing efficient DR market models in light of the SF lemma. Specifically, it starts from individual load models that can be non-convex in both cost and constraints, and shows how aggregates of such models can be considered approximately convex both in the set of feasible aggregate power profiles and total cost, and strictly convex under certain conditions. This relationship between the non-convex

individual DR models and corresponding aggregate convex load models and its implications has not previously been explored in the literature. This dissertation expands on the theory and provides sufficient conditions for exact convexity and guidelines on how to come close to strict convexity.

With the understanding that the aggregate models are approximately convex, the problem of delivering these aggregate models to the ISO remains, as the they may be quite complex even though they are convex, both in terms of constraints and cost. To this end, this dissertation attempts to close the loop from the individual to the ISO by exploring several approaches to construct low-order aggregate bids that both describe the capabilities of the population in terms of aggregate power consumption but also aggregate cost, with the aggregate cost derivation not previously found in the literature. Similar to some of the existing literature, a polytopic inner approximation is suggested to describe the aggregate power capabilities, but unlike existing methods, the proposed approach has a configurable complexity, even when aggregating large numbers of dissimilar resources, and is not confined to a particular shape of polytopes as in [Zhao *et al.*(2017), Müller *et al.*(2017)] which may lead to considerable losses when aggregating diverse resources. Inspired by the Central Limit Theorem (CLT) from probability theory, this dissertation further suggest a novel ellipsoidal model for describing DR aggregates that are derived in a distributed fashion, and extend this notion to suggest what the true aggregate cost function of DR aggregates is.

1.3 Organization

The remainder of the dissertation is split into three main chapters.

Chapter 2 proposes the Continuous-Time Multi-Stage Unit Commitment (CT-MSUC) and its implications for system operators.

Chapter 3 shifts to Demand Response (DR) modeling, and introduces aggregate models for storage, EVs DAs and TCLs based on convex or convexified individual resources. It explores several metrics of flexibility for such aggregates and looks at integration into an Economic Dispatch (ED) formulation before showing some numerical results.

Chapter 4 starts from a generic flexible load description and revisits the individual load models to consider non-convex ignored in Chapter 3. It considers how an aggregate of such loads can be considered approximately convex and how to pass such models on to the ISO market models as a computationally simple yet flexible model, considering the economic aspects of pricing and costs.

1.4 Notation

Slanted lower case variables (x) denote scalars, bold lower or upper case variables (i.e. $\boldsymbol{x}$ or $\boldsymbol{X}$) indicate vectors and matrices (or tensors). Boldface calligraphic letters ($\boldsymbol{\mathcal{A}}$) represent sets and $|\mathcal{A}|$ their cardinality. $\boldsymbol{A} \leq \mathbf{1}a$, $\boldsymbol{A} \geq \mathbf{1}a$ or $\boldsymbol{A} = \boldsymbol{a}$ with vectors on both sides are element-wise operations with $\mathbf{1}$ denoting a vector of ones of the appropriate size. The notation $\boldsymbol{A}_{\boldsymbol{v}}$ where $\boldsymbol{A}$ is a tensor and $\boldsymbol{v}$ a set of indexes, is the element of $\boldsymbol{A}$ corresponding to the tuple of indexes $\boldsymbol{v}$, examples are $\boldsymbol{x}_i$ or $\boldsymbol{X}_{i,j}$ which refer to the corresponding elements. The notation vec$(\boldsymbol{A})$ is an abbreviation for the vectorize operator that stacks all the entries of tensor $\boldsymbol{A}$ into a vector. Transpose is denoted by $\top$ while $\boldsymbol{A}^T$ is simply $\boldsymbol{A}$ to the power of T.

CHAPTER 2

MULTI-STAGE STOCHASTIC CONTINUOUS-TIME UNIT COMMITMENT

2.1 Continuous Time Modeling

This section lays out the most significant modeling changes between a conventional UC formulation and the proposed stochastic polynomial continuous-time representation of load and generation. Some aspects are similar to existing literature [Parvania and Scaglione(2016)], but are also summarized here to make the dissertation self-contained.

2.1.1 Polynomial Representation Synopsis

Polynomial splines (piece-wise polynomials) are chosen as a way to describe continuous load and generation trajectories with a relatively small number of coefficients. Figure 2.1.1 (c) shows a continuous (cubic spline) load trajectory, and contrasts it with how load is conventionally modeled as a discontinuous zero order piecewise polynomial in Figure 2.1.1 (a). Rather than approximating the trajectories with a constant value over each interval, in the proposed model, the trajectories are linear combinations of a set of polynomials. These polynomials are *vectors* in the Hilbert space of functions and the coefficients are the coordinates of the analog function with respect to the basis. A collection of polynomials of degree n can form a basis that spans a vector space of dimension at most $n+1$ and all signals in the sub-space have $n+1$ coordinates (i.e. the coefficients that multiply the polynomials in the linear combination). Because of Weierstrass approximation theorem, one can approximate any analog trajectory on a finite interval onto this subspace with error that vanishes as the order n goes to infinity. In this chapter, extensive use is made of the Bernstein polynomials basis functions, which for degree n are:

$$\mathrm{b}_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i},\ t \in [0,1], i \in [0,n] \tag{2.1.1}$$

Further, defining a vector of these basis functions $\mathbf{b}_n(t) = (\mathrm{b}_{0,n}(t), \ldots, \mathrm{b}_{n,n}(t))^\intercal$; any n-th order polynomial function in $t \in [0,1]$ can be expressed as

$$\mathrm{x}(t) = \sum_{i=0}^{n} x^{(i)} \mathrm{b}_{i,n}(t) \equiv \boldsymbol{x}^\intercal \mathbf{b}_n(t) \tag{2.1.2}$$

where $\boldsymbol{x} = [x^{(0)}\ x^{(1)} \ldots\ x^{(n)}]^\intercal \in \mathbb{R}^{n+1}$ is the corresponding transposed vector of coefficients.

Bernstein polynomials have several useful properties that facilitate the representation of derivatives and integrals in one common basis and allow to bound the analog functions in the space. Firstly, due to the linearity of the operation of derivative and integral of a function:

$$\dot{\mathrm{x}}(t) = \sum_{i=0}^{n} x^{(i)} \dot{\mathrm{b}}_{i,n}(t), \quad \int_a^t \mathrm{x}(u)du = \sum_{i=0}^{n} x^{(i)} \int_a^t \mathrm{b}_{i,n}(u)du \tag{2.1.3}$$

Thus, the derivative of a basis function can be written as the finite difference of lower order basis functions,

$$\dot{\mathrm{b}}_{i,n}(t) = n(\mathrm{b}_{i-1,n-1}(t) - \mathrm{b}_{i,n-1}(t)),$$

and $\boldsymbol{x}$, the coefficients of $\mathrm{x}(t)$, and those of $\dot{\mathrm{x}}(t)$, denoted by $\dot{\boldsymbol{x}}$ are interchangeable:

$$\dot{\boldsymbol{x}} = \boldsymbol{M}_n \boldsymbol{x} \in \mathbb{R}^n \tag{2.1.4}$$

where $\boldsymbol{M}_n \in \{-n, 0, n\}^{n \times n+1}$ is a bi-diagonal matrix that corresponds to the change of coordinates from $\dot{\mathrm{b}}_{i,n}(t)$ to $\mathrm{b}_{i,n}(t)$. Both derivatives and integrals of the basis are polynomials of degrees $n-1$ and $n+1$ respectively. Secondly, as was noted in [Parvania and Scaglione(2016)], the so called *convex hull property* implies that $\min_i x^{(i)} \leq \boldsymbol{x}^\intercal \mathbf{b}_n(t) \leq \max_i x^{(i)}$ for $t \in [0,1]$ (see also Figure 2.2.1). Thirdly, the function $\mathrm{x}(t) = \boldsymbol{x}^\intercal \mathbf{b}_n(t)$ passes through $(0, x^{(0)})$ and $(1, x^{(n)})$, at the edges of the interval $0 \leq t \leq 1$. This is particularly useful when defining continuous splines; to enforce C^0 continuity across splines covering the kth and $k+1$th intervals, whose coefficient vectors are $\boldsymbol{x}_k$ and $\boldsymbol{x}_{k+1}$ respectively, by simply enforcing the equality $x_k^{(n)} = x_{k+1}^{(0)}$. Denote by $[\boldsymbol{x}]^{(i)}$ the ith entry of the vector $\boldsymbol{x}$ in the brackets, to enforce C^1 continuity, it suffices to enforce $[\boldsymbol{M}_n \boldsymbol{x}_k]^{(n)} = [\boldsymbol{M}_n \boldsymbol{x}_{k+1}]^{(0)}$, and by adding further mappings $\boldsymbol{M}$ one can enforce continuity of higher order derivatives.

Formally, having a piece-wise function x(t) with t $\in [0, Kh]$, divided into n intervals of unit length ($h = 1$), defining each segment $\mathrm{x}_k(t)$ on $t \in [0, 1]$ and denoting by rect(t) the rectangular pulse with unit length, the relationship between the function and the segments is:

$$\mathrm{x(t)} = \sum_{k=0}^{K-1} \mathrm{x}_k(\mathrm{t} - k)\mathrm{rect}(\mathrm{t} - k).\ \ \mathrm{t} \in [0, K] \tag{2.1.5}$$

Hence, polynomial expansion can be used for each segment, in lieu of a single constant x_k used in discrete time, i.e. $\mathrm{x}_k(t) = \boldsymbol{x}_k^{\intercal}\mathbf{b}_n(t)$ with the Bernstein basis of order n. The above properties will be leveraged in both spline representations of net-load trajectories and further in net-load scenario tree generation.

2.1.2 Stochastic Formulation

As discussed in the introduction, modern electricity portfolio management models are best represented by multi-stage stochastic programs. The input of such programs consists of a *scenario tree* that models the probabilistic information of the underlying uncertainty (net-load, spot prices, etc.). The scenario tree is a directed graph and shows the stage-wise evolution of information structure (*filtration*). The tree structure clusters the realizations of underlying stochastic process into a set of branches with associated probabilities of occurrence. A key feature in multi-stage stochastic programs is known in the literature as *non-anticipativity*, meaning that the actions or optimal decisions are taken only based on the information up to the present time, independent of the future. Scenario trees are often constructed to meet a minimum approximation error relative to some distance metric. There is a fair amount of literature on how to generate a scenario tree which optimally and accurately represents the probability model of the underlying stochastic process in discrete time (see e.g. [Pflug and Pichler(2015)]). Here, a straightforward scenario reduction algorithm is adopted, with additional continuity constraints to derive the desired continuous-time net-load scenario tree as the baseline model for proposed CT-MSUC. Algorithm 1 describes the proposed

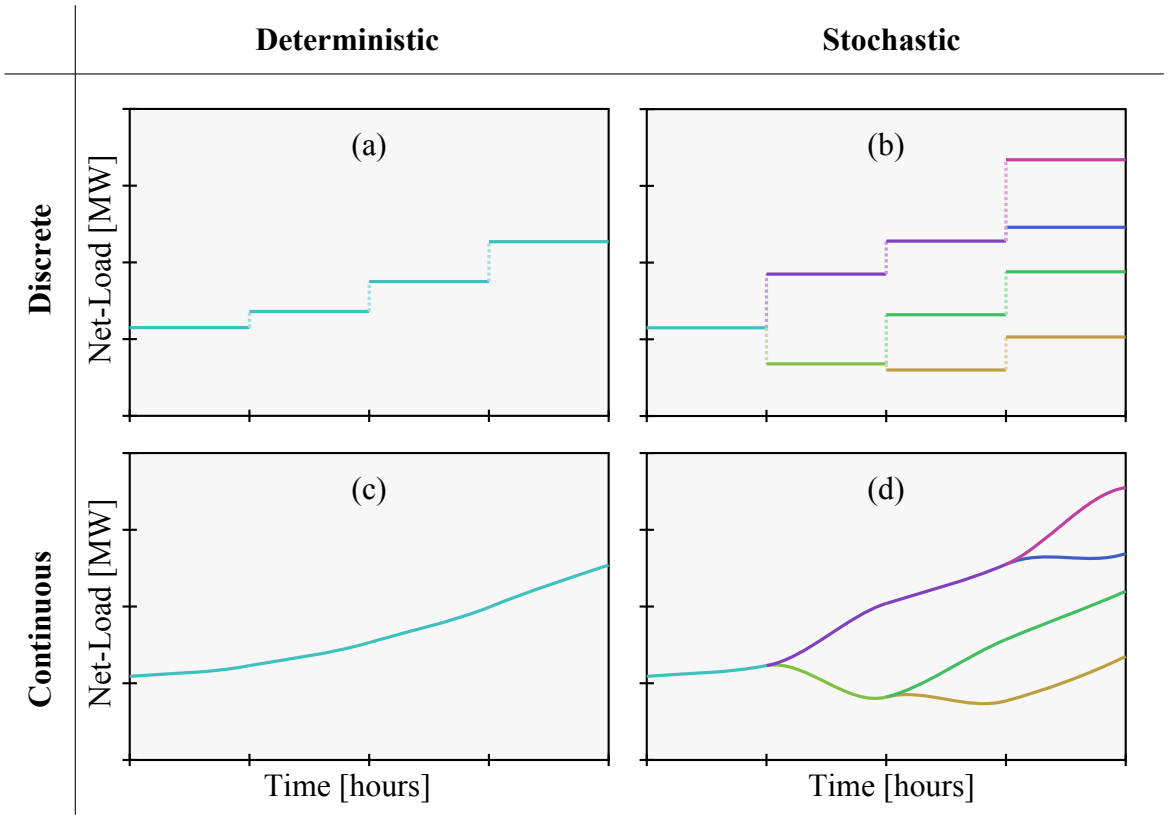


Figure 2.1.1: Highlight are the different ways of estimating load in UC; (a) conventional load as deterministic discrete-time, (b) a stochastic discrete-time load, (c) a continuous-time deterministic load and (d) a stochastic continuous-time load.

approach using k-means clustering to reduce a set of S input scenarios $\Xi \in \mathbb{R}^{S\times K\times|\mathcal{B}|\times(n+1)}$, where it is assumed that the original net-load trajectories are mapped to their nearest $n+1$ Bernstein coefficients, into a desired tree structure with fixed the number of nodes per time step through vector $\boldsymbol{c}$.

2.1.2.1 Net-Load Scenario Tree Construction

As mentioned above, existing scenario construction algorithms do not represent the continuous-time nature of load/demand stochastic process and therefore the corresponding generation decisions. By employing the spline representations of the net-load stochastic process, this aspect can be represented up to the desired level of precision. In the other words, the pro-

posed solution can numerically approximate the variational solutions asymptotically, when the order of the spline representation grows.

A continuous time scenario tree $\mathcal{T}$ is built, with set of nodes $\mathcal{V}$, where associated with each tree node $v \in \mathcal{V}$ a vector of control points contains the Bernstein coefficients of net-load: $\boldsymbol{\xi}_v^b = [\xi_v^{b(0)}\ \xi_v^{b(1)}\ \dots\ \xi_v^{b(n)}]$. The continuous-time net-load functions at bus $b \in \mathcal{B}$ leading to node $v \in \mathcal{V}$ is thus $\xi_v^b(t) = (\boldsymbol{\xi}_v^b)^\intercal \mathbf{b}_n(t)$ with π_v representing the joint probability of the free parameters in the polynomial trajectory. To navigate the tree, the following notions are used:

- the children of node v are contained in the set $\mathcal{C}(v)$,
- the parent (immediate ancestor) of node v is given by the function $\alpha^1(v)$,
- the grand-parent by $\alpha^2(v)$ and so on, with $\alpha^0(v) = v$,
- the shorthand $v^- = \alpha^1(v)$ for brevity,
- $\boldsymbol{\xi} \sim \mathcal{T}$ refers to the complete construction.
- the nodes pertaining the possible outcomes at a certain hour k are contained in the set $\mathcal{V}(k)$ and the hour a node v corresponds to is denoted by $\tau(v)$.

The proposed scenario tree construction approach is based on a *recursive scenario reduction*. This strategy consists of modifying a given *fan* of trajectories by bundling them according to the *k-means* clustering algorithm. An assumption is made, that the original input trajectories have already been mapped to their nearest Bernstein coefficients. It is evident that constructed trees are much smaller than the given fan of trajectories and nevertheless, they represent a viable approximation with respect to the appropriate norm. More elaborate description on the load data and distribution of the load across the system topology is incorporated in Section 2.4.1. Figure 2.1.1 shows how a scenario tree for net-load may look

like, both in discrete time (b) and continuous time (d). Time is quantized into *one hour* intervals, such that each node reflects actions taken over a single hour, thus simplifying the notation (the extension to arbitrary time intervals is straight-forward, and skipped for brevity).

2.1.2.2 Continuous-time Cost

For a stochastic continuous time formulation, the continuous time expected cost is expressed; that is, given a continuous time dispatch function for a particular interval (the subscript k is omitted for clarity), $\mathrm{x}(t) = \boldsymbol{x}^\intercal \mathbf{b}(t)$, a cost function $\mathrm{C}(\mathrm{x})$ and $t \in [0, 1]$:

$$\int_0^1 \mathbb{E}[\mathrm{C}(\mathrm{x}(t))]\, dt = \int_0^1 \mathbb{E}\left[\mathrm{C}(\boldsymbol{x}^\intercal \mathbf{b}(t))\right]\, dt \tag{2.1.6}$$

For a linear cost function $\mathrm{C}(\mathrm{x}) = c_1\mathrm{x} + c_0$ this becomes:

$$\int_0^1 \mathbb{E}[\mathrm{C}(\mathrm{x}(t))]\, dt = c_1 \int_0^1 \mathbb{E}[\boldsymbol{x}^\intercal]\mathbf{b}(t)\, dt + c_0 = c_1 \frac{\mathbf{1}^\intercal \cdot \mathbb{E}[\boldsymbol{x}]}{n+1} + c_0 \tag{2.1.7}$$

since Bernstein basis functions of the same order have the same definite integral $\frac{1}{n+1}$ over interval $[0, 1]$.

Various power system costs are considered quadratic, with cost of generation often approximated with a function $\mathrm{C}(\mathrm{x}) = c_2\mathrm{x}^2 + c_1\mathrm{x} + c_0$. In practice, for a convex cost function, it is often implemented as a piece-wise linear function $\mathrm{C}(\mathrm{x}) = c_1^i\mathrm{x} + c_0^i$ if $\mathrm{x}_{i-1} < \mathrm{x} \leq \mathrm{x}_i$ (with appropriate boundary conditions), which in a linear optimization program is achieved through a set of constraints $\mathrm{C} \geq c_1^i\mathrm{x} + c_0^i$ for all i. Unfortunately, for piece-wise linear cost functions, it is not possible to express the *continuous cost* in terms of the polynomial coefficients[1]. Instead, considering a generic power profile $\mathrm{x}(t) = \boldsymbol{x}^\intercal \boldsymbol{b}(t)$ expressed in terms of the Bernstein basis, as well as a generic quadratic cost function $C(\mathrm{x}) = c_2\mathrm{x}^2 + c_1\mathrm{x} + c_0$

[1]This would require splitting the integral on crossings between function pieces at instants in time that are not linearly determined from the coefficients, and integrating arbitrary ranges of the basis functions.

Algorithm 1 Constructing scenario trees in continuous time from empirical net-load trajectories

$K \leftarrow 24$ *(time horizon length)*

$\boldsymbol{c} \leftarrow \mathbb{R}^K$ *(number of nodes per stage)*

$m \leftarrow 0$ *(index of root node)*

$\boldsymbol{\xi}_m$ *(root node value)*

$\boldsymbol{s}_m$ *(indices of all input scenarios)*

ϵ_m *(Root Mean Squared Error (RMSE) between root node value and the set of input scenario)*

$\mathcal{V} = \{0\}$ *(initialize the node set)*

for $k \in \{1, 2, \ldots, K\}$ **do** *(loop over all time/stages)*

  for $v \in \mathcal{V}[k-1]$ **do** *(loop through all nodes of past stage)*

    $m \leftarrow m + 1$ *(increment node counter)*

    $\boldsymbol{\xi}_m, \epsilon_m \leftarrow k$-means$(\boldsymbol{\Xi}[\boldsymbol{s}_v], 1)$ *(apply kmeans to obtain a single centroid – function returns the centroid and corresponding error)*

    $\mathcal{V} \leftarrow \mathcal{V} \cup \{m\}$ *(add m to the set $\mathcal{V}$)*

  end for

  (now the algorithm has moved forward one stage with no splitting)

  while $|\mathcal{V}[k]| < c_k$ *(cardinality of nodes in stage k is smaller than desired number of nodes)*

    $m \leftarrow m + 1$ *(increment node counter*) **do**

    $v \leftarrow \text{argmax}_j \boldsymbol{\epsilon}_j,\ j \in \mathcal{V}[k]$ *(which ff has the largest RMSE)*

    $[\boldsymbol{\xi}_v, \boldsymbol{\xi}_m], [\epsilon_v, \epsilon_m] \leftarrow k$-means$(\boldsymbol{\Xi}[\boldsymbol{s}_v], 2)$ *(apply kmeans to obtain two centroids – replace existing node and create a new node at m, also track corresponding error)*

    $\mathcal{V} \leftarrow \mathcal{V} \cup \{m\}$ *(add m to the set $\mathcal{V}$)*

  end while

end for

the following expression is obtained for the continuous cost:

$$\int_0^1 \mathbb{E}[\mathrm{C}(\mathrm{x}(t))]\,dt = c_2 \int_0^1 \mathbb{E}\left[(\boldsymbol{x}^\intercal \mathbf{b}_n(t))^2\right] dt + c_1 \int_0^1 \mathbb{E}[\boldsymbol{x}^\intercal \mathbf{b}_n(t)]dt + c_0$$
$$= \frac{c_2}{2}\mathbb{E}[\boldsymbol{x}^\intercal \boldsymbol{U}_\beta \boldsymbol{x}] + c_1 \boldsymbol{u}_\beta^\intercal \mathbb{E}[\boldsymbol{x}] + c_0 \tag{2.1.8}$$

Using the Bernstein basis, a dense matrix $\boldsymbol{U}_\beta$ is obtained resulting in bi-linear coefficient cross terms. However, if an *orthogonal* basis (e.g. Legendre polynomials) is chosen, the corresponding matrix $\boldsymbol{U}_\delta$ is diagonal (the cross terms fall out) allowing one to circle back to the aforementioned linear approximation of the quadratic term. Defining a new basis $\boldsymbol{\delta}^\intercal \mathbf{d}_n(t) = \boldsymbol{x}^\intercal \mathbf{b}_n(t) = \mathrm{x}(t)$ where $\mathbf{d}_n(t)$ are orthogonal basis functions in the range $[0, 1]$, $\boldsymbol{\delta} = \boldsymbol{D}\boldsymbol{x}$ and $\Delta_i \approx \delta_i^2$, the quadratic objective can be approximated within a minimizing optimization program as:

$$\min \int_0^1 \mathbb{E}[\mathrm{C}(\mathrm{x}(t))]\,dt \approx \frac{c_2}{2}\sum_{i=0}^{n} \mathbb{E}[\Delta_i][U_\delta]_{i,i} + c_1 \boldsymbol{u}_\delta^\intercal \mathbb{E}[\boldsymbol{\delta}] + c_0$$
$$\text{s.t. } \Delta_i \geq a_1^j \delta_i + a_0^j \quad \forall i \in \{0, \ldots, n\}, j \in \mathcal{J}_i \tag{2.1.9}$$

where $\mathcal{J}_i$ reflects a set of constraints used for the approximation $\Delta_i \approx \delta_i^2$. This allows one to apply a piece-wise linear approximation to the continuous quadratic costs, introducing a trade-off between number of constraints (problem complexity) and accuracy. For the remainder of the chapter the Bernstein basis is used, but for simulations this mapping is utilized to approximate the quadratic cost terms just described. This approach resolves one of the limitations of past work in [Parvania and Scaglione(2016)].

2.1.3 Energy Storage Model

The conventional power system model is expanded to track energy output of dispatchable resources, allowing one to include constraints on energy, particularly suited for storage or storage-like resources such as dispatchable demand. This is presently done to varying

degrees by some ISOs to accommodate storage resources [California ISO(2018)]. In discrete time, tracking the energy output e^g of generator g producing x_k^g of average power in period k becomes a simple sum $e_k^g = \sum_{p=0}^{k} x_p^g$, approximating the corresponding continuous time formulation $\mathrm{e}^g(\mathrm{t}) = \int_0^{\mathrm{t}} \mathrm{x}^g(\tau)\, d\tau$.

In continuous-time, it is possible to continually track the energy and constrain it to honor capacity limits. Recalling the relationship between a continuous function defined in terms of Bernstein coefficients and its derivative, the relationship between generators g coefficients of power $\boldsymbol{x}_k^g$ and energy $\boldsymbol{e}_k^g$ for interval k are:

$$\boldsymbol{x}_k^g = \boldsymbol{M}_{n+1} \boldsymbol{e}_k^g \tag{2.1.10}$$

The energy, being an integral, has an additional degree of freedom. However, as the sum energy output of generators is inherently continuous (a discontinuous jump in the energy profile would require infinite power), the coefficient $e_k^{g(0)}$ can be assumed to be known, leading to the reverse the relationship (2.1.10). Splitting $\boldsymbol{M}_{n+1}$ into $\boldsymbol{m}_{n+1}$ denoting the first column, and $\boldsymbol{M}'_{n+1}$, a square matrix containing the remaining columns, and derive:

$$\left(e_k^{g(1)}, \ldots, e_k^{g(n+1)}\right) = \left(\boldsymbol{M}'_{n+1}\right)^{-1} \left(\boldsymbol{x}_k^g - \boldsymbol{m}_{n+1} e_k^{g(0)}\right) \tag{2.1.11}$$

Given an initial value $\mathrm{e}^g(0)$ and a piece-wise power profile $\boldsymbol{x}_k^g$ for all k, one can thus iteratively calculate all coefficients $\boldsymbol{e}_k^g$ and the corresponding continuous energy profile $\mathrm{e}^g(\mathrm{t})$, which will be C^{C+1} continuous given a C^C continuity of the power profile. To honor energy storage devices minimum and maximum energy, the convex hull property is used:

$$\underline{e}^g \leq \mathrm{e}^g(\mathrm{t}) \leq \overline{e}^g \iff \underline{e}^g \leq e_k^{g(i)} \leq \overline{e}^g \quad \forall i, k \tag{2.1.12}$$

2.2 The Continuous-Time Multi-State Stochastic Unit Commitment

The underlying methods used to improve upon a conventional SUC to formulate the CT-MSUC are explained in Section 2.1; to summarize, the options are:

(a) The continuous time formulation of power (load and generation),

(b) the multi stage scenario tree used to characterize the underlying uncertainty (both (a) and (b) are visualized in Figure 2.1.1,

(c) tracking and constraining of energy to ease modeling of storage-like resources.

As the formulation is no longer a deterministic plan in time, it takes on the nodal form of the scenario tree, with most variables being indexed w.r.t. a certain path on the tree, instead of time.

2.2.1 Considerations Regarding Optimizing in Continuous-Time

In continuous time, similar to the continuous load $\xi^b(\mathrm{t})$ at bus b, generator power, ramping and energy are defined as continuous-time functions $\mathrm{x}^g(\mathrm{t})$, $\dot{\mathrm{x}}^g(\mathrm{t})$, $\mathrm{e}^g(\mathrm{t})$ for all generators $g \in \mathcal{G} = \{1, \ldots, G\}$, further specifying $\mathcal{G}(b)$ as the set of generators at bus b. By ensuring that both the polynomial of load (power) and generation power are of order n and have the same continuity, Consumption and generation is continually balanced by ensuring that, at every time instant, the sum coefficients of consumption and generation are equal. Introducing:

$$\mathbf{e}_v^g = (e_v^{g(0)}, e_v^{g(1)}, \ldots, e_v^{g(n)}, e_v^{g(n+1)})^\intercal \tag{2.2.1}$$

$$\boldsymbol{x}_v^g = (x_v^{g(0)}, x_v^{g(1)}, \ldots, x_v^{g(n-1)}, x_v^{g(n)})^\intercal \tag{2.2.2}$$

$$\dot{\boldsymbol{x}}_v^g = (\dot{x}_v^{g(0)}, \dot{x}_v^{g(1)}, \ldots, \dot{x}_v^{g(n-2)}, \dot{x}_v^{g(n-1)})^\intercal \tag{2.2.3}$$

denoting the Bernstein coefficients of the corresponding continuous functions of generator energy, power and ramp, for the interval corresponding to the scenario tree node $v \in \mathcal{V}$. Utilizing (2.1.4) to find the derivative coefficients, the generator continuous-time energy,

power and ramp functions for node $v \in \mathcal{V}$ are:

$$\mathrm{e}_v^g(t) = (\boldsymbol{e}_v^g)^\intercal \mathbf{b}_{n+1}(t) \tag{2.2.4}$$

$$\mathrm{x}_v^g(t) = (\boldsymbol{x}_v^g)^\intercal \mathbf{b}_n(t) = (\boldsymbol{M}_{n+1}\boldsymbol{e}_v^g)^\intercal \mathbf{b}_n(t) \tag{2.2.5}$$

$$\dot{\mathrm{x}}_v^g(t) = (\dot{\boldsymbol{x}}_v^g)^\intercal \mathbf{b}_{n-1}(t) = (\boldsymbol{M}_n\boldsymbol{x}_v^g)^\intercal \mathbf{b}_{n-1}(t) \tag{2.2.6}$$

where ramp is the derivative of power, and power the derivative of energy. Generator commitment y^g, start-up $\overline{s}^g$ and shut-down $\underline{s}^g$ indicators are hourly as in conventional models, but indexed using the tree nodal notation, allowing different commitment schedules depending on the tree path traveled. A particular realization corresponds to a certain path in the scenario tree[2] $\mathcal{H} \subseteq \mathcal{V}$ and the corresponding trajectory can be reconstructed as:

$$x_{\mathcal{H}}^g(t) = \sum_{v \in \mathcal{H}} (\boldsymbol{x}_v^g)^\intercal \mathbf{b}(t - \tau(v))\mathrm{rect}(t - \tau(v)), \tag{2.2.7}$$

and similarly for the energy and ramp trajectories.

2.2.2 Stochastic Unit Commitment Modeling Changes

As the input to the problem is a scenario tree of future load, the output will be a tree of generation dispatch and commitment profiles. As discussed in Section 2.1.2, load is considered to be a continuous-space random process, but approximate by a tree structure. As in traditional power system operation, where a difference between actual and forecasted load is observed, the load will not follow any path of the tree precisely, the tree is constructed to capture the various major load trends. This mismatch is compensated for in a traditional fashion, by allocating certain reserve power capacity around each path of the tree, reflecting certain units that can change their dispatch to meet the load realization. Nodal reserve functions are defined along with their corresponding coefficients $\hat{\mathrm{r}}_v^g(t)$, $\check{\mathrm{r}}_v^g(t)$, $\hat{\boldsymbol{r}}_v^g$ and $\check{\boldsymbol{r}}_v^g$, such

[2]The only case in which the path is equal to the set of nodes is when the problem is deterministic, i.e. there is a single future forecast.

that a unit g produces between $[\mathrm{x}_v^g(t) + \check{\mathrm{r}}_v^g(t), \mathrm{x}_v^g(t) + \hat{\mathrm{r}}_v^g(t)]$ if the realization passes through node v.

2.2.3 Optimization Objective

The stochastic and continuous-time aspects of the formulation add some complications to the objective of the UC formulation. Conventionally, the goal is to minimize cost of generation; to this end there exists a vast literature, e.g. [Zhou and Botterud(2014), Liu *et al.*(2017), Hreinsson *et al.*(2015)] in including the reserve power allocation in the UC problem, in order to increase feasibility and lower cost compared to solving the two problems separately. In the proposed formulation, cost functions are added to the energy and ramp dimensions. The rationale is that for energy, storage devices can show their lean towards buying or selling depending on their storage level, and for ramping, units can try to incorporate the long term cost of wear and tear by penalizing ramps. The cost of commitment, startup and shutdown is defined as Y^g, $\overline{S}^g$, $\underline{S}^g$. The power, energy and ramping and reserves costs are:

$$X^g(\mathrm{x}) = X_2^g \mathrm{x}^2(t) + X_1^g \mathrm{x}(t) \tag{2.2.8}$$

$$E^g(\mathrm{e}) = E_2^g \mathrm{e}^2(t) + E_1^g \mathrm{e}(t) + E_0^g \tag{2.2.9}$$

$$\dot{X}^g(\dot{\mathrm{x}}) = \dot{X}_2^g \dot{\mathrm{x}}^2(t) + \dot{X}_1^g \dot{\mathrm{x}}(t) \tag{2.2.10}$$

$$R^g(\mathrm{r}) = R_2^g \mathrm{r}^2(t) + R_1^g \mathrm{r}(t) \tag{2.2.11}$$

Note that the cost terms could vary depending on the interval k, but for simplicity the same cost curves are used throughout the time horizon. The cost of node $v \in \mathcal{V}$ is formulated as:

$$\begin{aligned} J_v = \sum_{g \in \mathcal{G}} \Bigg[& \overline{S}^g \overline{s}_v^g + \underline{S}^g \underline{s}_v^g + \int_0^1 \Big(R^g\big(\hat{\mathrm{r}}_v^g(\tau)\big) + R^g\big(\check{\mathrm{r}}_v^g(\tau)\big) \\ & + E^g\big(\mathrm{e}_v^g(\tau)\big) + X^g\big(\mathrm{x}_v^g(\tau)\big) + \dot{X}^g\big(\dot{\mathrm{x}}_v^g(\tau)\big) \Big)\, d\tau \Bigg] \end{aligned} \tag{2.2.12}$$

where the integrals are mapped into functions of coefficients as explained in Section 2.1.2.2. The objective becomes the total expected real-time cost becomes:

$$\min \sum_{v \in \mathcal{V}} \pi_v J_v \tag{2.2.13}$$

where π_v is the probability of node v.

2.2.4 Program Formulation

This section describe the constraints of the SUC formulation, some being previously discussed in [Parvania and Scaglione(2016)] and [Hreinsson *et al.*(2018)].

2.2.4.1 Continuity and Derivative Constraints

Energy is mapped to power and to ramp through the following constraints:

$$\boldsymbol{x}_v^g = \boldsymbol{M}_{n+1} \boldsymbol{e}_v^g \qquad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.14}$$

$$\dot{\boldsymbol{x}}_v^g = \boldsymbol{M}_n \boldsymbol{x}_v^g \qquad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.15}$$

C^C continuity is enforced for the continuous-time power polynomials (C^{C+1} for energy, C^{C-1} for ramp).

$$\left[\boldsymbol{M}_{n+1-c} \cdots \boldsymbol{M}_{n+1} \boldsymbol{e}_{v^-}^g\right]^{(n+1-c)} = \left[\boldsymbol{M}_{n+1-c} \cdots \boldsymbol{M}_{n+1} \boldsymbol{e}_v^g\right]^{(0)}$$
$$\forall c \in \{-1, \ldots, C\}, g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.16}$$

where $[\cdot]^{(i)}$ denotes vector element i. Note that for the discontinuous case, $c = -1$, $\boldsymbol{M}_{n+2} \cdots \boldsymbol{M}_{n+1} = \boldsymbol{I}$ (identity) and the constraints reduce to $\boldsymbol{e}_{v^-}^{g(n+2)} = \boldsymbol{e}_v^{g(0)}$; for $c = 0$, $\boldsymbol{x}_{v^-}^{g(n+1)} = \boldsymbol{x}_v^{g(0)}$ is added, and so on.

Reserve capacity must also be continuous, to ensure it is deliverable in real-time. For $c \in \{0 \ldots, C\}, g \in \mathcal{G}, v \in \mathcal{V}$:

$$\left[\boldsymbol{M}_{n-c+1} \cdots \boldsymbol{M}_n \hat{\boldsymbol{r}}_{v^-}^g\right]^{(n-c)} = \left[\boldsymbol{M}_{n-c+1} \cdots \boldsymbol{M}_n \hat{\boldsymbol{r}}_v^g\right]^{(0)} \tag{2.2.17}$$

$$\left[\boldsymbol{M}_{n-c+1} \cdots \boldsymbol{M}_n \check{\boldsymbol{r}}_{v^-}^g\right]^{(n-c)} = \left[\boldsymbol{M}_{n-c+1} \cdots \boldsymbol{M}_n \check{\boldsymbol{r}}_v^g\right]^{(0)} \tag{2.2.18}$$

Remark 1 (Implementation Note). Of the coefficients $\boldsymbol{e}_v^g$, $\boldsymbol{x}_v^g$ and $\dot{\boldsymbol{x}}_v^g$, ignoring boundary conditions, only $n - C$ are free variables, with all other having a direct dependence through the derivative constraints or from preceding or following nodes. The presentation is simpler to follow if (as done throughout the chapter) constraints are defined in terms of the appropriate variable $\boldsymbol{e}$, $\boldsymbol{x}$ or $\dot{\boldsymbol{x}}$. However, by defining the dependent variables as expressions of the free variables, the numerical optimization complexity can change. Experience indicates that defining $n - C$ of the $\boldsymbol{x}$ coefficients as free (decision) variables, instead of $\boldsymbol{e}_v^g$ reduces the computational cost.

2.2.4.2 Balance Constraints

To balance generation $\boldsymbol{x}$ and load $\boldsymbol{\xi}$ the sum of the vectors of coefficients of generation and those of the load must be equal. Defining the vector of the coefficients of power injection at bus b, node v as $\boldsymbol{h}_v^b = \sum_{g \in \mathcal{G}(b)} (\boldsymbol{x}_v^g) - \boldsymbol{\xi}_v^b$; then the balance constraint is:

$$\sum_{b \in \mathcal{B}} \boldsymbol{h}_v^b = \mathbf{0} \quad \forall v \in \mathcal{V} \tag{2.2.19}$$

Denoting by $\epsilon_v^b(t) = [\boldsymbol{\epsilon}_v^b]^\intercal \boldsymbol{b}_n(t)$ the root mean square error between a scenario tree segment $\xi_v^b(t)$ and the bundle of scenarios it approximates, a nodal reserve proportional (by ρ) to this error is required for all nodes $v \in \mathcal{V}$:

$$\sum_{g \in \mathcal{G}(b)} (\mathbf{x}_v^g + \hat{\mathbf{r}}_v^g) - \hat{\boldsymbol{h}}_v^b \geq (\boldsymbol{\xi}_v^b + \rho\, \boldsymbol{\epsilon}_v^b) \tag{2.2.20}$$

$$\sum_{g \in \mathcal{G}(b)} (\mathbf{x}_v^g - \check{\mathbf{r}}_v^g) - \check{\boldsymbol{h}}_v^b \leq (\boldsymbol{\xi}_v^b - \rho\, \boldsymbol{\epsilon}_v^b) \tag{2.2.21}$$

$$\sum_{b \in \mathcal{B}} \hat{\boldsymbol{h}}_v^b = \mathbf{0}\ ,\ \sum_{b \in \mathcal{B}} \check{\boldsymbol{h}}_v^b = \mathbf{0} \tag{2.2.22}$$

If $\xi_v^b(t)$ is unbiased, then $\epsilon_v(t)$ is the standard deviation of the sample paths. As such, the solution is feasible for ρ times the conditional standard deviation, which can use as a bound for the deviation from $\xi_v^b(t)$ to a likely interval.

2.2.4.3 Energy Limits

To limit the continuous value of energy, power and ramping, the convex hull property is used (as explored in detail in [Parvania and Scaglione(2016)]), stating that a function lies in between the convex hull of its coefficients. In the case of energy, to ensure lower ($\underline{e}$) and upper ($\overline{e}$) energy limits are continuously honored, it is sufficient to ensure that the coefficients never go outside these limits:

$$\underline{e}^g \leq \boldsymbol{e}_v^g \leq \overline{e}^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.23}$$

Remark 2. Any deviations in energy levels that are a result of reserve power activation are not included in this constraint. This follows a common practice of UC formulations, where e.g. only ramping constraints between expected power dispatch levels are included, with nothing preventing excessive ramps during reserve events. One could include such constraints; in this case it would suffice to track the upper energy bound with respect to $x(t) + \hat{r}(t)$ and the lower w.r.t. $x(t) - \check{r}(t)$. To unburden the presentation this is omitted.

