On Rate Design in Modern Electricity Sectors

by

Felipe Ignacio Castro Altamirano

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Energy & Resources

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Duncan Callaway, Chair
Professor Daniel Kammen
Professor Pravin Varaiya
Professor Shmuel Oren

Summer 2017

On Rate Design in Modern Electricity Sectors

Abstract

On Rate Design in Modern Electricity Sectors

by

Felipe Ignacio Castro Altamirano

Doctor of Philosophy in Energy & Resources

University of California, Berkeley

Associate Professor Duncan Callaway, Chair

This dissertation focuses on the problem of designing rates in the utility sector. It is motivated by recent developments in the electricity industry, where renewable generation technologies and distributed energy resources are becoming increasingly relevant. Both technologies disrupt the sector in unique ways. While renewables make grid operations more complex, and potentially more expensive, distributed energy resources enable consumers to interact two-ways with the grid. Both developments present challenges and opportunities for regulators, who must adapt their techniques for evaluating policies to the emerging technological conditions.

The first two chapters of this work make the case for updating existing techniques to evaluate tariff structures. They also propose new methods which are more appropriate given the prospective technological characteristics of the sector. The first chapter constructs an analytic tool based on a model that captures the interaction between pricing and investment. In contrast to previous approaches, this technique allows consistently comparing portfolios of rates while enabling researchers to model with a significantly greater level of detail the supply side of the sector. A key theoretical implication of the model that underlies this technique is that, by properly updating the portfolio of tariffs, a regulator could induce the welfare maximizing adoption of distributed energy resources and enrollment in rate structures. We develop an algorithm to find globally optimal solutions of this model, which is a nonlinear mathematical program. The results of a computational experiment show that the performance of the algorithm dominates that of commercial nonlinear solvers. In addition, to illustrate the practical relevance of the method, we conduct a cost benefit analysis of implementing time-variant tariffs in two electricity systems, California and Denmark. Although portfolios with time-varying rates create value in both systems, these improvements differ enough to advise very different policies. While in Denmark time-varying tariffs appear unattractive, they at least deserve further revision in California. This conclusion is beyond the reach of previous techniques to analyze rates, as they do not capture the interplay between an intermittent supply and a price-responsive demand.

While useful, the method we develop in the first chapter has two important limitations. One is the lack of transparency of the parameters that determine demand substitution patterns, and demand heterogeneity; the other is the narrow range of rate structures that could be studied with the technique. Both limitations stem from taking as a primitive a demand function. Following an alternative path, in the second chapter we develop a technique based on a pricing model that has as a fundamental building block the consumer utility maximization problem. Because researchers do not have to limit themselves to problems with unique solutions, this approach significantly increases the flexibility of the model and, in particular, addresses the limitations of the technique we develop in the first chapter. This gain in flexibility decreases the practicality of our method since the underlying model becomes a Bilevel Problem. To be able to handle realistic instances, we develop a decomposition method based on a non-linear variant of the Alternating Direction Method of Multipliers, which combines Conic and Mixed Integer Programming. A numerical experiment shows that the performance of the solution technique is robust to instance sizes and a wide combination of parameters. We illustrate the relevance of the new method with another applied analysis of rate structures. Our results highlight the value of being able to model in detail distributed energy resources. They also show that ignoring transmission constraints can have meaningful impacts on the analysis of rate structures. In addition, we conduct a distributional analysis, which portrays how our method permits regulators and policy makers to study impacts of a rate update on a heterogeneous population. While a switch in rates could have a positive impact on the aggregate of households, it could benefit some more than others, and even harm some customers. Our technique permits to anticipate these impacts, letting regulators decide among rate structures with considerably more information than what would be available with alternative approaches.

In the third chapter, we conduct an empirical analysis of rate structures in California, which is currently undergoing a rate reform. To contribute to the ongoing regulatory debate about the future of rates, we analyze in depth a set of plausible tariff alternatives. In our analysis, we focus on a scenario in which advanced metering infrastructure and home energy management systems are widely adopted. Our modeling approach allows us to capture a wide variety of temporal and spatial demand substitution patterns without the need of estimating a large number of parameters. We calibrate the model using data of appliance ownership, census household counts, weather patterns, and a model of California's electricity network. The analysis shows that the average gains of implementing time-varying rates with respect to a simple flat rate program are rather mild, not greater than 2 dollars per month, even in the scenario in which volumetric charges are allowed to vary freely from hour to hour. Our results also show that factors such as the presence of an air conditioning system and the exterior temperature profile can have a meaningful impact on the surplus gains that different rates generate on households. These two results combined suggest that defaulting all residential customers into a time-of-use rate structure, which is the current path California is following for the residential sector, may not be an optimal strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

To my wife, my best friend. And to my parents.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my adviser, Duncan Callaway, for his help and unconditional support during all these years. I would also like to thank my dissertation committee and researchers at my department, The Energy and Resources Group. Without their insightful comments and significantly diverse perspectives, this work would be only a thin shadow of what it is today.

Chapter 1

Introduction

This dissertation focuses on the problem of designing rates in the utility industry. It develops a technique to conduct applied analysis of rate structures and uses this method to study retail pricing in California. This work is highly motivated by recent developments in the electricity sector. The industry is experiencing two mayor changes: one corresponds to the increasing importance of renewables, such as wind or solar power, in generation mixes; the other is the the emergence of distributed energy resources (DERs) (IEA, 2013). While the electricity sector has experienced many waves of technological transformation, the advent of these technologies poses unique challenges.

High-voltage grid-connected intermittent generation technologies, such as windmills or solar photovoltaic systems, challenge grid operations. The output of these technologies is driven mainly by weather conditions, such as wind speed, wind direction, cloud cover or haze, which change considerably across time. As a result, systems operators cannot dispatch these resources at will. In absence of economically feasible storage, the intermittentcy of wind or solar requires additional reserves to ensure reliability and more capacity to meet demand. These requisites could translate into higher production costs and even undermine carbon emission reductions (Frondel et al., 2015). Rate structures that reflect closely system conditions could work in sync with renewable resources, incentivizing consumption when these resources are available, and bringing demand down when they are scarce.

The increasing penetration of distributed energy resources, including advanced metering infrastructure (AMI), energy management systems (HEMS), rooftop solar photovoltaic and battery storage systems are enabling residential consumers, or “prosumers”, to interact bidirectionally with electric grid. Customers can not only purchase electricity but also provide the system with energy and reliability services (Schleicher-Tappeser, 2012). In regulated retail sectors, residential rate structures greatly influence this interaction. Rates can impact adoption decisions, by changing the economic value of the technology, and can influence the usage of this resource. Rates may also increase distributional disparities, creating cross subsidies between those who can and cannot adopt the distributed technology (Eid et al., 2014).

Despite the crucial role of rate design, the body of techniques to discern in a systematic manner among many alternatives is rather scant. Factors explaining this reality include limitations of the theory and the focus of the empirical work. While the theory asserts that real-time pricing (RTP), a structure with charges that vary freely across time, is optimal from an economic efficiency perspective (Joskow and Tirole, 2006), regulators must balance efficiency with other public goals, such as rate simplicity, equity or meeting environmental directives (Stanton, 2015). Absent a comprehensive quantification of the impacts associated with implementing RTP, it is difficult for regulators to judge the value of this alternative, especially when it may compromise other regulatory goals.

The empirical literature has tried to quantify impacts, however, the scope has been somewhat limited. Researchers have focused on estimating price responsiveness to quantify changes in efficiency in the short-run. Examples include the work of Allcott (2011), Caves et al. (1984a), Faruqui and Sergici (2010), Faruqui and Sergici (2011) and Herter (2007). Other relevant metrics such as long-run welfare effects, equity or environmental implications have been explored either in isolation or with stylized analyses, but never with an applied approach, within a unified framework.[1]

1.1 A Technique to Evaluate Rate Structures

The first two chapters of this dissertation focus on developing a comprehensive technique to evaluate rate structures. Building upon a mature theory of pricing for public utilities—Peak-Load Pricing, the first chapter constructs an analytic tool based on a model that captures the interaction between pricing and investment. In contrast to previous approaches, this technique allows consistently comparing portfolios of rates while capturing complexities emerging in modern electricity sectors. Welfare analyses conducted with the method can account for interactions between intermittent renewable generation, distributed energy resources and tariff structures.

We explore the theoretical and practical implications of the model that underlies the technique. Our analysis shows that a regulator can induce the welfare maximizing configuration of the demand by properly updating the portfolio of tariffs. We exploit the structure of the model to construct a simple algorithm to find globally optimal solutions of the associated nonlinear optimization problem. A computational experiment suggests that the specialized procedure can outperform standard nonlinear programming techniques. This dominance in performance is specially stark for large-sized instances.

To illustrate the practical relevance of the rate analysis method, we conduct a numerical study of the deployment of advanced metering infrastructure. Previous cost-benefit evaluations considered this technology in isolation and did not study how pricing could affect its value. In the chapter, we take a different approach. We consider customers with and without AMI as resources competing with technologies at the supply side. Customer with AMI can enroll in a time-varying rate and thus be a source of demand responsiveness. Our analysis numerically explores the optimal mix of demand and supply side resources. We consider

[1] See, for instance, Borenstein (2006), Borenstein (2013), Crew et al. (1995) and Joskow and Tirole (2006).

two hypothetical systems constructed with data from California and Denmark. The analysis shows that pricing greatly influences the optimal deployment of AMI. If customers with this technology enrolled in real-time pricing, the optimal deployment would be up to ten times greater than if they enrolled in a time-of-use (TOU) rate structure. In addition, as one would expect with supply side technologies, the optimal deployment of AMI depends upon the specific characteristics of the systems, in this exercise, the correlation between the time series of baseline consumption and renewable production. This correlation is greater in Denmark's than in California, which makes demand responsiveness more valuable in the latter system. As a result, with data from California the optimal roll-out of AMI could be up to five times higher than with Denmark's data. The results suggest fairly different policy prescriptions. In the case of Denmark, there seems to be little value in implementing time-varying rates, such as TOU or RTP. On the other hand, in California regulators should seriously consider time-varying rates, as they could create value, even after netting out AMI costs.

1.2 Developing a Comprehensive Approach to Evaluate Rates

While useful, the model that underlies the method that we develop in the first chapter has two important limitations. One is the lack of transparency of the parameters that determine demand substitution patterns, and demand heterogeneity; the other is the narrow range of rate structures that could be studied with the technique. All models based upon the theory of Peak-Load Pricing inherit these limitations. The theory assumes that the demand, the solution of the consumer utility maximization problem, is a singleton. Thus, the models based upon this theory take this consumer demand as a primitive, instead of considering the consumer problem as the fundamental building block. While simplifying mathematical analysis, this assumption considerably restricts modeling flexibility. Important determinants of household electricity consumption behavior such as weather patterns or appliance ownership cannot be transparently included in models where a representative demand function is the starting point. In addition, while a researcher could represent a wide variety of time-varying rate structures, more complex designs such as combining a time-varying rate with a demand charge or a block rate cannot be captured. Modeling these rates requires to modify the consumer problem, breaking the basic assumptions of Peak-Load Pricing.

In the second chapter, we develop a model that tackles both limitations by considering as basic building block the consumer maximization problem. As a result, the technique allows researchers to consistently benchmark a large class of tariffs. The literature of utility pricing is divided into two branches: non-linear and time-varying pricing. Non-linear pricing encompasses prices that differ depending on the levels of aggregated consumption. On the other hand, in time-varying pricing volumetric charges are contingent upon time. Researchers have studied these branches independently, comparing structures within the branches, but never between. The model we develop provides a unified analytic framework to compare most structures relevant in practice, including non-linear and time-varying prices, and also structures with a charge for peak-consumption. It also makes possible capturing an unprecedented level of detail, making direct and transparent the inclusion of important technological aspects emerging in the utility industry. Our model finds the Ramsey-Boiteux prices while simultaneously determining production schedules, supply and demand side investments and consumption allocations.

This flexibility comes at a cost, however. The model becomes more complex since, in general, the demand function does not have analytic expression. We formulate the utility pricing problem as a Bilevel Program and develop a distributed algorithm to solve it. We first cast the Bilevel Problem as a Mathematical Program with Equilibrium Constraints, and solve it using conic and mixed-integer programming, as well as a non-convex variant of the Alternating Direction Method of Multipliers. Our solution approach is suitable to distributed computation and, thus, appropriate for large-scale applications. A numerical experiment shows that the performance of the technique is robust to instance sizes and a wide combination of parameters. With a simple application, we illustrate its value for the emerging utility industry.

The analysis highlights the value of being able to model DERs. It shows how the economics of rooftop PV systems impact the rate comparison, and that time-varying rates and inexpensive DERs could complement each other; the former can improve the relative value of the latter while DERs being inexpensive increases the gains of switching from a time invariant rate. In addition, the exercise shows that ignoring transmission constraints can have meaningful impacts on the analysis of rate structures. In the exercise, the presence of network constraints translates into significantly higher welfare gains when switching from a flat rate to more sophisticated structures. Omitting this element underestimates the benefits by about 3000 times. Finally, we are able to conduct a distributional analysis, which portrays how our method permits regulators and policy makers to study impacts of a rate update on a heterogeneous population. While a switch in rates could have a positive impact on the aggregate of households, it could benefit some more than others. It could even harm some customers, which can be particularly problematic if those harmed were low income households. A technique such as the one introduced in this chapter permits to anticipate these impacts, letting regulators decide among rate structures with considerably more information than what would be available with alternative techniques.