2.2.4.4 Generation Limits

Similar to energy, ensuring that generator coefficients sit between the minimum and maximum generation $\underline{x}^g$, $\overline{x}^g$ is sufficient to ensure that the generator limits are honored. The handling of commitment variables for continuous trajectories is more involved. As an example, for a unit that is online during a particular hour $\tau(v)$, but offline during the previous hour $\tau(v^-)$, stating that $\boldsymbol{x}_{v^-}^g = 0$ and $\boldsymbol{x}_v^g \geq \underline{x}^g$ is a violation of the continuity constraints if $C > -1$. To allow for smooth on/off transitions that do not violate continuity the enforcement of these constraints is shifted in time. The power plot of Figure 2.2.1 attempts to visualize this, where the effective min/max constraints are shown with dotted lines. Mathematically, these constraints are expressed as follows. For the first $i \in \{0, \ldots, C\}$, they are

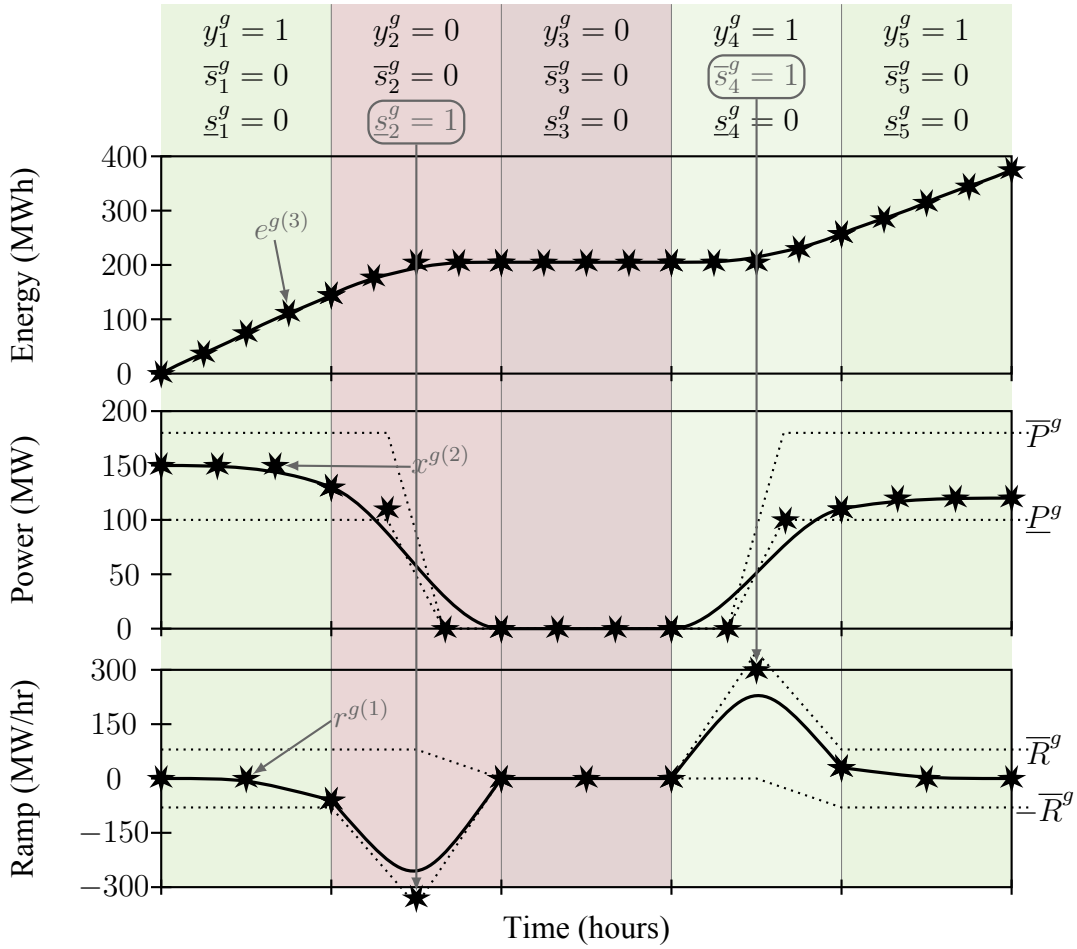


Figure 2.2.1: The relationship between continuous energy, power and ramping (continuous paths), along with the Bernstein coefficients (stars). Additionally the figure shows generator's commitment (on/off) status, power and ramping limits and how those limits change on startup/shutdown.

bounded by the preceding commitment variable:

$$\boldsymbol{x}_v^{g(i)} + \hat{\boldsymbol{r}}_v^{g(i)} \leq \overline{x}^g y_{v^-}^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \setminus \mathcal{V}(1) \tag{2.2.24}$$

$$\boldsymbol{x}_v^{g(i)} - \check{\boldsymbol{r}}_v^{g(i)} \geq \underline{x}^g y_{v^-}^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \setminus \mathcal{V}(1) \tag{2.2.25}$$

However, for the remaining coefficients $i \in \{C+1, \ldots, n\}$:

$$\boldsymbol{x}_v^{g(i)} + \hat{\boldsymbol{r}}_v^{g(i)} \leq \overline{x}^g y_v^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.26}$$

$$\boldsymbol{x}_v^{g(i)} - \check{\boldsymbol{r}}_v^{g(i)} \geq \underline{x}^g y_v^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.27}$$

Further discussion of these constraints, as well as the following ramping constraints can be found in [Parvania and Scaglione(2016)].

2.2.4.5 Ramping Limits

To enable changes in commitment for units that have slow ramp rates, the ramping limits are relaxed on coefficients during on-off transitions, With generator g ramp limit $\overline{\dot{x}}^g$, for $v \in \mathcal{V} \setminus \mathcal{V}(1)$:

$$\dot{\boldsymbol{x}}_v^{g(C)} \leq \overline{\dot{x}}^g + n\overline{x}^g \overline{s}_v^g \ , \ \dot{\boldsymbol{x}}_v^{g(C)} \geq -\overline{\dot{x}}^g - n\overline{x}^g \underline{s}_v^g \tag{2.2.28}$$

For the first ramping coefficients the preceding interval commitment is considered. For $i \in \{0, \ldots, C-1\}, v \in \mathcal{V}$:

$$\dot{\boldsymbol{x}}_v^{g(i)} \leq \overline{\dot{x}}^g y_{v^-}^g \ , \ \dot{\boldsymbol{x}}_v^{g(i)} \geq -\overline{\dot{x}}^g y_{v^-}^g \tag{2.2.29}$$

For the last ramping coefficients the present interval commitment is considered. For $i \in \{C+1, \ldots, n-1\}, v \in \mathcal{V}$:

$$\dot{\boldsymbol{x}}_v^{g(i)} \leq \overline{\dot{x}}^g y_v^g \ , \ \dot{\boldsymbol{x}}_v^{g(i)} \geq -\overline{\dot{x}}^g y_v^g \tag{2.2.30}$$

Figure 2.2.1 provides insights into these constraints, where the ramping constraints are visualized with a dotted line.

2.2.4.6 Minimum On/Off Constrains

Denoting by T_{on}^g and T_{off}^g the minimum on- and off-time respectively, the minimum on and off constraints are conventional [Hedman *et al.*(2009)]; recalling that $\alpha^i(v)$ is the i-th ancestor

of node v:

$$\overline{s}_v^g - \underline{s}_v^g = y_v^g - y_{v\_}^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.31}$$

$$y_v^g \geq \sum_{u=0}^{\min(T_{\text{on}}^g, \tau(v))-1} \overline{s}_{\alpha^u(v)}^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.32}$$

$$1 - y_v^g \geq \sum_{u=0}^{\min(T_{\text{off}}^g, \tau(v))-1} \underline{s}_{\alpha^u(v)}^g \quad \forall g \in \mathcal{G}, v \in \mathcal{V} \tag{2.2.33}$$

2.2.4.7 Transmission Line Flow Constraints

Following continuous load and generation, transmission line flows will certainly change in continuous fashion. Assuming all operations are element-wise, for all transmission lines $l \in \mathcal{L}$, buses $v \in \mathcal{V}$ and coefficients $i \in \{0, \ldots, n\}$:

$$-\overline{F}^l \leq \sum_{b \in \mathcal{B}} F_{b,l} \boldsymbol{h}_v^{b(i)} \leq \overline{F}^l \tag{2.2.34}$$

$$-\overline{F}^l \leq \sum_{b \in \mathcal{B}} F_{b,l} \hat{\boldsymbol{h}}_v^{b(i)} \leq \overline{F}^l \tag{2.2.35}$$

$$-\overline{F}^l \leq \sum_{b \in \mathcal{B}} F_{b,l} \check{\boldsymbol{h}}_v^{b(i)} \leq \overline{F}^l \tag{2.2.36}$$

where $F_{b,l}$ is the Power Transmission Distribution Factor (PTDF) of bus b to line l, and $\overline{F}^l$ is the maximum power the line can transfer.

2.3 Formulation Complexity and Solution Techniques

A few observations regarding the solution complexity are in order. First, compared to the MSUC the CT-MSUC formulation roughly doubles the number of continuous variables and adds continuity constrains. Even though the number of integer variables remains identical, the large number of binary variables due to the multi-stage formulation, along with the increase in continuous variables and constraints, one quickly suffers from the curse of dimensionality, where hours or days become necessary to obtain a solution for cases that can

be viewed as being relatively small for a conventional UC solver. For stochastic formulations, scalability is an issue in general, and in practice, decomposition and other advanced solving methods are necessary for reasonable solving times. Here two prominent decomposition techniques are discussed along with how to apply them to the proposed formulation.

2.3.1 Progressive Hedging

Progressive Hedging (PH) [Watson and Woodruff(2011)] falls into the category of *Dual Ascent* methods, more specifically those that employ an *Augmented Lagrangian* function with an augmented penalty term which tries to force a sub-set of decision variables across sub-problems to be equal. There are different ways to approach a SUC problem with PH, but a straightforward way would be to consider each path $\mathcal{H}$ along the scenario tree as an independent (think deterministic) sub-problem, and then through the augmented Lagrangian, introduce penalties for any difference in those nodal decision variables that should be identical because two or more sub-problems share a node of the scenario tree.

PH has been shown to be successful for mixed-integer programs such as SUC [Ordoudis *et al.*(2015), Ryan *et al.*(2013)]. However, the problem with PH, as well as other methods involving augmented Lagrangians, is that even though they have impressive convergence results for *convex problems*, there are no guarantees of convergence for mixed-integer problems, and obtaining convergence involves heuristics and precise tuning, for which different approaches may not translate well across different types of problems.

2.3.2 Stochastic Dual Dynamic Integer Programming

Stochastic Dual Dynamic Integer Programming (SDDiP) [Zou *et al.*(2017)] builds on Stochastic Dual Dynamic Programming (SDDP) [Pereira and Pinto(1991)], a well known decomposition algorithm commonly employed for hydro-scheduling. SDDP utilizes the duality of linear programs and the recursive problem structure of multi-stage decision

problems, allowing *Benders cuts* [Benders(1962)] to "flow" up the decision tree such that the root formulation captures the most interesting vertices of the sub-problems. SDDP further describes how this approach can be performed using Monte Carlo sampling of the scenario tree paths, allowing arbitrarily large scenario trees, while converging to the optimal value with certain confidence guarantees.

SDDiP extends the fundamental idea of SDDP, but introduces other types of cuts, that for binary state variables *guarantees convergence*. This has previously been shown to be successful for the application of SUC [Zou *et al.*(2018)].

First, denoting certain variables as *local* variables:

$$\boldsymbol{\phi}_v = [\boldsymbol{e}_v, \boldsymbol{x}_v, \dot{\boldsymbol{x}}_v, \boldsymbol{h}_v, \hat{\boldsymbol{h}}_v, \check{\boldsymbol{h}}_v, \hat{\boldsymbol{r}}_v, \check{\boldsymbol{r}}_v] \tag{2.3.1}$$

while also defining a set of *binary state* variables, in this case, a vector derived from the local and binary variables:

$$\boldsymbol{\chi}_v = [\boldsymbol{\chi}_v^y, \boldsymbol{\chi}_v^{\overline{s}}, \boldsymbol{\chi}_v^{\underline{s}}, \boldsymbol{\chi}_v^e, \boldsymbol{\chi}_v^x, \boldsymbol{\chi}_v^{\hat{r}}, \boldsymbol{\chi}_v^{\check{r}}] \in \{0,1\}^{|\boldsymbol{\chi}_v|} \tag{2.3.2}$$

The first three components of $\boldsymbol{\chi}_v$ are related to the binary variables. First, $\boldsymbol{\chi}_v^y = [y_v^1, \ldots, y_v^G]$ is the commitment of all generators, defining $I_g(v) = \min(T_{\text{on}}^g, \tau(v)) - 1$:

$$\boldsymbol{\chi}_v^{\overline{s}} = \left[\left[\overline{s}_{\alpha^0(v)}^1, \ldots, \overline{s}_{\alpha^{I_1(v)}(v)}^1 \right], \ldots, \left[\overline{s}_{\alpha^0(v)}^G, \ldots, \overline{s}_{\alpha^{I_G(v)}(v)}^G \right] \right]$$

is a vector of all generators start-up indicators as far back as is relevant for the minimum on constraints, and similarly for turn-off:

$$\boldsymbol{\chi}_v^{\underline{s}} = \left[\left[\underline{s}_{\alpha^0(v)}^1, \ldots, \underline{s}_{\alpha^{I_1(v)}(v)}^1 \right], \ldots, \left[\underline{s}_{\alpha^0(v)}^G, \ldots, \underline{s}_{\alpha^{I_G(v)}(v)}^G \right] \right]$$

where $I_g(v) = \min(T_{\text{off}}^g, \tau(v)) - 1$. The remaining four components of $\boldsymbol{\chi}_v$ relate to the continuous problem variables.

To maintain a C^C continuity of $\mathrm{x}(t)$, $\hat{\mathrm{r}}(t)$ and $\check{\mathrm{r}}(t)$ between nodes v^- and v, the last $C+1$ coefficients of the respective vectors of node v^- need to be available when defining

the constraints of node v. Additionally, to track the state of energy across nodes $\mathrm{e}(t)$, last coefficient of $\boldsymbol{e}$ must be provided as an initial condition for child nodes. SDDiP only guarantees convergence when the state variables are binary. Defining a quantization step-size q MW, the variables $x_v^{g(i)} = \underline{x}^g + q\left[\mathbf{x}_v^{g(i)}\right]^\intercal \mathbf{b}$, where $\mathbf{b}$ is a vector $[2^0, 2^1, \ldots, 2^{B^g-1}]^\intercal$ and $\mathbf{x} \in \{0,1\}^{B^g}$ is a *binary* vector of length B^g, where B^g is sufficiently large to cover the operating range of the generator: $B^g = \left\lceil \log_2\left((\overline{x}^g - \underline{x}^g)/q + 1\right)\right\rceil$. The same quantization is performed for the coefficients $\hat{\boldsymbol{r}}$, $\check{\boldsymbol{r}}$ and $\boldsymbol{e}$, though the necessary length of the corresponding binary vector varies between generators and type of variable. The quantization step-size should be identical across variables of power and energy, due to the finite difference relationship between the two. The implications of the quantization only affect the balance constraint (2.2.19), where a deviation of up to $q/2$ must be permitted (but penalized). Now, the remaining parts of $\boldsymbol{\chi}_v$ are:

$$
\begin{aligned}
\boldsymbol{\chi}_v^e &= \left[\mathbf{e}_v^{1(n+1)}, \ldots, \mathbf{e}_v^{G(n+1)}\right] \\
\boldsymbol{\chi}_v^e &= \left[\left[\mathbf{x}_v^{1(n-C)}, \ldots, \mathbf{x}_v^{1(n)}\right], \ldots, \left[\mathbf{x}_v^{G(n-C)}, \ldots, \mathbf{x}_v^{G(n)}\right]\right] \\
\boldsymbol{\chi}_v^{\hat{r}} &= \left[\left[\hat{\mathbf{r}}_v^{1(n-C)}, \ldots, \hat{\mathbf{r}}_v^{1(n)}\right], \ldots, \left[\hat{\mathbf{r}}_v^{G(n-C)}, \ldots, \hat{\mathbf{r}}_v^{G(n)}\right]\right] \\
\boldsymbol{\chi}_v^{\check{r}} &= \left[\left[\check{\mathbf{r}}_v^{1(n-C)}, \ldots, \check{\mathbf{r}}_v^{1(n)}\right], \ldots, \left[\check{\mathbf{r}}_v^{G(n-C)}, \ldots, \check{\mathbf{r}}_v^{G(n)}\right]\right]
\end{aligned}
$$

where the sans-serif font indicates the quantized binary representation of the corresponding slanted variables.

Define $\mathcal{C}$ as the constraints (2.2.14)-(2.2.36), and $\mathcal{C}_v$ to denote the constraints relevant for a particular node v (possibly depending on a previous node v^-). Assigning the root node

$v = 0$, a recursive problem can be formulated using the objective function J from (2.2.12):

$$Q_v\left(\boldsymbol{\chi}_{v^-}\right) = \min_{\boldsymbol{\chi}_v, \boldsymbol{\phi}_v, \boldsymbol{z}_v} J_v\left(\boldsymbol{\chi}_v, \boldsymbol{\phi}_v, \boldsymbol{z}_v\right) + \varphi_v(\boldsymbol{\chi}_v) \tag{2.3.3a}$$

$$\text{s.t. } \left(\boldsymbol{\chi}_v, \boldsymbol{\phi}_v, \boldsymbol{z}_v\right) \in \mathcal{C}_v \tag{2.3.3b}$$

$$\boldsymbol{z}_v = \boldsymbol{\chi}_{v^-} \tag{2.3.3c}$$

$$\boldsymbol{\chi}_v \in \{0,1\}^{|\boldsymbol{\chi}_v|} \tag{2.3.3d}$$

where, in a dynamic programming fashion, $\varphi_v(\boldsymbol{\chi}_v)$ represents the *cost to go*, whose exact value is:

$$\varphi_v(\boldsymbol{\chi}_v) = \sum_{v' \in \mathcal{C}(v)} \frac{\pi_{v'}}{\pi_v} Q_{v'}\left(\boldsymbol{\chi}_v\right) \tag{2.3.4}$$

SDDiP is an iterative algorithm that approximates the cost to go function with a number of *cuts*, such that:

$$\varphi_v\left(\boldsymbol{\chi}_v\right) \approx \psi_v^i\left(\boldsymbol{\chi}_v\right) = \min\left\{\theta_v : \theta_v \geq \underline{\theta}_v,\ \theta_v \geq \sum_{v' \in \mathcal{C}(v)} \pi_{v'}^v \left[\upsilon_{v'}^l + (\lambda_{v'}^l)^\intercal \boldsymbol{\chi}_v\right],\ l = 1, \ldots, i\right\} \tag{2.3.5}$$

where $\pi_{v'}^v = \pi_{v'}/\pi_v$ is the marginal probability of the nodal transition $v \to v'$.

2.3.2.1 Benders Cuts

By relaxing the binary constraints of problem (2.3.3), the conventional Benders cuts can be computed. Assigning λ_v to be the dual of constraint (2.3.3c), (2.3.3) can be solved for a given $\boldsymbol{\chi}_v^i$, adding to the parent problem the cut:

$$\theta_v \geq \sum_{v' \in \mathcal{C}(v)} \pi_{v'}^v \left[Q_{v'}^i + (\boldsymbol{\lambda}_{v'}^i)^\intercal (\boldsymbol{\chi}_v - \boldsymbol{\chi}_v^i)\right] \tag{2.3.6}$$

2.3.2.2 Lagrangian Cuts

Defining the Lagrangian of (2.3.3) with constraint (2.3.3c) relaxed:

$$\mathcal{L}_v^i\left(\overline{\boldsymbol{\lambda}}_v^i\right) = \min_{\boldsymbol{\chi}_v, \boldsymbol{\phi}_v, \boldsymbol{z}_v} J_v\left(\boldsymbol{\chi}_v, \boldsymbol{\phi}_v, \boldsymbol{z}_v\right) + \varphi_v(\boldsymbol{\chi}_v) + \overline{\boldsymbol{\lambda}}_v^i(\boldsymbol{\chi}_{v^-}^i - \boldsymbol{z}_v) \tag{2.3.7a}$$

$$\text{s.t. } (\boldsymbol{\chi}_v, \boldsymbol{\phi}_v, \boldsymbol{z}_v) \in \mathcal{C}_v \tag{2.3.7b}$$

$$\boldsymbol{\chi}_v \in \{0,1\}^{|\boldsymbol{\chi}_v|} \tag{2.3.7c}$$

The Lagrangian dual problem thus becomes:

$$\overline{\boldsymbol{\lambda}}_v^i = \arg\max_{\overline{\boldsymbol{\lambda}}_v^i} \mathcal{L}_v^i \tag{2.3.8}$$

which can then solve with e.g. subgradient methods [Held *et al.*(1974)], obtaining a cut:

$$\theta_v \geq \sum_{v' \in \mathcal{C}(v)} \pi_{v'}^v \left[\mathcal{L}_{v'}^i(\overline{\boldsymbol{\lambda}}_{v'}^i) + (\overline{\boldsymbol{\lambda}}_{v'}^i)^\intercal(\boldsymbol{\chi}_v - \boldsymbol{\chi}_v^i)\right] \tag{2.3.9}$$

The subgradient method works as follows:

(a) Start by initializing $\overline{\boldsymbol{\lambda}} = \mathbf{0}$.

(b) Solve the inner problem, obtaining the argument $\boldsymbol{z}(\overline{\boldsymbol{\lambda}})$.

(c) Move $\overline{\boldsymbol{\lambda}}^+ = \overline{\boldsymbol{\lambda}} + \alpha(\boldsymbol{\chi}_{v^-} - \boldsymbol{z}(\overline{\boldsymbol{\lambda}}))$.

(d) Assign $\overline{\boldsymbol{\lambda}} \leftarrow \overline{\boldsymbol{\lambda}}^+$ and go to step (b), unless $\overline{\boldsymbol{\lambda}}$ has sufficiently converged.

2.3.2.3 Strengthened Benders Cuts

This cut family is essentially a combination of the prior two. Starting by solving the relaxed problem (as for a Benders Cut) and obtaining the dual variable $\boldsymbol{\lambda}_v^i$ corresponding to constraint (2.3.3c), then using that dual vector to solve $\mathcal{L}_v^i(\boldsymbol{\lambda}_v^i)$ the cut becomes:

$$\theta_v \geq \sum_{v' \in \mathcal{C}(v)} \pi_{v'}^v \left[\mathcal{L}_{v'}^i(\boldsymbol{\lambda}_{v'}^i) + (\boldsymbol{\lambda}_{v'}^i)^\intercal(\boldsymbol{\chi}_v - \boldsymbol{\chi}_v^i)\right] \tag{2.3.10}$$

Table 2.1: Generator operational properties, cost coefficients, minimum and maximum energy, power and ramp speeds, as well as minimum on and off times. Note that $\mathrm{e}(0) = \mathrm{e}(K)$ for the storage devices S1000. Omitted cost coefficients are zero.

Unit	$[\underline{x}, \overline{x}]$ MW	$\overline{\dot{x}}$ MW/h	$[\underline{e}, \overline{e}]$, e(0) MWh	$(T_{\text{on}}, T_{\text{off}})$ h	Y	$\tilde{Y}$	$\overline{S}$	E_1	E_0	X_2	X_1	$\dot{X}_2$	R_1
U12	$[2, 12]$	12	$[0, \infty], 0$	$(4, 2)$	86	22	100	0.00	0	0.328	56.56	0.10	7.07
U20	$[16, 20]$	20	$[0, \infty], 0$	$(1, 1)$	400	100	100	0.00	0	0.000	130.00	0.05	16.25
U50	$[10, 50]$	50	$[0, \infty], 0$	$(1, 1)$	1	0	0	0.00	0	0.000	0.10	0.03	0.01
U76	$[15, 76]$	76	$[0, \infty], 0$	$(8, 4)$	212	53	200	0.00	0	0.014	16.08	0.10	2.01
U100	$[25, 100]$	100	$[0, \infty], 0$	$(8, 8)$	781	195	200	0.00	0	0.053	43.66	0.07	5.46
U155	$[54, 155]$	155	$[0, \infty], 0$	$(8, 8)$	382	96	250	0.00	0	0.008	12.39	0.07	1.55
U197	$[69, 197]$	180	$[0, \infty], 0$	$(12, 10)$	832	208	400	0.00	0	0.007	48.58	0.07	6.07
U350	$[140, 350]$	240	$[0, \infty], 0$	$(24, 24)$	665	166	700	0.00	0	0.005	11.85	0.30	1.48
U400	$[100, 400]$	60	$[0, \infty], 0$	$(24, 24)$	395	99	1500	0.00	0	0.000	4.42	0.90	0.55
S1000	$[-90, 90]$	90	$[0, 1000], 500$	$(1, 1)$	10	0	0	0.06	-30	0.009	1.80	0.01	0.50

2.4 Numerical Simulations

For numerical simulations the single area (24 buses) of the Institute of Electrical and Electronics Engineers (IEEE) IEEE Reliability Test System (RTS) [Grigg *et al.*(1999)]. The RTS system was used as a starting point for topology, generator placement and properties, but as the test case data does not provide the full set of parameters needed for the proposed model, certain missing parameters were derived from existing parameters; the readers should refer to Table 2.1 for all relevant generator properties and cost coefficients used in the simulations. Units prefixed with U are the conventional RTS generators, while S1000 describes a 1000 MWh storage unit. For simulations including storage, two S1000 devices were placed on buses 109 and 110. The rationale here was to place the storage devices on the boundary between the well-connected and the weakly-connected parts of the system (in terms of transmission capacity). In the test setup, storage devices offer load-shifting and additional flexibility to compensate for variability and stochasticity of loads, by participating as conventional units with an energy component to their bids in the UC cost minimization. Faced with the curse of dimensionality, both 12 and 24 hour commitment horizons were considered, allowing experimentation with more complicated scenario trees in the 12 hour case. The continuous-time simulations use cubic splines ($n = 3$) that are C^1 continuous (ramp is continuous), whereas the discrete-time formulations provide trajectories that are piece-wise constant ($n = 0$) and discontinuous (C^{-1}). Programming the proposed algorithm, plotting and solving was done with Python [Van Der Walt *et al.*(2011), Jones *et al.*(2014), Hunter(2007)] and Gurobi 8 [Gurobi Optimization(2015)] using a 60 core Intel Xeon processor running at 2.30 GHz. Decomposition algorithms such as aforementioned SDDiP and/or PH were not used for this numerical results, but are under development as part of future work.

2.4.1 Load Data

Electric load and generation data comes from California ISO (CAISO) [California ISO(2017)]: they include ten time series, five for different utility company load profiles, and five showing the solar and wind infeed from different California regions, all with a 5 minute resolution. To capture the temporal trends observed, and to map them onto the RTS grid topology, time series were linearly combined and scaled such that the total net-load (load minus renewable infeed) is a fixed ratio of the aggregate CAISO net-load, scaled by a factor of $1/15$, such that the load is within the generating capacity of the RTS system.

The processing pipeline from the original CAISO data to the eventual net-load trees used in simulations is visualized in Figure 2.4.1, along with one output tree aggregated over all the buses. To reduce the dimensionality during the tree generation, principal component analysis were applied and extracted $F = 20$ features to capture the most significant trends across all buses for each hour. The tree size was predetermined to have a certain number of nodes per hour, and recursive k-means was used in such a way as to prioritize branching on those nodes where the error between the centroid and the corresponding bundle of sample paths was high. After constructing the tree, features were mapped back to per-bus load trajectories, before mapping those to the desired polynomial; the coefficients were computed solving a least squares regression problem.

Remark 3. The tree construction algorithm is sub-optimal; it was chosen as an heuristic to obtain a multi-bus tree that was a suitable input to the proposed formulation. While these details are useful for reproducibility of the results, with the exception for the continuous-time load tree representation, how the data are curated and compressed is not the main focus of this chapter.

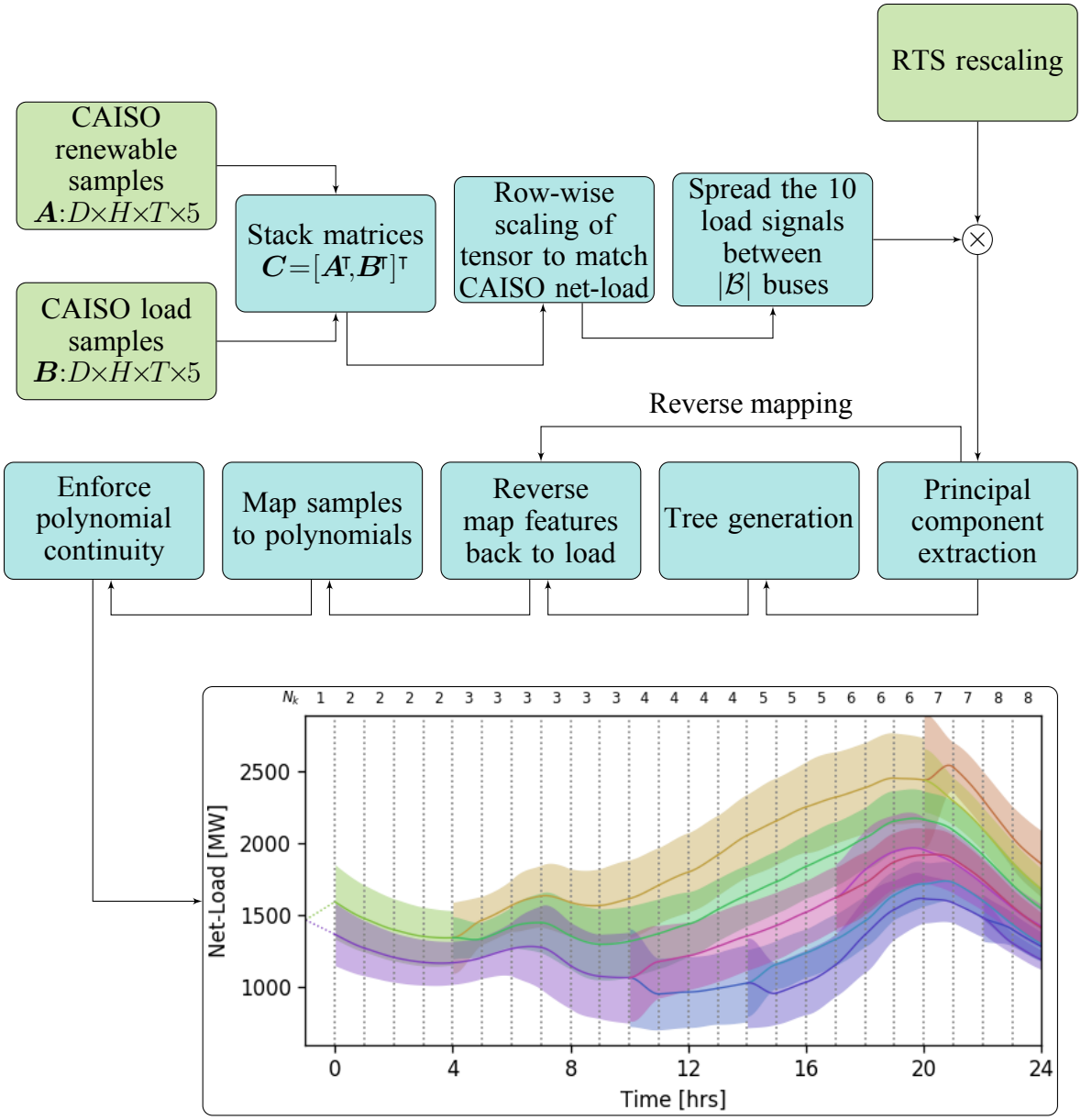


Figure 2.4.1: The steps taken from CAISO net-load data to a multivariate scenario tree with the load per bus as a piece-wise continuous polynomial. D is number of input sample days, H the number of hours, T the samples per hour, $|\mathcal{B}|$ number of buses, F number of extracted Principle Component Analysis (PCA)/Singular Value Decomposition (SVD) features, n is the polynomial degree and $\mathcal{T}(\cdot)$ denotes the entire tree structure, with each nodes value dimension $\cdot$.

2.4.2 Comparison

Comparing all combinations of formulations that are (i) continuous/discontinuous, (ii) deterministic/stochastic (iii) including/excluding storage devices and (iv) several different values of ρ. To capture what happens in real time, results were validated against a set of 109 test load trajectories that were not used as part of the tree construction as follows. Optimization was performed as described in Section 2.2 with a deterministic cost, but fix the commitment variables to the values found in the day-ahead solution, such that the model effectively is solving an OPF. While for the deterministic solutions there is only one set of commitment decisions, for the multi-stage formulations, to choose these integer variables the day-ahead tree was traversed *one* hour at a time, choosing the branch that is closest (in the L_1 norm) to the test load. The real load trajectory was assumed to be continuous cubic spline, which is a good approximation of the 5 minute resolution sample data [Hreinsson *et al.*(2018)] and certain constraints violations were allowed, prioritizing them through penalties. Specifically, to account for deviations from the precise forecast/tree paths, violations of min/max power and ramp constraints were allowed with the following penalties:

1) A unit that is asked to deviate from its day-ahead schedule is compensated with an additional 30% of its marginal cost, calculated at the day-ahead schedule set-point.

2) A unit that is asked to produce outside the day-ahead determined operation range (reserve range) gets an additional 100 per MWh. This is implemented with slack variables on (2.2.26)-(2.2.27) where $\overline{P}/\underline{P}$ reflects the reserve region.

3) A unit that is asked to produce outside of its min/max generation gets 200 per MWh compensation. This is implemented with slack variables on another set of (2.2.26)-(2.2.27), where $\overline{P}/\underline{P}$ retain their original meaning.

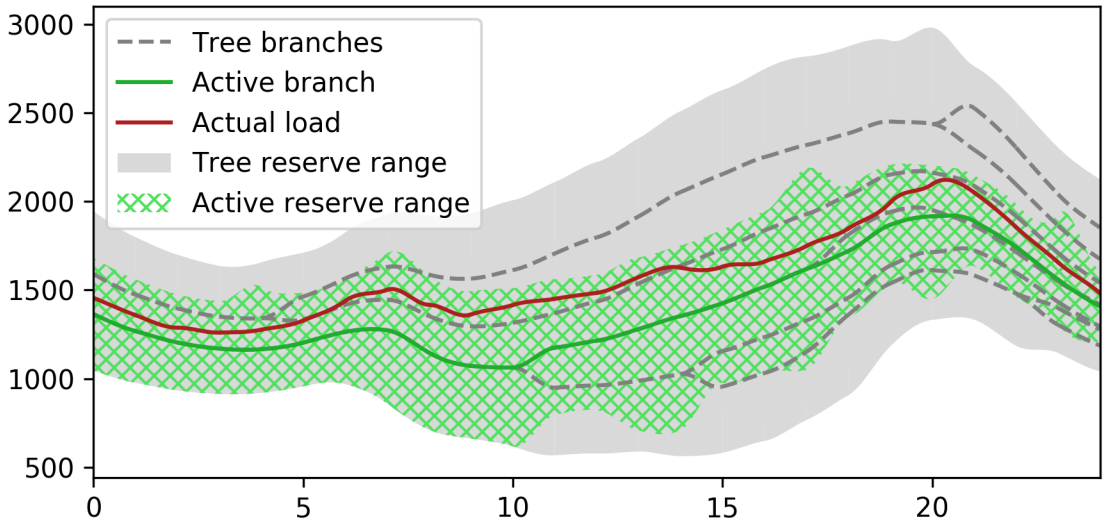


Figure 2.4.2: A sample scenario, showing test load with red line, the active tree branch with a green line, the corresponding reserve allocation with cross hatch, the inactive tree branches as dashed lines and the entire tree's reserve range in gray.

4) A unit that is asked to violate its ramping constraints gets compensated with additional 100% of the marginal ramp cost at the upper ramp limit. This is implemented through slack variables on (2.2.28)-(2.2.30).

Remark 4. The penalty terms can be additive, and the unit is always getting paid the underlying cost (the objective). Also, given the somewhat arbitrary choice of penalties, in comparing the results real-time penalties incurred are not shown but only the number of violations.

2.4.3 Results

To get a sense for what a solution to the CT-MSUC formulation looks like, Figure 2.4.2 shows a sample test profile (red), along with the underlying day-ahead solution tree (dotted gray lines), the scheduled reserve range (shaded gray area), the active reserve range (green hatch), which is similar to the scheduled reserve range but takes into account what units are committed on the closest path on the tree (green line). Figure 2.4.3 compares a sample

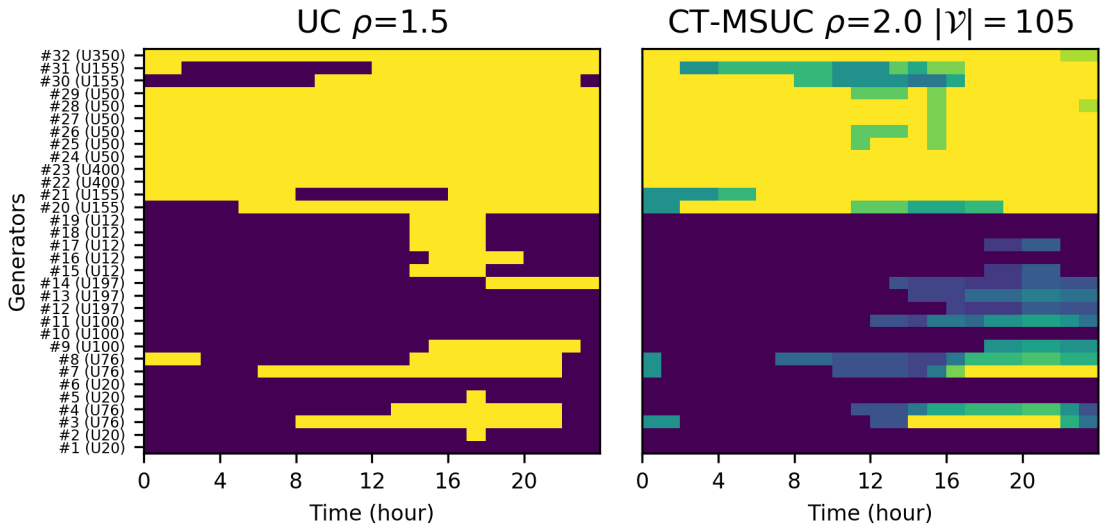


Figure 2.4.3: Unit Commitment for the CT-MSUC simulation. Purple indicates that unit is off for all tree paths of that hour, yellow is on for all tree paths, and the color in-between indicates some paths have the unit committed while others have it offline.

CT-MSUC commitment profile against a standard UC commitment solution, where yellow indicates units that are on, purple units that are off and for the continuous stochastic case the in-between shade indicates units that are online during some tree paths but not others.

For a more meaningful quantitative comparison, Figure 2.4.4 shows the total cost plotted against the number of constraint violations in real-time, for various solutions. The violations are further detailed in Figure 2.4.5 where the split between the various constraints can be seen, and as is expected, one can observe that the least penalized constraints are the ones more frequently violated. Figure 2.4.6 compares the number of hours committed for each solution, showcasing the flexibility of the generators chosen. There are no notable differences in the flexibility of units that can be attributed to the solution being continuous or not, but there is a clear tendency toward fewer committed hours for the solutions that are stochastic and/or include storage. This is interesting, as it points out that the modeling of uncertainty has more significant impact than modeling the inter-hourly variability associated to the trajectories.

A clear separation can be seen between solutions with and without storage, as the added storage allows for load shifting and better use of inexpensive generation, thus reducing costs. Testing the storage solutions reveals that, even though adding storage reduces the number of pmin/pmax violations as well as the time spent outside generator reserve range, there is a sharp increase in ramp limit violations not seen in the solutions without storage. This is a side-effect of the lower number of hours committed (as evident in Figure 2.4.6), requiring the remaining generators to move around more, particularly in conjunction with the charging and discharging of the storage units.

The stochastic formulations outperform the deterministic solutions, as it is clear from Figure 2.4.4. In fact, for a particular cost there is always a stochastic solution not far away that violates significantly fewer constraints than its deterministic counterpart. For the 12 hour case larger trees were modeled (in that they have more branching). Clearly, moving to a larger tree did not yield much benefit and, everything else being equal, the difference between the various stochastic solutions is small. It is useful to recognize, however, that the tree construction a heuristic was used and, thus, this test may not necessarily provide a good benchmark for the effect of the tree size on performance. These simulations are more useful to understand how complexity scales with the size of the tree (see Figure 2.4.7).

As for the continuous vs discrete formulations, Figure 2.4.5 suggests strongly that the continuous time solutions lead to fewer constraint violations in real-time, compared with their discrete counterparts. However, they also seem to be slightly more expensive in that they commit more units and demand more reserves. This is partially due to the complexity of the problem formulation; in the tests the Mixed-Integer Linear Program (MILP) solver in general left a larger MIP-gap for the continuous time formulations than the discrete ones, meaning that there is some room for improvement in the day-ahead solution for the continuous cases. In practice, depending on how constraint violations are penalized, the cost benefit of the continuous time formulation can range from small to significant. Irrespective

of the cost, the benefit of less ambiguity in the continuous solution remains an advantage of the continuous formulation.

Figure 2.4.7 shows the time required by the MILP solver (Gurobi) to obtain a solution for each of the problems. The most significant take-away message from this figure is that solving a multi-stage stochastic formulations, including moderately sized scenario trees that embed binary variables does not scale well, without utilizing some form of advanced solving techniques and the additional variables in the continuous formulation only exacerbate the situation. A continuous-time formulation with ca. 6000 binary variables could be solved, but beyond that the solver took more than the maximum of a few days to solve.