1.3 An Applied Analysis of Utility Pricing in California

In the third chapter, we use the technique developed in the second in an applied study. This work contributes with analysis of residential rate structures, in the context of California's electricity sector. The state of California is in the process of transitioning all residential customers from a default increasing block rate structure to a time-varying tariff. In July of

2015, decision $D.15-07-001$ set a schedule and a series of steps to transition all residential customers to a time-of-use rate by 2019. In this chapter, we embark on an exploration of the long-run consequences of this rate redesign.

Our bottom up approach to demand modeling allows us to capture a wide variety of temporal and spatial substitution patterns without the need of estimating a large number of parameters. We calibrate the demand model using data of appliance ownership, census household counts and weather patterns. We supplement this data with a realistic model of the western interconnection, including a simplified representation of the network's topology and a comprehensive set technical parameters for the power production technologies. A distinctive element of our analysis is the variety of perspectives it offers. We are able to asses the impacts of diverse rate updates from an efficiency point of view, while at the same time we are able to grasp the distributional and environmental consequences.

Besides the time-of-use structure, we compare a set of rates which are plausible future alternatives, a TOU with a charge for peak-consumption (TOU& DC) and an RTP tariff. Our analysis indicates considerable welfare gains when switching to an RTP structure. These gains are an order of magnitude greater than those obtained when switching to a TOU or a TOU& DC program. The total gains are so small, in this second case, that they cannot cover the costs of AMI. Importantly, the fact that these gains are small is because under TOU or TOU& DC customers with central air conditioning are worse off. On hot summer days, when the system is not under peak conditions, these customers will pay much more than what they would have paid under RTP. Even though customers with no central air conditioning benefit from switching, the welfare loss of those with these devices is such that the total efficiency gains are small when transferring all customers to the TOU or TOU& DC rates. This suggests that defaulting all customers to TOU, which is the path that California's regulators are following, might not be an optimal strategy. Instead, targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

Chapter 2

Optimal Rate Design in Modern Electricity Sectors

2.1 Introduction

Electricity sectors are experiencing two major shifts: accelerating deployment of *intermittent renewable generation*[1] and pervasive information and communications technologies (ICT) (IEA, 2013). The management of the demand, or *demand response*, is broadly seen as an important resource in the presence of intermittent renewables, and ICT has a clear role to play in providing information or control signals to end-use devices to enable this resource (Joskow and Wolfram, 2012). There are two types of demand response programs. One includes programs in which a utility can directly alter the demand level of a customer.[2] The other encompasses programs where price signals are the means the utility uses to influence its demand (FERC, 2006). In this paper we focus on the latter type of programs, commonly known as rates or tariffs.

In many jurisdictions, a regulated utility distributes the electricity to most end customers. In these instances, this agent collects its revenue through a set of rates or tariffs, which are under the oversight of the regulatory body. A central element defining these tariffs is their structure or design, which specifies what charges compose these instruments (Joskow, 2007, p. 1276). Given the new reality of the electricity industry, where renewable generation and distributed energy resources (DERs) play an increasingly important role, regulators are exploring more sophisticated tariff structures (Stanton, 2015). Advanced rate designs that align better with marginal costs (e.g., time-varying rates) can reduce generation fuel costs and decrease investment in distribution, transmission and generation infrastructure (Joskow and Wolfram, 2012). However, when other public goals—such as carbon emission reductions—are at stake and the presence of intermittent renewables is significant, the case

[1] Intermittent renewables include wind and solar power plants.

[2] In the US, an example of this type of program is *direct load control*. Utilities can remotely control some of the devices of a customer under this program.

for advanced rates is less clear-cut. Depending on the characteristic of the system, these rate designs may or may not decrease emissions (Holland and Mansur, 2008); simpler tariffs structures, such as a flat rate (FR), may produce greater environmental benefits and even improve consumer surplus (Kök et al., 2016).

The emergence of distributed energy resources further complicates the analysis of rate structures. Advanced metering infrastructure (AMI), a set of technologies allowing utilities to collect and transmit granular consumption information, and home energy management systems (HEMS) enable the implementation of sophisticated rate designs, and can boost consumer price-responsiveness (Faruqui and Sergici, 2010). However, these DERs are not necessarily cost-effective. While AMI comes along with operational savings, including a decrease in meter reading or fault detection expenditures, these are insufficient to cover its capital costs (Faruqui et al. 2010, Faruqui and Sergici 2010). Improved tariff designs in combination with HEMS could fill this gap. But they are not exempt of costs either. Implementing complex rate structures requires at least creating awareness in the population and educating retail customers on the benefits these rates; home energy management systems, that can enlarge the benefits of advanced rates, such as smart thermostats or in-home displays (see Faruqui and Sergici 2013), require meaningful capital outlays as well.

Given these benefits and costs, to what extent it is beneficial for a system to implement more sophisticated rate structures? While this policy-relevant question has been traditionally approached by economists, we believe that the OR community has an important role to play in improving past answers, specially in the face of the complexities unfolding in the electricity industry. Sioshansi (2012) gives a step in this direction. The paper investigates the impacts of different tariff structures on the efficient operation of Plug-In Hybrid Electric Vehicles (PHEVs). Using a Unit Commitment and Vehicle-Charging models, the paper finds that a simple flat rate can outperform time-varying rate designs. Sioshansi (2012) detailed model illuminates the inefficiencies that can emerge at the retail level when using time-varying structures in sectors with non-convex costs—where production technologies have costs with terms independent of the level of production (O'Neill et al., 2005). This finding contradicts economists common wisdom, which asserts that rate designs that reflect marginal costs better are more desirable (see, e.g., Joskow and Wolfram 2012).

Kök et al. (2016) is another example showing that capturing new complexities, characteristic of modern electricity sectors, can challenge intuition. The paper investigates the interaction between two rate designs, a flat and a time-varying structure, and investment in the presence of renewable generation. Using a model that captures the intermittentcy of renewables and the pricing behavior of the utility, the paper finds that the flat rate design leaves consumers always better-off relative to the time-varying structure. Although this conclusion rests upon the characteristics of the setting that Kök et al. (2016) analyze, the result does provide additional evidence of the importance of detailed modeling when evaluating tariff structures.

The present work adds to the contributions by developing a general technique to evaluate rate structures. We take as a starting point empirical methods which have been used to conduct welfare analysis of rate changes (e.g., Acton and Bridger M. 1983, Caves et al.

1984b, Gallant and Koenker 1984, Howrey and Varian 1984, Lillard and Aigner 1984, Parks and Weitzel 1984, Borenstein 2005, Borenstein and Holland 2005, Taylor et al. 2005 and Allcott 2011). We add layers of detail that capture salient aspects of modern electricity sectors. Specifically, we extend previous techniques in four significant ways. First, our method improves the consistency of the comparison among rates. Existing approaches compare tariff structures either specifying ex-ante *rate levels* (the specific values of each of the charges composing the rate structure), or imposing unnecessary constraints on them (e.g., Acton and Bridger M. 1983, Borenstein 2005, Sioshansi 2012). In contrast, our technique computes rate levels following an optimality criterion. Researchers compare the best case of a rate structure against the best case of another (see Subsection 2.3 for more details).

In addition, we improve the supply cost representation of previous methods. With the exception of Sioshansi (2012), past work either represents these costs with stylized models (e.g., Borenstein 2005, Kök et al. 2016) or omits the supply side, only assessing welfare impacts on end customers (e.g., Lillard and Aigner 1984, Allcott 2011). The present technique permits researchers to include complex representations of the supply side. The main difference with the approach of Sioshansi (2012) is that we consider long term costs while this author focuses on the short-run.

A third extension allows comparing multiple rate structures simultaneously. Except for Borenstein (2005) and Borenstein and Holland (2005),

The last element that distinguish our method from others is the endogenous computation of an optimal *demand mix*. That is, the model we use for comparing rates finds the socially optimal fraction of customers enrolled in different programs. Borenstein and Holland (2005) explores the welfare implications of alternative demand mixes. Specifically, the paper studies the impacts of exogenously varying the fraction of customers under an advanced tariff. It shows that while the marginal benefit of increasing this fraction decreases, the marginal cost remains constant. Consequently, the authors observe that the optimal fraction ultimately depends upon the specific characteristics of the system under study. The endogenous computation of an optimal demand mix simplifies the analysis in Borenstein and Holland (2005). More importantly, in combination with the other improvements we introduce, it allows comparing portfolios of rate structures making far less assumptions than one would have to if using other approaches.

Our technique uses a nonlinear optimization model that we build based on the *Peak-Load Pricing* theory. This strand of utility pricing was developed in the seventies and eighties with the work of Steiner (1957), Boiteux (1960), Drèze (1964), Crew and Kleindorfer (1976), Panzar (1976), Joskow (1976), Carlton (1977) and Chao (1983), and more recently was revisited by Zöttl (2010) and Chao (2011). While it originally intended to provide theoretical guidelines for the optimal pricing of public utility services (Crew et al., 1995), other authors have used this theory as a framework to analyze a range of regulatory issues. Joskow and Tirole (2006) uses Peak-Load Pricing to construct a benchmark to understand the implications of competition at the retail level in the electricity industry; Zöttl (2010) explores the incentives to invest of strategic firms participating in markets where demand is fluctuating and storage is prohibitively expensive; and Chao (2011) uses the theory to explore the interaction between time-varying rates and intermittent renewables. Following the approach of these papers, we use Peak-Load Pricing as a basis for our model which we modify to meet our purposes.

2.2 Peak-Load Pricing: An Overview

The problem that this theory addresses is how to price the set of commodities that a regulated monopoly provides. It answers this question taking the perspective of a regulator. Optimal prices are such that the societal welfare is maximized subject to the revenue sufficiency and technical constraints of the regulated monopoly (Crew et al., 1995). We formalize the Peak-Load Pricing problem following Crew et al. (1995) and Joskow and Tirole (2006).

Before starting, we introduce some notation. Let Ω be a discrete sample space, q_ω the probability that $\omega \in \Omega$ occurs, and $E[\cdot]$ the associated expectation operator. We refer to an element in Ω as outcome or state of nature, and distinguish a random from a deterministic variable placing a bar on top of the former. Given a random variable $\bar{y}$, we denote y_ω the realization of this variable when ω occurs. The symbol $\top$ indicates the transpose of a vector.

The theory considers a monopolist offering a set $\{1, \ldots, T\}$ of goods. Customers are of different types $i \in I$, and distribute according to a frequency function δ_i, denoting the number of types i. A quasi-linear utility $U^i_\omega(d) + m_\omega$ characterizes the preferences of the customers of type i over consumption bundles $d \in \mathbb{R}^T_+$. The scalar m_ω is her expenditure in all other goods. For a vector of prices $p_\omega \in \mathbb{R}^T_+$, a customer with an income M_i consumes according to the demand function $D^i_\omega(p_\omega) := arg\,max_{d\geq 0}\left\{U^i_\omega(d) + M_i - p_\omega^\top d\right\}$.[3] It is customary to assume that $U^i_\omega(\cdot)$ is strictly concave and, thus, $D^i_\omega(p)$ is a singleton. The *gross surplus* of this customer is $S^i_\omega(p_\omega) := U^i_\omega\left(D^i_\omega(p_\omega)\right)$.

We define $\bar{D}^{I'}(\bar{p})$ as the aggregated consumption of types $i \in I'$, where $I' \subseteq I$. That is, $\bar{D}^{I'}(\bar{p}) = \int_{I'} \bar{D}^i(\bar{p})\delta_i di$. Defined similarly, $S^{I'}_\omega(p_\omega)$ represents the aggregated gross surplus.

The monopolist offers a two-part rate structure. That is, a contract which has a fixed charge l and a vector of volumetric charges $\bar{p}$. In this arrangement, the monopolist charges $p_{t\omega}$ per unit of consumption of good t under ω. The corresponding consumer surplus is

$$CS(l, \bar{p}) = E\left[\bar{S}^I(\bar{p}) - \bar{p}^\top \bar{D}^I(\bar{p})\right] - l \cdot \nu_I, \tag{2.1}$$

where $\nu_{I'} := \int_{I'} \delta_i di$ for any $I' \subseteq I$.

The monopolist produces with a set of technologies that we index with the letter $k \in K$. Each technology differs from others on its variable costs per unit of production, $c_{\omega k} \in \mathbb{R}^T_+$, its fixed costs $\hat{r}_k$, and its availability factor, $\rho_{\omega k} \in \mathbb{R}^T_+$, capturing the variability in the technology's availability—e.g., due to the intermittent output of some renewables or the occurrence of outages. Before uncertainty realizes, the monopolist decides the installed capacity of each technology, x_k. After, the firm determines a production vector for each

[3] The problem the customer solves is $\left\{U^i_\omega(d) + m_\omega : M_i \geq p_\omega^\top d + m_\omega\right\}$. Because peak-load pricing assumes $m_\omega > 0$ in any optimum, one can simplify the problem.

technology, $y_{\omega k} \in \mathbb{R}^T_+$. For a consumption vector $\bar{d}$, the monopolist's cost function satisfies

$$C\left(\bar{d}\right) = \min_{(x,y)} \sum_{k \in K} E\left[\bar{y}_k^\top \bar{c}_k + x_k \hat{r}_k\right] \tag{2.2}$$

$$\text{subject to}$$

$$\bar{d} \leq \sum_{k \in K} \bar{y}_k, \tag{2.3}$$

$$0 \leq \bar{y}_k \leq x_k \bar{\rho}_k, \; k \in K \tag{2.4}$$

and the firm's profit is

$$\Pi(l, \bar{p}) = E\left[\bar{p}^\top \bar{D}^I(\bar{p})\right] + l \cdot \nu_I - C\left(\bar{D}^I(\bar{p})\right) - \Pi_0, \tag{2.5}$$

where Π_0 captures transmission and distribution costs, overhead expenses and the opportunity cost of the monopolist.

The welfare maximization problem or, as it is referred to in Peak-Load Pricing (e.g., Crew et al. 1995, Joskow and Tirole 2006), the *Ramsey* problem is

$$\max_{(l,\bar{p})} \left\{CS(l, \bar{p}) : \Pi(l, \bar{p}) \geq 0, (l, \bar{p}) \in \mathcal{L} \times \mathcal{P}\right\}. \tag{2.6}$$

Henceforth, we refer to $\mathcal{L} \times \mathcal{P}$ as rate structure, and to an element of this set as rate level.

The literature theoretically explores optimal pricing rules for alternative structures. For instance, Joskow and Tirole (2006) consider the case where $\mathcal{L} = \mathbb{R}$, and compare a real-time pricing structure (RTP), in which $\mathcal{P} = \mathbb{R}_+^{T \cdot |\Omega|}$, with a flat rate structure, where $\mathcal{P} = \left\{\bar{p} \in \mathbb{R}_+^{T \cdot |\Omega|} : p_{t\omega} = p_{t'\omega'} \; \forall (t, \omega)\right\}$; Crew et al. (1995), on the other hand, focus on the case where $\mathcal{L} = \{0\}$, and also review the real-time and flat rate cases.

2.3 A Method to Compare Rate Structures

We propose using the model of Peak-Load Pricing as a framework to compare rate structures. For a set of tariffs under analysis, the comparison requires solving the Ramsey problem for each of the corresponding structures, and then comparing the optimal values of the problem.

This method is closely related to the approaches used by Acton and Bridger M. (1983), Caves et al. (1984b), Gallant and Koenker (1984), Howrey and Varian (1984), Lillard and Aigner (1984), Parks and Weitzel (1984), Taylor et al. (2005) and Allcott (2011). For a change in rates, these studies compute a change in welfare (W) using that $\Delta W = \Delta \Pi - \Delta r + \Delta Y$, where $\Delta \Pi$ is the change in producer surplus, Δr the variation in demand side infrastructure costs, and ΔY is the compensating variation—the money that when taken away from an individual leaves him with the same level of welfare he had before the price change (Mas-Colell et al., 1995, pp. 80–91). Researchers can use the optimal value of the Ramsey problem to quantify ΔW. With the following Lemma, we formalize this claim.

Lemma 2.3.1 *Suppose prices change from $(l_1, \bar{p}^1)$ to $(l_2, \bar{p}^2)$, and let v_1 and v_2 be the optimal value of the Ramsey problem for the rates 1 and 2, respectively. Then,*

$$\Delta W = v_2 - v_1 - \Delta r. \tag{2.7}$$

The proof of this and all other results in this section are in the Appendix A.1.

Consistently Comparing Rate Structures

One advantage of using (2.7) is that it offers a consistent criterion to compare rate structures. The solution of the Ramsey problem corresponds to the optimal rate levels. Thus, researchers compute the change in welfare associated to the best case of each structure.

This criterion offers an alternative to previous approaches. Allcott (2011), Howrey and Varian (1984), Lillard and Aigner (1984), Parks and Weitzel (1984) and Taylor et al. (2005) compute the welfare changes using predefined rates. Acton and Bridger M. (1983) compare two time-of-use (TOU)[4] with a flat rate structure. The authors determine the TOU assuming a difference between the peak and off-peak charges and imposing revenue neutrality. Caves et al. (1984b) follows a similar approach except that the authors search for an optimal TOU rate evaluating various peak to off-peak ratios. The method that is closest to ours is the one that Gallant and Koenker (1984) use. This work computes welfare changes from a flat rate to a TOU and to an RTP structures. As in our method, the paper numerically finds optimal rate levels for each structure. The key difference is that the authors use a simplified representation of the production costs. While a simplification, the function do captures an important trade-off present in electricity industries: a more capital intensive production mix, commonly associated to an increased average cost of capital, will tend to have lower short-run marginal costs at all levels of production.

A Flexible Cost Function

The difficulty of using the approach of Gallant and Koenker (1984) is that the cost function cannot be easily customized. It is not possible to use this function, for instance, in systems with potentially high penetrations of intermittent renewables. A crucial determinant of the value of these technologies is how their output correlates with consumption. This aspect is missing in the cost representation of Gallant and Koenker (1984). Using the cost function of Peak-Load Pricing, on the other hand, avoids this problem, without missing the trade-off between capital and short-run marginal costs.

Furthermore, a researcher can easily modify (2.2)-(2.4) to increase the realism and suitability of this cost representation depending on the data available. Indeed, the technical results of Subsections 2.3 and 2.3 hold if the objective function is convex, and (2.4) is a

[4] A TOU structure charges differently depending on the hour of the day, day of the weak and possibly season.

general convex set. In particular this allows modeling a transmission system, rationing and disruption costs,[5] and technologies with storage and ramping constraints.

Comparing Portfolios of Rate Structures

An element characteristic of previous tariff analyses is the pairwise comparison of rate structures. With the exception of Borenstein (2005) and Borenstein and Holland (2005), past work assesses welfare changes resulting from the whole population being in one rate and then switching to another.

In practice, a utility recovers its costs offering a portfolio of tariffs. Each rate in the portfolio applies to an specific class—a fraction of the customer base with particular cost characteristics (RAP, 2011, pp. 47–50). It is not uncommon for utilities to distinguish various classes (e.g., industrial, commercial and residential customers) and, thus, offer portfolios with several rates (RAP, 2011, pp. 47–50). It seems then appropriate to have methods that allow measuring welfare changes when changing more than one rate at a time.

These approaches can also help improving analyses focusing on just one rate structure. Changing the tariff of one class can impact other classes as well. For instance, if a new rate reduces the contribution of the class to the aggregated peak consumption, this will cause a reduction in the overall production costs.[6] If this effect is systematic, in the medium to long term, it will translate into lower bills for all customers, not just those in the class with the new rate.

To the best of our knowledge, only Borenstein (2005) and Borenstein and Holland (2005) explore welfare changes while simultaneously adjusting multiple rates. These papers study a setting with two type of customers, those enrolling in a flat rate structure and those in an RTP tariff. The studies analyze various scenarios in terms of the fraction of the population under each rate. For each scenario they compute rate levels that satisfy market equilibrium conditions, and calculate the corresponding welfare metric. Our method builds upon the idea of comparing multiple rates which are adjusted simultaneously to new systems' conditions. We expand the approach of Borenstein and Holland (2005) with a model that allows comparing general portfolios of rate structures.

In order to achieve this goal, we introduce a slight modification to the Ramsey problem. Let $\mathcal{L} := \mathcal{L}_1 \times \ldots \times \mathcal{L}_n$, $\mathcal{P} := \mathcal{P}_1 \times \ldots \times \mathcal{P}_n$ and $\mathcal{L} \times \mathcal{P}$ be the constraint set for the rate revels, that we call *portfolio of rate structures*. We consider $\mathcal{P}$ being a convex set. Besides, we let $l = (l_1, \ldots, l_n)^\top$ be a vector of fixed charges and $\bar{p} = ((\bar{p}^1)^\top, \ldots, (\bar{p}^n)^\top)^\top$ a block vector of volumetric charges. Now the duple $(l, \bar{p})$ corresponds to a portfolio of rate levels. The subindex h identifies a particular rate and the partition of population under this tariff (for instance, a class). The set I_h contains the types under partition h, and α_h is the number of customer under this partition. We focus on the case where $I_h \cap I_{h'} = \emptyset$, $\forall h \neq h'$. This is the

[5] Following the technique of Crew and Kleindorfer (1976).

[6] In the electricity sector more efficient units have priority. This implies that during peak periods the more inefficient plants are used, which increases the marginal cost of production.

relevant case since classes distinguish customers with different characteristics. We modify the consumer surplus function of Peak-Load Pricing as follows

$$CS(l,\bar{p}) = \sum_{h=1}^{n} E\left[\bar{S}^h(\bar{p}^h) - \bar{D}^h(\bar{p}^h)^\top \bar{p}^h\right] - l^\top \alpha, \tag{2.8}$$

and similarly update the profit function

$$\Pi(l,\bar{p}) = \sum_{h=1}^{n} E\left[\bar{D}^h(\bar{p}^h)^\top \bar{p}^h\right] + (l+r)^\top \alpha - C\left(\sum_{h=1}^{n} \bar{D}^h(\bar{p}^h)\right) - \Pi_0, \tag{2.9}$$

where, for simplicity, we replaced $\bar{S}^{I_h}(\bar{p}^h)$ and $\bar{D}^{I_h}(\bar{p}^h)$ by $\bar{S}^h(\bar{p}^h)$ and $\bar{D}^h(\bar{p}^h)$, respectively. In (2.9) the vector $r = (r_1, \ldots, r_n)^\top$ has in its components fixed costs which are directly associated with each rate. For example, if a utility decides to implement a time-varying rate for customers in h, it needs to upgrade its metering equipment. The parameter r_h is the cost of the new capital, and may also account for program implementation and marketing costs (all expressed as annuities).

Observe that (2.8) and (2.9) do not alter the structure of the Ramsey problem, only increase its dimensionality. If it is possible to solve (2.6) efficiently, then the same applies for the new problem.

Optimal Demand Mix

Distributed energy resources change the way in which customers interact with the electricity grid. Households and businesses become more responsive to complex price signals (Faruqui and Sergici, 2013), and they may even sell energy and provide reliability services to the grid (IEA, 2016, pp. 198–202). In an effort to adapt to this emerging reality, regulators are rethinking the design of existing rates (Stanton, 2015). The challenge involves developing tariffs that (i) provide the right long term incentives, so that DERs are adopted efficiently (NARUC, 2016, p. 41),[7] and (ii) uncover the operational value of these resources (Lazar and Gonzalez, 2015). To enable researchers to explore the extent in which a given rate structure meets these goals, we introduce a final modification to the Peak-Load Pricing framework.

This modification builds upon the observations of Borenstein and Holland (2005) and Joskow and Tirole (2006). These papers analyze the value of advanced metering infrastructure (AMI) as enabler of real-time pricing. They observe that, depending on the capital cost of this distributed energy resource, it is optimal to deploy AMI in only a fraction of the customer base. The reason is that the marginal benefit of an increasing number of customers enrolled in RTP decreases (Borenstein and Holland, 2005), while the marginal cost (the capital cost of AMI) remains constant. This result suggests treating the number of customers

[7] In this context, the word adoption refers to households and businesses acquiring a resource relevant for the grid operation, for instance, a solar photovoltaic panel or an electric vehicle. Efficient adoption refers to the deployment of these resources at the right place and time.

with the same rate structure and DER in a similar fashion as one treats the installed capacity of a production technology. Making an analogy with the supply side, one can think the distribution of customers across tariffs and DERs as a demand mix. The final modification we introduce to the Peak-Load Pricing model, allows comparing tariff structures under optimal supply and demand mixes.

Beyond improving the internal consistency of the rate evaluation method we propose, this symmetrical treatment improves the accuracy of the technique. Comparing rates assuming arbitrary long-run configurations for the demand can produce misleading results. Past rate analyses, such as Caves et al. (1984b) or Gallant and Koenker (1984), concluded that time-of-use rates were not cost-effective for residential customers. Metering costs would have outweighed efficiency gains if the program would have been implemented for the whole customer base. The observations of Borenstein and Holland (2005) and Joskow and Tirole (2006) weaken the conclusions of these studies. Time-varying rates could have passed the cost-benefit test if the authors would have focused on the optimal subset of the population. In addition, computing a demand mix is important for a reason not directly related with rate analysis. An optimal supply and demand mix provides regulators and policymakers with a snapshot of the long term configuration of the system, given a portfolio of rate structures. This perspective can be used as a benchmark and also to set targets for DERs adoption or rate enrollment.

To model a demand mix, we now assume that I is a discrete set, with ν_i being the number of customers under type $i \in I$ (e.g., $I = \{\text{industrial}, \text{commercial}, \text{residential}\}$). For customers that adopt a technology h the rate that applies is $(l_h, \bar{p}^h)$; and α_h continues to be the number of customers in h. Defining the matrix Γ such that $[\Gamma]_{ih} = 1$ if a type i can enroll in h and 0 otherwise, the feasible region for α, henceforth the demand mix, is $\mathcal{A} := \left\{\alpha \in \mathbf{R}^n_+ \colon \Gamma\alpha \leq \nu\right\}$.

As in the setting of Borenstein and Holland (2005), we focus on the case where $\sum_i [\Gamma_{ih}] = 1$. That is, we consider that one rate applies to only one customer type. In our setting the type i is a representative customer. Though interesting, we postpone the development of the general case, in which different customer types can enroll in the same rate, for future work.