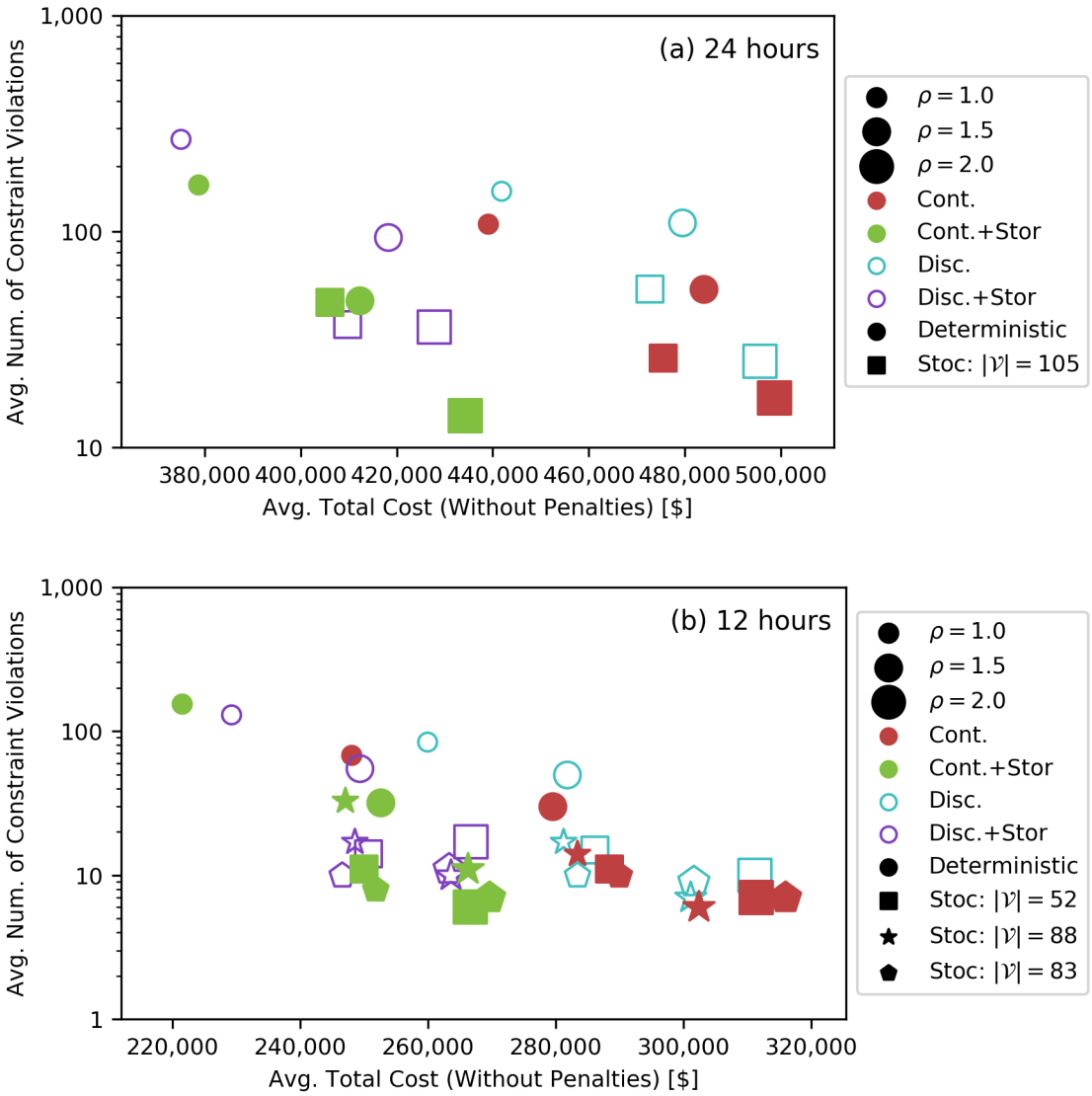


Figure 2.4.4: Average total cost (day-ahead and real-time costs without any penalties) plotted against the number of constraint violations on a log scale for the 24 hour (a) and 12 hour (b) case. Note that here off-schedule is not considered as a constraint violation, though in practice that would incur operational costs.

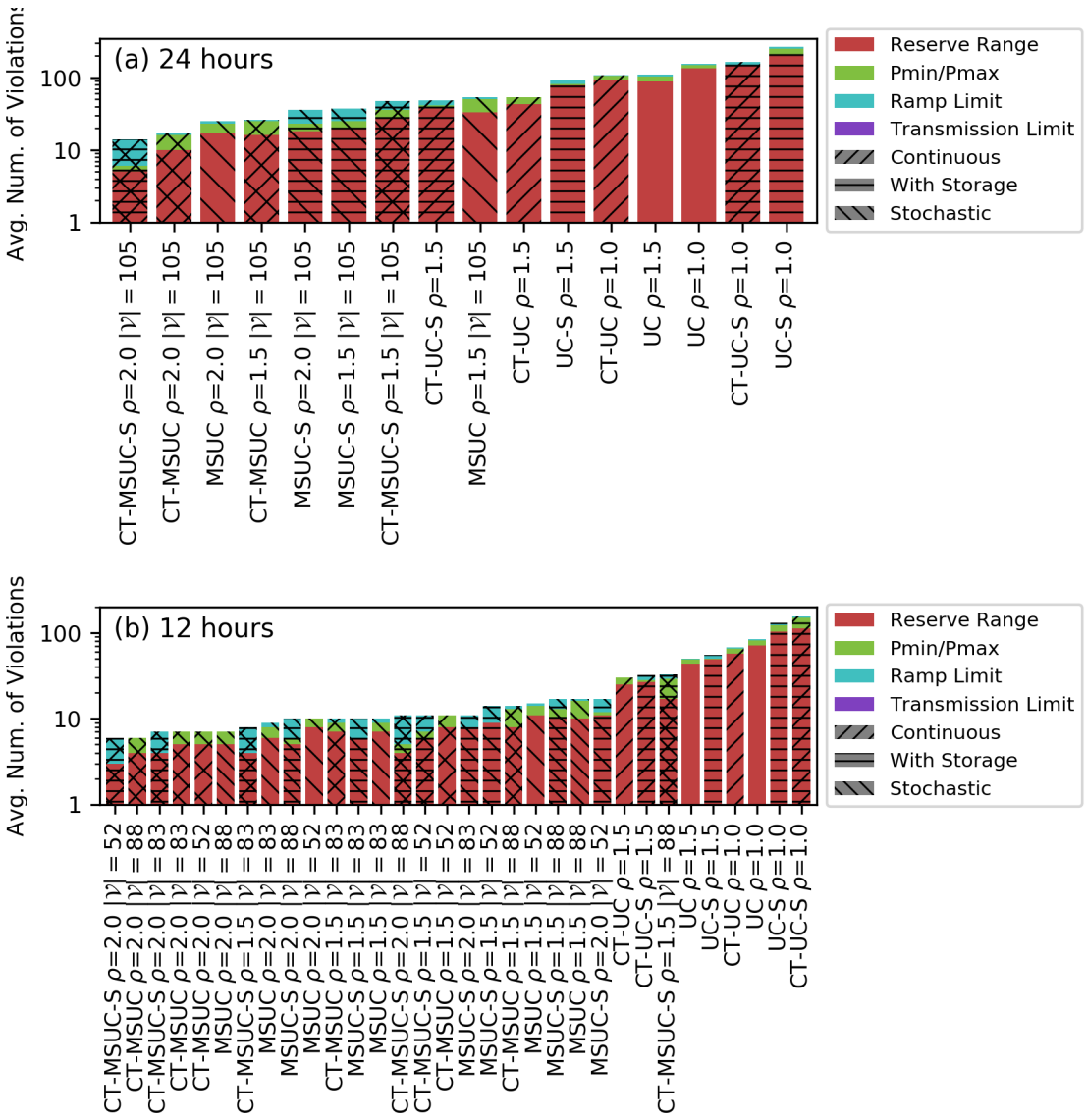


Figure 2.4.5: A sorted chart showing the number of constraint violations each solution incurs for the 24 hour (a) and 12 hour (b) case, categorized by violation type.

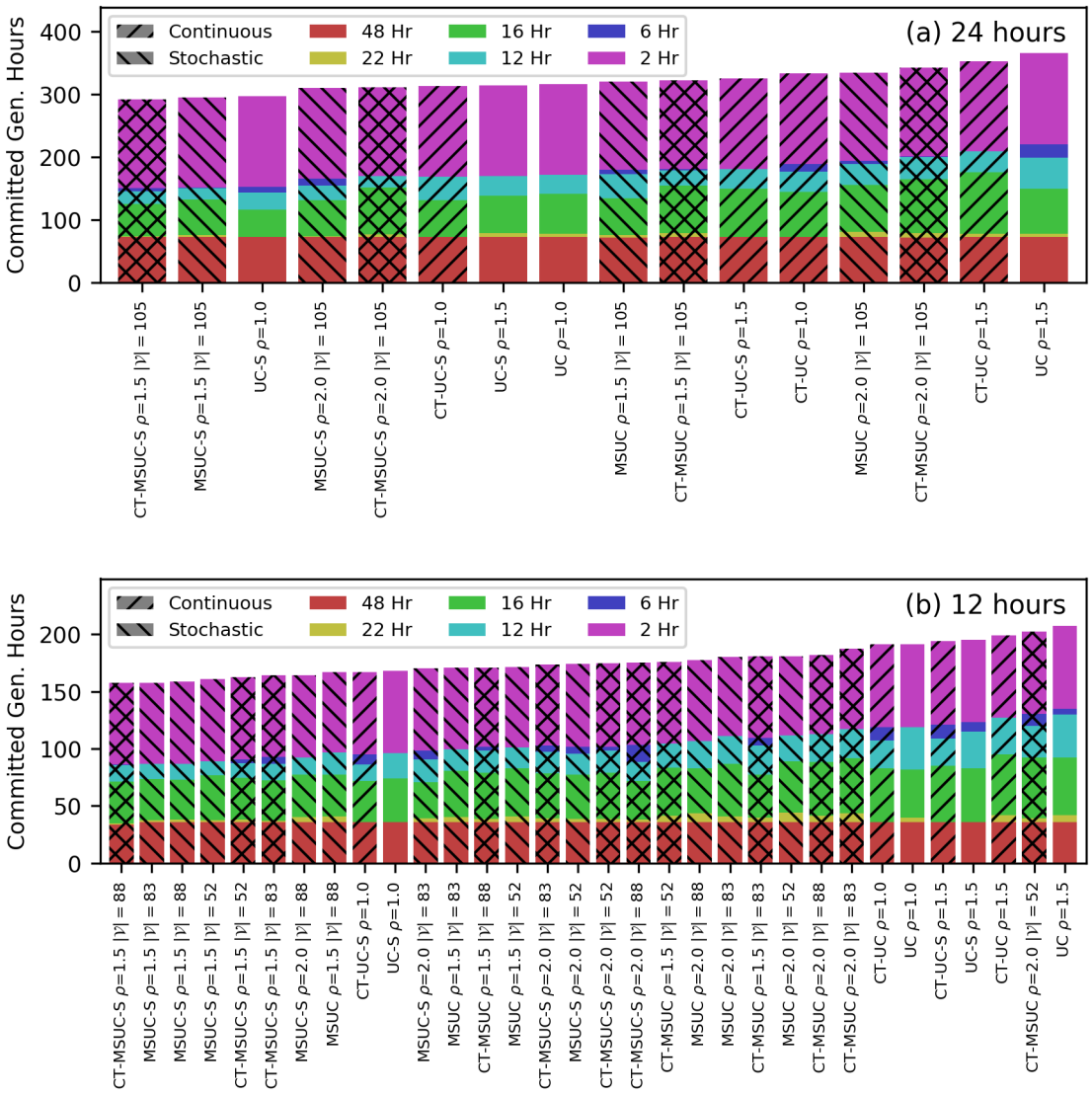


Figure 2.4.6: The number of hours committed for the different solutions of the 24 hour (a) and 12 hour (b) cases, with the colors indicating the flexibility of units through the length of their on/off cycle ($T_{\text{on}} + T_{\text{off}}$).

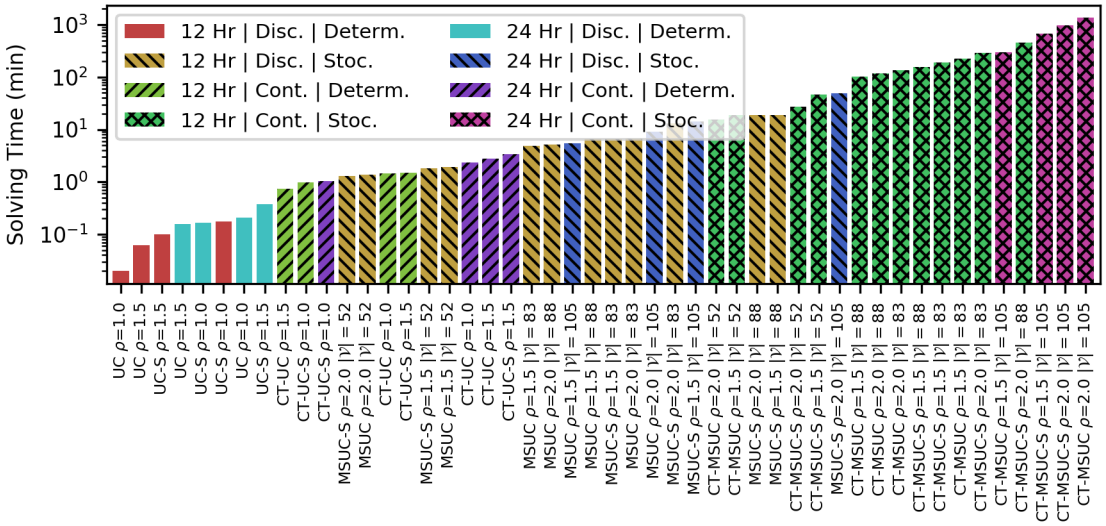


Figure 2.4.7: The required solving time plotted for the various tested solutions. Note that the vertical axis is plotted on a logarithmic scale.

CHAPTER 3

AGGREGATION AND DISAGGREGATION FOR DIRECT LOAD CONTROL ALGORITHMS

3.1 Aggregate Modeling Background

Lets start with some background definitions:

Definition 1 (Prosumer)**.** A prosumer is a generalization of a consumer or a producer, that is someone that either consumes or produces or energy, or is capable of both e.g. batteries or homes with solar generation.

Throughout the remaining chapters the word prosumer will be used when discussing generic entities that could either be producers or consumers. Let $\mathcal{K} = \{0, \ldots, K\}$ be the set containing the $K + 1$ discrete time indexes of equally spaced intervals of duration h in the decision horizon including the times $k_0 + kh$ with $k \in \mathcal{K}$, where $k = 0$ indicates initial conditions. The devices and appliances considered in the context of DR are naturally flexible (responsive), having a variety of different possible profiles of energy consumption.

Definition 2 (Feasible Region for a flexible prosumer)**.** Let $\boldsymbol{p}^i = (p[1], \ldots, p[K]) \in \mathbb{R}^K$ be the samples of the ith generation trajectory (positive for production, negative for consumption) and $\mathcal{P}^i \subseteq \mathbb{R}^K$ its *feasible region*, that is the set of power profiles that can be chosen from for that particular prosumer.

Definition 3 (Convexity)**.** A set is *convex* if a line drawn between any two points in the set lies entirely within the set. Similarly, a *convex function* is a function whose epigraph is a convex set. Further, the *convex hull* of a set $\mathcal{X}$, Conv$(\mathcal{X})$, is the smallest *minimal set* that contains all the points from $\mathcal{X}$.

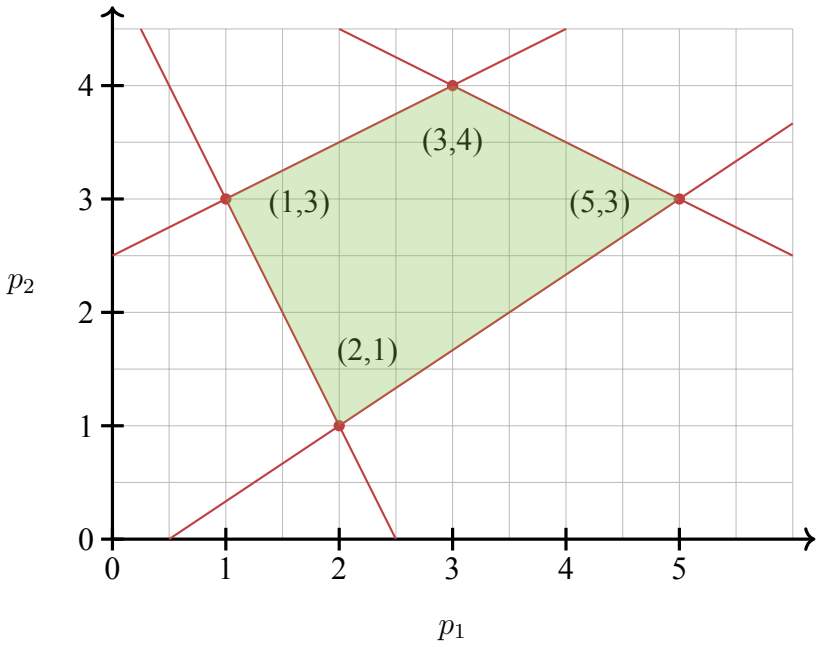


Figure 3.1.1: Convex polytopes showing the vertices and half-space cuts.

A Polytope, visualized in Figure 3.1.1 can be seen as a set with a particular geometric structure:

Definition 4 (Polytope [Boyd and Vandenberghe(2004)])**.** A *polytope* is a geographic shape (a set) with flat edges (facets) in an n dimensional space. A *convex polytope* has the additional property that a line between any two points within the polytope resides entirely within the polytope itself.

The feasible set of all sum load profiles of two or more feasible regions is described by the Minkowski sum of sets.

Definition 5 (Minkowski Sum)**.** A *Minkowski sum*, denoted by $\oplus$, is the set $\mathcal{X}$ containing all possible sums of elements from sets $\mathcal{P}^i$, $i \in \mathcal{I}$. More precisely:

$$\mathcal{X} = \bigoplus_{i \in \mathcal{I}} \mathcal{P}^i = \mathcal{P}^1 \oplus \cdots \oplus \mathcal{P}^N \tag{3.1.1}$$

$$= \left\{ \boldsymbol{x} = \boldsymbol{p}^1 + \cdots + \boldsymbol{p}^N \middle| \boldsymbol{p}^1 \in \mathcal{P}^1, \ldots, \boldsymbol{p}^N \in \mathcal{P}^N \right\} \tag{3.1.2}$$

A Minkowski sum of two sets is visualized in Figure 3.1.

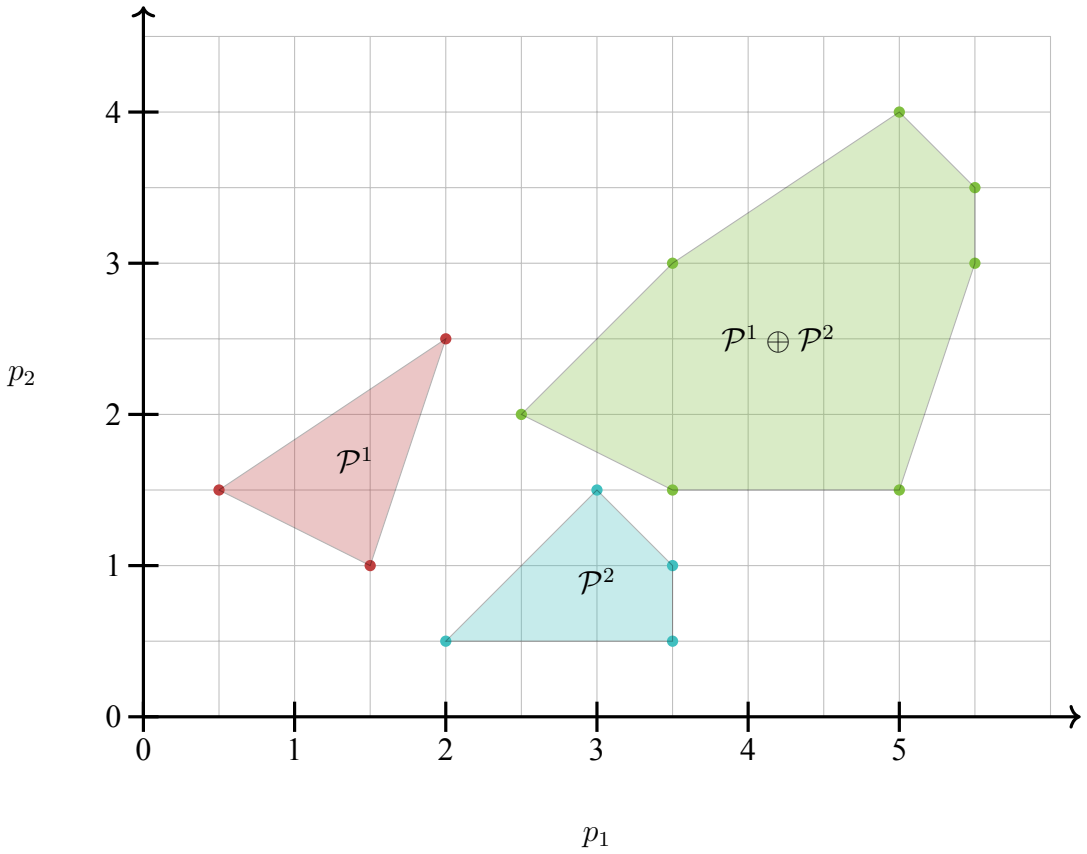


Figure 3.1.2: Minkowski sum of two sets $\mathcal{X} = \mathcal{P}_1 \oplus \mathcal{P}_2$.

Finding the exact Minkowski sum of a large number of sets is hard [Trangbæk *et al.*(2011)], and even though it can be approximated for convex sets [Barot and Taylor(2014)], in this chapter an alternate approach is proposed. Starting from a simple example, one can think of an *ideal* battery i of capacity Z^i with an initial charge of I^i at time k_0. Its *state of energy* at evenly spaced discrete time instants $k_0 + kh$ with $k \in \mathcal{K}$ can be described by the vector $\boldsymbol{\zeta}^i \in \mathbb{R}^{K+1}$ and the set of possible load profiles is:

$$\mathcal{P}^i = \left\{ \boldsymbol{p}^i \,|\, \boldsymbol{p}^i = -\dot{\boldsymbol{\zeta}}^i, \zeta_0^i = I^i, 0 \leq \zeta^i[k] \leq Z^i \; \forall k \in \mathcal{K} \right\} \subset \mathbb{R}^K \tag{3.1.3}$$

where $\boldsymbol{p}$ is the feasible load profile, expressed as the amount of energy [J] consumed in the interval between two consecutive time instants h seconds apart, which can be mapped into the piece wise constant average power profile $\boldsymbol{p}/h$ [W], finally $\dot{\boldsymbol{\zeta}} \in \mathbb{R}^K$ is the vector of

finite differences between consecutive energy states for load i,

$$\dot{\boldsymbol{\zeta}}^i = \left[\zeta^i[1], \ldots, \zeta^i[K]\right] - \left[\zeta^i[0], \ldots, \zeta^i[K-1]\right] \tag{3.1.4}$$

3.2 State-Space Model

Building on existing literature [Alizadeh *et al.*(2015)], a state space quantized into N_u states is considered, while introducing the two mappings:

$$U(\zeta) : \mathbb{R} \to \mathcal{U}, \;\; Z(u) : \mathcal{U} \to \mathbb{R} \tag{3.2.1}$$

where $\mathcal{U} = \{0, \ldots, N_u - 1\} \subset \mathbb{N}$, so that $Z(U(\zeta))$ is the quantized value corresponding to ζ and $U(\zeta)$ is the integer index of the corresponding quantization interval. For batteries that can be charged and discharged at constant rate $\pm\rho$ [W], considering that their state of charge in each unit of time h [s] can remain the same or change by $\pm\rho h$ [J], the quantized values are $Z(u) = \rho \cdot h \cdot u$ [J]. Applying the mapping on an entry by entry basis, the vector $\boldsymbol{u}^i = U(\boldsymbol{\zeta}^i) \in \mathcal{U}^{K+1}$ is introduced, whose entries are the indexes of the quantization intervals of the ith battery state vector $\boldsymbol{\zeta}^i$. Note that $Z(\boldsymbol{u}^i)$ is the quantized version of $\boldsymbol{\zeta}^i$.

Now, (3.1.3) can be expressed in terms of the quantized state space:

$$\mathcal{P}^i = \left\{\boldsymbol{p}^i \,|\, p^i[k] = Z(u^i[k-1]) - Z(u^i[k]), u_0^i = U(I^i), u^i[k] \in \mathcal{U}\; \forall k \in \mathcal{K}\setminus\{0\}\right\} \tag{3.2.2}$$

Going from (3.2.2) to the Minkowski sum for an aggregate of batteries $\mathcal{X}$ with homogeneous capacity Z^i can be done directly, as in [Alizadeh *et al.*(2015)]:

$$\mathcal{X}_{\text{storage}} = \{\boldsymbol{p} \,|\, \boldsymbol{p} = \sum_{u=1}^{N_u}\sum_{v=1}^{N_u}(Z(u)-Z(v))\dot{\boldsymbol{D}}_{u,v}, \boldsymbol{D}_{u,u} = \boldsymbol{0}\,\forall u, \boldsymbol{D} \in \mathbb{N}^{N_u^2\times K}, \sum_{v=1}^{N_u}\dot{\boldsymbol{D}}_{u,v} \leq \boldsymbol{n}_u\,\forall u\} \tag{3.2.3}$$

where $\boldsymbol{D}$ is a tensor whose element $D_{u,v}[k]$ is the number of batteries that have moved from state u to state v up to time k, $\dot{\boldsymbol{D}}[k] = \boldsymbol{D}[k] - \boldsymbol{D}[k-1]$, and $\boldsymbol{n} \in \mathbb{N}^{N_u\times K}$ is a matrix denoting the population in each of the states over time. The dynamics of $\boldsymbol{n}_u$ are as follows:

$$\boldsymbol{n}_u = \sum_{i\in\mathcal{I}}\delta(U(I^i) - u) + \sum_{v=1}^{N_u}\boldsymbol{D}_{v,u} - \sum_{v=1}^{N_u}\boldsymbol{D}_{u,v} \tag{3.2.4}$$

Here, $\mathcal{I}$ denotes the set of all participating batteries and the first sum in (3.2.4) represents the initial state of $\boldsymbol{n}_u$ at k_0[1], while the second and third terms are vectors over time indicating the movement of devices *to* (second term) and *from* (third term) state u for time instants $0 < k \leq K$. It is important to note here that $\boldsymbol{D}$ is a decision variable describing the schedule for future consumption.

3.3 Service/Slack Load Models for Electric Vehicles and Deferrable Appliances

In this section, a unified model for EVs and DAs is built. Electric power is assumed to be normalized with the length of each time step h, so effectively all variables are in terms of energy: $\rho\ [J] = \frac{\rho_{\text{actual power}}\ [W]}{h\ [s]}$. Both EVs and DAs have in common that, not only do they need to receive energy from the grid for a certain *service time*, but incentivized customers may be willing to have devices available to the system for longer than this minimum *service time*, offering flexibility. Examples of this would be to turn on dishwashers or plug in EVs before going to bed, leaving plenty of *slack time* between the service time required to charge the vehicle or run the dishwasher and the *deadline*, at which the dishwasher will be emptied or the EV unplugged. To be precise, define:

Definition 6 (Service Time)**.** The **service time** for a device is the minimum time the device needs to receive the required energy.

Definition 7 (Slack Time)**.** The **slack time** of a device is the difference between the time a device is available in the system, ready to be serviced, and the required service time.

This shared property is used to define a unified model, in which a load state is characterized at each time instant $k \in \mathcal{K}$ by the pair $(u_r, u_s) \in \mathcal{U}_{rs} = \{0, \ldots, N_r - 1\} \times \{0, \ldots, N_s - 1\} \subset \mathbb{N}^2$, where u_r denotes the remaining required service time, while u_s accounts for the

[1]Using the Kronecker delta notation, $\delta(U(I^i) - u) = 1$ only if a battery came with initial state of charge I^i such that $U(I^i) = u$ and zero otherwise.

remaining slack time. The reader is cautioned that in the following $\boldsymbol{u} = (u_r, u_s) \in \mathcal{U}_{rs}$ but also $\boldsymbol{u}^i$ is a vector with entries $\boldsymbol{u}_k^i,\ k \in \mathcal{K}$ representing the evolution of the state of the load in the decision horizon.

In (3.2.4) the first sum, as mentioned, initializes the vector $\boldsymbol{n}_u$. While storage batteries would always be available online, in general flexible devices come and go. To describe the term for EVs, an *arrival process* is introduced, that captures the fact that vehicles are only available for charge after they plug in. More specifically, assume that there are $|\mathcal{I}|$ vehicles arriving in the decision horizon; the ith EV is plugged in (arrives) at a random discrete time k_i^a and is unplugged (departs) at discrete time $k_i^d > k_i^a$. Naturally, EV i requires a certain charge $C^i = Z^i - I^i$ [J] while it is plugged in, and is assumed to be chargeable at discrete evenly spaced power levels[2] $P = [0, \dots, \hat{\rho}] \in \mathbb{R}^{m+1}$. Recalling the meaning of the adimensional state value u introduced before, the ith EV arriving in the system will have initial state coordinates pair (u_r, u_s) equal to:

$$\boldsymbol{u}^i[k < k_i^a] = \mathbf{0},\ \boldsymbol{u}^i[k_i^a] = \left(\frac{\lfloor m\, C^i/\hat{\rho} \rfloor}{m}\ ,\ k_i^d - k_i^a - \frac{\lfloor m\, C^i/\hat{\rho} \rfloor}{m} \right). \tag{3.3.1}$$

Observe that the ith EV starts at a distance $\|\boldsymbol{u}^i[k_i^a]\|_1 = k_i^d - k_i^a$ from the origin $(u_r, u_s) = (0, 0)$, and at every discrete time step $k > k_i^a$ (i.e. in an interval of duration h) reaches a new state $\boldsymbol{u}^i[k+1]$ such that $\|\boldsymbol{u}^i[k+1] - \boldsymbol{u}^i[k]\|_1 = 1$, that is it moves by exactly one position closer to the origin. This means that for $m = 1$ either the service time has decreased, since the EV was charged, or the slack time has decreased, since the EV did not charge. For $m > 1$ a combination can be seen of the two that adds up to one, which can be interpreted as either charging at $\rho < \hat{\rho}$ throughout the period, or charging at full capacity but only for a fraction of the interval length h.

[2] This either assumes a smart charger that supports different power levels (similar to how Tesla allows single or dual chargers), or allows the model to support full power for a fraction of h, giving more resolution and less quantization errors.

To compute the Minkowski sum of $\mathcal{P}^i$ for EVs, define the dynamics of the population state matrix $\boldsymbol{n}$. For this, the matrix $\boldsymbol{a}$ for the arrival process must be introduced. Specifically, in each possible state pair $\boldsymbol{u} = (u_r, u_s)$ the arrival vector in $\mathbb{N}^K$ is:

$$\boldsymbol{a}_{\boldsymbol{u}}[k] = \sum_{i \in \mathcal{I}} \delta(\boldsymbol{u}^i[k_i^a] - \boldsymbol{u}) \cdot \mathrm{u}(k - k_i^a) \tag{3.3.2}$$

where u is the unit step function and again $\mathcal{I}$ denotes the set of all possible participating devices. Formulating the unified model with a state space of dimension $N_r \times N_s$, with all possible coordinates as members of $\boldsymbol{\mathcal{U}}$:

$$\begin{aligned} \mathcal{X} = \big\{ \boldsymbol{p} \,|\, \boldsymbol{p} = -\sum_{\boldsymbol{u} \in \boldsymbol{\mathcal{U}}} \sum_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}(\boldsymbol{u})} R(\boldsymbol{u}, \boldsymbol{v}) \dot{\boldsymbol{D}}_{\boldsymbol{u},\boldsymbol{v}},\ \boldsymbol{D}_{\boldsymbol{u},\boldsymbol{u}} = \boldsymbol{0}\ \forall \boldsymbol{u} \in \boldsymbol{\mathcal{U}}, \\ \boldsymbol{D} \in \mathbb{N}^{(N_r \times N_s)^2 \times K},\ \sum_{\boldsymbol{v} \in \mathcal{U}} \dot{\boldsymbol{D}}_{\boldsymbol{u},\boldsymbol{v}} = \boldsymbol{n}_{\boldsymbol{u}}\ \forall \boldsymbol{u} \in \boldsymbol{\mathcal{U}} \big\} \end{aligned} \tag{3.3.3}$$

where the state population tensor $\boldsymbol{n} \in \mathbb{N}^{N_r \times N_s \times K}$ dynamics are:

$$\boldsymbol{n}_{\boldsymbol{u}} = \boldsymbol{a}_{\boldsymbol{u}} + \sum_{\boldsymbol{v} \in \boldsymbol{\mathcal{U}}} \boldsymbol{D}_{\boldsymbol{v},\boldsymbol{u}} - \sum_{\boldsymbol{v} \in \boldsymbol{\mathcal{U}}} \boldsymbol{D}_{\boldsymbol{u},\boldsymbol{v}}. \tag{3.3.4}$$

For EVs, the states energy values are uniformly spaced, thus:

$$R(\boldsymbol{u}, \boldsymbol{v}) = \hat{\rho} \cdot h(u_r - v_r)/m. \tag{3.3.5}$$

The set $\boldsymbol{\mathcal{V}}(\boldsymbol{u})$ of indexes in the summation over $\boldsymbol{v}$ denotes the possible moves from $\boldsymbol{u}$, again for EVs:

$$\mathcal{V}_{\mathrm{EV}}(\boldsymbol{u}) = \big\{ \boldsymbol{v} \,|\, \|\boldsymbol{v} - \boldsymbol{u}\|_1 = \min(\|\boldsymbol{u}\|_1, 1), (\boldsymbol{u} - \boldsymbol{v})_r \geq 0, (\boldsymbol{u} - \boldsymbol{v})_s \geq 0, v_r, v_s \geq 0 \big\} \tag{3.3.6}$$

where the equality constraint only considers movements to states at distance 1 unless it is close to the origin, and the inequality constraints ensure that it stays in the upper right quadrant of the state space and can only move towards either of the two axis.

A sample initial state space for EVs is depicted in Figure 3.3.1 where the diameter of the circles indicates a states population. Note that the model can incorporate arbitrary EV

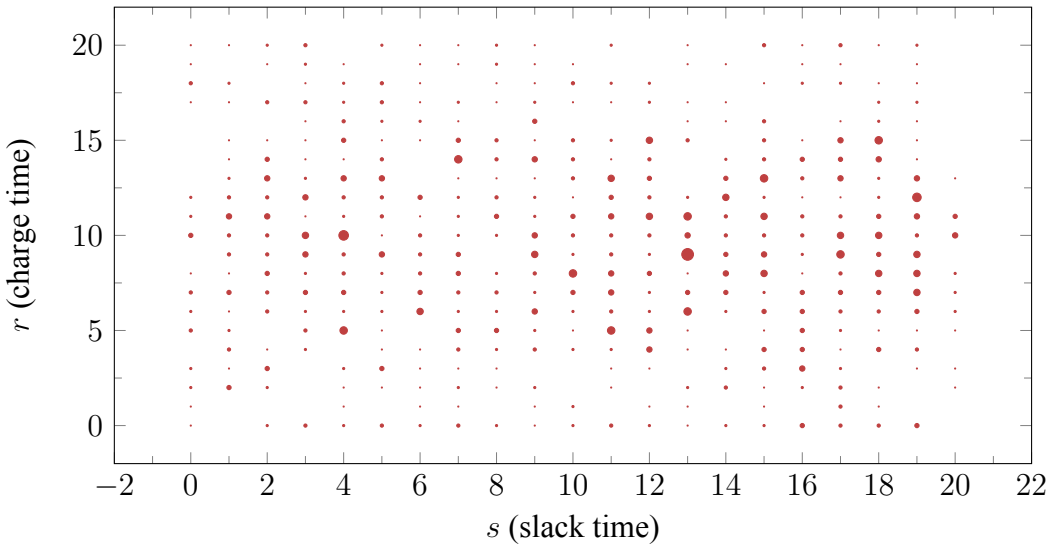


Figure 3.3.1: Sample state space for EVs. The state-space is dimensionless, with coordinate $\boldsymbol{u}$ corresponding to $\hat{\rho} \cdot h \cdot u_r$ [J] of energy requirement and $h \cdot u_s$ [s] of slack time.

loads that are interruptible with any initial state of charge or capacity for the EV battery in a single aggregate space, unlike [Alizadeh *et al.*(2015)]. Only EVs with different charging levels need to be separated in a different state space.

Deferrable appliances such as washers, dryers and dishwashers have similar characteristics of being turned on (not started) at k_i^a with a requirement of being finished by k_i^d. Further they have a fixed power profile that can not be paused and is described by the vector $\boldsymbol{p} = [p_1, \ldots, p_L]$, giving a slack time of $k_i^d - k_i^a - L$ and a service time L. As only devices with similar $\boldsymbol{p}$ can be aggregated, they all arrive requiring exactly L of service time, providing the initial coordinates for the two dimensional state:

$$\boldsymbol{u}^i[k < k^a] = 0, \;\; \boldsymbol{u}^i[k^a] = (L \,,\; k_i^d - k_i^a - L) \tag{3.3.7}$$

For DAs the Minkowski sum aggregate is the same as for EVs (3.3.3) except that $R(\cdot)$ depends on $\boldsymbol{p}$, seeing as once a device has been started ($u_r < L$) it can not be interrupted

and this behavior must be captured by $\mathcal{V}_{\text{DA}}(\boldsymbol{u})$:

$$R(\boldsymbol{u}, \boldsymbol{v}) = \sum_{w=v_r}^{u_r-1} p_{L-w+1} \tag{3.3.8}$$

$$\mathcal{V}_{\text{DA}}(\boldsymbol{u}) = \boldsymbol{\mathcal{V}}_{\text{EV}}(\boldsymbol{u}) \cap \left\{\boldsymbol{v} \,|(\boldsymbol{v}-\boldsymbol{u})_r = \min(u_r, 1)\ \forall \boldsymbol{u} : u_r < L\right\}$$

3.3.1 Reserve Capacity

Besides the set of possible power profiles $\mathcal{X}$, there are other metrics to evaluate the state-space. The future energy requirement of an aggregate is:

$$\boldsymbol{E} = -\sum_{\boldsymbol{u}\in\boldsymbol{\mathcal{U}}} R(\boldsymbol{u}, \boldsymbol{0})\boldsymbol{n}_{\boldsymbol{u}}. \tag{3.3.9}$$

Another noteworthy quantity is the aggregated slack that is stored in the system. This is analogous to (3.3.9):

$$\boldsymbol{S} = \sum_{\boldsymbol{u}\in\boldsymbol{\mathcal{U}}} u_s \boldsymbol{n}_{\boldsymbol{u}}. \tag{3.3.10}$$

As for other metrics, it may be important to know how fast one can ramp down or up for a single period given a schedule $\boldsymbol{D}$:

$$\boldsymbol{p}^+ = \sum_{(\boldsymbol{u},\boldsymbol{v})\in\boldsymbol{\mathcal{W}}^+} R\big(\boldsymbol{u}, \boldsymbol{v}\big)\boldsymbol{n}_{\boldsymbol{u}} \tag{3.3.11}$$

$$\mathcal{W}^+ = \{(\boldsymbol{u}, \boldsymbol{v}) \,|\, \boldsymbol{u} \in \mathcal{U}, v_r = \max(u_r - 1, 0), v_s = \max(u_s + u_r - v_r - 1, 0)\} \tag{3.3.12}$$

$$\boldsymbol{p}^- = \sum_{(\boldsymbol{u},\boldsymbol{v})\in\boldsymbol{\mathcal{W}}^-} R(\boldsymbol{u}, \boldsymbol{v})\boldsymbol{n}_{\boldsymbol{u}} \tag{3.3.13}$$

$$\mathcal{W}^- = \{(\boldsymbol{u}, \boldsymbol{v}) \,|\, \boldsymbol{u} \in \mathcal{U}, v_s = \max(u_s - 1, 0),\ v_r = \max(u_r + u_s - v_s - 1, 0)\} \tag{3.3.14}$$

The sets $\mathcal{W}$ denote the most favorable moves from any point $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$ with respect to either of the objectives.

3.4 Control Strategies

With the state-space model in place, how does one schedule consumption and come up with a $\boldsymbol{D}$? The answer is tightly coupled with the aggregators goal, whether it is to counteract

stochasticity of other loads or generation, provide ancillary services or to get cheaper energy on the market. To start off, looking closer at Figure 3.3.1 two *extreme* strategies (or policies) can be described:

Policy 1 (energy first): Moving all the dots along the r axis (down on the figure) until they reach the s axis and are forced to move along the s dimension. Here, the devices consume energy as fast as they can, similar to inflexible loads.

Policy 2 (slack first): The complementary strategy is to first move all devices along the s axis, consuming slack time and delaying service as long as possible.

Policy 3 (prioritize reserves): A strategy with the goal of maximizing short-term reserve power potential, positive or negative. More precisely, it aims to find the load curve $\boldsymbol{p}$ that maximizes the minimum $\boldsymbol{p}^+ - \boldsymbol{p}$ and $\boldsymbol{p} - \boldsymbol{p}^-$.

Figure 3.4.1 shows for a certain initial state, the resulting load, stored energy and slack for the three policies. It is clear that policies 1 and 2 bound policy 3 w.r.t. stored energy and slack, and in fact, by design, they bound any other feasible policy. The edge strategies (1 and 2) are simple and can be expressed as a linear transform:

$$\boldsymbol{n}[k] = \boldsymbol{\xi}(\boldsymbol{n}[k-1]) = \boldsymbol{A}_0\boldsymbol{n}[k-1]\boldsymbol{A}_1 + \boldsymbol{A}_2\boldsymbol{n}[k-1]\boldsymbol{A}_3 + \boldsymbol{A}_4\boldsymbol{n}[k-1]\boldsymbol{A}_5 \qquad (3.4.1)$$

where $\boldsymbol{A}_i$ are simple shifting matrices (all zeros except for ones on upper or lower diagonal) and picking matrices (all zeros except for $\{\boldsymbol{A}_i\}_{1,1} = 1$). A property of the vec operator applied to the product of three matrices, $\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{A})$, allows defining $\boldsymbol{A} = (\boldsymbol{A}_2^\top \otimes \boldsymbol{A}_1 + \boldsymbol{A}_4^\top \otimes \boldsymbol{A}_3 + \boldsymbol{A}_6^\top \otimes \boldsymbol{A}_5)$ where $\otimes$ is the Kronecker product. Thus, one can express (3.4.1) as a single linear dynamical model $\text{vec}(\boldsymbol{n}[k-1]) = \boldsymbol{A}\,\text{vec}(\boldsymbol{n}[k-1])$. Policy 3 is formulated as a mixed-integer linear program and solved with Gurobi [Gurobi Optimization(2015)].