The new consumer surplus function is now

$$CS(l, \bar{p}) = \sum_{h=1}^{n} \alpha_h E\left[\bar{S}^h(\bar{p}^h) - \bar{D}^h(\bar{p}^h)^\top \bar{p}^h\right] - l^\top \alpha, \tag{2.10}$$

and the utility surplus

$$\Pi(l, \bar{p}) = \sum_{h=1}^{n} \alpha_h E\left[\bar{D}^h(\bar{p}^h)^\top \bar{p}^h\right] + (l + r)^\top \alpha - C\left(\sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right), \tag{2.11}$$

where r_h not only includes the costs associated to the implementation of the rate but also those related to the technology of the customers enrolled in h. For simplicity, henceforth we consider that r_h includes the costs $\Pi_0 / \sum_i \nu_i$.[8]

[8] In practice utilities and customers share technology costs, for instance, utilities may own metering

We simplify the Ramsey problem noting that $\Pi(l, \bar{p}) = 0$ in the optimum. Using this condition, and equations (2.10) and (2.11), the version of the Ramsey problem we propose for rate analysis is

$$\max_{(\alpha,\bar{p})} \sum_{h=1}^{n} \alpha_h E\left[\bar{S}^h(\bar{p}^h) - r_h\right] - C\left(\sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right) \tag{2.12}$$

$$\text{subject to}$$

$$\bar{p} \in \mathcal{P}, \tag{2.13}$$

$$\alpha \in \mathcal{A}. \tag{2.14}$$

Researchers evaluating portfolios of rate structures can solve (2.12)-(2.14) for alternative definitions of $\mathcal{P}$ and compare the optimal value of this problem. Note that it is no longer necessary to add exogenously the variation in demand side capital costs (Δr) to the variation in optimal values of the Ramsey problem, as in (2.7). These costs are part of the objective of (2.12)–(2.14).

Enhancing the Applicability of the Framework

The Ramsey problem is nonlinear, and non-convex. This poses two important challenges to analysts using this model. First, in general it is not possible to guarantee that a solution of (2.12)-(2.14) is globally optimal. Thus, despite finding solutions for two competing portfolios of rate structures their comparison could be inconsistent. The analyst could benchmark a sub-optimal against an optimal solution. Second, the non-convexity of the problem limits its scalability as the problem size greatly decreases the performance of non-convex solvers. To enhance the practical applicability of the method, we make the following assumption:

Assumption 2.3.1 *The gross surplus function $S^h_\omega(\cdot)$ is strictly concave and the demand function $D^h_\omega(\cdot)$ is convex.*[9]

Under Assumption 2.3.1 (2.12)–(2.14) remains nonlinear and not necessarily convex or concave. However, its specific structure allows us to develop an efficient solution method. The following proposition suggests a suitable approach.

Proposition 2.3.1 *Let P_α refer to the problem* (2.12)–(2.14) *with α entering as a fixed parameter, and let $g(\alpha)$ be the optimal value of P_α. Under Assumption 2.3.1, for any $\alpha \in \mathcal{A}$, $g(\alpha)$ is concave.*

infrastructure and customers rooftop solar panels. However, ownership is not relevant from a social planning perspective when the focus is efficiency. Besides, utilities can always pass along this cost with the fixed charge l_h.

[9] Convex demand functions are mappings whose components are all convex.

Indeed, even though $g(\cdot)$ does not have an explicit functional form, we can leverage an iterative procedure solving

$$\max\{g(\alpha)\colon \alpha \in \mathcal{A}\} \tag{2.15}$$

to find a solution to the Ramsey problem. Because (2.15) is a convex problem, such an approach could potentially outperform non-convex solvers. In Section 2.4 we present evidence suggesting that this is in fact the case.

Optimal Long Term Incentives

An implicit assumption of the Ramsey problem is that regulators can control the demand mix α, i.e., the adoption of technologies and rate enrollment. While these authorities could have certain influence on the roll-out of distribution equipment, for many demand side technologies and rates customers are the agents making the adoption and enrollment decisions. At best, regulators could design long term incentives consistent with a desired demand configuration. Thus, the solution of (2.12)–(2.14) should be interpreted as the demand mix given optimal long term incentives.

For concreteness, we now propose one approach for setting these incentives. Let $\bar{p}(\alpha)$ be the optimal solution of P_α. Define, in addition, $\bar{\lambda}(\alpha)$ as the dual variable of (2.3) when the demand parameter is the aggregated demand (the argument of $C(\cdot)$ in (2.11)) evaluated at $(\alpha, \bar{p}(\alpha))$. Further, define the vector of fixed charges $l(\alpha)$ as follows,

$$l_h(\alpha) = E\left[(\bar{\lambda}(\alpha) - \bar{p}^h(\alpha))^\top \bar{D}^h(\bar{p}^h(\alpha))\right] + r_h. \tag{2.16}$$

The incentive rule we propose is *set the rate levels at* $(l(\alpha), \bar{p}(\alpha))$ *when the demand mix is* α.

This rule has appealing theoretical properties. It replicates the incentive structure of an industry with a distribution utility and a competitive wholesale market (see e.g., Joskow and Tirole (2006)). More importantly, one can show that that these incentives align customers' individual choices with the maximization of societal welfare. The three results that follow formalize this property.

First, we observe that if the system is at a socially optimal configuration, under the rule we propose no customer has incentives to switch.

Proposition 2.3.2 *Let* $(\alpha^*, \bar{p}(\alpha^*))$ *be an optimal solution of* (2.12)-(2.14)*, and* $l(\alpha^*)$ *the vector of fixed charges. Then, the portfolio of rate levels* $(l(\alpha^*), \bar{p}(\alpha^*))$ *is such that no customer has an incentive to switch to an alternative rate.*

Assuming that the utility updates rates relatively fast compared with how customers switch, it is possible to show a result akin to Theorem 5 in Borenstein and Holland (2005). The result provides an intuition about the welfare effects of customers switching, including effects on these customers and on the population as a whole.

Proposition 2.3.3 *If immediately after a rate update the first group of customers switching leave $h1$ to enroll in $h2$, then (i) the surplus of switchers increases, (ii) the difference between surplus of customers in $h1$ and $h2$ decreases, (iii) the aggregated welfare increases, and (iv) the marginal benefit of the switch decreases.*

The final proposition shows that a regulator could induce the optimum of the Ramsey problem instructing an incentive rule such as the one we propose. Interestingly, this incentive rule does not require the regulator knowing the optimal demand mix in advance. The current conditions fully determine the incentive.

Proposition 2.3.4 *The incentive rule that sets rate levels at $(l(\alpha), \bar{p}(\alpha))$, when α describes the demand mix, induces the Ramsey optimum in the long-run.*

Beyond being compelling from a theoretical perspective, this result enhances the consistency of the method introduced in this paper. When using (2.12)–(2.14) to compare structures, a researcher compares among best cases which could be implemented. The result also improves the value of the Ramsey problem as a benchmark case for planning studies.

2.4 Solution Method

In principle, one could solve (2.12)–(2.14) with a non-convex optimization package. Here we describe an alternative approach that exploits Proposition 2.3.1. We propose solving (2.15), thereby solving implicitly the Ramsey problem. Because $g(\alpha)$ does not have an explicit functional form, we use an iterative procedure comprising an inner and outer routines. The inner routine solves P_α, a task that, in general, a convex optimization solver can handle. The outer routine searches an optimum for (2.15). At each iteration it uses information of the optimal solution of P_α in order to compute a search direction and a step size. The outer routine can be implemented with a variety of nonlinear optimization algorithms; in this paper we implement it with a *Barrier* method (Boyd and Vandenberghe, 2009).

The barrier method works by dropping the problem's inequality constraints and augmenting the objective with a *barrier function* $\phi(\cdot)$. The new objective is $z_t(\alpha) := tg(\alpha) + \phi(\alpha)$, where t is a weighting parameter that changes across iterations. We define $\phi(\cdot)$ as $\phi(\alpha) := \sum_{i\in I} \ln(\nu_i - \Gamma_{i\bullet}\alpha) + \sum_{h=1}^{n} \ln(\alpha_h)$, where $\Gamma_{i\bullet}$ is the ith row of the matrix Γ. The new problem is an unconstrained nonlinear program, thus, it can be handled with a Newton-like method. Algorithm 1 describes the outer routine.

In practice, one cannot use a standard Newton method in order to maximize $z_t(\cdot)$. Because this function does not have an analytic formula, there is no analytic expression for its inverse Hessian. A suitable approach is using a quasi-Newton algorithm (Luenberger and Ye, 2008). These methods use an approximation of the inverse Hessian when computing a Newton step. The more sophisticated versions calculate improved approximations of this matrix using first order information gathered as the optimization procedure progresses. In this paper we use the *Limited Memory Broyden-Fletcher-Goldfard-Shanno* method (L-BFGS),

Algorithm 1 Outer routine solved via Barrier method

```
Initialization: Given a strictly feasible $\alpha$, $t \leftarrow t^0 > 0, \mu > 1$ and $\epsilon > 0$
while $(|I| + n)/t \geq \epsilon$ do
    $t \leftarrow t + \mu$
    Find with Newton method $\alpha^* := \mathrm{a}rg\, max_{\alpha}\, z_t(\alpha)$
    $\alpha(t) \leftarrow \alpha^*$
```

which guarantees R-linear convergence for uniformly concave problems. The performance of this algorithm improves considerably if $\nabla[z_t(\alpha)] = t\nabla g(\alpha) + \nabla\phi(\alpha)$ is available analytically. While deriving an analytic expression for the second term is straightforward, for the first we use a result from sensitivity analysis for nonlinear programs.[10] Its explicit formula is

$$[\nabla g(\alpha)]_h = E\left[\bar{S}^h\left(\bar{p}^h(\alpha)\right) - \bar{\lambda}(\alpha)^\top \bar{D}^h\left(\bar{p}^h(\alpha)\right)\right] - r_h. \tag{2.17}$$

Performance To test the performance of our approach, we constructed a simple experiment that is similar to the one described in Section 2.5 with two key differences. First, we used only two tariffs: (i) real-time pricing (RTP), for which the volumetric charge varies freely, and (ii) flat rate, for which the volumetric charge is constant across all time steps and outcomes. Second, in addition to examining a scenario with a sample spaces composed of 365 outcomes (as in Section 2.5), we also examined a smaller case with 50 outcomes. For each scenario, we implemented the preceding algorithm, which we will refer to as *parametric*, as well as a *first order* and *second order* procedures. The latter two methods solve the Ramsey problem directly with a nonlinear solver. While the first order passes to the solver the analytic formula of the gradients, the second order provides the analytic Hessian as well. We tested each algorithm-scenario combination for 20 different parameter choices corresponding to a range of technology costs and demand price-responsiveness parameters. The stopping criterion for all algorithms was the same: the lesser of the time to converge within a 10^{-6} duality gap, or $14,400$ seconds.

Overall, we found that the parametric method converged in 13±5s (for 50 states of nature, ± denotes standard deviation) and 226±79s (365 states of nature). The second order method converged in 49±25s and 3,132±1,762s (50 and 365 outcomes, respectively) and the first order method converged (or reached the maximum compute time) in 2,865±2,568s and 11,405±5,050s.

[10] For details see the proof of Proposition 2.

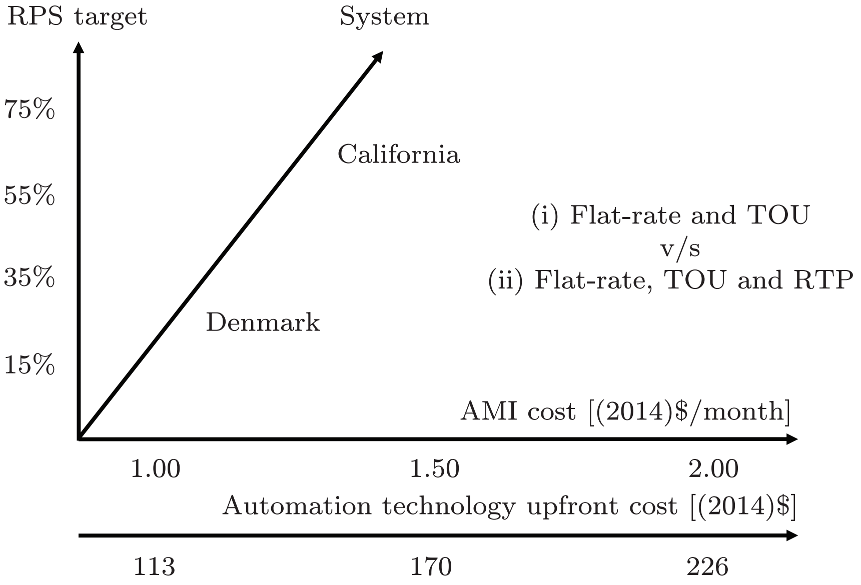

Figure 2.1: Structure of analysis

2.5 An Application: The Value of Real-Time Pricing

Analysis Design and Data Assumptions

We compared two portfolios of rate structures: (i) a portfolio that includes a flat rate and a time-of-use tariff[11] and (ii) a portfolio that adds an RTP tariff to the first portfolio. As Figure 2.1 shows, for each portfolio we considered a range of scenarios for demand-side technology costs, a range of *Renewable Portfolio Standard* (RPS)[12] targets, and load and renewable production data from two systems (Denmark and California; for simplicity we restrict renewable production to wind).

Table 2.1 shows the supply-side technologies we considered and their economic parameters. The only relevant supply-side technical parameter is the availability factor, which we set to 85% for all states of nature for non-wind technologies.[13] For wind, we use historical system-wide hourly capacity factors for 2014, available from California's Open Access Same-time Information System (OASIS) and the website of the Danish transmission system operator.

On the demand side, we consider one customer type[14] and three arrays of technologies, denoted as Tech 1, 2 or 3: (1) a standard meter, (2) advanced metering infrastructure (AMI) or (3) AMI plus automation technology. AMI enables customers to participate in

[11] Here we consider a TOU with different volumetric charges for each hour of the day.

[12] An RPS target mandates the utility to produce a fraction of its energy with renewables.

[13] The availability factor from NERC's Generating Availability Data System website.

[14] We chose one type due to data availability and to simplify the analysis.