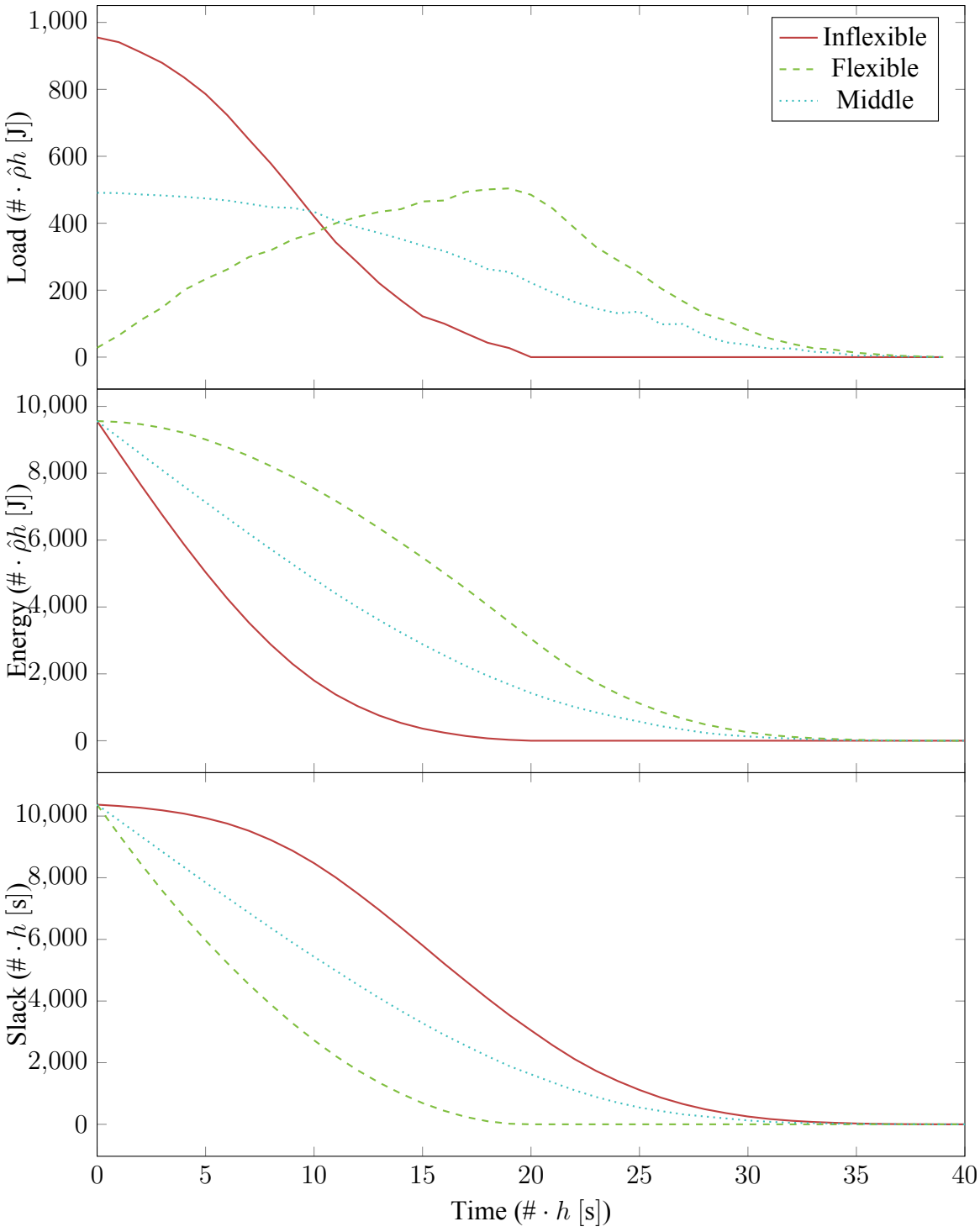


Figure 3.4.1: Sample strategy outcomes from the initial state in Fig. 3.3.1.

The take away message from Figure 3.4.1 is that a DR population is similar to storage; depending on the strategy that is chosen, different amounts of energy are "chosen" for future consumption. Thus, by planning to follow a certain schedule that leads down a certain trajectory with regard to the storage level, the schedule can be altered to reduce or increase consumption momentarily, as long as the trajectory remains within the bounds of the edge strategies.

While previously analyzing the possibilities based on a deterministic initial state, the arrivals captured by $\boldsymbol{a}$ are randomly replenishing the DR resources, keeping $\boldsymbol{E}$ and $\boldsymbol{S}$ from converging to zero. Predictions regarding the future arrivals allow for estimating the future storage potential less conservatively than assuming no future arrivals. In [Alizadeh *et al.*(2014b)] (Sec. V-E) it was found that the arrivals of cars that plug in at home fit the statistics of a non-stationary Poisson arrival process, which is consistent with the intuition that events of cars plugging in are independent. Denoting by $\boldsymbol{\lambda}$ the vector of expected number of arrivals, the random number of arrivals at quantized time t is $\dot{a}[k] = \sum_{\boldsymbol{u}\in\boldsymbol{U}}(\boldsymbol{a}_{\boldsymbol{u}}[k] - \boldsymbol{a}_{\boldsymbol{u}}[k-1]) \sim \text{Pois}(\lambda[k])$. A well known result in probability theory [Papoulis and Pillai(2002)] is that sampling randomly a Poisson process gives a Poisson process, therefore it can be stated that $\dot{\boldsymbol{a}}_{\boldsymbol{u}}[k] \sim \text{Pois}(\lambda[k] f_{\boldsymbol{U}}(\boldsymbol{u}, k))$. Thus, if $\dot{\boldsymbol{a}}[k] \in \mathcal{N}^{N_r \times N_s}$ denotes the matrix of state arrivals at time k, it follows that:

$$\boldsymbol{n}[k] = \dot{\boldsymbol{a}}[k] + \boldsymbol{\xi}_k(\boldsymbol{n}[k-1]) \tag{3.4.2}$$

but the two extreme policies 1-2 both correspond to linear dynamics $\text{vec}(\boldsymbol{n}[k]) = \text{vec}(\dot{\boldsymbol{a}}[k]) + \boldsymbol{A}\,\text{vec}(\boldsymbol{n}[k-1])$ and the state occupancy at time k is:

$$\text{vec}(\boldsymbol{n}[k]) = \text{vec}(\dot{\boldsymbol{a}}[k]) + \boldsymbol{A}\,\text{vec}(\dot{\boldsymbol{a}}[k-1]) + \cdots + \boldsymbol{A}^T\,\text{vec}(\dot{\boldsymbol{a}}[0]) + \boldsymbol{A}^T\,\text{vec}(\boldsymbol{n}[0]) \tag{3.4.3}$$

where $\boldsymbol{A}^K$ is $\boldsymbol{A}$ to the power of K.

To determine the DR Reserve Capacity storage potential, (3.3.9) can be rewritten:

$$\boldsymbol{E}[k] = \boldsymbol{e}^\intercal \cdot \boldsymbol{n}[k] \cdot \boldsymbol{1}_{N_s \times 1} = \text{vec}(\boldsymbol{e} \cdot \boldsymbol{1}_{1\times N_s})^\intercal \cdot \text{vec}(\boldsymbol{n}[k]) \tag{3.4.4}$$

where $\boldsymbol{e} = \boldsymbol{R}([0, \ldots, N_r - 1], 0)$ is a column vector mapping the state coordinates r to energy, and $\mathbf{1}$ is a vector of ones. By plugging (3.4.3) into (3.4.4) one obtains:

$$\boldsymbol{E}[k] = \sum_{i=0}^{k} \boldsymbol{v}_i \operatorname{vec}(\dot{\boldsymbol{a}}[k-i]) + \boldsymbol{v}_k \operatorname{vec}(\boldsymbol{n}[0]) \tag{3.4.5}$$

where $\boldsymbol{v}_i = \operatorname{vec}\left(\boldsymbol{e} \cdot \mathbf{1}_{1\times N_s}\right)^{\intercal} \cdot \boldsymbol{A}^i \in \mathbb{R}^{1\times N_r N_s}$. From this the expected value and variance can be derived:

$$\mathbb{E}\big[\boldsymbol{E}[k]\big] = \sum_{i=0}^{k} \boldsymbol{v}_i \mathbb{E}\big[\operatorname{vec}(\dot{\boldsymbol{a}}[k-i])\big] + \boldsymbol{v}_k \operatorname{vec}(\boldsymbol{n}[0]) \tag{3.4.6}$$

$$\operatorname{var}\big(\boldsymbol{E}[k]\big) = \sum_{i=0}^{k} \boldsymbol{v}_i \boldsymbol{C}(\operatorname{vec}(\dot{\boldsymbol{a}}[k-i]))\boldsymbol{v}_i^{\intercal} \tag{3.4.7}$$

where $\boldsymbol{C}(\operatorname{vec}(\dot{\boldsymbol{a}}[k-i]))$ is the covariance matrix of $\operatorname{vec}(\dot{\boldsymbol{a}}[k-i])$, which is only non-zero on the diagonal as the arrivals are independent. Knowing the mean and the variance, the central limit theorem can be applied to approximate the distribution of $\boldsymbol{E}[k] \sim \mathcal{N}\big(\mathbb{E}[\boldsymbol{E}[k]], \operatorname{var}(\boldsymbol{E}[k])\big)$. Based on this approximation and the assumption that the arrival rate $\boldsymbol{\lambda}$ is known, given a certain confidence parameter γ, the energy level of both edge strategies can be predicted, thus estimating the feasible future energy "storage" region, as shown in Figure 3.4.2.

This forecast can then be useful for several applications, such as estimating the amount of stochastic variability that can absorb by a pool of DR devices, attempting to sell this storage capacity for regulating services or simply use it to reduce the amount of energy purchases at peak prices by shifting them to off-peak hours.

3.5 Thermostatically Controlled Load Modeling

Modeling TCLs is a bit more intricate than e.g. EVs as TCLs are subject to the additional uncertainty of the ambient temperature. In this section a state-space TCL model is presented attempting to simultaneously allow accurate mapping onto electric load without sacrificing

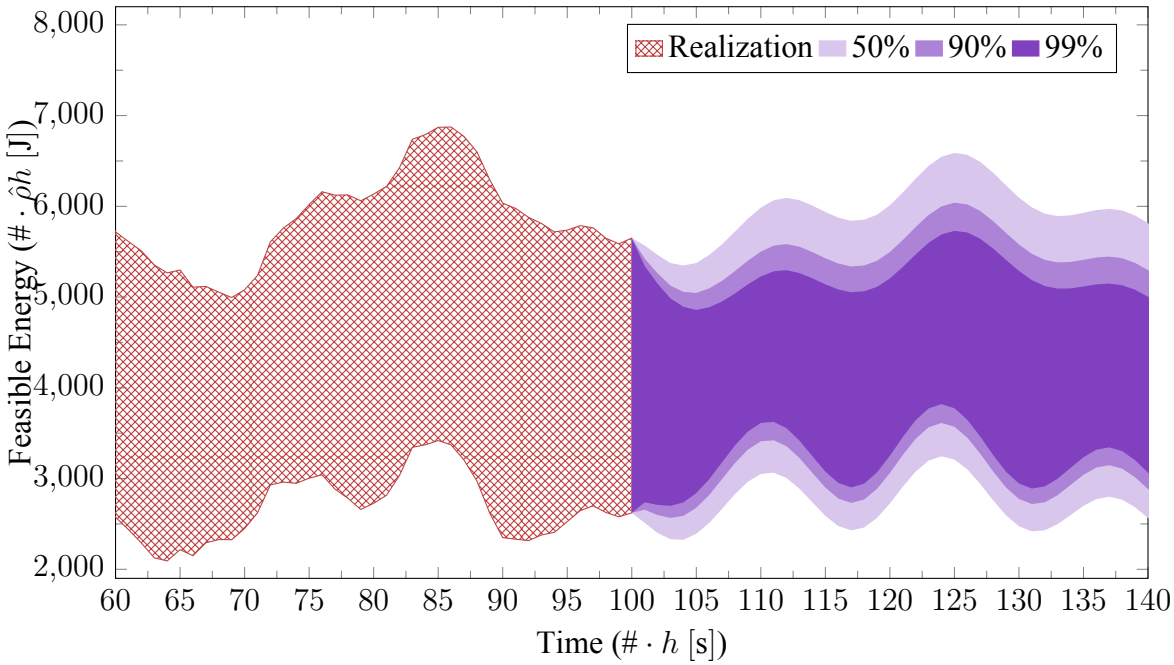


Figure 3.4.2: The (negative) energy stored in the system, with the area being bounded above and below by Policy 1 and 2 respectively. The crosshatched area ($60 \leq k \leq 100$) denotes a realization of an arrival process, while the solid region ($k > 100$) is the estimated future behavior of these bounds. The darkest area is bounded by 99% of the probability mass of the two policies, based on the normal distribution approximation with mean and variance (3.4.6)-(3.4.7). The lighter areas in succession denote the 90% and 50% regions.

what makes state-space models useful, e.g. predictability w.r.t. computational complexity and ease of inclusion in optimization problems.

Following most of the literature, this derivation starts from the assumption that the temperature dynamics of a heat-pump based TCL can be modeled as a first-order differential equation:

$$C\dot{\theta}(t) = \frac{\theta_{\text{o}}(t) - \theta(t)}{R} + m(t)\eta P - \frac{\hat{\varepsilon}(t)}{R} \tag{3.5.1}$$

with R being thermal resistance, C thermal capacitance, θ the inside temperature, θ_{o} the outside temperature, η the efficiency of the heat-pump, $m(t) \in \{-1, 0, +1\}$ the operational mode of the heat pump, P its continuous electrical power rating[3] and ε denoting any noise. The authors of [Mathieu *et al.*(2012)] have surveyed common values for such models. Extensions to this model include thermal mass temperatures (see e.g. [Zhang *et al.*(2013)]), adding an additional state dimension. Even though the proposed formulation could incorporate thermal mass, in order to streamline the presentation, the formulation will be based off of the simpler first order model (3.5.1).

3.5.1 Discrete-Time Model

Given a continuous time signal $s(t)$ and a sampling period h, in this chapter the convention is to denote samples as $s[k] \triangleq s(kh)$, and their finite difference and mid-point respectively, as:

$$\dot{s}[k] \triangleq s[k] - s[k-1]\ , \quad \overline{s}[k] \triangleq \frac{s[k] + s[k-1]}{2}. \tag{3.5.2}$$

[3]Water-heaters can be described using the same principles, with an additional energy loss component describing the hot water being replaced by cold water. However, in this dissertation, focus will remain on heat-pump based TCLs, primarily because they are more dependent on external temperatures than water-boilers, which means that handling their response requires more care.

From (3.5.1) two main observations are made. First, the energy stored in the thermal capacitance can be written w.r.t. the reference temperature $\theta_{\mathrm{r}}(t)$ as:

$$\mathfrak{e}(t) = \frac{(\theta(t) - \theta_{\mathrm{r}}(t))C}{\eta}. \tag{3.5.3}$$

The second observation made is from (3.5.1); the rate at which energy is gained (or lost) from the outside environment by a TCL circuit at any given time is $\frac{\theta_{\mathrm{o}}(t)-\theta(t)-\hat{\varepsilon}(t)}{\eta R}$. Unlike most of the literature which uses θ to describe the individual state, and track the operational mode (on/off) of the heat-pump $m(t)$ defined in (3.5.1), the joint energy and reference temperature $(\mathfrak{e}[k], \theta_{\mathrm{r}}[k])$ will be used as the individual TCL state. Contrary to previous models, this approach, as seen shortly, yields an explicit relationship between the electric load associated to changing state and the future random ambient temperature $\theta_{\mathrm{o}}[k]$. The energy $\Lambda[k]$ required by a TCL for a transition $\theta[k-1] \to \theta[k]$ ($\dot{\theta}[k]$) during the k-th interval $(k-1)h \leq t < kh$ can, using (3.5.3), be mapped into a change in state $\dot{\mathfrak{e}}[k]$ and reference temperature $\dot{\theta}_{\mathrm{r}}[k]$:

$$\begin{aligned} \Lambda[k] &= \frac{C(\theta[k] - \theta[k-1])}{\eta} + \frac{1}{R\eta}\int_{(k-1)h}^{kh} [\theta(t) - \theta_{\mathrm{o}}(t) + \hat{\varepsilon}(t)]\, dt \\ &= \dot{\mathfrak{e}}[k] + \frac{C\dot{\theta}_{\mathrm{r}}[k]}{\eta} + \frac{1}{R\eta}\int_{(k-1)h}^{kh} [\theta(t) - \theta_{\mathrm{o}}(t) + \hat{\varepsilon}(t)]\, dt. \end{aligned} \tag{3.5.4}$$

The first half of (3.5.4) captures the energy spent transitioning between different energy states, while the second half describes the energy gain over the thermal resistance due to the temperature difference. Note that since the integration is over an interval of length h, Λ is defined in terms of energy, like the stored energy $\mathfrak{e}$. Also, while $\Lambda[k]$ can be either positive or negative, indicating the direction of pumping (heating or cooling), the electric energy consumption is equal to $|\Lambda[k]|$ and it is always positive. From (3.5.4), define:

Proposition 1. Assume that, during the k-th interval:

1. the outside temperature is a non-stationary discrete time random process, i.e. $\theta_{\mathrm{o}}((k-1)h) \approx \theta_{\mathrm{o}}[k]$;

2. the reference temperature can only change between intervals, i.e. $\theta_{\mathrm{r}}(t) = \theta_{\mathrm{r}}[k]$ for $(k-1)h \leq t < kh$,

3. the leakage can be approximated using the average temperature, $\theta(t) \approx \overline{\theta}[k] = (\theta[k] + \theta[k-1])/2$, for $(k-1)h \leq t < kh$

4. the random error: $\varepsilon[k] \approx (R\eta)^{-1} \int_{(k-1)h}^{kh} \hat{\varepsilon}(t)\, dt$

then, the estimated energy cost is $|\Lambda[k]|$ where:

Approximate TCL energy expenditure $|\Lambda[k]|$*:*

$$\Lambda[k] \approx \dot{\mathfrak{e}}[k] + \frac{h}{RC}\overline{\mathfrak{e}}[k] + \frac{C}{\eta}\dot{\theta}_{\mathrm{r}}[k] + \frac{h}{R\eta}\Big(\overline{\theta}_{\mathrm{r}}[k] - \theta_{\mathrm{o}}[k]\Big) + \varepsilon[k]. \tag{3.5.5}$$

Proof: Using (3.5.3) in (3.5.4) the derivation is straightforward:

$$\begin{aligned}
\Lambda[k] &= \dot{\mathfrak{e}}[k] + \frac{C}{\eta}\dot{\theta}_{\mathrm{r}}[k] + \frac{1}{R\eta}\int_{(k-1)h}^{kh} [\theta(t) - \theta_{\mathrm{o}}(t) + \hat{\varepsilon}(t)]\, dt \\
&\approx \dot{\mathfrak{e}}[k] + \frac{C}{\eta}\dot{\theta}_{\mathrm{r}}[k] + \frac{h}{R\eta}\left(\overline{\theta}[k] - \theta_{\mathrm{o}}[k]\right) + \varepsilon[k] \\
&\approx \dot{\mathfrak{e}}[k] + \frac{C}{\eta}\dot{\theta}_{\mathrm{r}}[k] + \frac{h}{R\eta}\left(\frac{\eta}{C}\overline{\mathfrak{e}}[k] + \overline{\theta}_{\mathrm{r}}[k] - \theta_{\mathrm{o}}[k]\right) + \varepsilon[k]
\end{aligned}$$

which, grouping the terms, leads to (3.5.5). □

3.5.2 Responsive State-Space Model

In the model it is assumed that an Aggregator provides the users participating in the direct load control program with a choice of:

1. A limited set of reference temperatures $\mathcal{S}_{\theta_{\mathrm{r}}}$ and of possible transitions $\mathcal{E}_{\theta_{\mathrm{r}}} \subseteq \mathcal{S}_{\theta_{\mathrm{r}}} \times \mathcal{S}_{\theta_{\mathrm{r}}}$. For simplicity, it is assumed that the set contains contiguous values and allow only transitions between consecutive values, i.e.

$$\mathcal{S}_{\theta_{\mathrm{r}}} = \{\theta_{\mathrm{r}} | \theta_{\mathrm{r}} = i\Delta_\theta, i = I^{\min}, I^{\min}+1, \ldots, I^{\max}\} \tag{3.5.6}$$

$$\mathcal{E}_{\theta_{\mathrm{r}}} = \{\dot{\theta}_{\mathrm{r}} | \dot{\theta}_{\mathrm{r}} = (i-j)\Delta_\theta, |i-j| \leq 1, (i,j) \in \mathcal{S}_{\theta_{\mathrm{r}}}\} \tag{3.5.7}$$

Note that the sequence $\theta_{\mathrm{r}}[k]$ is assumed to be *user controlled*, and that a smart thermostat can translate the individual wish for a sharp reference temperature change to a gradual one, honoring $\mathcal{E}_{\theta_{\mathrm{r}}}$.

2. A temperature dead-band B, such that $|\theta(t) - \theta_{\mathrm{r}}(t)| \leq B$; correspondingly the energy states are bounded by:

$$|\mathfrak{e}(t)| \leq E \triangleq BC\eta^{-1}. \tag{3.5.8}$$

In controlling energy transitions, the energy state-space is discretized with a step-size of $\Delta_e = E/(2U+1)$, where U is the number of steps the energy can make upward or downward, as the state-space is assumed to be symmetric (extending equally far below and above the reference temperature that corresponds to $\mathfrak{e} = 0$). In addition, the transitions are assumed to only cover a certain number of contiguous energy states. The sets of states and state transitions are, respectively:

$$\mathcal{S}_{\mathfrak{e}} = \{\mathfrak{e} | \mathfrak{e} = \Delta_{\mathfrak{e}} u, u = 0, \pm 1, \ldots, \pm U\} \tag{3.5.9}$$

$$\mathcal{E}_{\mathfrak{e}} = \{(\dot{\mathfrak{e}}, \bar{\mathfrak{e}}) |\ |\Lambda[k]| \leq hP\} \tag{3.5.10}$$

where $\mathcal{E}_{\mathfrak{e}}$ is limited by the electrical power rating P.

Note that some choices of $\dot{\theta}_{\mathrm{r}} \neq 0$ may not be compatible with $\mathcal{E}_{\mathfrak{e}}$ since, clearly, the possible transitions for energy and reference temperature are both constrained by $|\Lambda[k]| \leq hP$, and the desired change $\dot{\theta}_{\mathrm{r}}$ may be impossible to achieve in one interval duration h, i.e. the set $\mathcal{E}_{\mathfrak{e}}$ as-is may be empty. To circumvent this limitation and allowing individuals to freely choose their reference temperature, an Aggregator is tracks the *target values* for reference temperature and stored energy, and correct the load response model to account for the delay required to meet the reference temperature change. As the reference temperature changes are considered to be relatively rare, the assumption is that eventually the actual temperature will catch up with the target value. This implies that a more general expression to (3.5.5) is

required, where the restriction $\mathcal{E}_{\mathfrak{e}}$ only applies when $\dot{\theta}_{\mathrm{r}}[k] = 0$ and is relaxed for $\dot{\theta}_{\mathrm{r}}[k] \neq 0$, while still capturing the effect of such transitions through an energy cost term that lasts multiple periods. This point will be re-visited after a few useful definitions.

3.5.2.1 Energy Cost Model

The following vectors of normalized target energy and reference temperatures are defined:

$$\boldsymbol{e}[k] = \frac{1}{\Delta_e}(e[k-1], e[k])^{\intercal}, \boldsymbol{\theta}_{\mathrm{r}}[k] = \frac{1}{\Delta_\theta}(\theta_{\mathrm{r}}[k-1], \theta_{\mathrm{r}}[k])^{\intercal} \tag{3.5.11}$$

where the true stored energy $\mathfrak{e}[k] \approx e[k]$. Clearly:

$$\begin{pmatrix} \dot{e}[k] \\ \overline{e}[k] \end{pmatrix} = \Delta_e \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \boldsymbol{e}[k] \ , \tag{3.5.12}$$

and similarly for $(\dot{\theta}_{\mathrm{r}}[k], \overline{\theta}_{\mathrm{r}}[k])^{\intercal}$. Further define:

$$\boldsymbol{T}[k] = (\boldsymbol{e}[k], \boldsymbol{\theta}_{\mathrm{r}}[k]), \tag{3.5.13}$$

refer to $\boldsymbol{T}[k]$ as the TCL state transition. To rewrite compactly (3.5.5), the following matrix is introduced:

$$\boldsymbol{\Gamma} = \begin{pmatrix} -1 & 1/2 \\ 1 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & \frac{C}{\eta} \\ \frac{h}{RC} & \frac{h}{R\eta} \end{pmatrix} \begin{pmatrix} \Delta_e & 0 \\ 0 & \Delta_\theta \end{pmatrix} \tag{3.5.14}$$

and also use the following notation:

$$\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_e, \boldsymbol{\gamma}_{\theta_{\mathrm{r}}}) \quad , \quad a = \frac{h}{R\eta}. \tag{3.5.15}$$

Now consider the case of a TCL having $\dot{\theta}_{\mathrm{r}}[k] = 0$. Simple algebra shows that (3.5.5) can be rearranged as:

$$\Lambda[k] = \boldsymbol{\gamma}_e^{\intercal}\boldsymbol{e}[k] + \boldsymbol{\gamma}_{\theta_{\mathrm{r}}}^{\intercal}\boldsymbol{\theta}_{\mathrm{r}}[k] - a\theta_{\mathrm{o}}[k] + \varepsilon[k] \tag{3.5.16}$$

$$= \mathrm{Tr}(\boldsymbol{\Gamma}^{\intercal}\boldsymbol{T}[k]) - a\theta_{\mathrm{o}}[k] + \varepsilon[k]. \tag{3.5.17}$$

The expression (3.5.16) breaks the energy spent in four terms:

1. the first term depends on $\boldsymbol{e}[k]$ and is subject to the control action of the Aggregator;

2. the second term, function of $\boldsymbol{\theta}_{\mathrm{r}}[k]$, is the operating mode chosen by the user;

3. the third term depends on the random $\theta_{\mathrm{o}}[k]$, and is due to Mother Nature;

4. the fourth is the random error $\varepsilon[k]$.

For $\dot{\theta}_{\mathrm{r}}[k] \neq 0$, the formulation assumes that TCLs changing their reference temperature can spread the energy expenditure for the transition over Q_k periods, as necessary to fulfill the total energy requirement, with each step not exceeding the power rating of P times h. Hence, the energy cost at time k for a TCL can, in general, be expressed as follows[4]:

Energy spent for transition $\boldsymbol{T}[k]$ at temperature $\theta_o[k]$:

$$|\Lambda[k]| = \left| \sum_{q=0}^{Q_k} H[k-q, q] + \varepsilon[k] \right| \tag{3.5.18}$$

$$H[k,q] = \begin{cases} Ph & 0 \leq q < Q_k \\ |\mathrm{Tr}(\boldsymbol{\Gamma}^{\intercal}\boldsymbol{T}[k]) - a\theta_{\mathrm{o}}[k]| - qPh & q = Q_k \\ 0 & \text{otherwise} \end{cases},$$

$$Q_k = \left\lfloor \frac{|\mathrm{Tr}(\boldsymbol{\Gamma}^{\intercal}\boldsymbol{T}[k]) - a\theta_{\mathrm{o}}[k]|}{Ph} \right\rfloor.$$

Clearly, when $Q_k = 0$ this reduces to the expression in (3.5.17). It is important to notice the explicit dependency of the energy spent (load response) from the random ambient temperature.

[4]Naturally, all the steps that correspond to a certain reference temperature transition will have the same sign.

3.5.2.2 Aggregate Quantized Population Model

For the remainder of this chapter the effect of the random noise $\varepsilon[k]$ will be ignored. The population model is built with two standard steps. The first consists of clustering of the parameters (R, C, η), so that the matrix $\mathbf{\Gamma}$ and coefficient a in (3.5.16) are chosen from a restricted set that approximates well the most common TCL characteristics, i.e. $\mathbf{\Gamma} \mapsto \mathfrak{Q}(\mathbf{\Gamma}) \in \mathcal{S}_{\mathbf{\Gamma}},\ a \mapsto \mathfrak{Q}(a) \in \mathcal{S}_a$:

$$\mathfrak{Q}(\mathbf{\Gamma}) = \arg \min_{\hat{\mathbf{\Gamma}}_s \in \mathcal{S}_{\mathbf{\Gamma}}} \|\hat{\mathbf{\Gamma}}_s - \mathbf{\Gamma}\| \,, \tag{3.5.19}$$

and similarly for $\mathfrak{Q}(a)$. Further, enumerate all pairs $(\hat{\mathbf{\Gamma}}_s, \hat{a}_s) \in \mathcal{S}_{\mathbf{\Gamma}} \times \mathcal{S}_a$ with an index $s = 1, \dots, S$. Note that this step yields errors in the load representation, which is assumed to be bounded. The second consists in quantizing the action space:

$$\mathfrak{q}(\boldsymbol{e}) \mapsto (u, v)^{\intercal} \ ,\ \mathfrak{q}(\boldsymbol{\theta}_{\mathrm{r}}) \mapsto (i, j)^{\intercal} \ ,\ \mathfrak{q}(\boldsymbol{T}) \mapsto \begin{pmatrix} u & i \\ v & j \end{pmatrix}. \tag{3.5.20}$$

In this way, the cost of the pth TCL in the population $\mathcal{I}$ is approximated as follows:

$$\hat{\Lambda}^{(p)}[k] = \mathrm{Tr}\Big(\mathfrak{Q}(\mathbf{\Gamma}^p)^{\intercal}\mathfrak{q}(\boldsymbol{T}^p[k])\Big) - \mathfrak{Q}(a^p)\theta_{\mathrm{o}}[k], \tag{3.5.21}$$

and similarly for $\hat{H}^{(p)}[k, q]$.

3.5.2.3 Population Model

To describe the expected load from the approximate aggregate model, it is convenient to break the population into the groups $\mathcal{I}_s, s = 1, \dots, S$ that belong to the different clusters for the TCL parameters:

$$p[k] = \sum_{s=1}^{S} p^{(s)}[k] = -\sum_{s=1}^{S} \sum_{\iota \in \mathcal{I}_s} \mathbb{E}\left[|\hat{\Lambda}^{(\iota)}[k]|\right] \tag{3.5.22}$$

For simplicity, momentarily ignore transitions in reference temperature. In this case:

$$p^{(s)}[k] = -\sum_{(u,v,i,j)} \left| \mathrm{Tr}\left((\hat{\mathbf{\Gamma}}_s)^{\intercal} \begin{pmatrix} u & i \\ v & j \end{pmatrix} \right) - \hat{a}_s \theta_{\mathrm{o}}[k] \right| \mathbb{E}\left[X_{uv}^{(s)ij}[k]\right]$$

where $X_{uv}^{(s)ij}[k]$ denotes the random variable equal to the population in cluster s that at time k has $\mathfrak{q}(\boldsymbol{T}^p[k]) \equiv \begin{pmatrix} u & i \\ v & j \end{pmatrix}$; using the $\delta(\boldsymbol{x})$ as an indicator function that is one only if the array $\boldsymbol{x} = \mathbf{0}$ and zero else, $X_{uv}^{(s)ij}[k]$ can be defined as the follows:

$$X_{uv}^{(s)ij}[k] = \sum_{\iota \in \mathcal{I}_s} \delta \left(\mathfrak{q}(\boldsymbol{T}^\iota[k]) - \begin{pmatrix} u & i \\ v & j \end{pmatrix} \right). \tag{3.5.23}$$

In a nutshell, the expression for $p^{(s)}[k]$ clarifies that the Aggregator can control the expected load profile by controlling the expected transitions of the population, but that the response is a function of the ambient temperature realization $\theta_{\mathrm{o}}[k]$.

To consider the general case in which customers can change their reference temperatures, the response has to be coded in terms of energy cost that corresponds to a particular array $\begin{pmatrix} u & i \\ v & j \end{pmatrix}$. That corresponds to:

Quantized profile for energy spent for transition $\begin{pmatrix} u & i \\ v & j \end{pmatrix}$*:*

$$H_{uv}^{(s)ij}[k,q] = \begin{cases} P^{(s)}h & 0 \leq q < Q_k^{(s)} \\ \lambda_{uv}^{(s)ij}(\theta_{\mathrm{o}}[k]) - qP^{(s)}h & q = Q_k^{(s)} \\ 0 & \text{otherwise} \end{cases} \tag{3.5.24a}$$

where the total energy spent for the transition is:

$$\lambda_{uv}^{(s)ij}(\theta_{\mathrm{o}}) = \left| \mathrm{Tr} \left((\hat{\boldsymbol{\Gamma}}_s)^\intercal \begin{pmatrix} u & i \\ v & j \end{pmatrix} \right) - \hat{a}_s \theta_{\mathrm{o}} \right| \tag{3.5.24b}$$

and the duration of the load response is:

$$Q_k^{(s)} = \left\lfloor \frac{|\lambda_{uv}^{(s)ij}(\theta_{\mathrm{o}}[k])|}{P^{(s)}h} \right\rfloor. \tag{3.5.24c}$$

leading to the general expression:

$$p^{(s)}[k] = - \sum_{(u,v,i,j)} \sum_{q=0}^{+\infty} H_{uv}^{(s)ij}[k-q,q] \mathbb{E}\left[X_{uv}^{(s)ij}[q] \right], \tag{3.5.25}$$

indicating that the model has memory.

3.5.3 Control Model

The key to scalability of the control lies in assuming that the Aggregator can broadcast commands that are cluster specific, but not customer specific, by controlling only the expected population trajectory as opposed to its exact values. Hence, the decision variable for the Aggregator are:

$$D_{uv}^{(s)ij}[k] \triangleq \mathbb{E}\left[X_{uv}^{(s)ij}[k]\right], \tag{3.5.26}$$

and the aggregator objective is to shape the expected load, which is a linear function of (3.5.26) (c.f. (3.5.25)):

Forecast of flexible TCL load response:

$$p[k] = \sum_{s=1}^{S} p^{(s)}[k] \tag{3.5.27a}$$

$$p^{(s)}[k] = -\sum_{(u,v,i,j)} \sum_{q=0}^{+\infty} H_{uv}^{(s)ij}[k-q,q] D_{uv}^{(s)ij}[q] \tag{3.5.27b}$$

In plain English, each $D_{uv}^{(s)ij}[k]$ is the expected number[5] of TCLs in cluster s transitioning from state u to state v and reference temperature i to j over period k.

3.5.3.1 Randomized Control Policy

The expectation in (3.5.26) can be made an explicit function of the transition probabilities $\pi_k^{(s)ij}(v|u)$ that are specified as commands by the Aggregator for a randomized policy. In fact, the instructions are the probabilities for changing normalized energy from value u to value v. This presumes that the execution of the commands consists of choosing at random to

[5] While not explicitly stated, since the policy is randomized, the case where the population participants vary can be handled with no change in the optimization formulation, by a scaling factor equal to the probability that nodes exit the control.

switch normalized energy u to v and reference temperatures i to j with probability $\pi_k^{(s)ij}(v|u)$. Let $\Pi_k^{(s)i}(u)$ denote the *state* probability, i.e. the probability that a TCL is in normalized energy state u and normalized reference temperature i. Under the randomized control policy $\delta\left(\mathfrak{q}(\boldsymbol{T}^p[k]) - \begin{pmatrix} u & i \\ v & j \end{pmatrix}\right)$ is a Bernoulli random variable with probability $\pi_k^{(s)ij}(v|u)\Pi_{k-1}^{(s)i}(u)$ of being equal to 1. This implies that:

$$D_{uv}^{(s)ij}[k] = |\mathcal{I}_s|\pi_k^{(s)ij}(v|u)\Pi_{k-1}^{(s)i}(u). \tag{3.5.28}$$

From Chapman-Kolmogorov theorem for Markov chains:

$$\Pi_k^{(s)j}(v) = \sum_{(u,i)} \pi_k^{(s)ij}(v|u)\Pi_{k-1}^{(s)i}(u), \tag{3.5.29}$$

which, in turn, implies that:

$$|\mathcal{I}_s|\Pi_k^{(s)j}(v) = \sum_{(u,i)} D_{uv}^{(s)ij}[k] \tag{3.5.30}$$

Combined with (3.5.26), this last equation (3.5.30) means that the Aggregator can evaluate the randomized policy values based on the optimum values of $D_{uv}^{ij}[k]$ as follows:

$$\pi_k^{(s)ij}(v|u) = \frac{D_{uv}^{(s)ij}[k]}{\sum_{(u',i')} D_{u'u}^{(s)i'i}[k-1]}. \tag{3.5.31}$$

The values of $\pi_k^{(s)ij}(v|u)$ are the instructions that are broadcast to the TCLs to plan their switching.

3.5.3.2 Feasible Action Space and its Representation

By deciding the values for $D_{uv}^{(s)ij}[k]$ over the horizon, the Aggregator can shape the expected aggregate load of (3.5.27a) within a feasible region determined by the constraints that exist on $D_{uv}^{(s)ij}[k]$. The following are known to the Aggregator before solving for D over a horizon Kh:

1. A temperature forecast or scenario, $\theta_{\text{o}}[k]$ for the horizon.

2. The initial population of each cluster, and $|\mathcal{I}_s|\Pi_0^{(s)i}(u)$, representing the initial spread of individuals across reference temperatures (i) and states (u), for all s, i, u.

3. The thermostat program of all individuals for the entire horizon; more specifically, the number of TCLs going from one reference temperature to another at time k:

$$\rho^{(s)ij}[k] = \sum_{(u,v)} X^{(s)ij}[k] = \sum_{\iota \in \mathcal{I}_s} \sum_{(u,v)} \delta \left(\mathfrak{q}(\boldsymbol{T}^{\iota}[k]) - \begin{pmatrix} u & i \\ v & j \end{pmatrix} \right) \tag{3.5.32}$$

The feasible action space is described, in order, by the following constraints on $D_{uv}^{(s)ij}[k]$: (3.5.33a) indicates that population transitions are non-negative; (3.5.33b) comes from the fact $\sum_{(v,j)} \Pi_k^{(s)j}(v) = 1$ and (3.5.29), and can be interpreted as the conservation of population mass; (3.5.33c) comes considering (3.5.31) and because $\sum_{(v,j)} \pi_k^{(s)ij}(v|u) = 1$; (3.5.33d) comes from accounting for the thermostat plans of the TCL population; (3.5.33e) forces the model to spread evenly the population transitioning from i to j throughout the departing reference temperatures state-space; finally, (3.5.33f) expresses the Aggregator restrictions on the reference temperature state transitions, as discussed in Section 3.5.2. Note that the set (3.5.33g) improves on (3.5.10) by forcing those individual changing $\theta_{\mathrm{r}}[k-1] \to \theta_{\mathrm{r}}[k]$ to arrive in the top/bottom of the dead-band, depending on the reference temperature change.

Feasible action space for TCL population:

$$D_{uv}^{(s)ij}[k] \geq 0 \quad \forall s, i, j, u, v \tag{3.5.33a}$$

$$\sum_{(v,j)} \sum_{(u,i)} D_{uv}^{(s)ij}[k] = |\mathcal{I}_s|; \ \ \forall s \tag{3.5.33b}$$

$$\sum_{(v,j)} D_{uv}^{(s)ij}[k] = \sum_{(u',i')} D_{u'u}^{(s)i'i}[k-1] \ \ \forall s, i, u \tag{3.5.33c}$$

$$\rho^{(s)ij}[k] = \sum_{(u,v)} D_{uv}^{(s)ij}[k] \ \forall k, i, j. \tag{3.5.33d}$$

$\forall k \in \mathcal{K}$, and $\forall s, u, i, j$ such that $i \neq j$ and $\rho^{(s)ij}[k] > 0$:

$$\sum_{(v)} D_{uv}^{(s)ij}[k] = \frac{\rho^{(s)ij}[k]}{\sum_{(j')} \rho^{(s)ij'}[k]} \sum_{(j',v)} D_{uv}^{(s)ij'}[k] \tag{3.5.33e}$$

$$D_{uv}^{(s)ij}[k] = 0 \ \ \forall \begin{pmatrix} u & i \\ v & j \end{pmatrix} \notin \mathcal{E}^{(s)}(\theta_{\mathrm{o}}[k]), \tag{3.5.33f}$$

where:

$$\mathcal{E}^{(s)}(\theta_{\mathrm{o}}) = \left\{ \begin{pmatrix} u & i \\ v & j \end{pmatrix} \middle| \left(\left| \mathrm{Tr}\left(\hat{\mathbf{\Gamma}}_s \begin{pmatrix} u & i \\ v & j \end{pmatrix} \right) - \hat{a}_s \right| \leq Ph \text{ and } i = j \right) \cup (v = (i-j)U \text{ and } |i-j| = 1) \right\} \tag{3.5.33g}$$

Alternatively, in vector form, the TCL load is referred to as $\boldsymbol{p} \in \mathbb{R}^K$, and use $\boldsymbol{D}$ for the tensor of the decision variables, i.e. all values $D_{uv}^{(s)ij}[k], \forall s, u, v, i, j, k$ in their respective sets. The constrained problem that includes (3.5.27) and (3.5.33) can be described generally as $\boldsymbol{p} \in \mathcal{P}$, where:

$$\mathcal{P}(\boldsymbol{\theta}_{\mathrm{o}}) = \{\boldsymbol{p} \,|\, \boldsymbol{p} = -\boldsymbol{M}(\boldsymbol{\theta}_{\mathrm{o}})\boldsymbol{D}\,,\ \boldsymbol{A}(\boldsymbol{\theta}_{\mathrm{o}})\boldsymbol{D} \leq \boldsymbol{b}\,,\ \boldsymbol{D} \geq 0\} \tag{3.5.34}$$

where $\boldsymbol{p} \in \mathbb{R}^K$, $(\boldsymbol{A}(\boldsymbol{\theta}_{\mathrm{o}}), \boldsymbol{b})$ represent (3.5.33), while the decision space of $\boldsymbol{D}$ is high dimensional, with $\dim(\boldsymbol{D}) = S \cdot \sum_k \left|\mathcal{E}^{(s)}(\theta_{\mathrm{o}}[k])\right|$.