Table 2.1: Economic parameters of supply-side technologies

		Base-load	Mid-merit	Peak	High-peak	Wind
Capital cost	k$/MW-yr	207	85	27	16	225
Fixed O&M		69	21	16	11	40
Total fixed		227	106	43	27	265
Fuel	$/MWh	11	27	43	66	0
Variable O&M		5	11	11	11	0
Total variable		16	38	54	77	0

Notes: Non-wind parameters taken from De Jonghe et al. (2012). Wind costs are the average of those from EIA (2013) for California and from Vitina (2015) for Denmark.

TOU or RTP tariffs, whereas automation technology enables customers to automate the price response of their appliances. We model the latter phenomenon assuming different price elasticities for customers with and without automation. Table 2.2 shows these elasticities whose range is taken from empirical estimates in Faruqui and Sergici (2013).

Table 2.2: Demand elasticities

No automation		Automation					
		Low increase		Medium increase		High increase	
own[a]	cross[b]	own	cross	own	cross	own	cross
(0.02)	0.07	(0.04)	0.14	(0.05)	0.20	(0.07)	0.27
(0.04)	0.14	(0.05)	0.20	(0.07)	0.27	(0.08)	0.33
(0.05)	0.20	(0.07)	0.27	(0.08)	0.33	(0.10)	0.40

[a] Own-price elasticity.

[b] Cross-price elasticity.

Note: Rows provide a range of own- and cross-price elasticities.

Tech 1 costs are normalized to zero; the costs to move to Tech 2 or 3 are incremental. AMI costs are based on U.S. DOE data (DOE, 2012a,c).[15] For automation technology we base costs on currently available advanced programmable controllable thermostats. The highest x-axis values in Fig. 2.1 agree with these incremental cost data; the two additional costs correspond to 25% and 50% reductions.

The baseline levels of consumption correspond to the hourly, system-wide load profile for the year 2014 for both, California and Denmark. While for California this data is publicly available in OASIS, the Danish transmission system operator makes the hourly, system-wide load profile available in its web page.

[15] These estimates are for capital and installation costs, net of any benefits to utility operations (e.g. meter reading and back-office staffing). The estimates do not include hypothetical reductions in energy or capacity expansion costs.

A.2 contains further details of our implementation for this exercise. It describes how we calibrated the demand parameters and makes explicit our treatment of the RPS target.

Results

As Figure 2.2 shows, welfare in Portfolio (ii) exceeds Portfolio (i) across the range of factors we explored. Additionally, higher elasticity levels increase the positive impact of adding RTP to (i). These findings are not surprising in light of previous work (e.g. Borenstein and Holland (2005), De Jonghe et al. (2012), Crew et al. (1995) and Joskow and Tirole (2006)). However, our framework allows a more comprehensive analysis.

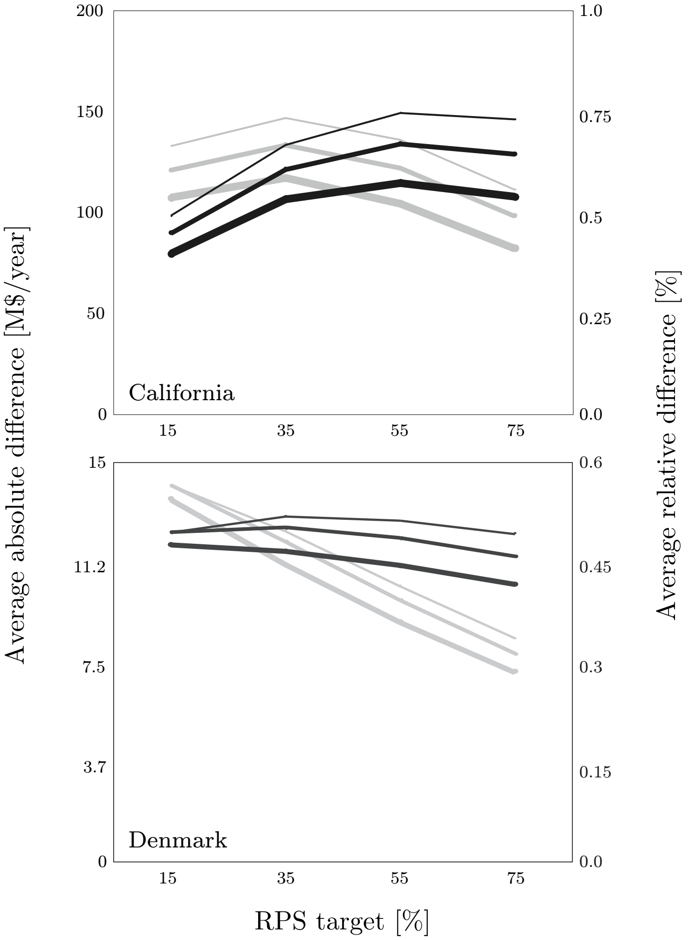

Figure 2.2: Average welfare differences between portfolios ((ii) - (i)) by RPS target. Absolute differences in black. Relative differences in gray. Line thickness represents different elasticity levels. Thicker lines correspond to lower own- and cross-price elasticities.

In terms of relative welfare differences, Portfolio (ii) does not seem particularly attractive.

Considering additional rate design factors such as simplicity or public acceptance, a regulator could conclude that neither in California nor in Denmark real-time pricing is valuable enough to justify its deployment. However, a significantly different picture emerges when one observes the results in Table 2.3.

Table 2.3: Demand mix by RPS target

			California				Denmark			
			RPS [%]							
Portfolio	Tech	Tariff	15	35	55	75	15	35	55	75
(i)	1	FR	97	97	96	97	96	97	99	99
	2	TOU	2	2	3	2	3	2	1	1
	3	TOU	1	1	1	1	1	1	0	0
(iI)	1	FR	87	82	79	77	96	95	95	95
	2	RTP	9	11	13	14	2	3	3	3
	3	RTP	4	7	8	9	2	2	2	2

Note: Values are percentages of population under each category.

The demand mix provides a complementary perspective on the relevance of demand alternatives for a system. While in California between 13 to 23 percent of the population would enroll in real-time pricing were it present, only a 4 to 5 percent would in Denmark. In equilibrium these differences appear not relevant because welfare results are similar, in relative terms. But, in view of Proposition 2.3.3, the roll-out of RTP will certainly accrue more benefits across time for customers in systems like California's than for those in systems like Denmark's. This policy-relevant perspective cannot be achieved without a modeling framework such as that developed in this paper.

An alternative way of inferring differences in the demand mix would be analyzing net-load duration curves.[16] Since demand responsiveness competes with peaking plants (Borenstein and Holland, 2005; De Jonghe et al., 2012), differences during peak-hours should translate into differences in the demand mix. Figure 2.3 shows, however, that small variations in the shape of the curves can imply significantly different demand mixes. Thus, using net-load duration curves in order to anticipate possible differences does not seem a suitable approach.

Finally, we point out two additional conclusions an analyst can derive from the demand mix. First, it allows simplifying portfolios of rate structures. For instance, Table 2.3 shows that when faced with Portfolio (ii) no customer enrolls in the time-of-use program, which

[16] Net load-duration curves are analysis tools used in the electricity sector to estimate the long term value of different production technologies. They plot hourly net loads (system demand minus the renewable production) for each hour of a period, say a year. Hours are sorted from left to right with the highest net load hour to the extreme left and the lowest to the extreme right. A point in this curve indicates the fraction of the time ($x-$coordinate) the net load is greater or equal to the net load in the curve (the ordinate). For more details, we refer the reader to Stoft (2002, pp. 40–45).

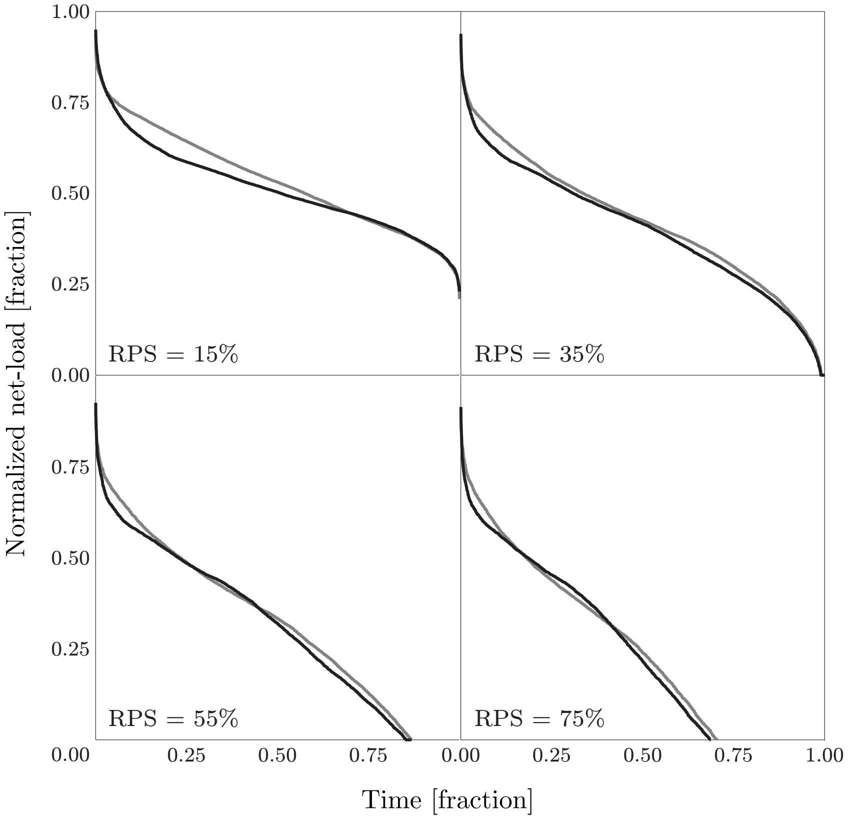

Figure 2.3: Net-load duration curves by RPS target. California net-load in black. Denmark net-load in gray. All curves reach a maximum of 100%.

indicates that a simpler portfolio achieves the same benefits. Second, the demand mix establishes targets for rolling-out demand technologies and rate structures. These targets are not trivial as they correspond to fractions of the population and are contingent on the many factors we explored. In particular, considering that the AMI costs we used are net of any benefits to utility operations (e.g. meter reading and back-office staffing), our results suggest that 100% smart meter rollouts are not cost-effective in the regions we investigated.

2.6 Conclusions

This work introduces an analytic method for helping planners and regulators in the design of rates in the electricity sector. It develops a nonlinear program that serves as tool to compare portfolios of rate structures, and proposes a suitable approach to find an optimal solution of this model.

The flexibility of our method allows consistently comparing tariff structures, and it enables researchers to model a richer set of trade-offs influencing the production costs in the sector. It also allows the comparison of portfolios of tariffs under optimal demand mixes. A theoretical exploration of the properties of the nonlinear program suggests that our method compares rates which are not only socially optimal but could also be implemented. Besides, the demand mix that the model computes offers a valuable perspective on the potential of competing demand alternatives. It helps planners and regulators to prioritize demand technologies and rates, and to establish appropriate levels of deployment for each demand option.

The application of the framework to the comparison of portfolios of rate structures in California and Denmark shows its practical value. As the theory predicts, real-time pricing increases welfare in both systems. But these benefits may not be enough to deploy it in either. The systems have, however, different demand mixes which indicate different policy prescriptions. While in Denmark RTP appears unattractive, it at least deserves further revision in California.

Chapter 3

A Mathematical Programming Approach to Utility Pricing

3.1 Introduction

The distribution segment of the supply chain of commodities such as gas or electricity has characteristics of natural monopoly. In these instances one firm, a distribution utility, serves all customers and maintains the entire distribution infrastructure. Absent competition, a regulator supervises the pricing, or rate design, of the utility's services. The resulting regulated prices serve two main purposes. They guarantee that the firm recovers its costs, and thereby sustains its operations, and sends proper economic signals to retail customers (Kahn, 1988, pp. 1–14).

Changes in the landscape of the utility industry are challenging prevailing rate designs. In some sectors, such as in the electricity industry, innovations in information and automation technologies enable utilities to implement more complex tariffs. For instance, advanced metering infrastructure (AMI) allows measuring and recording electricity consumption at the hourly timescale. With this technology utilities can implement time-varying rate structures, designs in which volumetric charges can change across time (Joskow and Wolfram, 2012).

The increasing penetration of distributed energy resources (DERs) and intermittent generation also incentivize innovation in rate design. The massive adoption of DERs, such as rooftop solar photovoltaic (PV) panels, home energy management systems (HEMS) or electric vehicles (EVs), has pushed regulators to rethink the way in which utilities should collect their revenue (NARUC, 2016). In a world of pervasive DERs, innovative tariffs design are expected to improve the efficiency of the sector, decreasing short- and long-run systems costs, and tackling distortions such as the cross-subsidization between customers with and without DERs (NARUC, 2016).

On the other hand, high-voltage grid-connected intermittent generation technologies, such as windmills or solar photovoltaic systems, bring new challenges to the grid operations. The output of these technologies is driven mainly by weather conditions, such as wind speed, wind direction, cloud cover or haze, which change considerably across time. As a result, systems operators cannot dispatch these resources at will. In absence of economically feasible storage, the intermittentcy of wind or solar requires additional reserves to ensure reliability and more capacity to meet demand. These requisites could translate into higher production costs and even undermine carbon emission reductions (Frondel et al., 2015). Rate structures that reflect closely system conditions could work in sync with renewable resources, incentivizing consumption when these resources are available, and bringing demand down when they are scarce.

In theory, there is one type of rate that could materialize all these benefits: a two-part real-time pricing (RTP). This variant of a time-varying rate, in which volumetric charges can differ from hour to hour, outperforms all other structures from an efficiency perspective (Joskow and Tirole, 2006). Despite of this fact, at the time of this writing, only a small number of jurisdictions have implemented this structure (Schwartz et al., 2017). Part of the reason is that designing rates not only involves economic considerations. Regulators must balance other objectives such as the simplicity, distributional impacts and stability of competing tariff designs (Bonbright, 1988). Having techniques that allow consistently quantifying the economic differences of various rates brings more clarity to the overall analysis. Regulators can balance economic gains of more efficient rates, such as a real-time pricing, versus other objectives. Even though RTP could maximize efficiency, the economic gains with respect to a more simple or stable rate could be small enough to advice against its implementation.

This paper contributes with a quantitative technique to evaluate rate structures. In contrast to previous approaches, our method allows comparing within a unified framework a large class of tariff designs. This includes time-varying structures, rates with demand charges, or charges for peak-consumption, and block rates—tariff with volumetric charges contingent on total consumption. A second distinctive characteristic of the present technique is that it enables modeling in a transparent manner important aspects emerging in the utility industry. Our framework allows representing in detail distributed energy resources, such as PV panels or battery storage systems, and flexible appliances, such as heating and cooling systems. This realistic representation of the demand side is embedded within a traditional capacity expansion setting (e.g., Murphy and Smeers 2005). This allows researchers to explore the long-run implications of different rate structures, taking into account their interaction with the full supply chain. More importantly, it enables to model the time-variant and stochastic nature of intermittent technologies, resources which are increasingly important in modern utility sectors.

Our model builds upon the theory of Peak-Load Pricing, a framework that captures the interaction between rate structures and investment decisions in the utility sector, and can be used to compare rate designs. In this setting, a regulator chooses the pricing of a monopolistic utility in order to maximize societal welfare, solving what is called the Ramsey-Boiteux problem (Joskow, 2007, ch. 6). In doing so, the regulator, or Ramsey planner, internalizes the consumer responses to different prices and the cost function of the monopolist, which together determine the optimal allocation (Crew et al., 1995). One can use this framework to compare rate structures. By solving the model with different constraints sets for the prices, researchers can compare the characteristics of the resulting equilibria.

The Ramsey-Boiteux problem is an instance of a Bilevel Model, a type of mathematical programming problem in which group of variables is constrained to be in the solution sets of subordinate, or lower level, problems (Dempe and Dutta, 2012). One way of approaching a Bilevel Problem is replacing these solution sets by the first order necessary conditions of the lower level problems. The resulting model is a Mathematical Program with Equilibrium Constraints (MPEC), which can be handled with specialized nonlinear algorithms.

Because of dimension of the type of problems we attempt to tackle with the present technique, we develop a decomposition approach for the MPEC version of the Ramsey-Boiteux problem (RMPEC). The algorithm is a nonlinear variation of the Alternating Directions Methods of Multipliers (ADMM), a distributed computation technique which blends the decomposabilty of dual subgradient methods and the convergence properties of the method of multipliers. ADMM algorithms solve optimization problems via successive iterations, each of which optimizes an augmented problem along two blocks of variables. Typically, the resolution of one block can be distributed, and the other uses as input parameters the distributed solutions. In our implementation, the distributed step involves solving low-dimensional MPECs, that we tackle using the mixed integer programing representation of Fortuny-Amat and McCarl (1981). To handle the problem of the other block of variables we use conic programming. While inexact, our variant of ADMM allows tackling large-scale instances of the Ramsey-Boiteux problem, greatly enhancing the applicability of our approach.

With a computational experiment, we explore the performance of the present technique. We test the algorithm on 200 hundred instances, which we construct varying the number of variables and parameters of the model. We find that the algorithm has desirable properties for practical applications. It vastly outperforms a popular commercial solver for Mathematical Programs with Equilibrium Constrains: Knitro. While this package is not able to find a solution within 24 hours even for instances of small size, our algorithm converges within 2 hours in 94% of the instances tested. Furthermore, the results suggest that the algorithm is suitable for distributed computation. The distributed step of our variant of ADMM increases close to linear with the size of the problem while the centralized step grows at a rate lower than linear.

In order to illustrate the value of the technique as a tool to compare rates in the the utility sector, we conduct an analysis of tariffs structures in a simplified setting. The analysis highlights the value of the modeling flexibility that the present technique provides. It shows how abstracting from network constraints can introduce significant distortions in the analysis of rates. In our exercise, omitting the existence of a network leads to underestimating the benefits of time-varying rate structures by close to 31 times. In addition, the analysis shows that DERs could potentially complement time-varying rates, increasing the value of this structures while at the same time these type of rates could incentivize greater adoption of distributed resources. Finally, the exercise highlights the value of being able to explore impacts of rate design on a population of heterogeneous customers. While in all cases switching from a flat rate to a more sophisticated rate structure improves the welfare of the aggregate of households, some rates benefit wealthier customers more. Moreover, there are time-varying structures that can even harm customers with low levels of wealth, being these worse off after the switch.

3.2 Pricing Utility Services

The pricing of utility services involves the determination of a *revenue requirement* and *rate structures*. While the former corresponds to the total compensation utilities receive from its customers, the latter are the instruments they use to collect such compensation.[1] To the extent that rate structures have an impact on consumption behavior which affects operational and capital expenses, determining the revenue requirement and the structure of the rates are not independent efforts. However, it has been the industry and academic practice to consider these steps independently, understanding the complexities involved in one task while simplifying the other (see, e.g., Joskow 2007, RAP 2011). In this work we adhere to this strategy, focusing on the determination of rate structures.

One can further divide the definition of a rate structure into selecting the set of charges that the tariff will include—what we call its design—and setting the actual values of these charges. The technique we contribute with allows regulators and policy makers to discern among competing rate designs. For simplicity and because it is the focus of this paper, we refer to the design of a rate as rate structure or tariff structure as well.

Quantitatives methods for evaluating rate structures

Applied analyses of rate structures in the utility sector have sought measuring welfare changes resulting from modifications in the design of the prevailing rates. The techniques developed in these studies differ in how they quantify a change on welfare, ΔW. All make use of the following identity,

$$\Delta W = \Delta Y + \Delta \Pi - \Delta r, \tag{3.1}$$

where Δr correspond to a change in customer related costs (e.g., capital cost of AMI), $\Delta \Pi$ change in production costs and ΔY is the compensating variation—the money that when taken away from individuals leaves them with the same level of welfare they had before the price change (Mas-Colell et al., 1995, pp. 80–91). Based on how the studies treat ΔY, we distinguish two groups. Caves et al. (1984b), Gallant and Koenker (1984), Howrey and Varian (1984), Lillard and Aigner (1984) and Parks and Weitzel (1984) start assuming a functional form for the indirect utility function—the optimal value of the consumer utility maximization problem, estimate the parameters of a theoretically consistent demand model, and use these estimates plus the relationship between the indirect utility and compensating

[1] For a more complete discussion on the subject of rate design, see (Joskow, 2007, ch. 6 - 7)

variation to derive the latter.[2] A second group uses the consumer surplus to approximate ΔY. Acton and Bridger M. (1983), Borenstein (2005), Borenstein and Holland (2005), Taylor et al. (2005) and Allcott (2011) start by assuming a functional form for a system of demand equations, estimate or calibrate its parameters, and integrate the system in order to compute the consumer surplus. Willig (1976) shows that when focusing on goods with an associated expenditure relatively small with respect to the customer's budget—such as the consumption of utility services, consumer surplus and compensating variation are equivalent.

The method that we propose in this paper falls within the second strand of approaches and extends previous techniques in significant ways. It allows the consistent comparison of a wide range of rate designs, including time-varying rate structures, rates with demand charges and increasing or decreasing block rates. With the exception of Gallant and Koenker (1984), previous studies have developed techniques to measure welfare differences between variations of time-varying rate designs, such as a flat rate (FR), a time-of-use rate (TOU) or a real-time pricing structure (RTP). These approaches have either compared tariff structures specifying ex-ante the value of the volumetric charges, or imposing unnecessary constraints on them (see, e.g., Acton and Bridger M. 1983 or Borenstein 2005). Gallant and Koenker (1984), on the other hand, develops a technique that permits researchers to compare time-varying rate structures with designs of the same kind supplemented with demand charges. In their setting volumetric charges are endogenous. This approach, however, does not permit to model other important groups of rate design such as block rate structures.

A second limitation of Gallant and Koenker (1984), which is shared by all other techniques with the exception of Caves et al. (1984b), is a simplified representation of the supply side. Previous work has either ignored the supply side (e.g. Allcott 2011, Howrey and Varian 1984 and Taylor et al. (2005)), or has simplified its representation with a cost function (e.g. Borenstein and Holland 2005 and Gallant and Koenker (1984)). Even though, Caves et al. (1984b) take a more comprehensive approach, using a detailed economic dispatch model, they use this model only to estimate a marginal production cost function. Our technique, on the other hand, finds rates while simultaneously optimizing short- and long-run production decisions for a wide variety of production technologies and in the presence of key infrastructure such as a transmission network.

While we believe these contributions are valuable, the most salient aspect of the present technique is that it allows researchers to construct a bottom up model of customer behavior. In our setting, the demand of a customer is the solution of a utility maximization problem subject to a set of constraints. This constraints can be defined by the researcher providing the flexibility to model a wide variety of devices, such as HVACs systems or refrigerators, or distributed energy resources, such as rooftop solar PV or battery storage systems. This modeling flexibility is fully unique to our method; all previous techniques use a top-down, aggregated representation of the demand. A bottom up model overcomes three important

[2] For an individual with an income I, the relationship between the compensating variation and the indirect utility function when prices change from p^0 to p^1 is $v(p^0, I) = v(p^1, I - \Delta Y)$, where $v(\cdot, \cdot)$ is the indirect utility function (Mas-Colell et al., 1995, pp. 80–91).

limitations of past approaches. One is that these implicitly consider a population of homogeneous customers. While heterogeneity may not be crucial in determining the aggregated benefits of a change in the design of rates, it will certainly help to understand the implications for different types of customers. One cannot use a model with a representative customer to study distributional impacts. A second drawback is that a top down approach precludes researchers from including in a transparent manner important determinants of consumption behavior, such as weather patterns. The third limitation, which is specially relevant today, is the impossibility of using top down approaches to study distributed energy resources. By their very nature, these technologies are sources of heterogeneity in the population, geographically, and in terms of consumption patterns. Given that rate design can not only influence how households use these resources but also how they adopt them, it is important to have tools that permit anticipate plausible outcomes.

We build our model considering as starting point the model we developed in the previous chapter, which is based on the theory of Peak-Load Pricing. For completeness and to introduce the notation that we will use throughout the chapter, we now succinctly review this theory.

Peak-Load Pricing: A Theoretical Framework to Compare Rate Structures

Let Ω be a discrete sample space, q_ω the probability that $\omega \in \Omega$ occurs, and $E[\cdot]$ the associated expectation operator. We refer to an element in Ω as outcome or state of nature, and distinguish a random from a deterministic variable writing the former in boldface. Given a random variable $\boldsymbol{y}$, we denote y_ω the realization of this variable when ω occurs. The symbol $\top$ indicates the transpose of a vector.

The theory of Peak-Load Pricing has as objective to provide guidelines for the pricing of public utility services (Crew et al., 1995). Its starting point is the problem faced by a regulator that set prices with the aim of maximizing societal welfare. At the same time, the regulatory body must guarantee that the regulated monopolist is able to cover its costs.

The monopolist serves a population of customers with different types $i \in I$. These distribute in the population according to the frequency function $\delta(\cdot)$, such that $\delta(i)$ is the number of customers with type i. The monopolist offers a set $\{1, \dots, T\}$ of commodities and the customers decide among consumption bundles $d \in \mathbb{R}^T_+$. A quasi-linear utility $U(d; \theta^i_\omega) + m_\omega$ characterizes the preferences of types i over these bundles, with $U : \mathbb{R}^T_+ \to \mathbb{R}$ and θ^i_ω a set of exogenous parameters. A type i has limited budget M_i. The customer's demand for each commodity results from her choosing optimally among bundles, i.e.,

$$D\left(p_\omega; \theta^i_\omega\right) := \underset{d \geq 0}{arg\,max} \left\{U(d; \theta^i_\omega) + M_i - p_\omega^\top d\right\}. \tag{3.2}$$

The theory assumes that U is strictly concave so that $D\left(p_\omega; \theta^i_\omega\right)$ is a singleton; and the gross surplus of this customer is $S(p_\omega; \theta^i_\omega) := U\left(D(p_\omega; \theta^i_\omega)\right)$. In addition for $I' \subseteq I$, we denote $\boldsymbol{D}^{I'}(\boldsymbol{p}) := \int_{I'} D(\boldsymbol{p};\boldsymbol{\theta}^i)\delta(i)di$ the aggregated demand of the types in I', define $\boldsymbol{S}^{I'}(\boldsymbol{p})$ in a similar fashion, and call $\nu_{I'} := \int_{I'} \delta(i)di$ the total number of customers with types in I'.