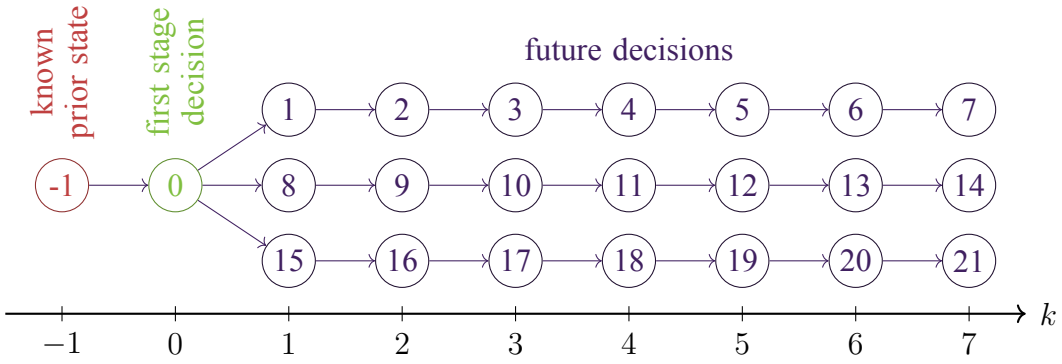


Figure 3.6.1: An example of a *scenario tree* that only branches at the 2^{nd} stage.

3.6 A Stochastic Security-Constrained Economic Dispatch with Responsive Loads

A straightforward way to incorporate uncertainty in power system models is to apply standard stochastic programming techniques, where one optimizes a collection of plans for different scenarios that reflect the underlying uncertainty on the future. More specifically, this section describes a rolling horizon *two stage* Security-Constrained Economic Dispatch (SCED) optimization, where the joint uncertainty of net-load and TCL outdoor temperature is considered, while being secure against the loss of a any single generator ($G-1$). In the first stage, decisions based on the realized uncertainty are made; these decisions account for a second stage consisting of several possible future trajectories. A time horizon of K intervals is considered, each of length h seconds such that it looks Kh seconds into the future. The formulation is described in *nodal form*, that is, instead of indexing variables and parameters by time k and scenario s they are simply indexed by the node number n, where each node has the parent node n^-. The set of nodes $\mathcal{V} = \{0, \ldots, N\}$ can thus be laid out on a graph as in Figure 3.6.1 where three future scenario trajectories are considered, that share the first stage parameters and decisions, with each node $n \in \mathcal{V}$ having an associated probability $\pi\{n\}$. Denoting the set of generators by $\mathcal{G}$, the set of generator outages by $\mathcal{G}^{\text{out}}$, and the set

of TCL aggregates as $\mathcal{M}$, the vector of decision variables is $\boldsymbol{\chi} = [\boldsymbol{x}, \hat{\boldsymbol{x}}, \overline{\boldsymbol{x}}, \boldsymbol{p}, \hat{\boldsymbol{p}}, \overline{\boldsymbol{p}}]$, and the expected cost is optimized by solving:

$$\min_{\chi} \sum_{g\in\mathcal{G}} \sum_{n\in\mathcal{V}} \pi\{n\} \left(\mathrm{C}_X^g(x^g\{n\}) + \overline{\mathrm{C}}_X^g(\overline{x}^g\{n\})\right) \tag{3.6.1a}$$

$$\sum_{g\in\mathcal{G}} x^g\{n\} + \sum_{m\in\mathcal{M}} p^m\{n\} = L\{n\} \quad \forall n \in \mathcal{V} \tag{3.6.1b}$$

$$\underline{P}^g \leq x^g\{n\} \leq \overline{P}^g - \overline{x}^g\{n\} \quad \forall n \in \mathcal{V}, g \in \mathcal{G} \tag{3.6.1c}$$

$$\left|x^g\{n\} - x^g\{n^-\}\right| \leq \overline{P}^g_{\text{ramp}} \quad \forall n \in \mathcal{V}, g \in \mathcal{G} \tag{3.6.1d}$$

$$\boldsymbol{p}^m \in \mathcal{P}^m(\boldsymbol{\theta}_{\text{o}}) \quad \forall m \in \mathcal{M} \tag{3.6.1e}$$

$$\boldsymbol{p}^m + \overline{\boldsymbol{p}}^m \in \mathcal{P}^m(\boldsymbol{\theta}_{\text{o}}) \quad \forall m \in \mathcal{M} \tag{3.6.1f}$$

$$\sum_{g\in\mathcal{G}} \hat{x}^g_{g'}\{n\} = x^g\{n\} - \sum_{m\in\mathcal{M}} \hat{p}^m_{g'}\{n\} \quad \forall n \in \mathcal{V}, g' \in \mathcal{G}^{\text{out}} \tag{3.6.1g}$$

$$0 \leq \hat{p}^m_{g'}\{n\} \leq \overline{p}^m\{n\} \quad \forall n \in \mathcal{V}, m \in \mathcal{M}, g' \in \mathcal{G}^{\text{out}} \tag{3.6.1h}$$

$$0 \leq \hat{x}^g_{g'}\{n\} \leq \overline{x}^g\{n\} \leq \overline{P}^g_{\text{resv}} \quad \forall n \in \mathcal{V}, g \in \mathcal{G}, g' \in \mathcal{G}^{\text{out}} \tag{3.6.1i}$$

$$x^g_g\{n\} = 0 \quad \forall n \in \mathcal{V}, g \in \mathcal{G}^{\text{out}} \tag{3.6.1j}$$

The objective (3.6.1a) is to minimize the cost of generation $x^g\{n\}$ and reserves $\overline{x}^g\{n\}$ subject to the cost functions C_X^g and $\overline{\mathrm{C}}_X^g$. Constraint (3.6.1b) describes the base-case (non-outage) balance between generation $x^g\{n\}$, TCL load $p^m\{n\}$ and the remaining net-load $L\{n\}$. Further, (3.6.1c) and (3.6.1d) describe conventional min/max power and ramp constraints, where $\overline{x}^g\{n\}$ is the reserve allocation of unit g at node n, with n^- indicating the parent of node n. Constraints (3.6.1e) and (3.6.1f) ensure that the base-case TCL load and its upper reserve allocation of TCL aggregate m are contained within the feasible set $\mathcal{P}^m$. Constraint (3.6.1g) ensures that enough reserves are available to replace any single generator outage, while (3.6.1h) and (3.6.1i) ensure that the maximum reserve allocation $\overline{x}^g$ and $\overline{p}^m$ is greater

or equal to the individual outage responses $\hat{x}^g$ and $\hat{p}^m$. Finally (3.6.1j) ensures an outaged unit can not contribute to its own replacement reserves.

Remark 5. Note that with a convex $\mathcal{P}^m$, constraints (3.6.1e) and (3.6.1f) are not sufficient to guarantee that any combination of the individual reserve responses $\hat{p}^m_{g'}\{n\}$ are feasible. However, due to the nature of the TCL model lacking inherent ramping constraints while having prominent energy constraints, (3.6.1f) and (3.6.1e) can be interpreted as bounds on not only power but also energy, and any power profile contained within these limits thus also respects this energy bound. This assumptions means that the probability of falling outside the feasible region $\mathcal{P}^m$ during some sequence of reserve events is small, but non-zero. This approximation is also apparent for generators; the standard SCED model does not properly bound generator ramps between any possible sequence of reserve events. The message here is that (a) these reserve bounds are approximations and (b) many $N-1$ events captured by conventional SCED/SCUC models are *rare*.

3.7 Numerical Results

3.7.1 Electric Vehicle Reserve Capacity Modeling

The 2009 National Household Travel Survey [Federal Highway Administration(2009)] is used as a source of data about personal travel during a "sample day"' in 2009. The dataset describes $223,000$ trips in cars, the time of departure, arrival and distance driven. This information is used to establish the energy consumption of each trip, the time cars get parked at home and how long they spend sitting before next usage. An EV with 35 kWh/100 miles efficiency and 10 kW charging capacity is chosen as a sample car, representative for an average Tesla Model S, whereas EV efficiency can range from 25-50 kWh/100 miles for different EV models. Figure 3.7.1 gives an overview of the dataset, showing the arrival rate throughout the day, the number of available (plugged in) EVs as well as the service and

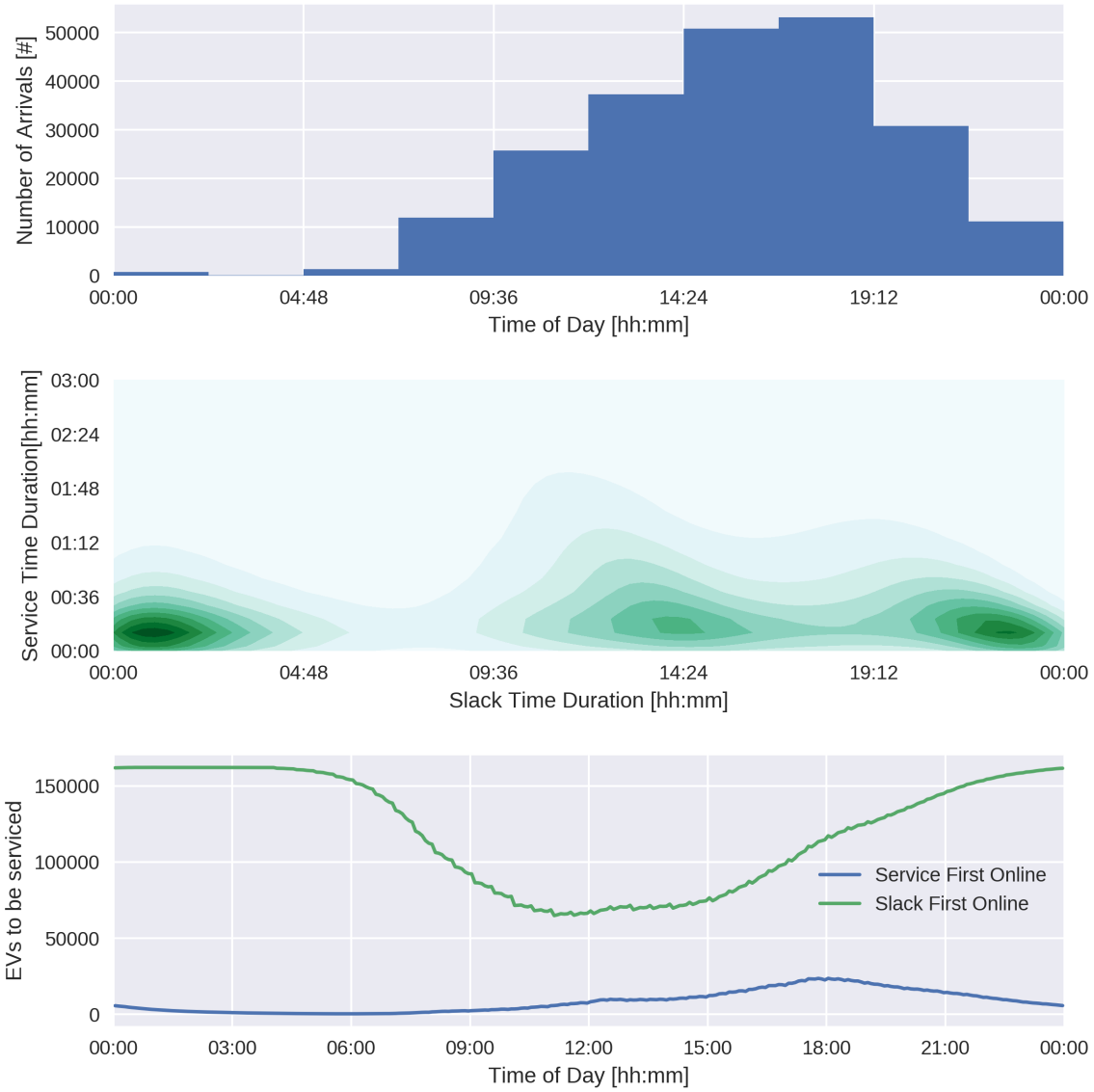


Figure 3.7.1: The top graph shows the arrival rate histogram for the survey data, the middle graph shows the distribution of service and slack time at any time of day, while the bottom graph shows the number of EVs plugged in and not fully charged

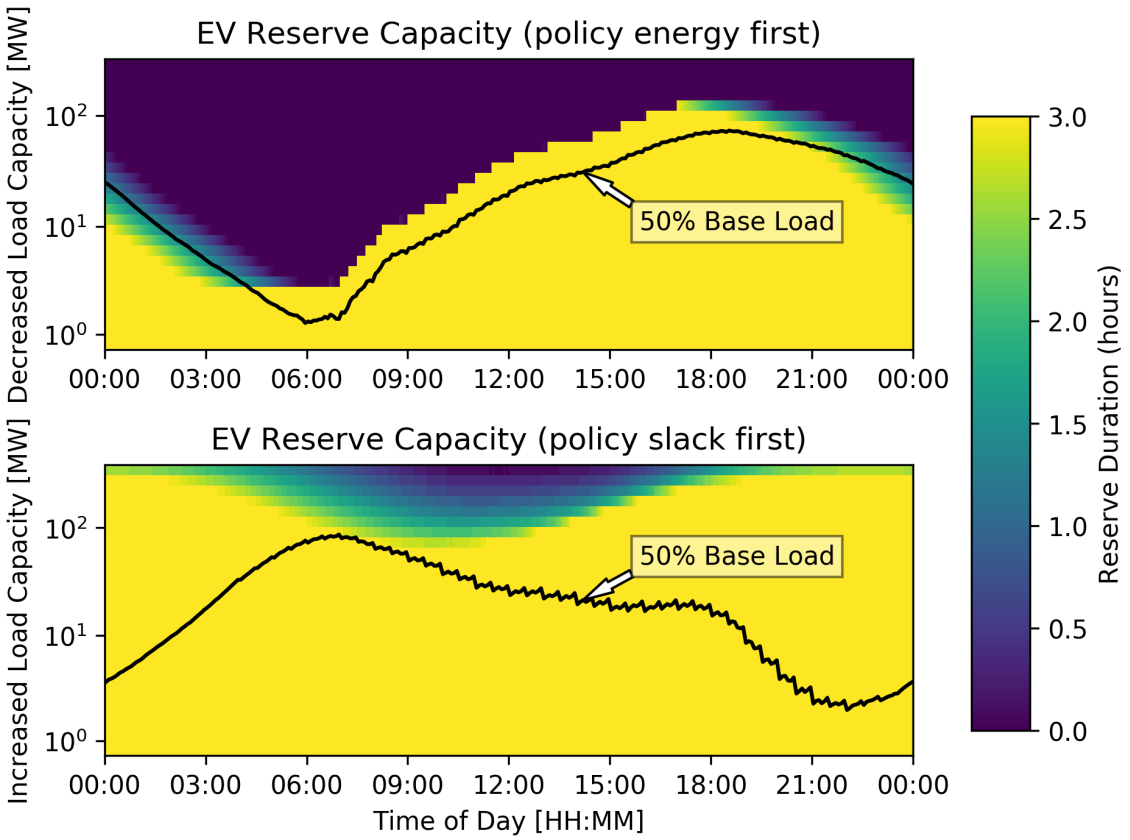


Figure 3.7.2: Reserve Capacity vs Base Load Capacity for two different consumption strategies.

slack time combinations, where service time denotes the time required for charging, and slack time is the leftover time before next usage.

Two default strategies to serve the EV load are considered:

Energy First serves electric vehicles as soon as possible, maximizing the curtailment potential at any time, while

Slack First waits as long as possible before serving the vehicles, thus allowing for sudden increase in consumption if necessary.

The reserve potential of these policies is visualized in Figure 3.7.2. Considering these $223,000$ arrivals on a typical week day, the "energy first" policy has a min/max load of

3/146 MW respectively, with the maximum 3 hour reserve capacity ranging from 2.6 MW to 100 MW at best. Between 5am and 9pm it provides more than 50% curtailment for 3 hours, while late evening/early night it drops down to 1-2 hours. The slack first policy has a min/max power of 4/173 MW but a much greater (down-spinning) reserve capacity in the range of 59-278 MW over the course of the day, easily offering consumption above 150% schedule for more than 3 hours. The power numbers grow proportionally with the number of arrivals, assuming similar service/slack requirements.

3.7.2 Distributed Storage Reserve Power

For Distributed Storage, the Tesla Powerwall battery was used as a reference, with a capacity of 13.5 kWh and maximum continuous power of 5 kW. To maximize the use of batteries for reserve power purposes, the mathematics become trivial, as the entire battery's capacity is dedicated to this purpose. If there are conflicting goals, such as minimizing distribution network energy transfer for households with renewable energy resources (consuming locally if possible), or if the battery is intended for load-shifting (buying cheaper), this naturally reduces the reserve capabilities.

Assuming the battery is pre-positioned at a charge level most suitable for the reserve capacity desired (full for curtailment, empty for "down-spinning"), Figure 3.7.3 shows the simple relationship between the number of batteries, the reserve capacity and duration (capped at 4 hours). The reserve target of supplying 100 MW of reserves for 4 hours is reached at roughly $30,000$ batteries.

3.7.3 Thermostatically Controlled Loads

Numerical simulations were performed using Python and a collection of scientific programming libraries [Van Der Walt *et al.*(2011), Jones *et al.*(2014), Hunter(2007), Pedregosa

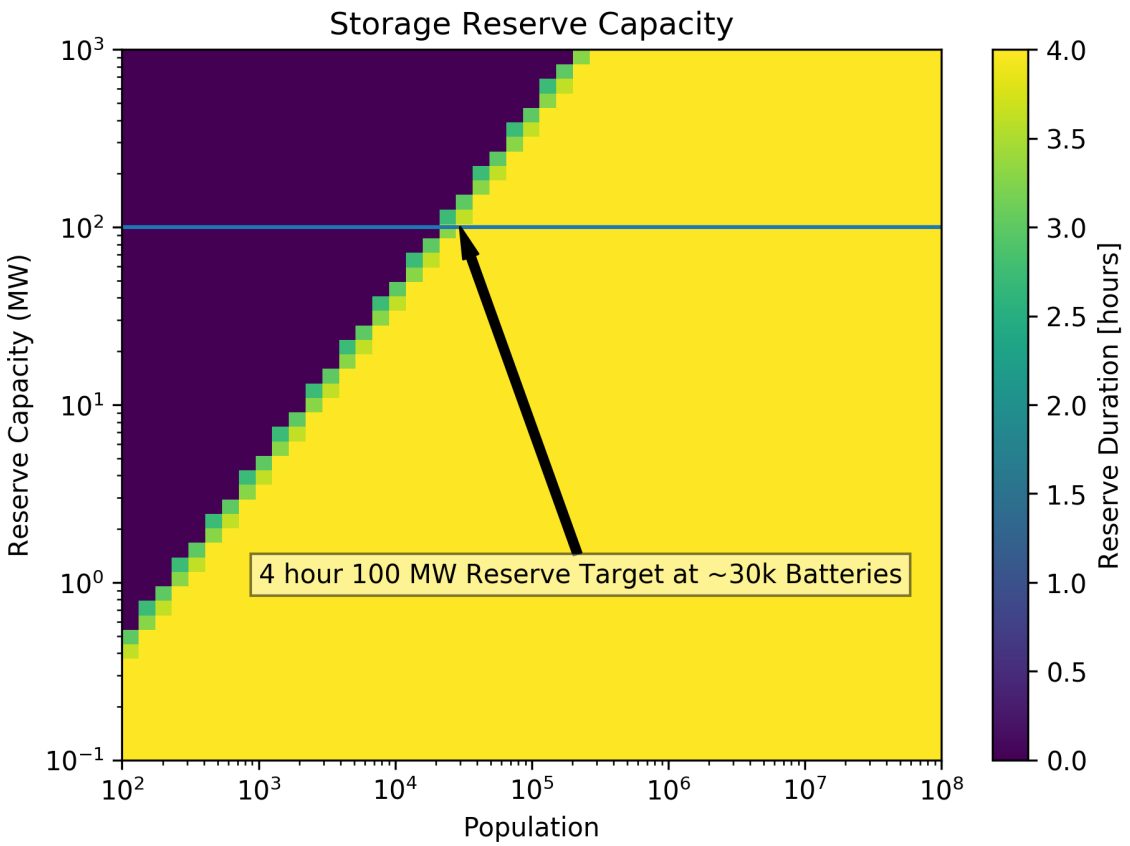


Figure 3.7.3: Reserve capacity/duration vs number of batteries.

et al.(2011)]. All optimization problems were solved with Gurobi 7.5 [Gurobi Optimization(2015)].

3.7.3.1 Thermostatically Controlled Load Cluster Population Estimation

In this section the model from Section 3.5 is applied in "reverse" to electric load data from CAISO and temperature data from National Oceanic & Atmospheric Administration (NOAA) to get an estimate for the contribution of TCLs to the overall CAISO load, as well as the distribution of the TCL population between S pre-defined clusters. Three years of 5 minute resolution data were gathered, where aggregate load consumption and forecasted wind and solar infeed came from CAISO, while from NOAA the temperature profiles for the seven California locations that are part of the NOAA Climate Reference Network [Dia-

mond *et al.*(2013)] were gathered. The U.S. Energy Information Administration's (EIA) Residential Energy Survey (RES) [Energy Information Administration, U.S. Department of Energy(2009)] states that California has an estimate of $7,000,000$ households with air conditioners, most of them falling into the "non-reversible" category, that is they are only capable of cooling but not heating, with households resorting to other energy sources for heating during the winter. The survey states that most households vary their reference temperature during the day, either manually or through programmable thermostats. Furthermore, the survey estimates that over the course of a year, a combined $\approx 3.2 \times 10^{16}$ Joules of energy are consumed by air conditioning loads. Incorporating data from [Mathieu *et al.*(2012)], a number of clusters are created for every combination of:

(a) the seven NOAA temperature observation locations,

(b) $R \in \{1.5, 2, 2.5\}$ [°C/kW],

(c) $C = \{1.5, 2, 2.5\}$ [kWh/°C] and

(d) $\theta_{\mathrm{r}} \in \{69, 73, 77, 81\}$ [°F].

Then the following least-squares problem is solved:

$$\min_{\delta,\phi,\upsilon} \sum_{\psi\in\Psi^{\varepsilon}} \sum_{k\in\mathcal{K}} \delta^2[\psi,k] \tag{3.7.1a}$$

$$\text{s.t. } \delta[\psi,k] = \omega[k] - p[\psi,k] - \beta[\psi,k] \tag{3.7.1b}$$

$$\sum_{s=1}^{S} |\mathcal{I}^s| \leq 7,000,000 \tag{3.7.1c}$$

$$(3.5.34), (3.5.33) \tag{3.7.1d}$$

where $[\psi, k]$ refers to a particular time k of sample day ψ of day-type Ψ^{ε}, with each day type ε capturing the season and weekday/weekend. The historical aggregate load is in β and $p[\psi, k]$ depends on historical records of temperature. The output of this least squares

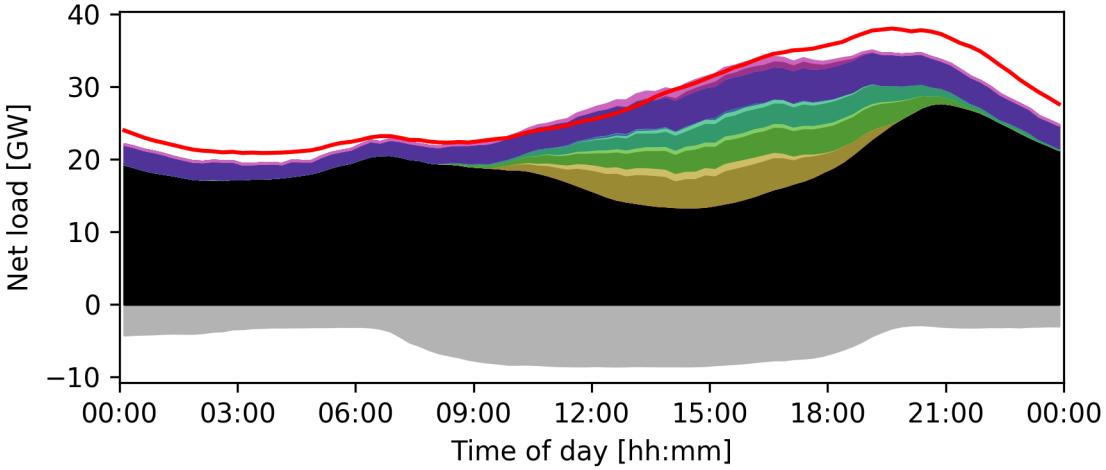


Figure 3.7.4: Disaggregated CAISO load for July 13th 2016: (a) the black mass is the base net-load signal $\beta[k]$, (b) the gray bottom layer is the renewable infeed, (c) the colored layers are the TCL consumption of different regions and (d) the red-line denotes the actual measured net-load of that day.

problem is the error δ and the number of individuals in each clusters and their transitions over the course of a day. Figure 3.7.4 shows the disaggregated CAISO load for July 13th 2016 (Wednesday), where the gray bottom layer shows renewable infeed, the black mass shows the estimated non-TCL base-net-load (β), the colored layers show the TCL baseline load of different regions, while the red line shows the actual net-load of that day. One can observe that the total net-load estimate is close to measured values, though the evening peak is slightly underestimated. Only five colored layers are visible, as two of the temperature regions are coastal and induced little cooling consumption on that day. The infamous CAISO "duck curve" is also clearly visible in the base-load, due to the significant renewable infeed during the middle of the day.

To get an indication of the quality of the results, this mechanism is used to calculate the annual energy consumption for the years 2010-2017, in order to compare it with the survey estimated value of 3.2×10^{16} J. For those years, the min, mean and max consumption is $5.6 \cdot 10^{16}$, $7.3 \cdot 10^{16}$ and $8.1 \cdot 10^{16}$ J respectively, indicating that the *residential* cooling energy is overestimated by a factor of approximately 2. Given the simplicity of the regression model,

this error is small, and the model gives a decent estimate of the order of magnitude and division of the TCL population between clusters across California. It is likely that the source of the error is that other large cooling loads are captured with this model, such as commercial and industrial buildings. It would be possible to further constrain (3.7.1) given the annual energy estimates, though certain variability between years should be expected. In any case, further numerical results are calculated using these population estimates, acknowledging these shortcomings.

3.7.3.2 Thermostatically Controlled Loads Reserve Power Estimation

Using the population profiles and the load control model, the reserve power capacity of California's air conditioners can be estimated for different days of the year. For this simulation, households are assumed to start at reference temperature, and the mean dead-band is fixed to 1° C. Given historical temperature recordings one can look at the potential both to increase and decrease (curtail) load. The results are depicted by Figure 3.7.5, showing over the course of a day, how much deviation from the base-line TCL consumption is possible and for how long (as the two are closely inter-dependent), with the base-line shown in red. If the yellow region touches the upper red line, one could double the TCL consumption for up to 3 hours, while if it touches the lower red-line, it could decrease the TCL consumption to zero for 3 hours. Looking at the numbers, one can see that on this summer day the aggregator could decrease the consumption by >1 GW for up to 3 hours, and reach a maximum consumption well above 20 GW for a brief time during the warmest part of the day, while keeping California customer comfortable. During the night the TCL load is smaller, so there is less energy to curtail (in the hundreds of MW), but if there are requirements for increased load (for example during loss of another load) consumption of TCLs can be doubled (additional 5 GW) for up to 15-30 minutes, or maintain close to a 1 GW increase for over an hour. This shows that during summer months, when cooling loads are a significant contributor to

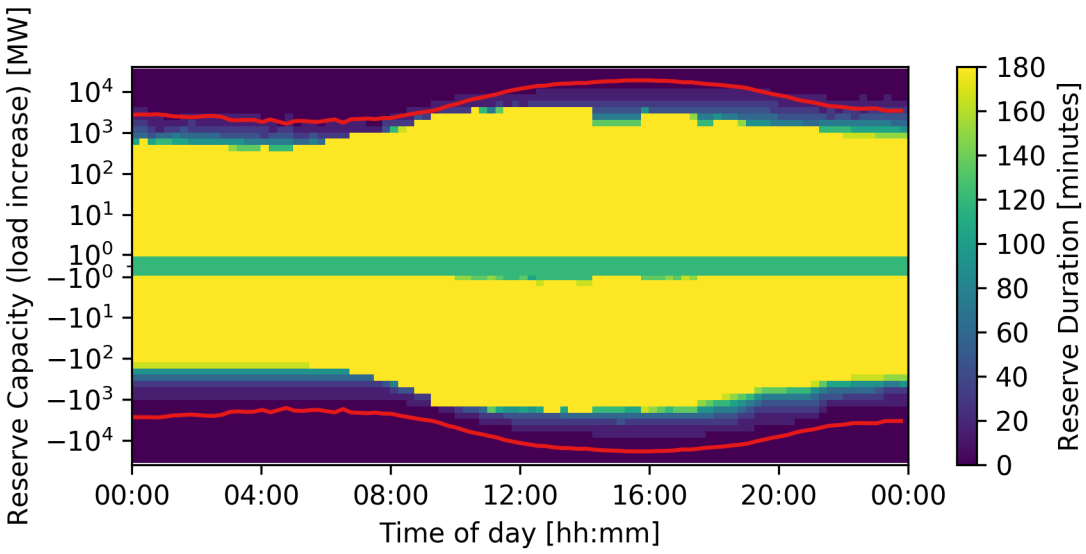


Figure 3.7.5: An example of summer day TCL reserve potential in the CAISO system. The red lines denote the base-TCL load, while the heatmap shows the duration and magnitude of available load increase and curtailment during the course of the day.

California's overall energy consumption, they can also play an important reserve capacity role, and are particularly well suited to take on fast unexpected ramps caused by renewable power sources.

CHAPTER 4

INSIGHTS ON CONVEXITY FROM THE SHAPLEY-FOLKMAN LEMMA AND MARKET INTEGRATION

4.1 Generic Individual Prosumer Models

As in [Alizadeh *et al.*(2014a),Barot and Taylor(2017),Zhao *et al.*(2017),Nazir *et al.*(2018), Müller *et al.*(2017)], this dissertation assumes that if the set of feasible power profiles $\mathcal{P}$ is convex, then it is a polytope. Such a set can be visualized over two intervals, something leveraged extensively throughout this chapter. Recall that convex polytopes are primarily described with (a) a set of half-spaces whose intersection is a polytope (H-rep) or (b) a set of vertices whose convex hull (the tightest convex region containing all the points) is a polytope (V-rep). As an example, Figure 3.1.1 visualizes a two-dimensional polytope, where the V-rep. is reflected by a set of vertices $\mathcal{V} = \{(2,1),(1,3),(3,4),(5,3)\}$, while the H-rep. is described by the constraint set for $\boldsymbol{p} \in \mathbb{R}^2$:

$$\begin{bmatrix} -2 & -1 \\ -1 & 2 \\ 1 & 2 \\ 2 & -3 \end{bmatrix} \boldsymbol{p} \leq \begin{bmatrix} -5 & 5 & 11 & 1 \end{bmatrix}^{\intercal}. \tag{4.1.1}$$

Looking back at Figure 3.1.1 it can be seen as a description of a flexible resource that for two consecutive hours ($K = 2$) can deliver any combination of power $[p_1, p_2]^{\intercal}$ that meets (4.1.1). In the figure, the horizontal axis indicate the power of period (e.g. hour) 1, and the vertical axis power of period 2, and the highlighted region visualizes all the (infinite) feasible power profiles available by the resource, i.e. each point $[p_1, p_2]$ on the figure denotes a particular power profile over the two intervals.

The set of feasible power profiles for individual resources is, in practice, often non-convex. In this chapter it is assumed that such a behavior can be modeled by introducing additional (binary) variables into the H-rep. for common flexible loads such as storage devices, EVs and TCLs. Adding binary variables means that the region $\mathcal{P}$ is composed of the union of several (possibly disjoint) polytopes. The following definition introduces a generic resource model for (possibly non-convex) resources.

Definition 8 (Individual Resource Model)**.** With a vector $\boldsymbol{p} \in \mathbb{R}^K$ denoting the power profile throughout the modeling horizon, and $\boldsymbol{\rho}$ a vector of auxiliary variables of length M, the set of all possible combinations of power and auxiliary variable values is defined as:

$$\overline{\mathcal{P}} = \{(\boldsymbol{p}, \boldsymbol{\rho}) | \boldsymbol{A} \begin{bmatrix} \boldsymbol{p}^\intercal & \boldsymbol{\rho}^\intercal \end{bmatrix}^\intercal \leq \boldsymbol{b}\} \subset \mathbb{R}^{K+M} \tag{4.1.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{J \times (K+M)}$ and $\boldsymbol{b} \in \mathbb{R}^J$ define both the constraints on, and the relationship between $\boldsymbol{p}$ and $\boldsymbol{\rho}$, where elements of $\boldsymbol{\rho}$ are continuous or discrete. Correspondingly, one can define $\mathcal{P}$ from $\overline{\mathcal{P}}$, removing the auxiliary variable dimensions as follows:

$$\mathcal{P} = \{\boldsymbol{p} \,|\, (\boldsymbol{p}, \boldsymbol{\rho}) \in \overline{\mathcal{P}}\} \subset \mathbb{R}^K. \tag{4.1.3}$$

For each feasible power profile and set of auxiliary variables in $\overline{\mathcal{P}}$ there is a corresponding cost function $\overline{\mathrm{C}}(\boldsymbol{p}, \boldsymbol{\rho}) : \mathbb{R}^{K+M} \to \mathbb{R}$ (roman font). Correspondingly, the cost function w.r.t. $\boldsymbol{p}$ alone, is the (roman font):

$$\mathrm{C}(\boldsymbol{p}) = \inf\{\overline{\mathrm{C}}(\boldsymbol{p}, \boldsymbol{\rho}) | (\boldsymbol{p}, \boldsymbol{\rho}) \in \overline{\mathcal{P}}\} : \mathbb{R}^K \to \mathbb{R} \tag{4.1.4}$$

that is, the lowest possible cost to procure $\boldsymbol{p}$. For negative values of p_k, $-\mathrm{C}$ indicates the utility (value of energy), while for positive p_k, $+\mathrm{C}$ indicates cost of generation.

In addition to the load models introduced in Chapter 3, the following sections will explore briefly several non-convex individual load models.

4.1.1 Storage Devices

Expanding on the ideal storage model introduced in (3.1.3), the relationship between stored energy u_k and charging/discharging power p_k in discrete time for a non-ideal battery is:

$$u_k = u_0 + \sum_{\kappa=0}^{k} \left[\eta_c(-p_\kappa)_+ - \eta_d^{-1}(p_\kappa)_+\right] \tag{4.1.5}$$

where $\eta_c \in [0,1]$ and $\eta_d \in [0,1]$ denote the known charging and discharging efficiencies, $(\cdot)_+ = \max\{0, \cdot\}$ and negative values of p_k indicate charging. This section assumes a converter/inverter that can charge/discharge continuously in the ranges $[-\overline{p}_-, -\underline{p}_-]$ (charging) and $[+\underline{p}_+, +\overline{p}_+]$ (discharging), where $\underline{p}_-, \underline{p}_+, \overline{p}_-$ and $\overline{p}_+$ are all positive. Storage is also energy constrained; $u_k \in [0, \overline{u}]$. Two integer variables are necessary to indicate off, charging ($\boldsymbol{y}_- \in \{0,1\}^K$) or discharging ($\boldsymbol{y}_+ \in \{0,1\}^K$) at each time k, whereas only one integer variable is needed to distinguish between charging/discharging if minimum constraints are omitted. With auxiliary variables $\boldsymbol{\rho} = [\boldsymbol{p}_+^\intercal, \boldsymbol{p}_-^\intercal, \boldsymbol{y}_+^\intercal, \boldsymbol{y}_-^\intercal]^\intercal$ where the first two indicate the positive and negative part of $\boldsymbol{p}$, is defined as:

$$\begin{aligned}\overline{\mathcal{P}}_{\text{NIS}} = \{(\boldsymbol{p},\boldsymbol{\rho}) \,|\, & 0 \le \boldsymbol{y}_- + \boldsymbol{y}_+ \le 1,\, \boldsymbol{p} = \boldsymbol{p}_+ + \boldsymbol{p}_-, (4.1.5), \\ & 0 \le \boldsymbol{u} \le \overline{u}, \boldsymbol{y}_+\underline{p}_+ \le \boldsymbol{p}_+ \le \boldsymbol{y}_+\overline{p}_+, -\boldsymbol{y}_-\underline{p}_- \ge \boldsymbol{p}_- \ge -\boldsymbol{y}_-\overline{p}_-\} \\ = \{(\boldsymbol{p},\boldsymbol{\rho}) \,|\, & \boldsymbol{A}_{\text{NIS}}[\boldsymbol{p}^\intercal, \boldsymbol{\rho}^\intercal]^\intercal \le \boldsymbol{b}_{\text{NIS}}\}\end{aligned} \tag{4.1.6}$$

This region is visualized in Figure 4.1.1c, where clear non-convexities can be observed due to the binary states. Using the same model an ideal storage device can be described (cf. Figure 4.1.1a), only $\underline{p}_- = \underline{p}_+ = 0$ and $n_c = n_d = 1$. The cost/reward function suggested for storage is assumed:

$$C(\boldsymbol{p}) = \sum_k (\alpha_k |p_k| + \beta_k p_k^2) \tag{4.1.7}$$

However, the storage owner could also be a price-taker in which case the market will leverage it for load-shifting.