The monopolist collects its revenue with a two-part rate structure, a contract $(l, \boldsymbol{p})$ where l is a fixed (or customer) charge and $\boldsymbol{p}$ a vector of charges per unit of consumption. The corresponding consumer surplus is

$$CS(l,\boldsymbol{p}) = E\left[\boldsymbol{S}^I(\boldsymbol{p}) - \boldsymbol{p}^\top \boldsymbol{D}^I(\boldsymbol{p})\right] - l \cdot \nu_I. \tag{3.3}$$

There is a set of production technologies that we index with the letter $k \in K$. Each technology differs from others on its variable costs per unit of production, $c_{\omega k} \in \mathbb{R}^T_+$, its fixed costs $\hat{r}_k$, and its availability factor, $\rho_{\omega k} \in \mathbb{R}^T_+$. The latter captures the variability in the technology's availability due to, for instance, the intermittent output of some renewables or the occurrence of outages. The installed capacity of technology k is x_k and $y_{\omega k} \in \mathbb{R}^T_+$ is its production vector. Wit this definitions, the production cost for a bundle $\boldsymbol{d}$ is

$$C(\boldsymbol{d}) = \min_{(x,y)} \sum_{k\in K} E\left[\boldsymbol{y}_k^\top \boldsymbol{c}_k + x_k \hat{r}_k\right] \tag{3.4}$$

$$\text{subject to}$$

$$\boldsymbol{d} \leq \sum_{k\in K} \boldsymbol{y}_k, \tag{3.5}$$

$$0 \leq \boldsymbol{y}_k \leq x_k \boldsymbol{\rho}_k, \ k \in K \tag{3.6}$$

The profit of the monopolist is

$$\Pi(l,\boldsymbol{p}) = E\left[\boldsymbol{p}^\top \boldsymbol{D}^I(\boldsymbol{p})\right] + l \cdot \nu_I - C\left(\boldsymbol{D}^I(\boldsymbol{p})\right) - \Pi_0, \tag{3.7}$$

where Π_0 captures transmission and distribution costs, overhead expenses and the opportunity cost of the monopolist. We note that this profit function can also represent the aggregated profits of a sector in which there is perfect competition at the wholesale level, and there is a regulated utility at the retail level (see Joskow and Tirole 2006, Chao 2011).

The welfare maximization problem or simply the Ramsey problem is

$$\max_{(l,\boldsymbol{p})} \left\{CS(l,\boldsymbol{p}) : \Pi(l,\boldsymbol{p}) \geq 0, (l,\boldsymbol{p}) \in \mathcal{L} \times \mathcal{P}\right\}. \tag{3.8}$$

Henceforth we refer to $\mathcal{L} \times \mathcal{P}$ as *rate structure*, and to an element of this set as *rate level.*

3.3 An Alternative Quantitative Technique

The model we developed in the previous chapter extends the basic setting Peak-Load Pricing. We now briefly review this model and discuss its limitations.

Let $h \in \{1, \ldots, H\}$ index the rates in the portfolio, and redefine $\mathcal{P} := \mathcal{P}_1 \times \cdots \times \mathcal{P}_H$, and $\mathcal{L} := \mathcal{L}_1 \times \cdots \times \mathcal{L}_H$ such that $\mathcal{P}_h$ is the feasible region for the volumetric charges of rate h and $\mathcal{L}_h$ constraints the fixed charges of the same rate. In this setting h also indicates the DER a customer adopts. The letter α_h denotes the number of customers enrolled in h, and the vector $\alpha \in \mathbb{R}_+^H$ represents the distribution of the population across rates, which we call demand mix. In this context I is a discrete set indexed by i, with ν_i being the number of customers with type i. Defining the matrix Γ such that $[\Gamma]_{ih} = 1$ if a type i can enroll in h and 0 otherwise, the feasible region for α is $\mathcal{A} := \left\{\alpha \in \mathbf{R}_+^n \colon \Gamma\alpha \leq \nu\right\}$. As in the basic Peak-Load Pricing framework, the functions U, D, S correspond to the direct utility, demand and gross surplus, respectively; and a set of endogenous parameters $\boldsymbol{\theta}^h$ determines them for each h.

The consumer surplus function in this model is

$$CS(l, \boldsymbol{p}) = \sum_{h=1}^{n} \alpha_h E\left[S(\boldsymbol{p}^h; \boldsymbol{\theta}^h) - D(\boldsymbol{p}^h; \boldsymbol{\theta}^h)^\top \boldsymbol{p}^h\right] - l^\top \alpha, \tag{3.9}$$

and the surplus of the regulated utility

$$\Pi(l, \boldsymbol{p}) = \sum_{h=1}^{n} \alpha_h E\left[D(\boldsymbol{p}^h; \boldsymbol{\theta}^h)^\top \boldsymbol{p}^h\right] + (l + r)^\top \alpha - C\left(\sum_{h=1}^{n} \alpha_h D(\boldsymbol{p}^h; \boldsymbol{\theta}^h)\right). \tag{3.10}$$

Because $\Pi(l, \boldsymbol{p}) = 0$ in the optimum, we can write the Ramsey problem as follows

$$\max_{(\alpha, \boldsymbol{p})} \sum_{h=1}^{n} \alpha_h E\left[S(\boldsymbol{p}^h; \boldsymbol{\theta}^h) - r_h\right] - C\left(\sum_{h=1}^{n} \alpha_h D(\boldsymbol{p}^h; \boldsymbol{\theta}^h)\right) \tag{3.11}$$

$$\text{subject to}$$

$$\boldsymbol{p} \in \mathcal{P}, \tag{3.12}$$

$$\alpha \in \mathcal{A}. \tag{3.13}$$

Limitations of the model

The Ramsey problem has some limitations as a model to compare rate structures. The main two relate to the realism of the demand representation and the type of rates that could be modeled with this framework. The demand representation in (3.11)–(3.13) is similar to that of the basic Peak-Load Pricing model in that it is the solution of a utility maximization problem akin to (3.2). The theory makes the assumptions needed so the solution set of this problem is a singleton. While this simplification makes the Ramsey problem amenable to mathematical analysis, it comes along with two mayor drawbacks. First, in many cases of interest, the solution set of the utility maximization problem may not be a singleton. Example 3.3.1 shows a case of great relevance in modern electricity systems.

Example 3.3.1 *Consider a household with an electric vehicle (EV) enrolled in a flat rate with volumetric charge $p \in \mathbb{R}_+$. For simplicity, we consider that the round-trip efficiency of* *the vehicle's battery is* 1*. While implausible, the reader can verify that this assumption does not alter the point we make with this example. The aggregated demand of the household is* $d + s$*, where* d *is the electricity consumption of the household and* s *that of the EV. The battery of the electrical vehicle has a maximum and minimum charge and discharge rates* $R^+ > R^-$*, and a maximum and minimum state of charge* $E^+ > E^-$*. Consistently, the feasible region for* s *is*

$$\mathcal{S} = \left\{ s \in \mathbb{R}^T : \; s0 + \sum_{\tau=1}^{t} s_\tau \in \left[E^-, E^+\right] \wedge s_t \in \left[R^-, R^+\right] \;\; \forall t \right\}, \tag{3.14}$$

with $s0$ *the initial state of charge. The utility maximization problem is*

$$\max_{(d,s)} \left\{ U(d;\theta) + M - p \sum_{t=1}^{T} (d_t + s_t) : \; d \geq 0, \;\; s \in \mathcal{S} \right\} \tag{3.15}$$

Let s^* *be optimal for* (3.15)*, suppose that there is a pair of consecutive periods where* $s_t^* < s_{t+1}^*$*, and define* $\Delta := (s_{t+1}^* - s_t^*) \cdot \psi$*. For any* $\psi \in (0,1)$*, we have that* $s^{**} = (s_1^*, \ldots, s_t^* + \Delta, s_{t+1}^* - \Delta, \ldots, s_T^*)$ *is also optimal for optimal for* (3.15)*.* □

A second problem with the demand representation of the Ramsey problems relates to the calibration of this function. Researchers have estimated price elasticities for electricity demand. One can classify the approaches in two groups. One group focuses on estimating own-price elasticties and peak-to-off-peak elasticities (e.g., Caves et al. 1984b, Howrey and Varian 1984 and Lillard and Aigner 1984). While these have been useful for analysis of time-off-use rate structures, they impose limits in terms of the range of rates that a researcher can analyze. To evaluate rate structures such as a real-time pricing, the approaches used in these papers fall short, as peak-to-off peak substitution is not a relevant concept in the case of RTP. To overcome these limitations other researchers have used techniques that permit to compute a full matrix of elasticities (e.g., Gallant and Koenker (1984) and Taylor et al. 2005). While this approach brings more flexibility to the analysis of rates, it has three limitations. One is that the demand system must be a linear function of the price, which limits the range of utility maximization problems one can consider in an analysis. A second drawback is the lack of transparency when introducing heterogeneity. In principle, one can capture heterogeneity estimating various elasticity matrices. Having these inputs, however, makes it hard to understand the role of fundamentals, such as weather patterns or new technological conditions, on the results of an analysis. A third limitation is range of rates that researchers can model. This framework permit analyzing any variations of a time-varying rate structures. However, other tariffs such as increasing block rates or rate structures supplemented with demand charges cannot be modeled.

A realistic demand model

We believe that a bottom up approach can tackle these limitations. It can simplify and make more transparent the calibration of the demand model. Further, a bottom up approach can enable the analysis of more sophisticated rate structures. We slightly modify the framework of the previous chapter to allow the bottom up modeling of the demand. Instead of considering the demand function as a primitive, we consider as fundamental inputs the elements defining the consumer maximization problem. In the present extension, the demand is simply some optimizer of this problem. The new version of the Ramsey problem follows

$$\max_{(\alpha,\boldsymbol{d},\boldsymbol{p})} \sum_{h=1}^{n} \alpha_h E\left[U(\boldsymbol{d}^h;\boldsymbol{\theta}^h)\right] - C\left(\sum_{h=1}^{n} \alpha_h \Psi_h \boldsymbol{d}^h\right) \tag{3.16}$$

$$\text{subject to}$$

$$\boldsymbol{d}^h \in \underset{\boldsymbol{d}}{arg\,max}\left\{E\left[U(\boldsymbol{d};\boldsymbol{\theta}^h) + M_h - \boldsymbol{d}^\top \Lambda^h \boldsymbol{p}^h\right] : b^h - A^h \boldsymbol{d} \geq 0\right\},\ \forall h \tag{3.17}$$

$$\boldsymbol{p} \in \mathcal{P}, \tag{3.18}$$

$$\alpha \in \mathcal{A}, \tag{3.19}$$

where for simplicity we have added the parameter r_h to the set of parameters $\boldsymbol{\theta}^h$. In this formulation, Λ^h, b^h, A^h permit modeling customers with a wide array of DERs as well as more complex rate structures. The parameter Ψ_h is a demand aggregation matrix. For concreteness, we now provide illustrative examples.

Examples

We start showing how to set the parameters Λ^h, b^h, A^h and Ψ_h to model a customer having DERs and enrolled in a simple flat rate structure. Next, we show how the parameters change to model the same customer under more complex rates. In both examples, we drop the subindex h since we focus on one customer.

Example 3.3.2 *Consider a household with a photovoltaic solar panel (PV), a battery storage systems (BS), and a thermostatically controlled load (TCL)—the latter regulates the temperature inside the customer premises (e.g., an AC system). Define $J = \{TCL, PV, BS, OD\}$ as the of devices the household owns, with OD representing all other devices of this customer. The demand vector of the household is $d = (d_{TCL}, d_{PV}, d_{BS}, d_{OD})$, and the customer's consumption choices are consistent with an additively separable utility function $U(d) = \sum_{j\in J} U_j(d_j)$. Neither the electricity consumption of the PV nor that of the BS produce any benefit for the household so $U_{PV}(d_{PV}) = U_{BS}(d_{BS}) = 0$. We leave the utility function associated to other devices (OD) as generic and concentrate in specifying the one corresponding to the TCL. As appendix B.3 shows, the inside temperature of the household w can be modeled as a linear function of the power consumption of the TCL*

$$w(d_{TCL};\xi,\hat{w}) = W_1(\xi)d_{TCL} + w_2(\xi,\hat{w}), \tag{3.20}$$

where ξ is the set of thermal characteristics of the dwelling and $\hat{w}$ is the outdoor temperature; W_1 and w_2 are a matrix and a vector depending on these parameters. We specify the utility *function associated to the TCL as the negative of the disutility that deviations from a desired indoor temperature, w_{target}, produce on the customer. That is,*

$$U_{TCL}(d_{TCL}) = -\beta \|w(d_{TCL}; \xi, \hat{w}) - w_{target}\|^2, \tag{3.21}$$

where β is a parameter characterizing how the household trades comfort for savings.

Let $\mathcal{S}_j$ be the constraint associated to the end use j. We have already defined S_{BS} in (3.14)*. Since this set is a polyhedron, we can write $\mathcal{S}_{BS} = \{d \in \mathbb{R}^T :\ b_{BS} - A_{BS}d \geq 0\}$. The constraint set for the photovoltaic system is simply $\mathcal{S}_{PV} = \{d \in \mathbb{R}^T :\ d_t \in [0, x\boldsymbol{\rho}_t]\ \forall t\}$, with x and $\boldsymbol{\rho}$ the nameplate capacity and availability factor of the PV, respectively. Again, $\mathcal{S}_{PV}$ is a polyhedron so it can be written as the intersection of halfspaces $b_{PV} - A_{PV}d \geq 0$. The final group of technological constraints corresponds to those imposing power limits one the TCL. Let $\mathcal{S}_{TCL} = \{d \in \mathbb{R}^T :\ d_t \in [0, d_{max}]\ \forall t]\}$; the corresponding inequality is $b_{TCL} - A_{TCL}d_{TCL} \geq 0$. The full set of constraints for the household demand follows*

$$\underbrace{\begin{bmatrix} b_{TCL} \\ b_{PV} \\ b_{BS} \\ z \end{bmatrix}}_{b} - \underbrace{\begin{bmatrix} A_{TCL} & Z & Z & Z \\ Z & A_{PV} & Z & Z \\ Z & Z & A_{BS} & Z \\ Z & Z & Z & -I_d \end{bmatrix}}_{A} \begin{bmatrix} d_{TCL} \\ d_{PV} \\ d_{BS} \\ d_{OD} \end{bmatrix} \geq 0, \tag{3.22}$$

where Z and z are, respectively, a matrix and a vector of zeros of the proper dimensions and I_d is an identity matrix. The corresponding demand aggregation matrix is $\Psi = [I_d\ {-I_d}\ I_d\ I_d]$.