4.1.2 Electric Vehicles

As shown in Chapter 3, EVs have some of the characteristics of energy storage. Here the approach to EVs is different in that it does not discretize the state (of charge), and the model can only charge at discrete power levels $\overline{p}$ ($p_k \in \{0, -\overline{p}\}$), and there is limited time κ to provide a pre-determined amount of energy $\overline{u}$. The state of charge is defined as u_k, which at plug in time t_a is $u(t_a) = 0$ and at plug-out time t_d should be $u(t_d) = \overline{u}$, also $t_d - t_a = \kappa = \kappa_c + \kappa_s$ where κ_c denotes the time required for charging, and κ_s the leftover slack. Due to the discrete charging levels this model is non-convex. Integer values are assumed for κ_c and κ_s, i.e. the energy requirement is an integer multiple of the charge rate, and define $\boldsymbol{\rho} \in \{0, 1\}^K$. The feasible set (cf. Figure 4.1.1b) is:

$$\begin{aligned}\overline{\mathcal{P}}_{\text{EV}} &= \left\{(\boldsymbol{p}, \boldsymbol{\rho}) \,\middle|\, \boldsymbol{p} = -\overline{p} \cdot \boldsymbol{\rho}, \mathbf{1}\boldsymbol{\rho} = \kappa_c, \sum_{k=1}^{t_a} \rho_k = 0, \sum_{k=t_d}^{K} \rho_k = 0\right\} \\ &= \{(\boldsymbol{p}, \boldsymbol{\rho}) \mid \boldsymbol{A}_{\text{EV}}[\boldsymbol{p}^\intercal, \boldsymbol{\rho}^\intercal]^\intercal \leq \boldsymbol{b}_{\text{EV}}\} \end{aligned} \tag{4.1.8}$$

In general consumers perceive more utility from earlier charging. For simulations the cost function used was:

$$\mathrm{C}^i(\boldsymbol{p}) = \sum_k (\alpha - k\beta) p_k \tag{4.1.9}$$

4.1.3 Thermostatically Controlled Loads

As in Section 3.5, TCL models start from a first order thermal circuit,

$$\dot{\theta} = \frac{\theta_{\text{o}} - \theta}{RC} + \frac{\rho\eta\tilde{p}}{C} + \epsilon \tag{4.1.10}$$

where θ and θ_{o} are the indoor and outdoor temperatures, R and C denote the buildings thermal resistance and capacitance, η, $\tilde{p}$ and ρ the heat-pump efficiency, power and state (heating, cooling, off) and ϵ noise. Again, power is assumed to be a continuous value in the range $[0, \tilde{p}]$. Instead of moving towards a state-space model, the model proposed in

this chapter considers the non-convexities that can arise from the binary heat-pump state of cool/heat. The progression of θ can be modeled with $\omega = (RC)^{-1}$, $\zeta = (1 - T\omega)$ and power terms substitute with vectors for direction $\boldsymbol{\rho} \in \{-1, 1\}^K$ and power $\boldsymbol{p} \in [0, \tilde{p}]^K$:

$$\theta_k = T \sum_{\tau=1}^{k} \zeta^{k-\tau} \left(\omega \theta_{\mathrm{o},\tau} + \eta C^{-1} \rho_\tau p_\tau \right) + \theta_0 \zeta^k \tag{4.1.11}$$

This relationship between heat-pump power and inside temperature $\boldsymbol{\theta} = [\theta(1), \ldots, \theta(K)]$ is a linear mapping $\boldsymbol{\theta} = \boldsymbol{A}_\theta(\boldsymbol{\rho} \circ \boldsymbol{p})$ where $\circ$ denotes element-wise multiplication and $\boldsymbol{A}_\theta$ contains the coefficients found in (4.1.11). For a TCL that can seamlessly transition between heating and cooling, $\boldsymbol{\rho} \in \{-1, 1\}^K$, the mapping mapping between thermal and electric power is non-convex, but for a single pumping direction, e.g. $\boldsymbol{\rho} = \mathbf{1}$, it becomes a convex affine mapping of electric power. The state transitions, limits on $\boldsymbol{p}$ and the maximum deviation from a reference temperature $|\theta(k) - \theta^{\mathrm{r}}(k)| \leq \theta_\Delta$ can be modeled as linear mappings (cf. Figure 4.1.1d):

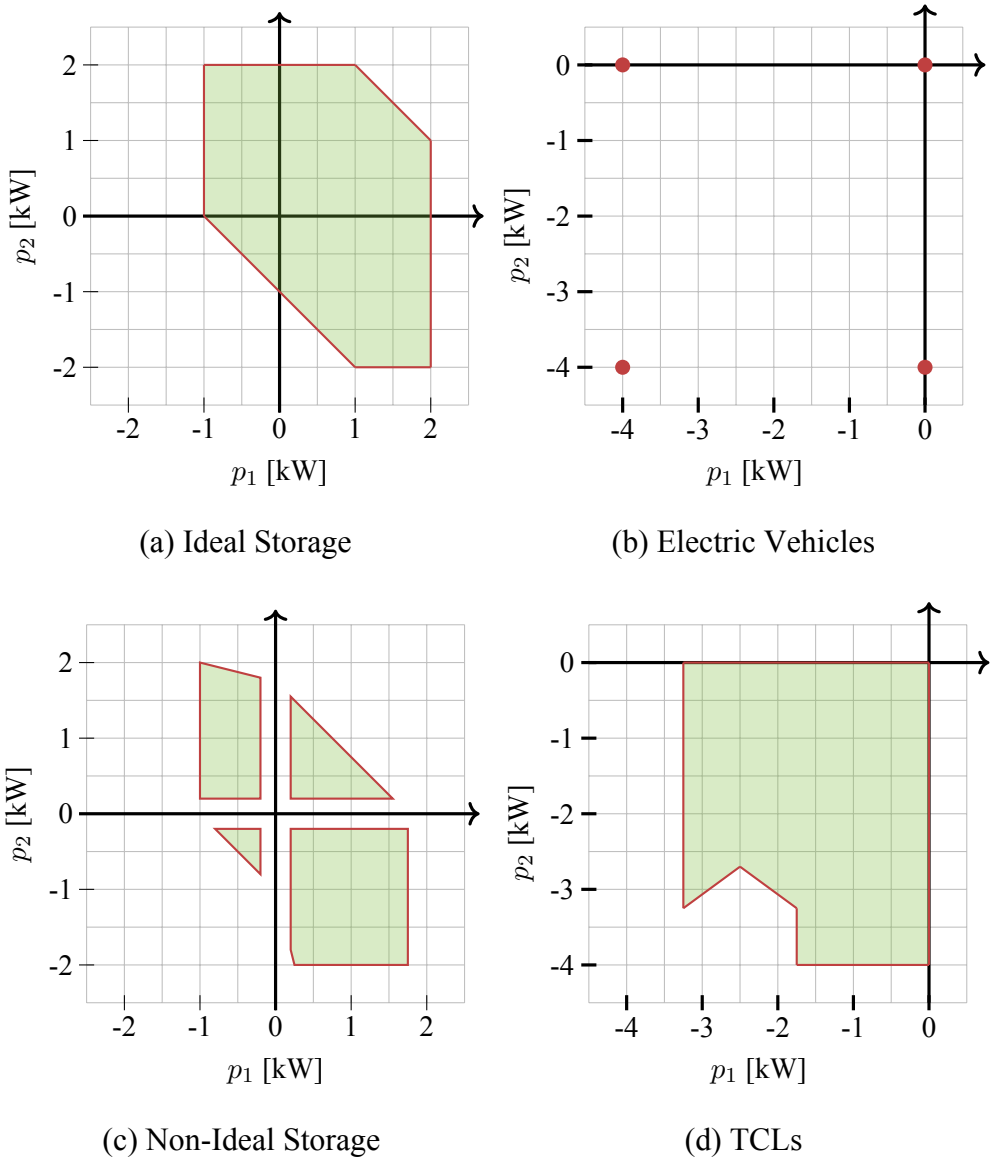
$$\overline{\mathcal{P}}_{\mathrm{TCL}} = \{(\boldsymbol{p}, \boldsymbol{\rho}) | \boldsymbol{A}_{\mathrm{TCL}} [\boldsymbol{p}^\intercal, \boldsymbol{\rho}^\intercal]^\intercal \leq \boldsymbol{b}_{\mathrm{TCL}}\} \tag{4.1.12}$$

Though not included here, lockout constraints can be incorporated into this model via additional constraints on the power profile (which relate back to min/max heat-pump on/off times), indicating that there is an additional inter-temporal relationship between the values of power between consecutive intervals. See relevant literature such as [Ziras *et al.*(2018)] for further details. The cost function of a TCL penalizes (pays consumers for) any deviation $\boldsymbol{\theta}^\delta = \boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{r}}$ from the reference temperature. For simulations the cost was modeled as a quadratic function with additional cost term related to $\boldsymbol{p}$:

$$\mathrm{C}_{\mathrm{TCL}}(\boldsymbol{p}, \boldsymbol{\rho}) = \sum_k \alpha \theta_k^d + \beta (\theta_k^d)^2 + \zeta p_k + \gamma \tag{4.1.13}$$

4.2 Aggregate Prosumer Modeling

This section establishes the conditions under which an aggregate of individual models can be considered convex.



(a) Ideal Storage

(b) Electric Vehicles

(c) Non-Ideal Storage

(d) TCLs

Figure 4.1.1: Feasible regions for individual resources, (a) ideal storage, (b) EVs, (c) non-ideal storage and (d) TCLs. The horizontal axis denotes power for period 1, while the vertical axis denote power for period 2. The set of possible profiles $[p_1, p_2] \in \mathcal{P}$ is the blue region whose edges are painted in red. The blue region is absent in (b) since the EV load considered can be only ON or OFF so that the feasible set contains four isolated points (red dots).

Let $\mathcal{I}$ denote the set of responsive loads (including storage) and N their number. Considering the finite decision horizon $k = 1, \ldots, K$, denote by p_k^i the average power of an individual load $i \in \mathcal{I}$ over the kth time interval. The entire profile corresponds to a vector in $\mathbb{R}^K$ denoted by $\boldsymbol{p}^i = [p_1^i, \ldots, p_K^i]^\intercal$. A flexible load can deliver more than one power profile $\boldsymbol{p}^i$ over the given time horizon, and the notation $\mathcal{P}^i \subset \mathbb{R}^K$ is used to describe the set of all feasible power profiles $\boldsymbol{p}^i$ for a given load over a horizon of length K. In practice, $\mathcal{P}^i$ captures all inter-temporal dependencies of possible power profiles throughout the modeling horizon.

Given feasible sets $\mathcal{P}^i$ and cost functions $\mathrm{C}^i(\boldsymbol{p})$ the goal is to study the feasible set $\mathcal{X}$ for the aggregate load $\boldsymbol{x} = \boldsymbol{p}^i + \cdots + \boldsymbol{p}^N$ and aggregate cost $C(\boldsymbol{x})$. The aggregate load set is defined by the Minkowski sum (Definition 5) of the individual loads feasible sets $\mathcal{P}^i$:

$$\mathcal{X} = \bigoplus_{i \in \mathcal{I}} \mathcal{P}^i = \mathcal{P}^1 \oplus \cdots \oplus \mathcal{P}^N \tag{4.2.1}$$

If all the individual models $\mathcal{P}^i$ are convex, then so is $\mathcal{X}$, facilitating its insertion into dispatch optimization problems. But what can be said about $\mathcal{X}$ composed of $\mathcal{P}^i$ which are in part (or all) non-convex, a common scenario in most real-world cases? The following section sheds light on this problem.

4.2.1 The Minkowski Sum of Non-convex Sets

For the upcoming derivation it is important to provide a clear definition of non-convexity:

Definition 9 (Non-Convexity)**.** The non-convexity of a set $\mathcal{P}$, denoted by the function $\mathrm{ncvx}(\mathcal{P})$ or symbol δ describes the maximum distance between a point $r \in \mathrm{Conv}(\mathcal{P})$ and the closest point $s \in \mathcal{P}$. More precisely:

$$\mathrm{ncvx}(\mathcal{P}) = \max_{r \in \mathrm{Conv}(\mathcal{P})} \min_{s \in \mathcal{P}} \|r - s\|_2 \tag{4.2.2}$$

Further define a set of *vectors of non-convexity* as the vectors $\overrightarrow{rs}$ where r and s are all possible pairs that meet (4.2.2):

$$\mathcal{N}(\mathcal{P}) = \left\{ \overrightarrow{rs} \,\middle|\, r \in \text{Conv}(\mathcal{P}), s \in \mathcal{P}, \|r-s\|_2 = \text{ncvx}(\mathcal{P}), \right.$$
$$\left. \|r-s\|_2 = \min_{s' \in \mathcal{P}} \|r-s'\|_2 \right\} \tag{4.2.3}$$

The problem of the convexity of aggregates (i.e., Minkowski sums) of non-convex sets was studied by Shapley (who later won the Nobel prize in Economics), Folkman and Starr, with initial findings published in [Starr(1969)]. The main result is as follows:

Lemma 1 (Shapley-Folkman Lemma [Starr(1969)])**.** *Considering all possible subsets $\mathcal{J}$ of $\mathcal{I}$ with cardinality at most K, the union of the Minkowski sums $\bigoplus_{i\in\mathcal{J}} \text{Conv}(\mathcal{P}^i) \oplus \bigoplus_{i\in\mathcal{I}\setminus\mathcal{J}} \mathcal{P}^i$ is a superset to the convex hull of the Minkowski sum $\bigoplus_{i\in\mathcal{I}} \mathcal{P}^i$:*

$$\text{Conv}\left(\bigoplus_{i\in\mathcal{I}} \mathcal{P}^i \right) \subseteq \bigcup_{\mathcal{J}\subseteq\mathcal{I}, |\mathcal{J}|\leq K} \left(\bigoplus_{i\in\mathcal{J}} \text{Conv}(\mathcal{P}^i) \oplus \bigoplus_{i\in\mathcal{I}\setminus\mathcal{J}} \mathcal{P}^i \right) \tag{4.2.4}$$

What the lemma means is that, by convexifying at most K sub-sets from the Minkowski sum at a time and overlaying (finding the union of) the feasible points of all such combinations on top of each other, the resulting union is a superset to the convex hull of the original Minkowski sum. The Shapley-Folkman Theorem that follows [Starr(1969)] provides more useful results, as it defines a bound on the non-convexity of $\mathcal{X}$, where they key message is that the bound *only depends on the dimension K and not on the number of sets N* (as long as $N \geq K$). Note that for the problems considered in this dissertation, the number of individual flexible loads $N = |\mathcal{I}|$ is much greater (in the thousands) than the dimension of the problem K (representing time periods, typically in the range 4-48), i.e. $N \gg K$. Starr's Corollary [Starr(1969)] tightens the bound provided by the SF theorem, expressing it in terms of the K largest non-convexities found among the sets $\mathcal{P}^i$; it concludes that the

relative distance between Conv($\mathcal{X}$) and $\mathcal{X}$ is inversely proportional to N, meaning that as $N \to \infty$, $\mathcal{X} \to$ Conv($\mathcal{X}$).

Starr [Starr(1969)] used the results to show that a competitive market economy could reach an equilibria using the convex hull of individual preferences, as long as the number of agents is larger than the dimension of the economy. Over the years vast literature of derived work has emerged in a variety of fields, including suggesting tighter bounds expressed in other metrics of non-convexity, see e.g. [Fradelizi *et al.*(2017)]. Looking at the Shapley-Folkman lemma from a market optimization viewpoint, recall that a prosumer has a total cost/utility function C which is a function of the quantity x produced/consumed. In conventional markets, the relationship between the optimal production quantity x^*, the price λ, and the cost function C is $\lambda = \frac{d\mathrm{C}}{dx}(x^*)$. This can be extended to higher dimensions with

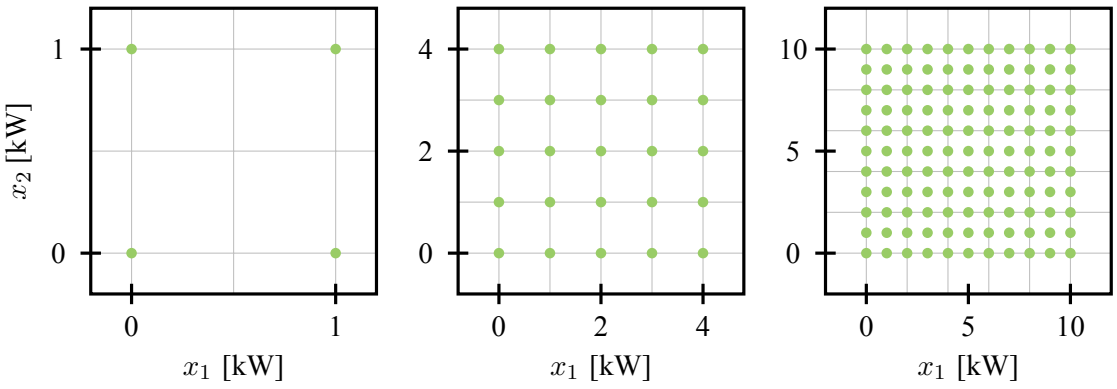
$$\boldsymbol{\lambda} = (\nabla \mathrm{C})(\boldsymbol{x}^*), \tag{4.2.5}$$

where $\boldsymbol{x}^* \in \mathbb{R}^K$ and $\boldsymbol{\lambda} \in \mathbb{R}^K$. With a population of prosumers, each having a feasible dispatch/load space $\mathcal{P}^i$ and corresponding cost/utility function $\mathrm{C}^i(\boldsymbol{p}^i)$, assume that these prosumers are *not controlled*, and instead that they behave independently of each other only considering their own constraints/cost functions given a price prediction $\boldsymbol{\lambda} \in \mathbb{R}^K$. If they behave rationally then, given a price $\boldsymbol{\lambda}$ the corresponding dispatch point $\boldsymbol{p}^i$ can be determined by solving:

$$\boldsymbol{p}^i = \underset{\boldsymbol{p}^i}{\arg\min}\ \mathrm{C}^i(\boldsymbol{p}^i) - [\boldsymbol{\lambda}]^\intercal \boldsymbol{p}^i \text{ s.t. } \boldsymbol{p}^i \in \mathcal{P}^i \tag{4.2.6}$$

For a given $\boldsymbol{\lambda}$ one can view this as a subproblem of a separable optimization problem, or alternately one can consider $\boldsymbol{\lambda}$ as the dual variable of a balance constraint and (4.2.6) to be a subproblem of dual decomposition, which alternates between a distributed (4.2.6) and an aggregate price update:

$$\boldsymbol{\lambda}^{j+1} = \boldsymbol{\lambda}^j + \alpha \left(\sum_{i \in \mathcal{I}} \boldsymbol{p}^i + \boldsymbol{y} \right) \tag{4.2.7}$$



(a) $N = 1$ (b) $N = 4$ (c) $N = 10$

Figure 4.2.1: Visualizing Starr's Corollary [Starr(1969)] for an aggregation of identical loads. As N grows, the absolute error stays the same, but the relative error (distance to the closest feasible point relative to the area/volume of the entire region) decreases.

where j indicates the iteration number and $\boldsymbol{y}$ the generation/consumption of other system participants. The authors of [Aubin and Ekeland(1976), Udell and Boyd(2016)] and related work show in a vein related to the SF lemma, that for separable problems with complicating constraints, the duality gap between the non-convex primal problem and its dual (which reflects the convexified problem) is bounded and proportional to the product of K and the largest non-convexity found in the resource set. Hence, as the population size increases, the relative duality gap goes to zero and the solution of the non-convex problem and its convexified counterpart converge. In [Bertsekas *et al.*(1983)] a similar argument is used to show that a UC problem has a vanishing duality gap as the number of generating units increases. In Section 4.3.2 a distributed method is proposed to aggregate individual loads based on sampling the population response from a price forecast, something only possible with the knowledge that the impact of non-convex individual behavior vanishes for aggregates of large populations.

4.2.2 The Shapley-Folkman Lemma and Demand Response

Coming back to the geometric interpretation of Lemma 1, how can it be applied to reason about the convexity of aggregates of DR resources? Starr's corollary [Starr(1969)] says that as $N \to \infty$, $\mathcal{X} \to \text{Conv}(\mathcal{X})$. To visually interpret this, imagine an aggregate of loads such as the one depicted in Figure 4.2.1(a), where $p_k \in \{0, 1\}$ kW. The aggregate set $\mathcal{X}$ composed of such loads, as seen in Figure 4.2.1, will be a lattice, with the maximum aggregate consumption $\max(x_k) = N$ kW, while the maximum non-convexity remains a constant $\text{ncvx}(\mathcal{X}) = \sqrt{K \cdot 0.5^2} = 0.707$. As N increases, the relative error (in the order of $0.707/N$) becomes insignificant relative to the aggregate load; in fact, only a single individual load needs to have its constraints violated for an aggregator to be able to dispatch continuously in the range $x_k \in [0, N]$ kW. What happens if in the aggregate there are loads that behave continuously (e.g. are convex)? If convex loads are absent, how many individual loads binary constraints need to be violated to treat the aggregate as convex? To answer these questions one needs to take a closer at what the approximation $\mathcal{X} \approx \text{Conv}(\mathcal{X})$ means. Start by defining a bound on the aggregate non-convexity in terms of individual non-convexity:

Lemma 2. *The non-convexity of the aggregate set is less than or equal to square root of sum squared of the K largest non-convexities from its composite members.*

$$\text{ncvx}(\mathcal{X}) \leq \max_{\mathcal{J} \subseteq \mathcal{I}, |\mathcal{J}| \leq K} \sqrt{\sum_{i \in \mathcal{J}} \left(\text{ncvx}(\mathcal{P}^i)\right)^2} \tag{4.2.8}$$

Proof. Start by sorting the sets $\mathcal{P}^i$ in order of non-convexity, such that $\mathcal{P}^1$ has the largest non-convexity and $\mathcal{P}^n$ the smallest. For an aggregate of two sets ($N = 2$) in a one dimensional space ($K = 1$), the aggregate convexity must be less than or equal to $\text{ncvx}(\mathcal{P}^1)$ as $\mathcal{P}^2$ simply adds points around the feasible points of $\mathcal{P}^1$ which only serve as to decrease the distance between any two points along the line ($K = 1$). Increasing the dimension to $K = 2$ the vectors of maximum non-convexity for the two sets could be orthogonal, in which case

the aggregate non-convexity is at most $\sqrt{(\text{ncvx}(\mathcal{P}^1))^2 + (\text{ncvx}(\mathcal{P}^2))^2}$. Adding another set $\mathcal{P}^3$ (with a smaller non-convexity) into the $K = 2$ dimensional space can only decrease the maximum non-convexity, as the vectors of maximum non-convexity $\mathcal{N}(\mathcal{P}^3)$ must be a linear combination of the vectors from $\mathcal{N}(\mathcal{P}^1) \cup \mathcal{N}(\mathcal{P}^2)$. By induction, one thus arrives at (4.2.8), that is, the maximum non-convexity is bounded by the K largest non-convexities (as Starr concluded), and (4.2.8) is an equality if all the first K sets have orthogonal vectors of maximum non-convexity. □

This non-convexity is visualized in Figure 4.2.2(e) and (f) where the maximum non-convexity (δ) is drawn with dashed lines. From the definition (4.2.1) of $\mathcal{X}$, several sets can be combined such that all that remains is the Minkowski sum of two sets, $\mathcal{Y}$ and $\mathcal{Z}$:

$$\mathcal{X} = \overbrace{\mathcal{P}^1 \oplus \cdots \oplus \mathcal{P}^Z}^{\mathcal{Z}} \oplus \overbrace{\mathcal{P}^{Z+1} \oplus \cdots \oplus \mathcal{P}^N}^{\mathcal{Y}} = \mathcal{Z} \oplus \mathcal{Y} \tag{4.2.9}$$

where $\mathcal{Z}$ is either the empty set or composed of Z convex members $\mathcal{P}^i$ (in which case $\text{ncvx}(\mathcal{Z}) = 0$), while $\mathcal{Y}$ contains the sum of the remaining $N - Z$ members not in $\mathcal{Z}$.

Further define the inner ($\text{innerr}(\mathcal{Z})$) and outer ($\text{outerr}(\mathcal{Z})$) radius of $\mathcal{Z}$ as the radius of the largest ball that can fit inside $\mathcal{Z}$ and the radius of the smallest ball that $\mathcal{Z}$ fits inside, as depicted in Figure 4.2.2(d). These are building blocks for the following lemma:

Lemma 3. *For a Minkowski sum* (4.2.9) *whose members $\mathcal{Z}$ and $\mathcal{Y}$ satisfy:*

$$\text{innerr}(\mathcal{Z}) \geq \text{ncvx}(\mathcal{Y}), \text{ncvx}(\mathcal{Z}) = 0 \tag{4.2.10}$$

the interior of $\mathcal{X}$ is strictly convex (has no holes). Here, the interior means any point at least $2 \cdot \text{outerr}(\mathcal{Z}) - \text{innerr}(\mathcal{Z})$ distance away (inside) from the surface of the object.

Proof. Given that the convex set $\mathcal{Z}$ (meeting the above conditions) super-imposed on every feasible point of $\mathcal{Y}$ fills up a surrounding spherical region of radius at least $\text{innerr}(\mathcal{Z})$, it fills

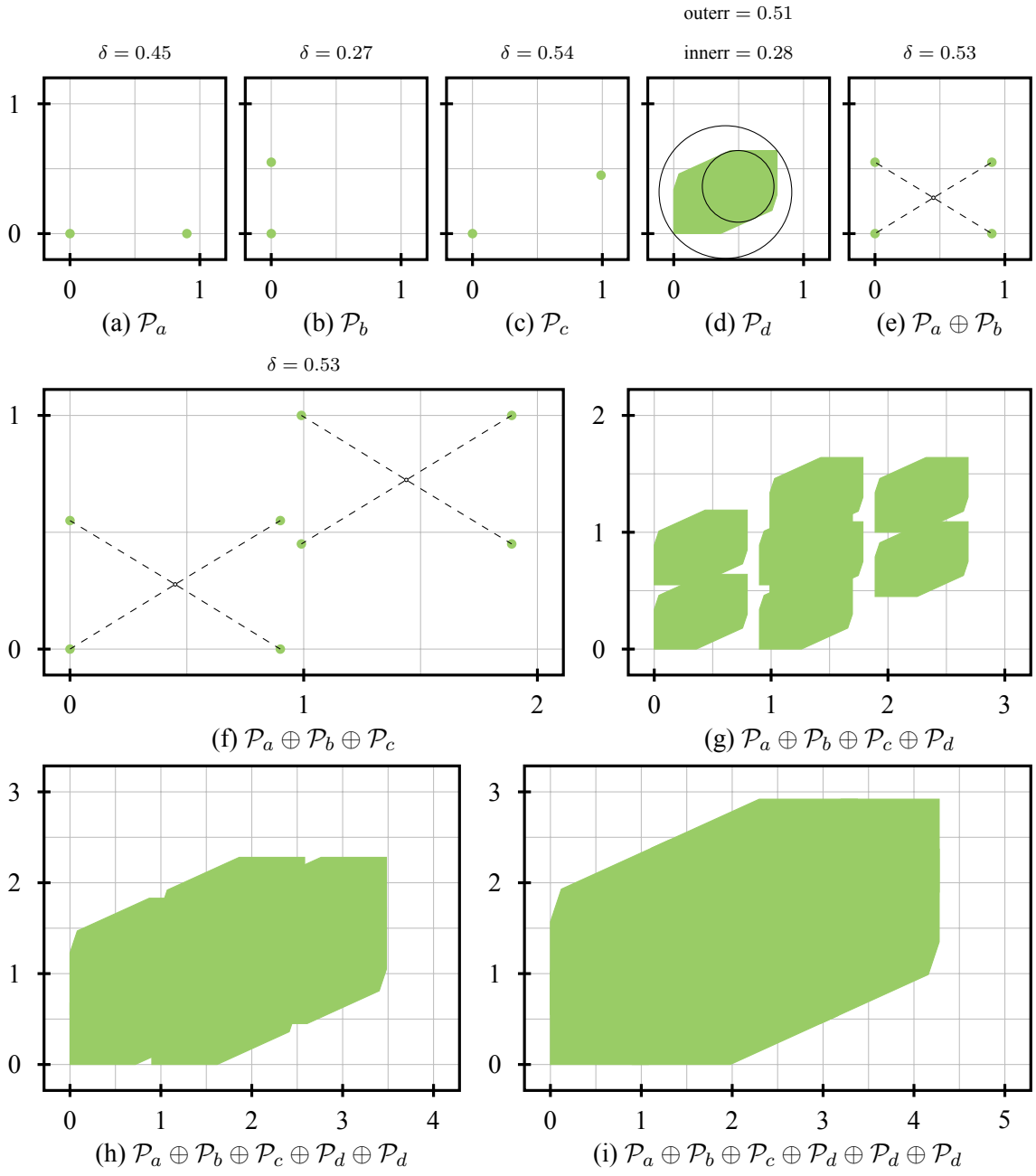


Figure 4.2.2: Two dimensional plots of the Minkowski sums of various loads. Figures (a) through (c) visualize non-convex loads $\mathcal{P}^a$, $\mathcal{P}^b$ and $\mathcal{P}^c$ with their measure of non-convexity ρ, while (d) shows a convex load $\mathcal{P}^d$ with its inner (innerr) and outer (outerr) radius. Figures (e) and (f) shows Minkowski sums of the non-convex sets, (g) adds $\mathcal{P}^d$ to the sum. Figure (h) adds another load with shape $\mathcal{P}^d$ to the aggregate, making the interior convex, while finally (i) adds yet another $\mathcal{P}^d$ making the Minkowski sum strictly convex.

in any non-convexities that are at most ncvx($\mathcal{Y}$) $\leq$ innerr($\mathcal{Z}$) away from the feasible points (it fills up all the holes of the lattice), i.e. Conv($\mathcal{Y}$) $\subseteq \mathcal{X}$. However, at the boundary of the region, depending on the specific shape of $\mathcal{Z}$ one may get a (non-convex) sawtooth-like behavior (see e.g. Figure 4.2.2(h)), meaning that $\mathcal{X}$ is not convex at the exterior. Considering a point x on the convex hull Conv($\mathcal{Y}$), for a $\mathcal{Z}$ that lies mostly orthogonal to the surface of Conv($\mathcal{Y}$) and is centered at x (the center of the maximum inner radius ball is at x), it may point outwards (from Conv($\mathcal{Y}$)) at most $2 \cdot$ outerr($\mathcal{Z}$) $-$ innerr($\mathcal{Z}$). With this, at a distance $2 \cdot$ outerr($\mathcal{Z}$) $-$ innerr($\mathcal{Z}$) from the exterior of the region, the remaining interior region is convex. $\square$

Remark 6. Note that this is a *sufficient condition* and often the interior becomes convex long before innerr($\mathcal{Z}$) $\geq$ ncvx($\mathcal{Y}$). Further, when arguing about the convexity, one would normally define $\mathcal{Z}$ as the *smallest* set satisfying innerr($\mathcal{Z}$) $\geq$ ncvx($\mathcal{Y}$) to minimize the depth of non-convexities $2 \cdot$ outerr($\mathcal{Z}$) $-$ innerr($\mathcal{Z}$).

What this means for DR aggregates is that with sufficient number of convex resources in the mix (devices that can operate continuously, e.g. storage devices) then any point in the interior of $\mathcal{X}$ is feasible without violating any individual constraints. This is visualized in Figures 4.2.2(g) and (h) where in the former case innerr($\mathcal{Z}$) $= 0.28 \leq 0.53 =$ ncvx($\mathcal{Y}$) whereas in the latter case another convex member has been added to $\mathcal{Z}$, meaning that innerr($\mathcal{Z}$) $= 0.56 \geq 0.53 =$ ncvx($\mathcal{Y}$) and the interior is convex. While in practice the non-convexities at the exterior are a non-issue as a large N means they are relatively shallow, is it possible to be more specific about what is happening at the exterior?

Define the operator $\mathfrak{F}(\mathcal{P})$ as returning the facets of $\mathcal{P} \subset \mathbb{R}^N$. If $\mathcal{P}$ is a convex polytope, then $\mathcal{F} \in \mathfrak{F}(\mathcal{P})$ is a polytope of dimension $n-1$ describing all points that lie on that hyperplane (facet). If $\mathcal{P}$ is non-convex, then $\mathcal{F} \in \mathfrak{F}(\mathcal{P})$ contains the points (usually vertices) that are members of $\mathcal{P}$ and lie on that particular facet of Conv($\mathcal{P}$). As such, $\mathcal{F} \in \mathfrak{F}(\mathcal{P})$

is simply a set of points on a hyperplane in $\mathbb{R}^n$ and associated normal vector (to know the "outward" direction), where the set has an associated non-convexity $\text{ncvx}(\mathcal{F})$ defined exactly as in (4.2.2). Further define the operator $\mathfrak{P}(\mathcal{F})$ as returning only the normal of the hyperplane, ignoring any shift in space or concept of size.

Lemma 4. *Consider an aggregate set $\mathcal{X}$. In order for $\mathcal{X}$ to be strictly convex ($\mathcal{X} = \text{Conv}(\mathcal{X})$), there must exist sets $\mathcal{Z}$ and $\mathcal{Y}$ such that $\mathcal{X} = \mathcal{Z} \oplus \mathcal{Y}$ (see* (4.2.9)*) that satisfy: (a) Lemma 3 and (b) for any facet $\mathcal{G} \in \mathfrak{F}(\mathcal{Y})$ there must exist a corresponding facet $\mathcal{F} \in \mathfrak{F}(\mathcal{Z})$ such that:*

$$\mathfrak{P}(\mathcal{F}) = \mathfrak{P}(\mathcal{G}), \quad \text{innerr}(\mathcal{F}) \geq \text{ncvx}(\mathcal{G}) \tag{4.2.11}$$

This means that for any facet $\mathcal{G}$ of $\mathcal{Y}$ there must exist a facet $\mathcal{F}$ of $\mathcal{Z}$ that faces in the exact same direction and whose inner radius exceeds that of the non-convexity of $\mathcal{G}$ ($\text{innerr}(\mathcal{F}) \geq \text{ncvx}(\mathcal{G})$). This effect is visualized in Figure 4.2.2(i) where three sets of the shape $\mathcal{P}^d$ are needed to obtain sufficiently large facets in $\mathcal{Z}$ to cover the non-convexities of the facets of $\mathcal{Y} = \mathcal{P}^a \oplus \mathcal{P}^b \oplus \mathcal{P}^c$. For DR aggregates this means that for $\mathcal{X}$ to be strictly convex, a sufficient number of resources of each type (having the same facets, but possibly scaled or stretched) need to be convex for the entire region to be convex. Actually, only the single largest resource of each type needs to be convex to make the entire aggregate convex, irrespective of the dimension K or number of devices N. Here largest means having facets whose surface covers any non-convexities found in other sets from the same type, as defined in Lemma 4. In Section 4.4.1 a small experiment is devised to showcase these results.

4.2.3 Convexity of Aggregate Cost

Prior sub-sections only looked at the convexity of the feasible region $\mathcal{X}$, but what can be said about the convexity of the aggregate cost? The *cost* corresponding to each $\boldsymbol{x}$ is:

$$C(\boldsymbol{x}) = \inf\left\{ \sum_{i=1}^{N} \mathrm{C}^i(\boldsymbol{p}^i) \middle| \boldsymbol{x} = \sum_{i=1}^{N} \boldsymbol{p}^i, \boldsymbol{p}^i \in \mathcal{P}^i \right\} \tag{4.2.12}$$

Here, the infimum ensures that the cheapest combination of individual loads is chosen to deliver a particular aggregate load $\boldsymbol{x}$. To argue about the aggregate cost convexity, one can build on the same theory as earlier sections:

Lemma 5. *By considering cost as an additional dimension to the set $\mathcal{P}$, the same logic applies as is done for the power profiles in Lemmas 1-4 to establish, under the same conditions, when the aggregate cost can be considered approximately or strictly convex, or not.*

Proof. Recall that $\mathrm{C}^i(\boldsymbol{p}^i)$ denotes the smallest cost of procuring $\boldsymbol{p}^i \in \mathcal{P}^i$ for individual i. Define epi f as the epigraph of a function f, (i.e. the set of points above the function). The following set adds the cost value as a dimension to $\mathcal{P}^i$:

$$\hat{\mathcal{P}}^i = \left\{ [p_1^i, \ldots, p_K^i, c]^\intercal \middle| \boldsymbol{p}^i \in \mathcal{P}^i, c \in \text{epi}\, \mathrm{C}^i(\boldsymbol{p}^i) \right\}. \tag{4.2.13}$$

Now the Minkowski sum can be applied to the set including the cost:

$$\hat{\mathcal{X}} = \bigoplus_{i \in \mathcal{I}} \hat{\mathcal{P}}^i. \tag{4.2.14}$$

The conditions under which $\hat{\mathcal{X}}$ is approximately convex are the same as for $\mathcal{X}$ following Lemmas 1-4. The aggregate cost function chooses the least expensive way to procure a particular $\boldsymbol{x}$ and is thus a surface of an approximately convex set $\hat{\mathcal{X}}$:

$$C(\boldsymbol{x}) \approx \inf\left\{ c \middle| [x_1, \ldots, x_K, c] \in \text{Conv}(\hat{\mathcal{X}}) \right\} \tag{4.2.15}$$

and therefore $C(\boldsymbol{x})$ is an approximately convex function. □

This means that without any assumptions on the individual cost functions or feasibility regions, *both the aggregate region and cost function are approximately convex*, and the number of resources that need to be convex or convexified is only a fraction of all the aggregated resources, assuming the dimension K is small compared to the number of resources N.

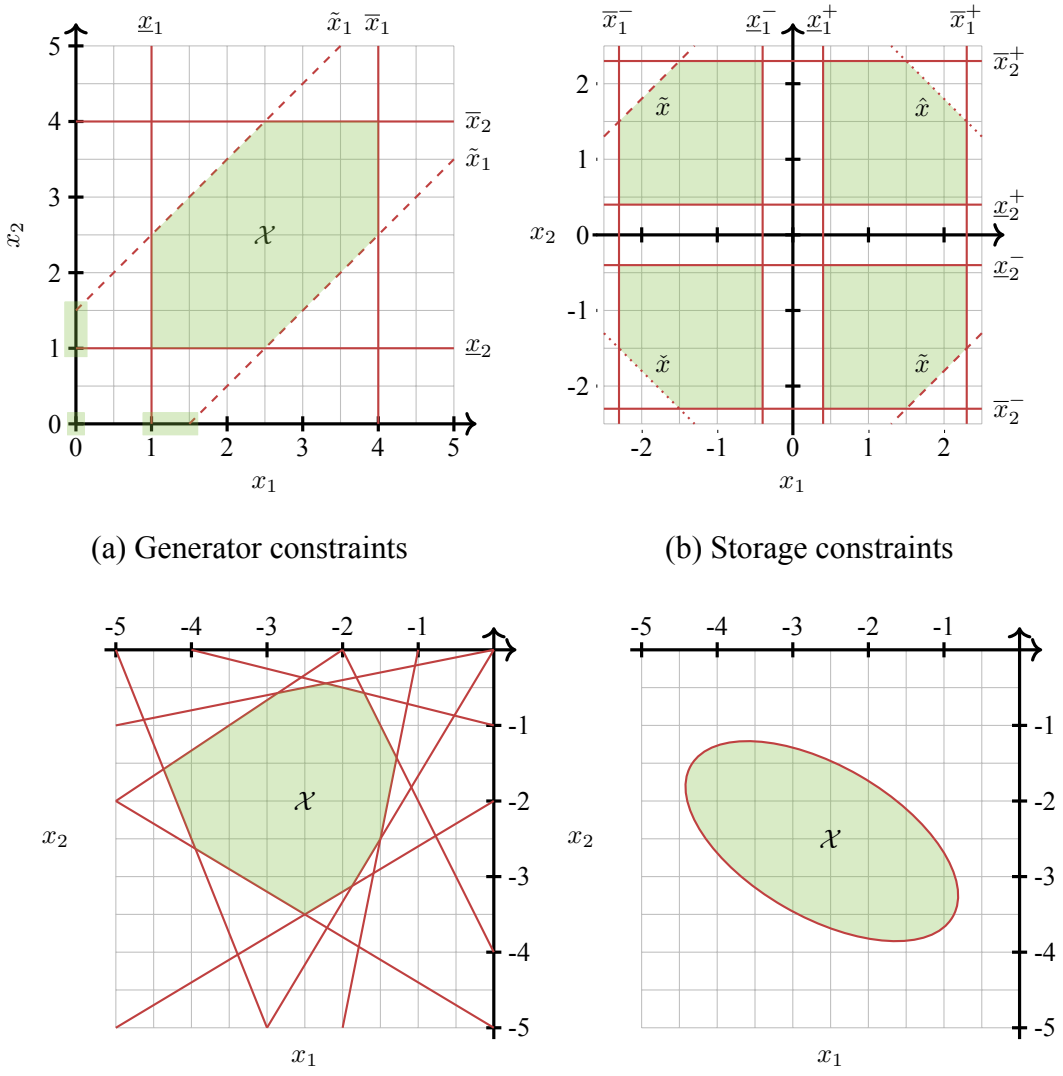
4.3 Interfacing with the System Operator

The previous section establishes the conditions under which the aggregate $\mathcal{X}$ and $C(\boldsymbol{x})$ can in practice be considered convex, showing that a competitive market with many complicated but "small" participants can be efficient. Nonetheless, passing the description of thousands or tens of thousand resources served by a single transmission system bus, one can still wind up with a very complex $\mathcal{X}/C(\boldsymbol{x})$, because of the curse of dimensionality. To include the aggregate DR in energy market models (where individual participants must have a capacity in the order of MW) a low-order approximation of $\mathcal{X}/C$ must be found that allows the ISO to leverage the DR flexibility without incurring excessive computational burden.

Here, the results of the previous section are leveraged, and the fact that the Minkowski sum and convex hull operators are commutative, to construct an approximation of $\mathcal{X}$ as if there is no residual non-convexity, i.e.:

$$\mathcal{X} = \text{Conv}\left(\bigoplus_{i\in\mathcal{I}} \mathcal{P}^i\right) = \bigoplus_{i\in\mathcal{I}} \text{Conv}(\mathcal{P}^i). \tag{4.3.1}$$

Note that the calculation of the Minkowski sum is intractable for very heterogeneous populations and, therefore, trying to compute it and then simplify it is not a viable approach. Instead, all the individual (relaxed) constraints can be included in a large Linear Program (LP). The following sub-sections look at how such an LP can be used to find low-order models. Even though this LP can be quite large for thousands of devices, it still solves reasonably fast with modern solvers and hardware.



(a) Generator constraints

(b) Storage constraints

(c) Custom constraints

(d) Elliptical constraints

Figure 4.3.1: Constraints (red) and corresponding feasible region (blue) of several proposed aggregate resource descriptions for two consecutive periods.

4.3.1 Reduced Order Polytopal Constraints

ISO optimization models are usually either linear programs (LPs) or mixed-integer LPs (MILPs), solvable with modern solvers in a reasonable amount of time (from seconds to tens of minutes depending on the application). The feasible region of a LP is a polytope which means that our polytopal model $\mathcal{X}$ from Section 4.2 can be placed directly in existing ISO models, either using H-rep. $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$, or alternatively, through a convex combination of vertices $\boldsymbol{v} \in \mathcal{V} = \text{Vert}(\mathcal{X})$:

$$\boldsymbol{x} = \sum_{i=1}^{|\mathcal{V}|} \mathbf{v}^i w^i \ , \ \sum_{i=1}^{|\mathcal{V}|} w^i = 1 \ , \ 0 \leq w^i \leq 1 \ \forall i. \tag{4.3.2}$$

If the complexity of including the full feasible region $\mathcal{X}$ directly in to the ISO optimization programs is prohibitive [Chen *et al.*(2016b)], the following sub-sections propose a few special cases of polytopes with a lower and more predictable complexity, of which the virtual generator and storage model can be found in existing literature, and are included here for comparison.

4.3.1.1 Virtual Generator

Market participation by aggregators may be limited to existing ISOs bid constructs designed for conventional generators. Generator models have constraints on (i) min and (ii) max power, (iii) max ramping, and min (iv) on and (v) off times. In this approximation of DR aggregates as virtual-generators, constraints minimum on/off times and maximum ramping are ignored. The reason is that individual DR devices are not large machines with considerable inertia (unlike conventional generators) which means that they can quickly (in a matter of seconds) go from minimum to maximum power; thus, such constraints are not useful. This leaves the min/max power constraints to work with. Figure 4.3.1a visualizes the constraints of a generator. By ignoring the ramping constraints, a generator in "power space" is a hyper-rectangle, a special case of a polytope. Equipped with an LP of stacked relaxed individual

constraints $\mathcal{X}$, the following optimization is one way to find the min/max power constraints:

$$\max_{\overline{\boldsymbol{x}},\underline{\boldsymbol{x}}} \ \|\overline{\boldsymbol{x}} - \underline{\boldsymbol{x}}\| \text{ s.t. } \overline{\boldsymbol{x}} \in \mathcal{X}, \underline{\boldsymbol{x}} \in \mathcal{X} \tag{4.3.3}$$

where $\overline{\boldsymbol{x}}$ and $\underline{\boldsymbol{x}}$ are vectors containing the min/max constraints the aggregator provides to the ISO and $\|\cdot\|$ is an appropriate metric. The simulations performed for this dissertation used $|\boldsymbol{x}| = \min \boldsymbol{x}$ so that (4.3.3) is an LP, meaning that the solver finds the largest hyper-cube within $\mathcal{X}$. Other reasonable norms include the L_1 norm for the largest sum flexibility or the geometric mean for the maximal volume. If the shape of $\mathcal{X}$ is unlike a hyper-rectangle, large parts of the flexibility available to the aggregator are lost in translation; the ISO will not know about it or have the possibility of leveraging it.

4.3.1.2 Storage Model

Another low-order polytopic model is that of a storage device. This has been gaining traction as Federal Energy Regulatory Commission (FERC) Order 841 mandates ISOs to include such models in energy markets (see Figure 4.3.1b for a visualization of the constraints mandated by FERC Order 841). For DR aggregates, the most significant improvement is the inclusion of a minimum/maximum constraint on energy (the sum of power).

In the literature there are several approaches to calculate the bounds of a virtual battery such as [Hao *et al.*(2014)] where virtual battery parameters are calculated directly for a particular DR population, or more general approaches such as [Zhao *et al.*(2017)] or [Müller *et al.*(2017)], though they may leave significant amount of feasible volume on the table if some individual resources are not well aligned with the storage constraints. Determining the min/max energy values from an optimization model such as (4.3.3) is not straight-forward and thus omitted here, and the reader is instead referred to existing models cited here. In Section 4.4 the from Hao et al. [Hao *et al.*(2014)] is used for comparison. In the following

sub-section a generic polytopal model is proposed instead, which has additional degrees of freedom compared to those limited to the constraints of storage devices.

4.3.1.3 Custom Constraints

The models discussed in Sections 4.3.1.1 and 4.3.1.2 have constraints with physical interpretations, but a bundle of DR resources may not conform well to these constraints [Barot and Taylor(2017), Zhao *et al.*(2017), Nazir *et al.*(2018), Müller *et al.*(2017)]. As an improvement, this section proposes a polytopic model that is of reduced complexity compared with the full $\mathcal{X}$. Again the polytope could be passed on to the ISO in H-form ($\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$) or V-form (4.3.2). The latter form adds K constraints and $|\mathcal{V}|$ continuous variables to the ISO model, while the former adds zero variables but as many constraints as there are rows (facets) in $\boldsymbol{A}$. In [Barot and Taylor(2017)], the authors, through an outer approximation that can grow in complexity given a diverse resource set (dissimilar rows of $\boldsymbol{A}$), build an aggregate polytope from the "ground up". Similarly [Zhao *et al.*(2017), Müller *et al.*(2017)] approximate the individual through prototype polytopes whose Minkowski sum can be calculated efficiently. Here, it is instead assumed that the complexity of the (convexified) aggregate $\mathcal{X}$ is manageable on its own, a reasonable assumption for some tens of thousands of convexified load models, but that the complexity prevents it from direct inclusion in other ISO models, particularly if many aggregates (for many buses or areas) need to be included. For that purpose, the complexity of $\mathcal{X}$ needs to be reduced, and in what follows two heuristics are proposed.

I: Cost Scenario Based Reduction: If one has a good mechanism to generate market price scenarios, the approximation can be tailored to capture the parts of the feasible region that are most significant given these predictions. Given a list of price scenarios $\boldsymbol{\lambda}^s \in \mathbb{R}^K$, $s \in \mathcal{S}$, start with an empty set $\mathcal{V} = \emptyset$ and solve for the optimal power profile given each price scenario:

$$\boldsymbol{x}^s = \underset{\boldsymbol{x}}{\arg\min}\ \boldsymbol{x}^\intercal \boldsymbol{\lambda}^s,\ \text{s.t. } \boldsymbol{x} \in \mathcal{X} \tag{4.3.4}$$

Adding the result $\boldsymbol{x}^s$ to $\mathcal{V}$ and determine whether desired complexity has been reached, in which case the algorithm stops. $\mathcal{X}'$ is then passed to the ISO in H- or V-form.

Remark 7. Unlike many algorithms that make predictions about future prices to build offer/bids to submit to the market (see Section 1.2.3) in an attempt to maximize some expected gain/minimize cost, the set $\mathcal{X}'$ will contain the optimal aggregate load/dispatch for *any* of the price scenarios considered; If the set of price predictions contains the correct future price, the approximation $\mathcal{X} \rightarrow \mathcal{X}'$ *will contain the corresponding optimal load profile*.

II: Geometric Elimination: The following heuristic is proposed to reduce the complexity of the aggregate region by removing vertices that are very close to others, thus decreasing the number of facets in the polytope. This algorithm is only feasible if one can do the translation from H to V-representation [Barot and Taylor(2017)] required for step 1). As such it is intractable for very complex $\mathcal{X}$ but can be used in conjunction with the Cost Scenario Based Reduction, where a low order representation of $\mathcal{X}$ has been obtained with a small number of vertices, to further eliminate vertices that are close to each other. The step-by-step algorithm is:

1) Assign $\mathcal{V} \leftarrow$ Vertices$(\mathcal{X})$.

2) Terminate if desired complexity of Conv$(\mathcal{V})$ is reached.

3) Calculate the distance between each pair of vertices $\boldsymbol{x}^u, \boldsymbol{x}^v \in \mathcal{V}$. Set $U = |\mathcal{V}|$.

4) For any $(u, v) \in \{(u, v)|u \in \{1, \ldots, U\}, v \in \{u+1, \ldots, U\}\}$, calculate the distance $\Delta^{u,v} = \|\boldsymbol{x}^u - \boldsymbol{x}^v\| \; \forall \boldsymbol{x}^u \in \mathcal{V}, \boldsymbol{x}^v \in \mathcal{V}$

5) Find the (u, v) pair that has the smallest distance $\Delta^{u,v}$.

6) Remove vertex $\boldsymbol{x}^v$ from $\mathcal{V}$, and go to step 2).

4.3.2 Low Order Elliptical Constraint

This section argues that a concise geometric shape to capture a large part of the feasible region is a K-dimensional ellipsoid. The insight leading to this consideration ties back to the SF lemma and the Central Limit Theorem (CLT). Left uncontrolled, individuals would base their economic decisions on operating their flexible resource *independently*. As stated in Section 4.2.3, the SF lemma and derived work [Aubin and Ekeland(1976), Udell and Boyd(2016)], show that the aggregate response to price has a vanishing duality gap relative to the primal problem in the limit, since the relative error between the primal problem and its convex counterpart converge as the population grows. A related basic fact is that the sample space of the *sum of random variables* is the Minkowski sum of the summands respective sample spaces, connecting the phenomenon tied to SF lemma with the CLT, stating that sum of a large number of loads can be approximated with a multi-variate normal distribution. The confidence (or level) surface around the mean of a K-dimensional multi-variate normal distribution is a hyper-ellipsoid, and its interior can be used as an approximation of such Minkowski sums. Two common descriptions for an ellipsoid are (where $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}$ and $\boldsymbol{B}' = \boldsymbol{Q}\boldsymbol{\Sigma}^{-1}$):

$$\{\boldsymbol{x}|(\boldsymbol{x}-\boldsymbol{d})^{\intercal}\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{d}) \leq 1\} \Leftrightarrow \{\boldsymbol{x}|\boldsymbol{x} = \boldsymbol{B}'\boldsymbol{u} + \boldsymbol{d}|\|\boldsymbol{u}\|_2 \leq 1\} \tag{4.3.5}$$

where $\boldsymbol{d}$ denotes the center of the ellipsoid and $\boldsymbol{B}/\boldsymbol{B}'$ describe its rotation and stretch. Finding the largest hyper-ellipsoid (by volume) that can be inscribed into a polytope described by a (linear) aggregate is a convex problem [Boyd and Vandenberghe(2004), Lin *et al.*(2018)]:

$$\max_{\boldsymbol{B}',\boldsymbol{d}} \log\det\left(\boldsymbol{B}'\right)^{-1} \text{ s.t. } \|\boldsymbol{B}'[\boldsymbol{A}]_i\|_2 + [\boldsymbol{A}]_i^{\intercal}\boldsymbol{d} \leq \boldsymbol{b}_i \; \forall i \tag{4.3.6}$$

where $[\boldsymbol{A}]_i$ is the i-th row of the polytope $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$. Note that (4.3.6), though convex, is in general an intractable problem for a complex $\mathcal{X}$ with a very large $\boldsymbol{A}$.

Taking this reasoning further, the ellipsoid can be found directly without first mapping out a polytope and solving (4.3.6), by using price scenarios $\boldsymbol{\lambda}^s$, $s \in \mathcal{S}$ to sample the region $\mathcal{X}$ and cost $C(\boldsymbol{x})$. Corresponding to each price scenario a set of individual dispatch profiles $\boldsymbol{p}^{i,s}$ and cost $\mathrm{C}^i(\boldsymbol{p}^{i,s})$ is obtained by solving (4.2.6) in parallel (and *privately* by the prosumers); the aggregate load profile and cost of each scenario $\boldsymbol{x}^s = \sum_{i\in\mathcal{I}} \boldsymbol{p}^{i,s}$ and cost $C(\boldsymbol{x}^s) = \sum_{i\in\mathcal{I}} \mathrm{C}^i(\boldsymbol{p}^{i,s})$, are samples for the aggregate dispatch profiles region $\mathcal{X}$ and cost $C(\boldsymbol{x})$. Assuming that:

(a) the number of realistic price scenarios is sufficiently large, and

(b) individual prosumer cost/utility function coefficients are distributed in a continuous range (i.e. not many individuals have the exact same cost parameters),

then, due to the CLT the density of aggregate load vectors $\boldsymbol{x}$ observed in different regions of $\mathbb{R}^K$ should match the density of a multi-variate distribution. Given the set of price scenarios $\mathcal{S}$ and corresponding load profiles $\mathcal{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be estimated and the feasible region directly defined as:

$$(\boldsymbol{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq Q(\phi) \tag{4.3.7}$$

where $Q(\phi)$ is the quantile function (inverse Cumulative Distribution Function (CDF)) of the Chi-squared distribution with K degrees of freedom, which reflects the (symmetric) probability mass ($\phi \in [0, 1]$) contained within the ellipsoid, a smaller ϕ makes the solution more robust against any uncertainty. It is clear that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ depend on the set of price scenarios $\mathcal{S}$. If the aggregate cost surface C is convex, the *price* function $\boldsymbol{\varphi}(\boldsymbol{x}) = C'(\boldsymbol{x})$ is monotonically increasing. If one thinks of the price scenarios being sampled from a distribution $\boldsymbol{\Lambda}$, there will be a corresponding distribution $\boldsymbol{X}$ of points in power space, which is obtained through the mapping $\boldsymbol{\varphi}^{-1}(\boldsymbol{\Lambda})$ assuming $\boldsymbol{\varphi}$ is strictly increasing (a one-to-one function). Given $\boldsymbol{\varphi}$, prices $\boldsymbol{\lambda}$ can be to obtain any distribution $\boldsymbol{X}$ (e.g. uniform) which is useful to e.g. fill the power space uniformly and build $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from that. However, as C

and $\boldsymbol{\varphi}$ are not known, obtaining a predictable distribution $\boldsymbol{X}$ calls for studying a sampling algorithm. Leaving that for future work, in the simulations provided here, samples $\boldsymbol{\lambda}$ are taken from a distribution describing anticipated market prices, and thus fill the power space and build the statistics using points reflecting those prices. As a DR aggregate could impact market prices, choosing the price scenarios becomes a circular problem. However, since the volume of the ellipsoid gives a margin of error in the price estimation, and the impact of the DR aggregate on market prices can be learned over time, this approximation is promising.

4.3.3 Aggregate Cost Approximation

Having discussed the various approximations for $\mathcal{X}$, now the attention is turned towards finding an aggregate bid that best reflects the sum cost of each responsive load. Continuing the reasoning of Section 4.3.2, this section argues that the true aggregate cost function C describes a part of the surface (boundary) of an ellipsoid. As such, the relationship between power $\boldsymbol{x}$ and cost c is, with $\hat{\boldsymbol{x}} = [\boldsymbol{x}^\intercal, c]^\intercal$:

$$(\hat{\boldsymbol{x}} - \hat{\boldsymbol{\mu}})^\intercal \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{x}} - \hat{\boldsymbol{\mu}}) = \hat{Q} \tag{4.3.8}$$

where the hat indicates augmented vectors/matrices with added entries for the cost term, $\hat{\mu} \in \mathbb{R}^{K+1}$, $\hat{\boldsymbol{\sigma}} \in \mathbb{R}^{(K+1)\times(K+1)}$ and $\hat{Q} \in \mathbb{R}$. By re-arranging (4.3.8) to a function $c = C(\boldsymbol{x})$ it is clear that C contains the square root of various quadratic and cross-terms of $\boldsymbol{x}$. However, as the function describes the surface of an ellipsoid, it is certainly convex.

Since passing such a function to the market is not current practice, this section looks at Piece-Wise Linear (PWL) cost functions, where independent PWL functions are used to describe the cost for each time interval k, that is, the total cost is decomposable as $C(\boldsymbol{x}) = C_1(x_1) + \cdots + C_K(x_K)$. This is the format used in present day markets for generator costs. What follows is a simple heuristic using a predetermined number of knots (function joints) along each dimension $k \in \mathcal{K}$ and finding the optimal PWL parameters

to describe the PWL bid. Figure 4.3.2 shows a two-dimensional case where the sum of the individual cost functions make up a cost surface. Point pairs $(\boldsymbol{x}^s, C(\boldsymbol{x}^s)), s \in \mathcal{S}$ are sampled from Conv($\mathcal{X}$) and then the following optimization program is suggested to find the PWL parameters that best approximate the surface. For a resolution of ν pieces per dimension, creating a grid of ν^K hyper-rectangles, evenly spaced cost function knots are at $\kappa_{k,i} = \underline{x}_k + \frac{i}{\nu}(\overline{x}_k - \underline{x}_k), i \in \{0, 1, \ldots, \nu\} \in \mathcal{U}$ where $\overline{x}_k$ and $\underline{x}_k$ denote the largest and smallest value of x_k, $\boldsymbol{x} \in \mathcal{X}$. The corresponding line segments are $\alpha_{k,i}x_k + \beta_{k,i}$ for $\kappa_{k,i-1} \leq x_k \leq \kappa_{k,i}, i \in \{1, 2, \ldots, \nu\}$, meaning that $C_k(x_k) \geq \alpha_{k,i}x_k + \beta_{k,i}$ for all segments $i \in \mathcal{U}$. Denote the samples as $\boldsymbol{x}^s$ with $s \in \mathcal{S}$, the approximated cost at $\boldsymbol{x}^s$ as E^s, and the mapping from $\boldsymbol{x}^s$ to the appropriate PWL segment i at time k as $\psi(x_k^s, k)$. Solve for $(E^s, \alpha_{k,i}, \beta_{k,i})$ with input parameters $(\boldsymbol{x}^s, C(\boldsymbol{x}^s), \kappa_{k,i}, \psi)$:

$$\min_{E,\alpha,\beta} \max_{s \in \mathcal{S}} (E^s - C(\boldsymbol{x}^s)) \tag{4.3.9a}$$

$$\text{s.t. } E^s = \sum_{k=1}^{K} \alpha_{k,\psi(x_k^s,k)} x_k^s + \beta_{k,\psi(x_k^s,k)} \; \forall s \in \mathcal{S} \tag{4.3.9b}$$

$$E^s \geq C(\boldsymbol{x}^s) \; \forall s \in \mathcal{S} \tag{4.3.9c}$$

$$\alpha_{k,i} \geq \alpha_{k,i-1} \; \forall k \in \mathcal{K}, i \in \{2, 3, \ldots, \nu\} \tag{4.3.9d}$$

$$\alpha_{k,i}\kappa_{k,i} + \beta_{k,i} = \alpha_{k,i+1}\kappa_{k,i} + \beta_{k,i+1} \forall k \in \mathcal{K}, \; i \in \{1, 2, \ldots, \nu - 1\} \tag{4.3.9e}$$

where the objective attempts to minimize the maximum error, (4.3.9b) calculates the sum cost from the PWL functions for each dimension, (4.3.9c) ensures that the approximation does not under-estimate the cost anywhere, (4.3.9d) ensures the convexity of the PWL functions (each segment has greater slope than the previous one), while (4.3.9e) ensure the functions are continuous across knots.

Remark 8. Even if C is convex, the approximation above is not likely to accurately capture its complexity, except when limited to a small subset of $\mathcal{X}$. Maximizing the size of the

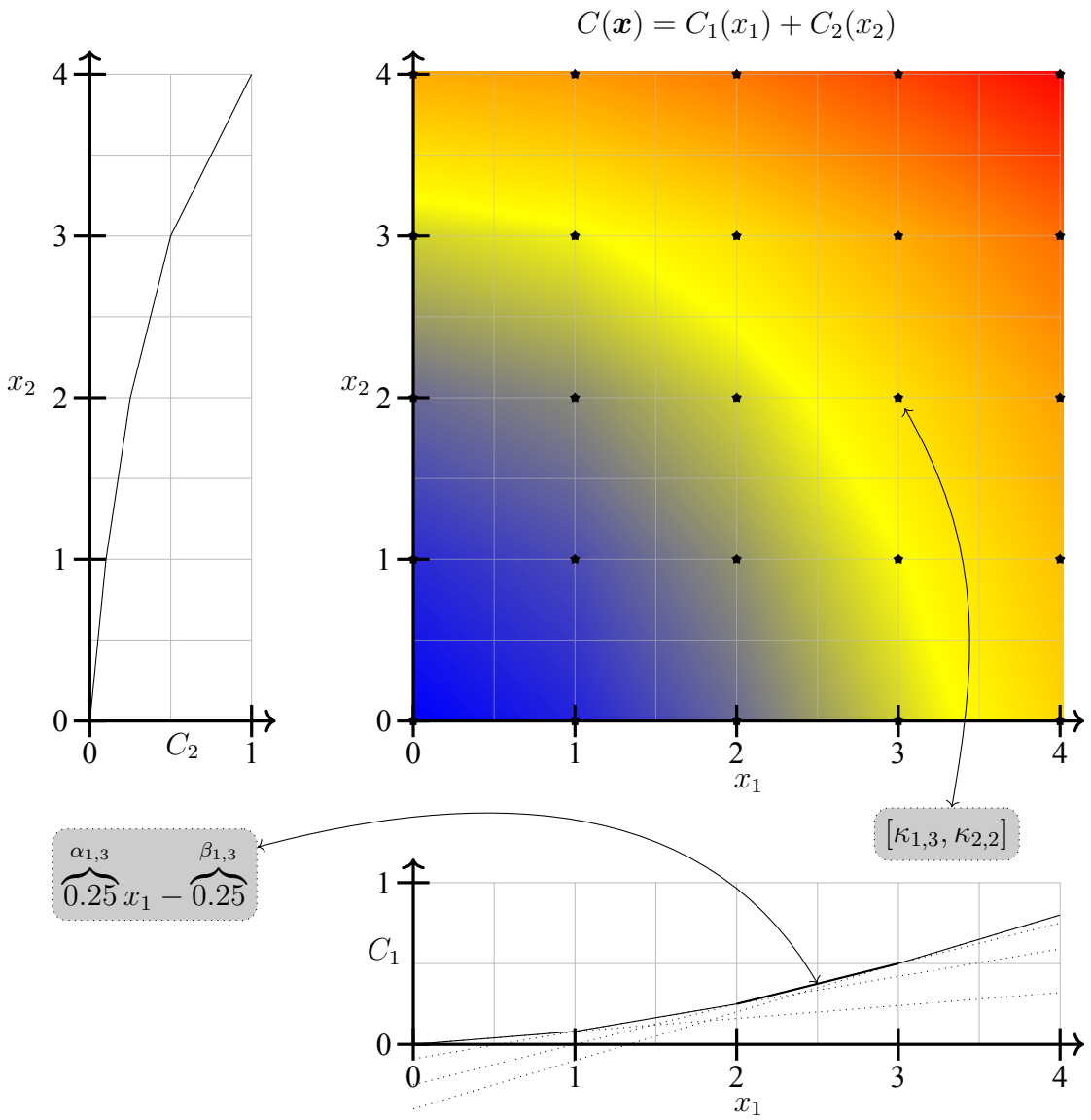


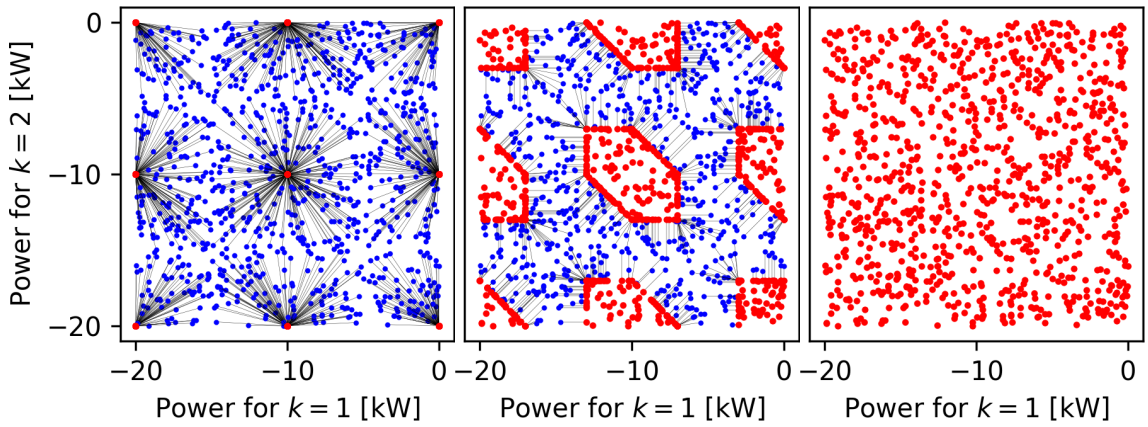
Figure 4.3.2: The PWL cost approximation visualized with $\nu = 5$, with C_1 and C_2 denoting the independent cost function for the first and second period, $C(\boldsymbol{x}) = C_1(x_1) + C_2(x_2)$ denoting the joint cost function and the stars indicating the knots κ in the surface, here spaced out on the major ticks.

feasible region and obtaining a good cost approximation may therefore often be opposing goals.

4.3.4 Disaggregation and Aggregator Revenue Models

In this dissertation it is assumed that the, similar to an ISO, is a neutral third party that simply works to aggregate the individual bids. For a cost surface approximation that accurately captures the aggregate cost, the aggregator only needs to pass on the market price to the individuals for them to dispatch themselves correctly. If the cost surface is not a good approximation of the actual aggregate cost, there are several ways to obtain the target dispatch. For the distributed elliptical model, one can solve a distributed iterative dual decomposition problem (4.2.6)-(4.2.7) to arrive at the target dispatch, similar to [Chang *et al.*(2013), Li *et al.*(2011b)]. For the other models one can for example solve the relaxed optimization problem given a fixed dispatch, and issue instructions to individuals based on that solution. For those decisions that are infeasible due to binary constraints, the individual can randomize their binary parameters weighted in accordance with the relaxed value. Alternately, the aggregator can employ randomized broadcast instructions to clusters of similar individuals to disaggregate the dispatch, as in [Alizadeh *et al.*(2014a), Hreinsson *et al.*(2020a)].

As far as the bid construction is concerned, it is assumed that the aggregator is a non-profit and its goal is to minimize the sum cost of the participating prosumers, which individually do not have market power. As for monetary settlement, the design of the aggregate cost curve is such that the what needs to be paid to the market (or comes from the market if net-producers) is less than what the individual loads are willing to pay, meaning there will be some left-over money at the aggregator. A possible service model is that the left-over money is used to operate the aggregator, or distributed back to the participants through a mechanism similar to the make-whole payments of the energy market. The design of an



(a) $N_{\text{CVX}} = 0$ (b) $N_{\text{CVX}} = 3$ (c) $N_{\text{CVX}} = 10$

Figure 4.4.1: Uniformly drawing sample points (blue) from $\text{conv}(\mathcal{X})$ and finding their closest corresponding point (red) in $\mathcal{X}$.

optimum bid that leverages the market position of the aggregator in the market goes beyond the scope of this paper.

4.4 Numerical Results

Numerical simulations were performed using Python and a collection of scientific programming libraries [Van Der Walt *et al.*(2011), Jones *et al.*(2014), Hunter(2007)], as well as utilizing Gurobi for solving optimization models [Gurobi Optimization(2015)]. For simulations a server with Intel Xeon E5-2680 v3 Central Processing Unit (CPU) was used.

4.4.1 Validation of the Shapley-Folkman Lemma

To validate results related to the SF lemma a simple case of $K = 2$ is considered with aggregates of two types of prosumers,

(a) non-convex loads where $p_k \in \{-10, 0\}$ kW, and

(b) storage like convex loads where $p \in [-1, 1]$ kW, $-1 \leq \mathbf{1} \cdot \boldsymbol{p} \leq 1$ kWh.

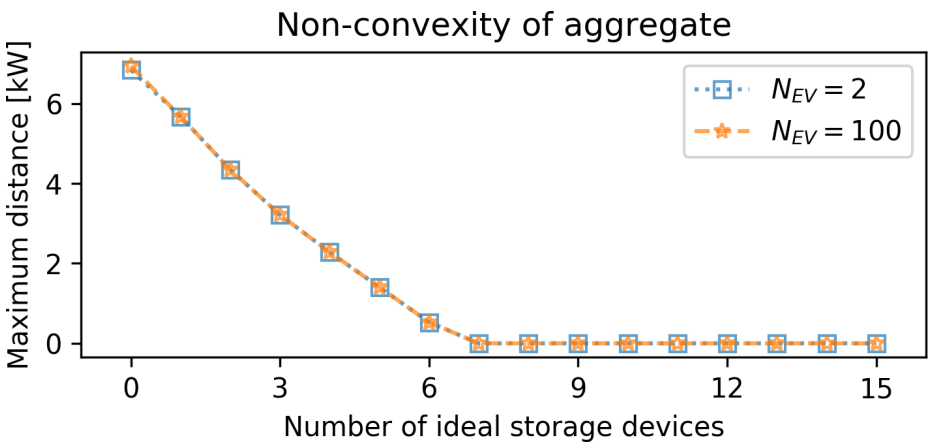


Figure 4.4.2: The gap between the full (binary) model $\mathcal{X}$ and the relaxed (continuous) one for vs the number of convex storage units.

Different combinations of non-convex loads $N_{\text{NCVX}} \in \{2, 100\}$ and convex loads $N_{\text{CVX}} \in \{0, 1, \ldots, 15\}$ are modeled. Figure 4.4.1 shows the feasible region $\mathcal{X}$ in the range $x_k \in [-20, 0]$ kW. In the $N_{\text{CVX}} = 0$ case, the non-convex loads form a lattice of feasible points (red dots), with the blue points showing $10,000$ random samples from $\text{Conv}(\mathcal{X})$ and the black lines indicating the closest feasible point. The maximum non-convexity here is $\sqrt{2 \cdot 5^2} = 7.07$ kW. As storage is added to the aggregate, a growing contiguous feasible region is observed around the lattice points, and at $N_{\text{CVX}} = 10$ it has filled in all the gaps. This is explained by Lemma 3, the inner radius of each storage device ($\underline{r} = \sqrt{2 \cdot 0.5^2}$) is $1/10$th of the observed non-convexity, meaning that an aggregate of 10 such devices has an aggregate inner-radius greater than the non-convexity of the non-convex loads. Figure 4.4.2 shows the maximum non-convexity as a function the number of continuous (ideal storage) devices.

4.4.2 Capacity Market Simulation

First model comparison is that of a capacity market, where payment is assumed to be proportional to offered capacity. The focus is on four models:

(a) The virtual generator model (Section 4.3.1.1).

(b) A set of custom constraints (Section 4.3.1.3).

(c) The storage model from Hao et al. [Hao *et al.*(2014)], where the three proposed variants (Table III, [Hao *et al.*(2014)]) of the sufficient (the inner approximation) model are included.

(d) The full model including all individual constraints.

The elliptical model is not included here as it builds on market price scenarios and is a better fit for energy market formulations. As [Hao *et al.*(2014)] is developed around an aggregate of TCLs this simulation is limited to include only TCLs for a more direct comparison. Starting with $K = 2$, Figure 4.4.3 visually compares the feasible regions of the three models for a population of $1,000$ TCLs. The custom constraints fill out the feasible region completely, and only five vertices are required to describe the entire region. The virtual generator rectangle sits on the bottom left where it can occupy the largest area and the three variants of the storage model from [Hao *et al.*(2014)] somewhat overlap with each other but emphasize different features.

Moving on to the $K = 24$ simulation, it includes a population of $15,000$ heterogeneous TCLs. As it is impossible to visualize such a high-dimensional region, the comparison is flattened in Figure 4.4.4 to two metrics averaged over 100 capacity market runs offering randomized price signals. Figure 4.4.4(a) shows the fraction of the capacity available by the different models, compared with the full region, essentially a comparison of the feasible region hyper-volume. Here, one can observe that the custom constraint set polytope performs well capturing about 75% of the region, the virtual generator coming in second at about 45% but the models from Hao et al. [Hao *et al.*(2014)] exposing between 10%-30% of the full feasible region. This is likely due to the fact that the parameters for the model from [Hao *et al.*(2014)] are partly a function of the TCLs with the smallest capacity, meaning that its volume suffers if the population it aggregates is diverse.

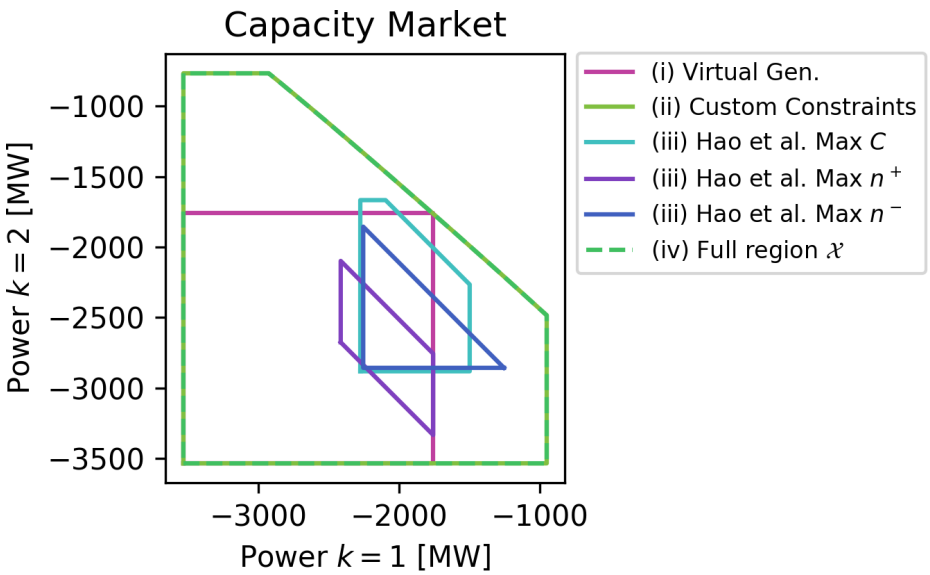


Figure 4.4.3: Visual comparison with $K = 2$ of the three models simulated in Section 4.4.2. Note that the models from Hao et al. [Hao *et al.*(2014)] suggest three different objectives when calculating the virtual battery parameters, with all of them included here for comparison.

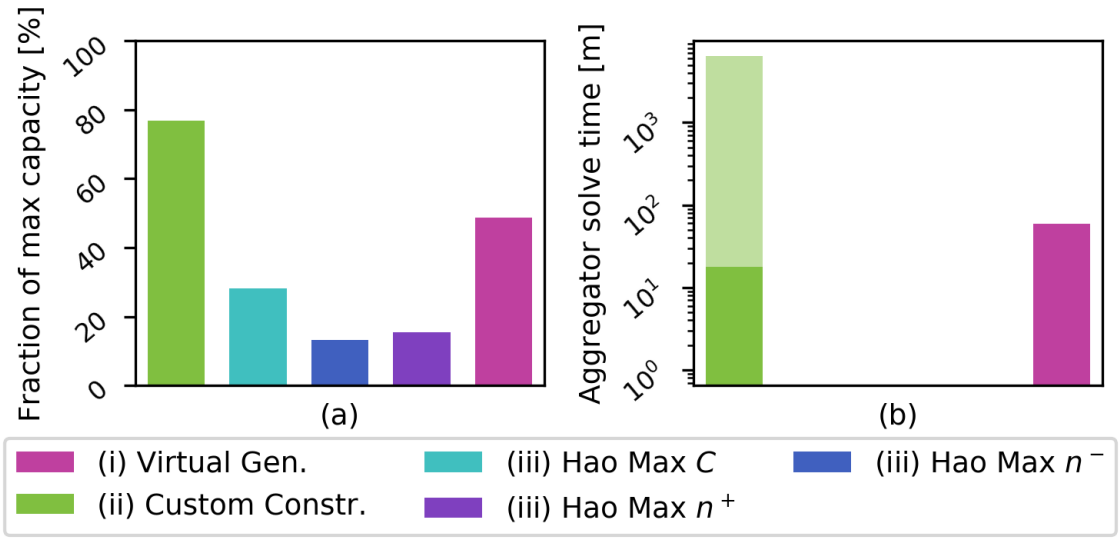


Figure 4.4.4: Capacity market comparison of the models listed in Section 4.4.2. Figure (a) shows the capabilities of the different models to provide reserve power, compared with the full model and can be interpreted as the difference in hyper-volume between the full region and the approximations (higher is better). Figure (b) shows the time required *by the aggregator* to build the model, with the darker shade indicating time that is not parallelizable, and light the time that can be reduced by solving across multiple CPUs.

Computationally the models from [Hao *et al.*(2014)] are extremely simple and fast, taking less than a second for the aggregator to compute, whereas the virtual generator takes tens of minutes and the custom constraints takes at least 20 minutes or possibly much longer depending on the number of CPUs available.

4.4.3 Energy Market Simulation

This comparison includes four models:

(a) The virtual generator model (Section 4.3.1.1).

(b) A set of custom constraints (Section 4.3.1.3).

(c) The approximate elliptical model (Section 4.3.2).

(d) The full model including all individual constraints.

A comparison with [Hao *et al.*(2014)] is not included as the paper does not include computations of aggregate costs. For $K = 2$, Figure 4.4.5 visually compares the feasible region of these models. This simulation also aggregated $1,000$ resources, of which 600 were TCLs, 260 EVs, and 120 and 20 non-ideal and ideal storage devices respectively. Here, one observes quite a different feasible region shapes compared with the population consisting of TCLs only, with the energy storage constraints being more evident given the added energy storage capacity. As a result of this changed region, the virtual generator captures less of the area, and the custom constraints also leave out the costlier parts of the region. The stars indicate for a sample market run the optimal dispatch point of the different models, which are heavily influenced by the cost function approximation quality. The approximate ellipsoids will be centered around the mean of the sampled dispatch points, where higher ϕ translates to a larger diameter.

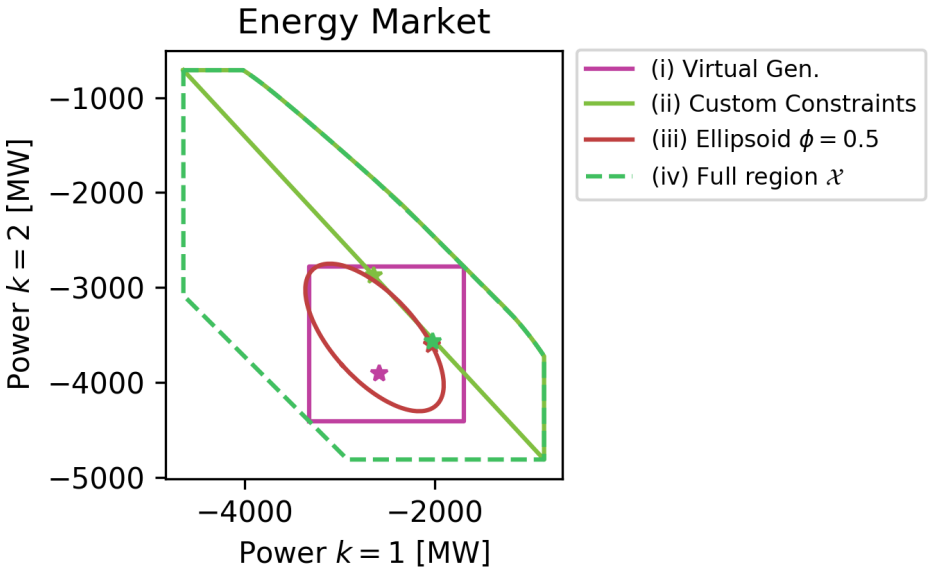


Figure 4.4.5: Visual comparison with $K = 2$ of the four models simulated in Section 4.4.3. The stars indicate the optimal dispatch points for a particular market solution.

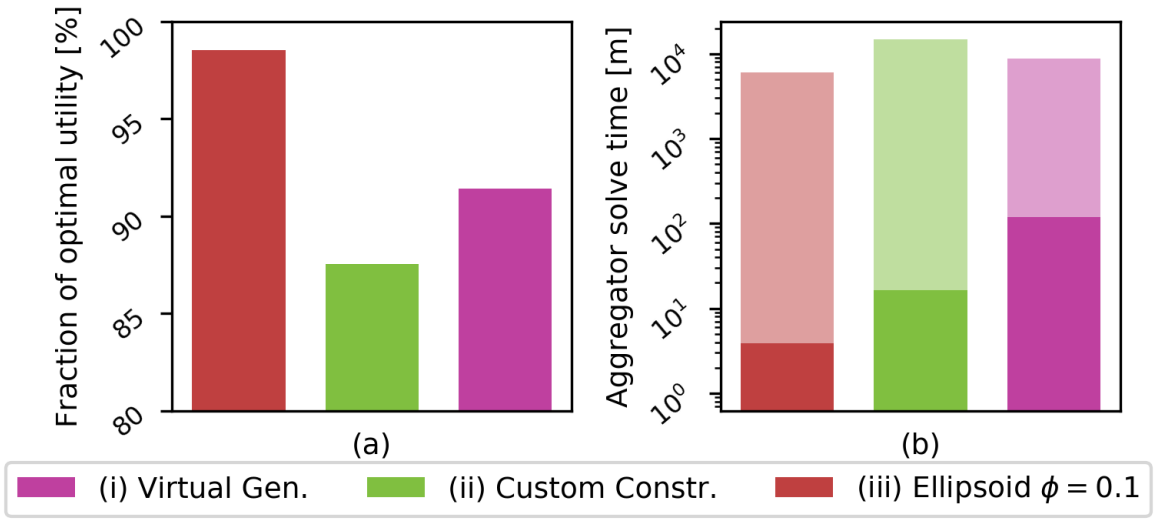


Figure 4.4.6: Energy market simulation of the models as described in Section 4.4.3. Figure (a) shows how close to the optimal solution (cost wise) the different models perform (higher is better), while (b) shows the aggregator solving time.

For the $K = 24$ simulation the population size is increased to $10,000$ TCLs, $4,450$ EVs, and 500 and 50 non-ideal and ideal storage devices, for a total of $15,000$ devices. Price scenarios spanning all the K intervals, both for model construction and for market experiments, were sampled from a multi-variate normal distribution with no cross correlation and a mean moving linearly from $\mu_1 = 9$ to $\mu_K = 5$ with standard deviation of $\sigma = 3$. For the approximate elliptical model $8,000$ samples were used to construct the model statistics, a trade-off between the accuracy of the statistics and the computational burden during model construction. Individual cost functions were structured as described in Section 4.1 and Figure 4.4.6 shows a comparison of methods modeled averaged over 100 price samples. Here the ellipsoidal model is a clear winner, both in terms of computational complexity and for how close it is to the optimal solution. Second best w.r.t. finding the optimal solution is the virtual generator and this can be explained by the poor cost approximation over the large region offered by the custom constraint model. Computationally however, assuming there is some parallel processing capabilities, the virtual generator performs the worst.

To summarize, different models have different strengths, the constraint set shines for capacity markets while the approximate elliptical shines for the energy market. In both cases the virtual generator performs poorly compared to the better candidate for each respective market.

4.4.4 Stochastic Security-Constrained Economic Dispatch with Thermostatically Controlled Loads

Like the simulations in Section 3.7.3.1, the load and renewable infeed data are from CAISO; the corresponding temperature data for seven California locations from NOAA [Diamond *et al.*(2013)]. From [Mathieu *et al.*(2012)] the population is assumed to be clustered with $R \in \{1.5, 2, 2.5\}$ °C/kW and $C \in \{1.5, 2, 2.5\}$ kWh/°C, allowing $\mathcal{S}_{\theta_r} = \{69, 72, 75, 78, 81\}$ °F. From [Energy Information Administration, U.S. Department of En-

ergy(2009)] a fixed number of $7,000,000$ households are assumed to use heat-pump air-conditioning. In order to obtain realistic parameters for the simulations, the model in (3.5.33) is reversed, using a simple linear regression to estimate how the population is spread between R/C clusters as well as how reference temperature changes over the course of the day. In practice, an Aggregator would know the individual $R/C/\theta_{\mathrm{r}}$ parameters of its population, and would cluster based on that knowledge. The way the scenario trees were generated is illustrated in Figure 4.4.7.

Simulations solve a rolling-horizon SCED with a two hour look-ahead window and a resolution of 15 minutes ($h = 900$s, $K = 8$). For 7 sample days scenario trees of net-load and temperature are built for each starting interval, using the closest (in terms of aggregate load) 50 sample days (excluding the target day) from the original data-set of 293 summer week-days (see Figure 4.4.7). Two variants of (3.6.1) are tested, in both cases considering any single generator outage $\mathcal{G}^{\mathrm{out}} \leftarrow \mathcal{G}$:

(i) A stochastic SCED (S-SCED) with 3 future scenarios.

(ii) A deterministic SCED (D-SCED) with a single forecast.

For the SCED $\leftrightarrow$ TCL interface, the following is simulated:

(a) An inflexible TCL aggregate, that simply consumes $\boldsymbol{p}^{\mathrm{b}}$.

(b) A virtual/negative generator model.

(c) A set of arbitrary constraints.

(d) The full set of constraints (3.5.34) included in the SCED.

All TCLs start at midnight in an energy neutral state ($u = 0$) and have the boundary condition (at the end of each rolling horizon window) to end in an energy neutral state. The RTS

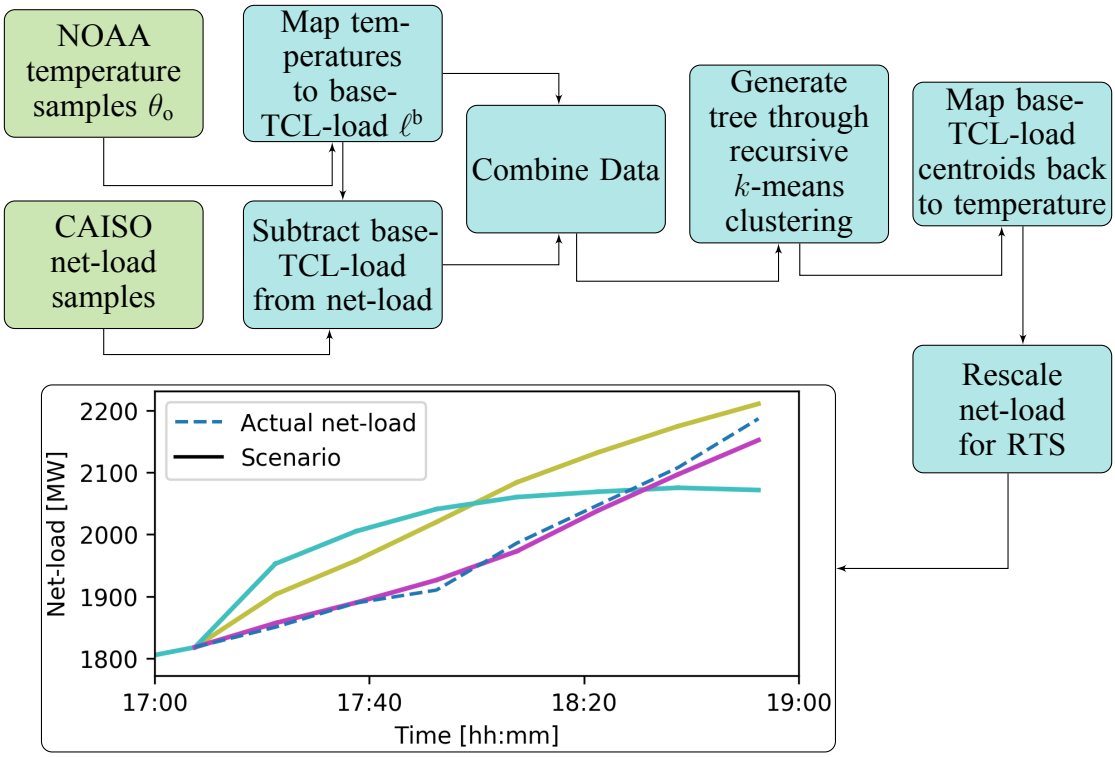


Figure 4.4.7: The process from input data to multi-variate scenario tree.

Table 4.1: Costs for the RTS case generators.

Unit Type	C_X	$\overline{C}_X$
Oil/Steam (U12)	$[55.15, 57.98]$	$[27.57, 28.99]$
Oil/CT (U20)	$[127.40, 132.60]$	$[63.70, 66.30]$
Hydro (U50)	$[0.00, 0.00]$	$[0.00, 0.00]$
Coal/Steam (U76)	$[15.76, 16.40]$	$[7.88, 8.20]$
Oil/Steam (U100)	$[43.01, 44.32]$	$[21.50, 22.16]$
Coal/Steam (U155)	$[12.14, 12.64]$	$[6.07, 6.32]$
Oil/Steam (U197)	$[47.85, 49.31]$	$[23.93, 24.65]$
Coal/Steam (U350)	$[11.79, 11.79]$	$[5.90, 5.90]$
Nuclear (U400)	$[4.38, 4.47]$	$[2.19, 2.23]$

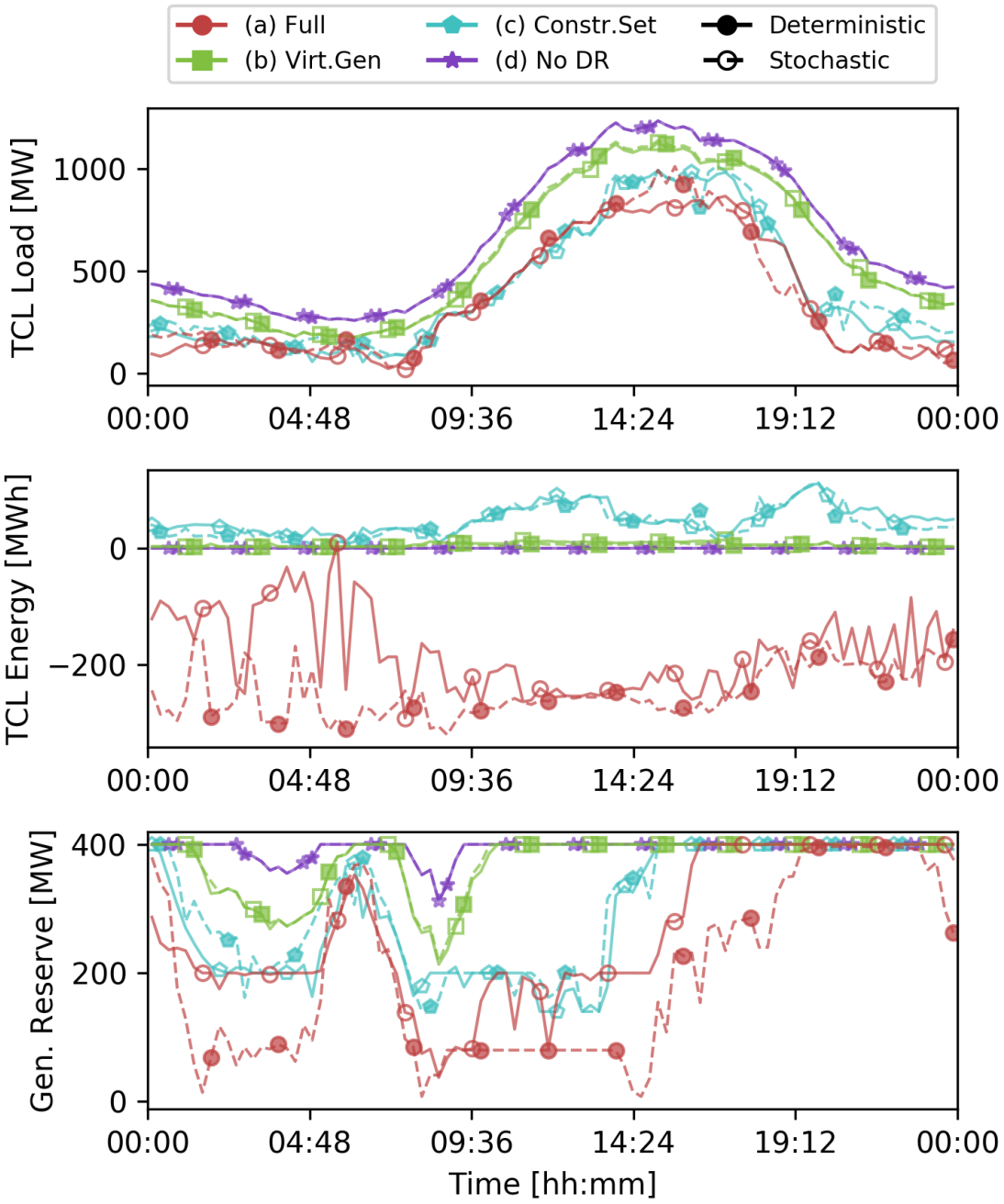


Figure 4.4.8: Using scaled CAISO data from July 20th 2017, the top figure shows the TCL load of the different models, the middle shows the energy stored or borrowed from the TCL aggregates, while the bottom figure shows the amount of reserves bought by generators.

Table 4.2: Summary of average costs for different solutions, and the percentage savings compared with the no DR solution.

Interface	(i) Stochastic SCED	(ii) Deterministic SCED
(a) No DR	$3,407,291$ (0.0%)	$3,407,028$ (0.0%)
(b) Virtual Gen	$3,300,076$ (3.1%)	$3,293,324$ (3.3%)
(c) Constr. Set	$3,101,865$ (9.0%)	$3,103,868$ (8.9%)
(d) Full Model	$3,004,401$ (11.8%)	$3,024,382$ (11.2%)

system is used as a starting-point for the simulations, but model it as a single bus system, with linear generator cost functions whose parameters are shown in Table 4.1.

Table 4.2 shows the average cost over the 7 sample days for different model combinations. Unsurprisingly, there is a clear trend where increased flexibility reduces costs, with the full model saving over 11% compared with no DR, and the constraint set tripling the savings from the virtual generator, from approximately 3% to 9%. Comparing the deterministic with the stochastic a 0.6% improvement is seen in the stochastic results, but this improvement is largely lost when using the constraint set approximation, and slightly negative for the virtual generator. The small improvement can be explained by the small uncertainty over the two-hour look-ahead horizon, along with the high flexibility of the TCLs allowing the operator to react quickly and cheaply to forecast deviations.

Figure 4.4.8 shows how the different models react for the sample day of July 20^{th} 2017. First, one can observe that all the solutions incorporating DR consume less power throughout the day compared with the non-controllable (no DR) counterpart. Although this can be explained, in part, by the full model decreasing the indoor temperature (during cooling) to reduce losses, this is not the case for the approximate approaches, which seem to keep the population at, or even slightly above, the reference temperature. Clearly having some foresight about future temperatures allows moderate energy savings with little to no impacts

on the population. As for the reduction in generator reserve requirements, the TCLs at times covers between 75% and 90% of the required reserve power. The generator reserve reduction varies considerably throughout the day, but interestingly, the stochastic full model seems to be able to save substantially more reserves compared with its deterministic counterpart.

CHAPTER 5
CONCLUSIONS

It is clear that the challenges presently facing power systems need to be approached from multiple angles, to enable a greener future that accommodates increased renewable generation, growing electrification of the transport sector and an overall boost in efficiency, without making sacrifices in reliability.

In this vein, the Continuous-Time Multi-Stage Unit Commitment (CT-MSUC) presented in Chapter 2 incorporates energy storage and allows system operators to better serve net-load with increasing inter-hour variability and uncertainty. Simulations show how such a formulation offers advantages over conventional deterministic Unit Commitment approaches, allowing for variations in commitment and dispatch depending on the specific realization of net-load. When analyzing various cost components, it is observed that less expensive solutions belong to those including storage, as expected.

On the Demand Response front, aggregate state-space models for EVs, DAs and TCLs were developed in Chapter 3. These models have certain characteristics of energy storage devices, particularly in the case of EVs and DAs, but less so for the temperature dependent TCL model. Several properties of these models were explored, control policies and the models were incorporated in a stochastic Economic Dispatch formulation to showcase their synergy with conventional power system models.

Furthermore, Chapter 4 showed how even aggregates of individual resources that have non-convex properties can, in aggregate be considered approximately convex. This is a key finding to allow aggregators to hide the inherent individual complexity when presenting the flexibility of large populations to system operators and energy markets. The chapter further explores different reduced order aggregate models for a concise but descriptive characteri-

zation of these resources, and weighs the trade-off between complexity and computational tractability.

This dissertation may provide answers to several questions, but it opens up even more new questions needing answers. There are plenty of research directions one can take from here, and the author hopes this dissertation and related work will provide a stepping stone for further research and development towards more efficient and flexible power systems.

REFERENCES

[Alizadeh and Scaglione(2013)] Alizadeh, M. and A. Scaglione, "Least laxity first scheduling of thermostatically controlled loads for regulation services", in "Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE", pp. 503–506 (IEEE, 2013).

[Alizadeh *et al.*(2014a)] Alizadeh, M., A. Scaglione, A. Applebaum, G. Kesidis and K. Levitt, "Reduced-order load models for large populations of flexible appliances", IEEE Transactions on Power Systems **30**, 4 (2014a).

[Alizadeh *et al.*(2015)] Alizadeh, M., A. Scaglione, A. Applebaum, G. Kesidis and K. Levitt, "Reduced-order load models for large populations of flexible appliances", IEEE Transactions on Power Systems **30**, 4, 1758–1774 (2015).

[Alizadeh *et al.*(2014b)] Alizadeh, M., A. Scaglione, J. Davies and K. S. Kurani, "A scalable stochastic model for the electricity demand of electric and plug-in hybrid vehicles", Smart Grid, IEEE Transactions on **5**, 2, 848–860 (2014b).

[Analui and Scaglione(2017)] Analui, B. and A. Scaglione, "A dynamic multistage stochastic unit commitment formulation for intraday markets", IEEE Transactions on Power Systems (2017).

[Aubin and Ekeland(1976)] Aubin, J.-P. and I. Ekeland, "Estimates of the duality gap in nonconvex optimization", Mathematics of Operations Research (1976).

[Bakirtzis *et al.*(2018)] Bakirtzis, E. A., C. K. Simoglou, P. N. Biskas and A. G. Bakirtzis, "Storage management by rolling stochastic unit commitment for high renewable energy penetration", Electric Power Systems Research **158**, 240–249 (2018).

[Barot and Taylor(2014)] Barot, S. and J. A. Taylor, "A concise, approximate representation of a collection of loads described by polytopes", arXiv preprint arXiv:1412.0939 (2014).

[Barot and Taylor(2017)] Barot, S. and J. A. Taylor, "A concise, approximate representation of a collection of loads described by polytopes", International Journal of Electrical Power & Energy Systems **84**, 55–63 (2017).

[Bashash and Fathy(2011)] Bashash, S. and H. K. Fathy, "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 (IEEE, 2011).

[Behboodi *et al.*(2018)] Behboodi, S., D. P. Chassin, N. Djilali and C. Crawford, "Transactive control of fast-acting demand response based on thermostatic loads in real-time retail electricity markets", Applied Energy **210**, 1310–1320 (2018).

[Benders(1962)] Benders, J. F., "Partitioning procedures for solving mixed-variables programming problems", Numerische mathematik **4**, 1, 238–252 (1962).

[Bertsekas *et al.*(1983)] Bertsekas, D., G. Lauer, N. Sandell and T. Posbergh, "Optimal short-term scheduling of large-scale power syst.", IEEE Transactions on Automatic Control **28**, 1, 1–11 (1983).

[Boyd and Vandenberghe(2004)] Boyd, S. and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).

[California ISO(2017)] California ISO, "California ISO Open Access Same-time Information System (OASIS)", URL `http://oasis.caiso.com` (2017).

[California ISO(2018)] California ISO, "Business Practice Manual for Market Operations", Tech. rep., URL `https://bpmcm.caiso.com/BPM%20Document%20Library/Market%20Operations/BPM_for_Market%20Operations_V55_redline.pdf`, [Online; accessed 22-March-2018] (2018).

[Callaway(2009)] Callaway, D. S., "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy", Energy Conversion and Management **50**, 5, 1389–1400 (2009).

[Callaway and Hiskens(2011)] Callaway, D. S. and I. A. Hiskens, "Achieving controllability of electric loads", Proceedings of the IEEE **99**, 1, 184–199 (2011).

[Campi *et al.*(2009)] Campi, M. C., S. Garatti and M. Prandini, "The scenario approach for systems and control design", Annual Reviews in Control **33**, 2, 149–157 (2009).

[Carpentier *et al.*(1996)] Carpentier, P., G. Cohen, J.-C. Culioli and A. Renaud, "Stochastic optimization of unit commitment: a new decomposition framework", IEEE Transactions on Power System **11**, 1067–1073 (1996).

[Chang *et al.*(2013)] Chang, T.-H., M. Alizadeh and A. Scaglione, "Real-time power balancing via decentralized coordinated home energy scheduling", IEEE Transactions on Smart Grid **4**, 3, 1490–1504 (2013).

[Chen *et al.*(2010)] Chen, L., N. Li, S. H. Low and J. C. Doyle, "Two market models for demand response in power networks", in "2010 First IEEE International Conference on Smart Grid Communications", pp. 397–402 (IEEE, 2010).

[Chen *et al.*(2016a)] Chen, S., Q. Chen and Y. Xu, "Strategic bidding and compensation mechanism for a load aggregator with direct thermostat control capabilities", IEEE Transactions on Smart Grid **9**, 3, 2327–2336 (2016a).

[Chen *et al.*(2016b)] Chen, Y., A. Casto, F. Wang, Q. Wang, X. Wang and J. Wan, "Improving large scale day-ahead security constrained unit commitment performance", IEEE Transactions on Power Systems (2016b).

[Chong and Malhamé(1984)] Chong, C.-Y. and R. P. Malhamé, "Statistical synthesis of physically based load models with applications to cold load pickup", IEEE transactions on power apparatus and systems , 7, 1621–1628 (1984).

[Coffman *et al.*(2019)] Coffman, A., N. Cammardella, P. Barooah and S. Meyn, "Aggregate capacity of tcls with cycling constraints", arXiv preprint arXiv:1909.11497 (2019).

[Contreras-Ocana *et al.*(2017)] Contreras-Ocana, J. E., M. A. Ortega-Vazquez and B. Zhang, "Participation of an energy storage aggregator in electricity markets", IEEE Transactions on Smart Grid **10**, 2, 1171–1183 (2017).

[Deane *et al.*(2014)] Deane, J., G. Drayton and B. Ó. Gallachóir, "The impact of sub-hourly modelling in power systems with significant levels of renewable generation", Applied Energy **113**, 152–158 (2014).

[Di Somma *et al.*(2018)] Di Somma, M., G. Graditi and P. Siano, "Optimal bidding strategy for a der aggregator in the day-ahead market in the presence of demand flexibility", IEEE Transactions on Industr. Electr. **66**, 2 (2018).

[Diamond *et al.*(2013)] Diamond, H. J., T. R. Karl, M. A. Palecki, C. B. Baker, J. E. Bell, R. D. Leeper, D. R. Easterling, J. H. Lawrimore, T. P. Meyers, M. R. Helfert, G. Goodge and P. W. Thorne, "US climate reference network after one decade of operations: Status and assessment", Bulletin of the American Meteorological Society **94**, 4, 485–498 (2013).

[Elghitani and Zhuang(2017)] Elghitani, F. and W. Zhuang, "Aggregating a large number of residential appliances for demand response applications", IEEE Transactions on Smart Grid (2017).

[Energy Information Administration, U.S. Department of Energy(2009)] Energy Information Administration, U.S. Department of Energy, "2009 Residential Energy Consumption Survey (RECS)", (2009).

[Federal Highway Administration(2009)] Federal Highway Administration, "2009 National Household Travel Survey", Tech. rep., Federal Highway Administration, U.S. Department of Transportation, Washington, DC, URL `https://nhts.ornl.gov` (2009).

[Foster and Caramanis(2013)] Foster, J. M. and M. C. Caramanis, "Optimal power market participation of plug-in electric vehicles pooled by distribution feeder", IEEE Transactions on Power Systems **28**, 3, 2065–2076 (2013).

[Fradelizi *et al.*(2017)] Fradelizi, M., M. Madiman, A. Marsiglietti and A. Zvavitch, "On the monotonicity of minkowski sums towards convexity", (2017).

[Gatsis and Giannakis(2013)] Gatsis, N. and G. B. Giannakis, "Decomposition algorithms for market clearing with large-scale demand response", IEEE Transactions on Smart Grid **4**, 4, 1976–1987 (2013).

[Grigg *et al.*(1999)] Grigg, C., P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty *et al.*, "The ieee reliability test system-1996. a report prepared by the reliability test system task force of the application of probability methods subcommittee", IEEE Transactions on power systems **14**, 3, 1010–1020 (1999).

[Gurobi Optimization(2015)] Gurobi Optimization, I., "Gurobi optimizer reference manual", URL `http://www.gurobi.com` (2015).

[Hao *et al.*(2013)] Hao, H., B. M. Sanandaji, K. Poolla and T. L. Vincent, "A generalized battery model of a collection of thermostatically controlled loads for providing ancillary service", in "Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing", pp. 551–558 (IEEE, 2013).

[Hao *et al.*(2014)] Hao, H., B. M. Sanandaji, K. Poolla and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads", IEEE Transactions on Power Systems **30**, 1, 189–198 (2014).

[Hao *et al.*(2015)] Hao, H., B. M. Sanandaji, K. Poolla and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads", IEEE Transactions on Power Systems **30**, 1, 189–198 (2015).

[Hedman *et al.*(2009)] Hedman, K. W., R. P. O'Neill and S. S. Oren, "Analyzing valid inequalities of the generation unit commitment problem", in "Power Systems Conference and Exposition, 2009. PSCE'09. IEEE/PES", pp. 1–6 (IEEE, 2009).

[Held *et al.*(1974)] Held, M., P. Wolfe and H. P. Crowder, "Validation of subgradient optimization", Mathematical programming **6**, 1, 62–88 (1974).

[Henríquez *et al.*(2017)] Henríquez, R., G. Wenzel, D. E. Olivares and M. Negrete-Pincetic, "Participation of demand response aggregators in electricity markets: Optimal portfolio management", IEEE Transactions on Smart Grid (2017).

[Hreinsson *et al.*(2018)] Hreinsson, K., B. Analui and A. Scaglione, "Continuous time multi-stage stochastic reserve and unit commitment", in "2018 Power Systems Computation Conference (PSCC)", pp. 1–7 (IEEE, 2018).

[Hreinsson and Scaglione(2017)] Hreinsson, K. and A. Scaglione, "On aggregating thermostatically controlled loads based on energy losses", in "Proceedings of the 2017 IEEE Power and Energy Societies General Meeting", (IEEE, 2017).

[Hreinsson *et al.*(2020a)] Hreinsson, K., A. Scaglione and M. Alizadeh, "An aggregate model of the flexible energy demand of thermostatically controlled loads with explicit outdoor temperature dependency", in "Proceedings of the 53rd Hawaii International Conference on System Sciences", (2020a).

[Hreinsson *et al.*(2020b)] Hreinsson, K., A. Scaglione, M. Alizadeh and Y. Chen, "New insights from the shapley-folkman lemma on dispatchable demand in energy markets", Submitted to the IEEE Transactions on Power Systems (2020b).

[Hreinsson *et al.*(2019)] Hreinsson, K., A. Scaglione and B. Analui, "Continuous time multi-stage stochastic unit commitment with storage", IEEE Transactions on Power Systems **34**, 6, 4476–4489 (2019).

[Hreinsson *et al.*(2016)] Hreinsson, K., A. Scaglione and V. Vittal, "Aggregate load models for demand response: Exploring flexibility", in "Proceedings of the 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)", pp. 926–930 (IEEE, 2016).

[Hreinsson *et al.*(2015)] Hreinsson, K., M. Vrakopoulou and G. Andersson, "Stochastic security constrained unit commitment and non-spinning reserve allocation with performance guarantees", International Journal of Electrical Power & Energy Systems **72**, 109–115 (2015).

[Hunter(2007)] Hunter, J. D., "Matplotlib: A 2d graphics environment", Computing in science and engineering **9**, 3, 90–95 (2007).

[Jiang and Low(2011)] Jiang, L. and S. Low, "Multi-period optimal energy procurement and demand response in smart grid with uncertain supply", in "2011 50th IEEE CDC and ECC", pp. 4348–4353 (2011).

[Jones *et al.*(2014)] Jones, E., T. Oliphant and P. Peterson, "{SciPy}: Open source scientific tools for {Python}", (2014).

[Kalsi *et al.*(2012)] Kalsi, K., M. Elizondo, J. Fuller, S. Lu and D. Chassin, "Development and validation of aggregated models for thermostatic controlled loads with demand response", in "System Science (HICSS), 2012 45th Hawaii International Conference on", pp. 1959–1966 (IEEE, 2012).

[Khatami *et al.*(2017)] Khatami, R., M. Parvania and P. Khargonekar, "Scheduling and pricing of energy generation and storage in power systems", IEEE Transactions on Power Systems (2017).

[Koch *et al.*(2011)] Koch, S., J. L. Mathieu and D. S. Callaway, "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services", in "Proceedings of the 2011 Power System Computation Conference", (2011).

[Kohansal and Mohsenian-Rad(2015)] Kohansal, M. and H. Mohsenian-Rad, "Price-maker economic bidding in two-settlement pool-based markets: The case of time-shiftable loads", IEEE Transactions on Power Systems **31**, 1, 695–705 (2015).

[Kowli and Meyn(2011)] Kowli, A. S. and S. P. Meyn, "Supporting wind generation deployment with demand response", in "Power and Energy Society General Meeting, 2011 IEEE", pp. 1–8 (IEEE, 2011).

[Kundu *et al.*(2011)] Kundu, S., N. Sinitsyn, S. Backhaus and I. Hiskens, "Modeling and control of thermostatically controlled loads", in "Proceedings of the 2011 Power System Computation Conference", (2011).

[Li *et al.*(2017)] Li, B., M. Vrakopoulou and J. L. Mathieu, "Chance constrained reserve scheduling using uncertain controllable loads part ii: Analytical reformulation", IEEE Transactions on Smart Grid (2017).

[Li *et al.*(2011a)] Li, G., J. Shi and X. Qu, "Modeling methods for genco bidding strategy optimization in the liberalized electricity spot market–a state-of-the-art review", Energy **36**, 8, 4686–4700 (2011a).

[Li *et al.*(2015a)] Li, N., L. Chen and M. A. Dahleh, "Demand response using linear supply function bidding", IEEE Transactions on Smart Grid (2015a).

[Li *et al.*(2011b)] Li, N., L. Chen and S. H. Low, "Optimal demand response based on utility maximization in power networks", in "2011 IEEE PES General Meeting", pp. 1–8 (2011b).

[Li *et al.*(2015b)] Li, S., W. Zhang, J. Lian and K. Kalsi, "Market-based coordination of thermostatically controlled loads—part i", IEEE Transactions on Power Systems (2015b).

[Lin *et al.*(2018)] Lin, C.-H., R. Wu, W.-K. Ma, C.-Y. Chi and Y. Wang, "Maximum volume inscribed ellipsoid: a new simplex-structured matrix factorization framework via facet enumeration and convex optimization", SIAM Journal on Imaging Sciences **11**, 2, 1651–1679 (2018).

[Liu *et al.*(2017)] Liu, C., A. Botterud, Z. Zhou and P. Du, "Fuzzy energy and reserve co-optimization with high penetration of renewable energy", IEEE Transactions on Sustainable Energy **8**, 2, 782–791 (2017).

[Lopez *et al.*(2018)] Lopez, I. D., D. Flynn, M. Desmartin, M. Saguan and T. Hinchliffe, "Drivers for sub-hourly scheduling in unit commitment models", in "2018 IEEE Power & Energy Society General Meeting (PESGM)", pp. 1–5 (IEEE, 2018).

[Lorca and Sun(2017)] Lorca, A. and X. A. Sun, "Multistage robust unit commitment with dynamic uncertainty sets and energy storage", IEEE Transactions on Power Systems **32**, 3, 1678–1688 (2017).

[Lu(2012)] Lu, N., "An evaluation of the hvac load potential for providing load balancing service", IEEE Transactions on Smart Grid **3**, 3 (2012).

[Lu and Chassin(2004)] Lu, N. and D. P. Chassin, "A state-queueing model of thermostatically controlled appliances", IEEE Transactions on Power Systems **19**, 3, 1666–1673 (2004).

[Mahdavi *et al.*(2016)] Mahdavi, N., J. H. Braslavsky and C. Perfumo, "Mapping the effect of ambient temperature on the power demand of populations of air conditioners", IEEE Transactions on Smart Grid (2016).

[Mahdavi *et al.*(2017)] Mahdavi, N., J. H. Braslavsky, M. M. Seron and S. R. West, "Model predictive control of distributed air-conditioning loads to compensate fluctuations in solar power", IEEE Transactions on Smart Grid **8**, 6, 3055–3065 (2017).

[Mathieu *et al.*(2012)] Mathieu, J., M. Dyson and D. Callaway, "Using residential electric loads for fast demand response: The potential resource and revenues, the costs, and policy recommendations", in "Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings", (2012).

[Mathieu *et al.*(2015)] Mathieu, J. L., M. Kamgarpour, J. Lygeros, G. Andersson and D. S. Callaway, "Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads", IEEE Transactions on Power Systems **30**, 2, 763–772 (2015).

[Mathieu *et al.*(2013a)] Mathieu, J. L., M. Kamgarpour, J. Lygeros and D. S. Callaway, "Energy arbitrage with thermostatically controlled loads", in "Proceedings of the 2013 European Control Conference", (2013a).

[Mathieu *et al.*(2013b)] Mathieu, J. L., S. Koch and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance", IEEE Transactions on Power Systems **28**, 1, 430–440 (2013b).

[Mathieu *et al.*(2013c)] Mathieu, J. L., S. Koch and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance", IEEE Transactions on Power Systems **28**, 1, 430–440 (2013c).

[Meyn *et al.*(2013)] Meyn, S., P. Barooah, A. Busic and J. Ehren, "Ancillary service to the grid from deferrable loads: The case for intelligent pool pumps in florida", in "2013 IEEE 52nd Annual Conference on Decision and Control (CDC)", pp. 6946–6953 (IEEE, 2013).

[Ming *et al.*(2017)] Ming, H., L. Xie, M. Campi, S. Garatti and P. Kumar, "Scenario-based economic dispatch with uncertain demand response", IEEE Transactions on Smart Grid (2017).

[Müller *et al.*(2017)] Müller, F. L., J. Szabó, O. Sundström and J. Lygeros, "Aggregation and disaggregation of energetic flexibility from distributed energy resources", IEEE Transactions on Smart Grid **10**, 2, 1205–1214 (2017).

[Nayyar *et al.*(2013)] Nayyar, A., J. Taylor, A. Subramanian, K. Poolla and P. Varaiya, "Aggregate flexibility of a collection of loads", in "Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on", pp. 5600–5607 (IEEE, 2013).

[Nazir *et al.*(2018)] Nazir, M. S., I. A. Hiskens, A. Bernstein and E. Dall'Anese, "Inner approximation of minkowski sums: A union-based approach and applications to aggregated energy resources", in "2018 IEEE CDC", (2018).

[Nowak and Römisch(2000)] Nowak, M. and W. Römisch, "Stochastic lagrangian relaxation applied to power scheduling in a hydro-thermal system under uncertainty", Annals of Operations Research **100**, 251 (2000).

[Ordoudis *et al.*(2015)] Ordoudis, C., P. Pinson, M. Zugno and J. M. Morales, "Stochastic unit commitment via progressive hedging—extensive analysis of solution methods", in "2015 IEEE Eindhoven PowerTech", pp. 1–6 (IEEE, 2015).

[Ortega-Vazquez *et al.*(2013)] Ortega-Vazquez, M. A., F. Bouffard and V. Silva, "Electric vehicle aggregator/system operator coordination for charging scheduling and services procurement", IEEE Transactions on Power Systems **28**, 2, 1806–1815 (2013).

[Ottesen *et al.*(2016)] Ottesen, S. Ø., A. Tomasgard and S.-E. Fleten, "Prosumer bidding and scheduling in electricity markets", Energy **94**, 828–843 (2016).

[Pandžić *et al.*(2013)] Pandžić, H., J. M. Morales, A. J. Conejo and I. Kuzle, "Offering model for a virtual power plant based on stochastic programming", Applied Energy **105**, 282–292 (2013).

[Papavasiliou *et al.*(2011)] Papavasiliou, A., S. Oren and R. O'Neill, "Reserve requirements for wind power integration:a scenario-based stochastic programming framework", IEEE Transactions on Power Systems **26**, 4, 2197–2206 (2011).

[Papoulis and Pillai(2002)] Papoulis, A. and S. U. Pillai, *Probability, random variables, and stochastic processes* (Tata McGraw-Hill Education, 2002).

[Parvania *et al.*(2013)] Parvania, M., M. Fotuhi-Firuzabad and M. Shahidehpour, "Optimal demand response aggregation in wholesale electricity markets", IEEE Transactions on Smart Grid **4**, 4, 1957–1965 (2013).

[Parvania *et al.*(2014)] Parvania, M., M. Fotuhi-Firuzabad and M. Shahidehpour, "Iso's optimal strategies for scheduling the hourly demand response in day-ahead markets", IEEE Transactions on Power Systems **29**, 6, 2636–2645 (2014).

[Parvania and Khatami(2017)] Parvania, M. and R. Khatami, "Continuous-time marginal pricing of electricity", IEEE Transactions on Power Systems **32**, 3, 1960–1969 (2017).

[Parvania and Scaglione(2016)] Parvania, M. and A. Scaglione, "Unit commitment with continuous-time generation and ramping trajectory models", IEEE Transactions on Power Systems **31**, 4, 3169–3178 (2016).

[Pedregosa *et al.*(2011)] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research **12**, 2825–2830 (2011).

[Pereira and Pinto(1991)] Pereira, M. V. and L. M. Pinto, "Multi-stage stochastic optimization applied to energy planning", Mathematical programming **52**, 1-3, 359–375 (1991).

[Pflug and Pichler(2015)] Pflug, G. C. and A. Pichler, "Dynamic generation of scenario trees", Computational Optimization and Applications **62**, 3, 641–668 (2015).

[Ramanathan and Vittal(2008)] Ramanathan, B. and V. Vittal, "A framework for evaluation of advanced direct load control with minimum disruption", IEEE Transactions on Power Systems **23**, 4, 1681–1688 (2008).

[Ruiz *et al.*(2009)] Ruiz, N., I. Cobelo and J. Oyarzabal, "A direct load control model for virtual power plant management", IEEE Transactions on Power Systems (2009).

[Ryan *et al.*(2013)] Ryan, S. M., R. J.-B. Wets, D. L. Woodruff, C. Silva-Monroy and J.-P. Watson, "Toward scalable, parallel progressive hedging for stochastic unit commitment", in "Power and Energy Society General Meeting (PES), 2013 IEEE", (2013).

[Samadi *et al.*(2015)] Samadi, P., V. W. Wong and R. Schober, "Load scheduling and power trading in syst. with high penetration of renewable energy resources", IEEE Transactions on Smart Grid **7**, 4, 1802–1812 (2015).

[Sanchez-Martin *et al.*(2012)] Sanchez-Martin, P., G. Sanchez and G. Morales-Espana, "Direct load control decision model for aggregated ev charging points", IEEE Transactions on Power Systems **27**, 3, 1577–1584 (2012).

[Scaglione(2016)] Scaglione, A., "Continuous-time marginal pricing of power trajectories in power systems", in "Information Theory and Applications Workshop (ITA), 2016", pp. 1–6 (IEEE, 2016).

[Shiina and Birge(2004)] Shiina, T. and J. R. Birge, "Stochastic unit commitment problem", International Transaction on Operations Research **11**, 19–32 (2004).

[Sojoudi and Low(2011)] Sojoudi, S. and S. H. Low, "Optimal charging of plug-in hybrid electric vehicles in smart grids", in "Power and Energy Society General Meeting, 2011 IEEE", pp. 1–6 (IEEE, 2011).

[Sortomme and El-Sharkawi(2012)] Sortomme, E. and M. A. El-Sharkawi, "Optimal combined bidding of vehicle-to-grid ancillary services", Smart Grid, IEEE Transactions on **3**, 1, 70–79 (2012).

[Starr(1969)] Starr, R. M., "Quasi-equilibria in markets with non-convex preferences", Econometrica: journal of the Econometric Society pp. 25–38 (1969).

[Subramanian *et al.*(2012)] Subramanian, A., M. Garcia, A. Dominguez-Garcia, D. Callaway, K. Poolla and P. Varaiya, "Real-time scheduling of deferrable electric loads", in "American Control Conference (ACC), 2012", pp. 3643–3650 (IEEE, 2012).

[Subramanian *et al.*(2013)] Subramanian, A., M. J. Garcia, D. S. Callaway, K. Poolla and P. Varaiya, "Real-time scheduling of distributed resources", IEEE Transactions on Smart Grid **4**, 4, 2122–2130 (2013).

[Takriti *et al.*(1996)] Takriti, S., J. R. Birge and E. Long, "A stochastic model for unit commitment problem", IEEE Transactions on Power Systems **11**, 3, 1497–1508 (1996).

[Taylor(2015)] Taylor, J. A., "Financial storage rights", IEEE Transactions on Power Systems **30**, 2, 997–1005 (2015).

[Tindemans *et al.*(2015)] Tindemans, S. H., V. Trovato and G. Strbac, "Decentralized control of thermostatic loads for flexible demand response", IEEE Transactions on Control Systems Technology **23**, 5, 1685–1700 (2015).

[Totu *et al.*(2016)] Totu, L. C., R. Wisniewski and J. Leth, "Demand response of a tcl population using switching-rate actuation", IEEE Transactions on Control Systems Technology (2016).

[Trangbæk *et al.*(2011)] Trangbæk, K., M. Petersen, J. Bendtsen and J. Stoustrup, "Exact power constraints in smart grid control", in "Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on", pp. 6907–6912 (IEEE, 2011).

[Udell and Boyd(2016)] Udell, M. and S. Boyd, "Bounding duality gap for separable problems with linear constraints", Computational Optimization and Applications (2016).

[Van Der Walt *et al.*(2011)] Van Der Walt, S., S. C. Colbert and G. Varoquaux, "The numpy array: a structure for efficient numerical computation", Computing in Science & Engineering **13**, 2, 22–30 (2011).

[Vrakopoulou *et al.*(2017)] Vrakopoulou, M., B. Li and J. L. Mathieu, "Chance constrained reserve scheduling using uncertain controllable loads part i: Formulation and scenario-based analysis", IEEE Transactions on Smart Grid (2017).

[Watson and Woodruff(2011)] Watson, J.-P. and D. L. Woodruff, "Progressive hedging innovations for a class of stochastic mixed-integer resource allocation problems", Computational Management Science **8**, 4, 355–370 (2011).

[Wiebking(1977)] Wiebking, R., "Stochastische modelle zur optimalen lastverteilung in einem kraftwerksverbund", Mathematical Methods of Operations Research **21**, 6, B197–B217 (1977).

[Yao *et al.*(2013)] Yao, W., J. Zhao, F. Wen, Y. Xue and G. Ledwich, "A hierarchical decomposition approach for coordinated dispatch of plug-in electric vehicles", IEEE Transactions on Power Systems **28**, 3, 2768–2778 (2013).

[Yoon *et al.*(2014)] Yoon, J. H., R. Baldick and A. Novoselac, "Dynamic demand response controller based on real-time retail price for residential buildings", IEEE Transactions on Smart Grid **5**, 1, 121–129 (2014).

[Zhang *et al.*(2013)] Zhang, W., J. Lian, C.-Y. Chang and K. Kalsi, "Aggregated modeling and control of air conditioning loads for demand response", IEEE Transactions on Power Systems **28**, 4, 4655–4664 (2013).

[Zhao *et al.*(2013)] Zhao, C., J. Wang, J.-P. Watson and Y. Guan, "Multi-stage robust unit commitment considering wind and demand response uncertainties", IEEE Transactions on Power Systems **28**, 3, 2708–2717 (2013).

[Zhao *et al.*(2017)] Zhao, L., W. Zhang, H. Hao and K. Kalsi, "A geometric approach to aggregate flexibility modeling of thermostatically controlled loads", IEEE Transactions on Power Systems **32**, 6, 4721–4731 (2017).

[Zheng *et al.*(2015)] Zheng, Q. P., J. Wang and A. L. Liu, "Stochastic optimization for unit commitment—a review", IEEE Transactions on Power Systems **30**, 4, 1913–1924 (2015).

[Zhong *et al.*(2013)] Zhong, H., L. Xie and Q. Xia, "Coupon incentive-based demand response: Theory and case study", IEEE Transactions on Power Systems **28**, 2, 1266–1276 (2013).

[Zhou and Botterud(2014)] Zhou, Z. and A. Botterud, "Dynamic scheduling of operating reserves in co-optimized electricity markets with wind power", IEEE Transactions on Power Systems **29**, 1, 160–171 (2014).

[Ziras *et al.*(2018)] Ziras, C., S. You, H. W. Bindner and E. Vrettos, "A new method for handling lockout constraints on controlled tcl aggregations", in "2018 Power Systems Computation Conference", pp. 1–7 (IEEE, 2018).

[Zou *et al.*(2017)] Zou, J., S. Ahmed and X. A. Sun, "Stochastic dual dynamic integer programming", Mathematical Programming pp. 1–42 (2017).

[Zou *et al.*(2018)] Zou, J., S. Ahmed and X. A. Sun, "Multistage stochastic unit commitment using stochastic dual dynamic integer programming", IEEE Transactions on Power Systems (2018).