Given that the household is under a flat rate, $\mathcal{P} = \{\boldsymbol{p} \in \mathbb{R}^T :\ p_{\omega t} = p_{\omega' t'}\ \forall(\omega', t')\}$ and $\Lambda = \Psi^\top$. □

In the example that follows, we consider the same household enrolled in a flat rate structure supplemented with a demand charge—a charge per unit of peak consumption.

Example 3.3.3 *Redefine the households demand as follows $d \leftarrow (d, d_{DC})$, where $d_{DC} \in \mathbb{R}$ is the consumption on peak. We also update the parameters of the constraint* (3.22) *assigning*

$$b \leftarrow \begin{bmatrix} b \\ z \end{bmatrix} \quad \text{and} \quad A \leftarrow \begin{bmatrix} A & z \\ \Psi & -e \end{bmatrix}, \tag{3.23}$$

with e a vector of ones of dimension T. The price-demand multiplication matrix $\Lambda \leftarrow [\Lambda\ z;\ z^\top\ 1]$ and the new demand aggregation matrix is now $\Psi \leftarrow [\Psi\ z]$. Finally, the constraint for the prices updates as follows $\mathcal{P} \leftarrow \mathcal{P} \times \mathbb{R}_+$. □

In our last example we also consider as starting point the definition of the parameters and variables in Example 3.3.2. We show how to update such parameters and variables in order to model a customer enrolled in a increasing block (IB) structure.

Example 3.3.4 *In the case of an IB design a customer pays a volumetric charge which differs depending on the level of her total consumption over a certain horizon. Here we consider that the relevant horizon is $\{1, \ldots, T\}$. There are N consumption blocks with upper bounds $\{q_n\}_{n=1}^{N}$, corresponding to the components of the vector $q \in \mathbb{R}_+^N$. In an increasing block structure if the total consumption is within block n, i.e., if it falls in the interval $[q_{n-1}, q_n]$, then the per unit charge is at least as high as that of the block n', for any $n' < n$.*

We start redefining the household demand $d \leftarrow (d, d_{IB})$, where $d_{IB} \in \mathbb{R}^N$ has in its n-th component the total consumption if it falls in the n-th block. The new parameters of the constraints follow

$$b \leftarrow \begin{bmatrix} b \\ 0 \\ q \\ z_N \end{bmatrix} \quad \textit{and} \quad A \leftarrow \begin{bmatrix} A & Z \\ e_T^\top \Psi & -e_N^\top \\ Z' & I_N \\ Z' & -I_N \end{bmatrix}, \tag{3.24}$$

with Z, Z' matrices of zeros of the proper dimensions, z_N an N-dimensional zeros vector, e_T and e_N vectors of ones with T and N components, respectively, and I_N the identity of dimension N. The price-demand multiplication matrix changes to $\Lambda \leftarrow [Z' \;\; I_N]^\top$, and the demand aggregation matrix becomes $\Psi \leftarrow [\Psi \;\; I_N \cdot 0]$. Finally, we redefine the constraint set for the price vector as

$$\mathcal{P} \leftarrow \left\{ p \in \mathbb{R}_+^N : \; p_n \geq p_{n-1} \; \forall n \in \{2, \ldots, N\} \right\}. \tag{3.25}$$

□

3.4 Solution Method

The solution method that we describe here is well suited for the class of problems (3.16)–(3.19) where U is quadratic. We leave the more general case for future research. Even with this simplification, we believe that the development of a solution technique for (3.16)–(3.19) is a valuable endeavor.

The mathematical program (3.16)–(3.19) is an instance of a Bilevel Programming problem. The first problem of this class, introduced by von Stackelberg (1954), modeled the interaction of two firms. The leader firm, which moves first, selects its production quantity knowing that the follower will observe its decision and respond accordingly. That is, in defining its strategy the leader takes into account the reaction of the follower, which in turn depends on the leader's decision. Bilevel optimization problems generalize this setting. A program in this class has an upper level (leader's problem) that has a set of constraints which are the solution set of subordinate (follower) problems.

When a solution set has more than one element, one can take two approaches. One, called pessimistic, assumes that followers do not cooperate with the leader; the other, often referred to as optimistic, assumes the opposite. Under this approach, the leader can select any element of the solution set of a follower. In this paper, we take the optimistic approach.

Because in our setting the subordinate problems (which (3.17) describes) are convex, their first order necessary conditions are also sufficient. Thus, we can use these conditions to express the solution sets of the subordinate problems analytically. In doing so we are effectively casting the Bilevel Problem (3.16)–(3.19) as a Mathematical Program with Equilibrium Constraints (MPEC).

Formulating the Ramsey problem as an MPEC

Let $\boldsymbol{\nu}^h$ be the Lagrange multiplier of the constraint of problem (3.17). For those customers under rate h, the Lagrangean of this problem is

$$L = E\left[U(\boldsymbol{d}^h; \boldsymbol{\theta}^h) + M_h - (\boldsymbol{d}^h)^\top \Lambda^h \boldsymbol{p}^h + (b^h - A^h \boldsymbol{d}^h)^\top \boldsymbol{\nu}^h\right], \tag{3.26}$$

and the first order necessary conditions are

$$\nabla U(d^h_\omega; \theta^h_\omega) - \Lambda^h p^h_\omega - (A^h)^\top \nu^h_\omega = 0, \tag{3.27}$$

$$0 \leq \nu^h_\omega \perp b^h - A^h d^h_\omega \geq 0 = 0 \tag{3.28}$$

for all $\omega \in \Omega$ and $h = 1, \ldots, n$, with the symbol $\perp$ indicating that the vectors must be orthogonal. We reformulate the Bilevel Model (3.16)–(3.19) replacing (3.17) with conditions (3.27)–(3.28). Henceforth we refer to this reformulation as RMPEC. The new problem falls within a class of non-linear programs which are particularly difficult to solve: MPEC. Because these problems fail to satisfy the Mangasarian-Fromovitz constraint qualifications at every feasible point, there are no convergence guarantees for standard non-linear methods. As a consequence, researchers have developed specialized algorithms to find stationary points for these programs (Giallombardo and Ralph, 2008). While these algorithms may work well for small- to middle-sized instances, due to the combinatorial nature of MPECs (Luo et al., 1996, ch. 1), finding an exact solution for large-scale instances becomes impractical. Because the instances that we are interested to tackle are large-scale, in what follows we develop a specialized approximation method to find stationary points for large-scale instances of the RMPEC. Our technique is based upon the Alternating Direction Method of Multipliers (ADMM).

Decomposing the problem

The key idea behind the ADMM algorithm is to combine the decomposability of the dual ascent or dual subgradient methods with the superior convergence of the method of multipliers. The former group of methods aims to find a solution to the original (primal) problem by solving its dual with an iterative procedure. This is often useful when the dual problem has a structure that permits simplifying its resolution. One important example is when the primal has coupling constraints—constraints that link a group of variables together. One can construct a dual relaxing these constraints. By doing so, the computation of a dual step can be decoupled and distributed, improving the scalability of the problem. A dual step is the multiplication of a step direction by a step size. In the case of the dual ascent method the step direction is the gradient of the dual function at the current point. Its computation requires the Lagrangian having a unique optimum at such a point. While this holds in many problems in some important cases, such as for mixed integer programs, the condition does not hold. A researcher can overcome this difficulty with a dual subgradient method. These use subgradients as steps directions. While consecutive iterations may not improve the objective function of the dual problem, provided that the step size is selected properly, the subgradient method reduces the distance between the current dual solution and the dual optimum at each iteration. One problem when using this method is that the computation of the step size is not straightforward. The correct magnitude depends on the optimal value, which is not known. As a result, the performance of this method is highly dependent on the specific structure of the problem at hand, being very effective in some cases and showing very slow convergence or an oscillating behavior in others.

A notable approach that overcomes the limitations of the dual ascent or subgradient methods is the method of multipliers. The crucial idea behind this technique is to strengthen the convexity of the objective function of the original problem to facilitate the search of an optimal solution. One way of achieving this is by adding a term to the objective which penalizes the violation of a set of constraints. Since what is penalized are constraint violations, the original and the problem with the penalty—henceforth the augmented problem—are equivalent. However, under mild conditions the latter's structure permits the application of a dual ascent method, with good global convergence properties. The main drawback of this technique is that the penalty term makes it impossible to decouple and distribute the computation of the dual steps, which hinders the practical use of this approach for large-scale problems.

The alternating direction method of multipliers seeks to combine the best properties of the dual ascent (or subgradient) method and the method of multipliers. It does so by introducing a slight modification to the dual step computation of the associated augmented problem. For concreteness and because we will use it later, we now provide a description of the algorithm ADMM for a class of problems that, as we show in the next subsection, contains the RMPEC.

Let F and G_j be multivariate smooth functions, with $j \in \{0, 1, \ldots, J\}$ and J possibly large. Let $Z = Z_0 \times Z_1 \times \ldots \times Z_J$ be the domain of F, with Z_j a convex set contained in $\mathbb{R}^{m_j}$. The domain of each G_j is $Z_j \times \bar{\mathcal{W}}_j$, also a convex set, and we denote $(z_j, \bar{w}_j)$ a generic element of this set. The range of F is $R_F \subseteq \mathbb{R}$ and that of G_j is $R_{G_j} \subseteq \mathbb{R}^{\iota_j}$, with ι_j some positive integer. The problem of interest is

$$\min_{(z,\bar{w}) \in Z \times \bar{\mathcal{W}}} \{F(z) : \; G_j(z_j, \bar{w}_j) \leq 0 \;\; \forall j = 1, \ldots, J\}, \tag{3.29}$$

where $\bar{\mathcal{W}} = \bar{\mathcal{W}}_1 \times \cdots \times \bar{\mathcal{W}}_J$.

Suppose we need to decouple this problem, for instance, because the constraints that the functions G_j define make the problem hard to solve. To this end, consider the equivalent program

$$\min \; F(z) \tag{3.30}$$

$$\text{subject to}$$

$$G_j(\hat{w}_j, \bar{w}_j) \leq 0 \quad \forall j = 1, \ldots, J, \tag{3.31}$$

$$z_j = \hat{w}_j \quad \forall j = 1, \ldots, J, \tag{3.32}$$

$$(z, \bar{w}) \in Z \times \bar{\mathcal{W}}, \tag{3.33}$$

We could use a dual subgradient method relaxing the constraints (3.32) to decouple this problem. But, as we discussed before, this approach can have poor performance. Instead, we use the ADMM algorithm. As the method of multipliers, this technique also defines an Augmented Lagrangean

$$L_\rho(w, z; \gamma) = F(z) + \sum_{j=1}^{J} \gamma_j^\top (z_j - \hat{w}_j) + \frac{\rho}{2} \|z_j - \hat{w}_j\|^2, \tag{3.34}$$

where $w = (\hat{w}_1^\top, \cdots, \hat{w}_J^\top, \bar{w}_1^\top, \cdots, \bar{w}_J^\top)^\top$, $\gamma = (\gamma_1^\top, \cdots, \gamma_J^\top)^\top$ is the block vector of the dual variables of (3.32) and $\rho > 0$ a penalty. The key difference between ADMM and the method of multipliers is that the former do not minimizes at each iteration $L_\rho(w, z; \gamma)$. Instead, it decreases this function in two steps, a w- and z-minimization step. Before presenting this algorithm, we define the matrices B_1, B_2 such that (3.32) is equivalent to

$$B_1 w + B_2 z = 0. \tag{3.35}$$

Algorithm 2 describes ADMM applied to problem (3.29).

Algorithm 2 Alternating Direction Method of Multipliers

```
Initialization: Given $z^0 \in Z$, $\gamma^0 = 0$, two tolerances $e^{pri} > 0$ and $e^{dual} > 0$, and a primal and a dual residual $r^0, s^0$, such that $\|r^0\| > e^{pri}$ and $\|s^0\| > e^{dual}$
while $(\|r^k\| > e^{pri}) \wedge (\|s^k\| > e^{dual})$ do
    w-minimization: $w^{k+1} \leftarrow arg\,min_w \left\{ L_\rho(w, z^k; \gamma^k) : \; G(\hat{w}_j, \bar{w}_j) \leq 0, \; \bar{w}_j \in \bar{\mathcal{W}}_j, \; \forall j \right\}$
    z-minimization: $z^{k+1} \leftarrow arg\,min_z \left\{ L_\rho(w^{k+1}, z; \gamma^k) : \; z \in Z \right\}$
    $\gamma^{k+1} \leftarrow \gamma^k + \rho(B_1 w^{k+1} + B_2 z^{k+1})$
    $r^{k+1} \leftarrow B_1 w^{k+1} + B_2 z^{k+1}$
    $s^{k+1} \leftarrow \rho B_1^\top B_2 (z^{k+1} - z^k)$
```

The *w-minimization* step can be distributed noting that if z is fixed then the Augmented Lagrangian can be decoupled. Thus, another way of obtaining w^{k+1} is by assigning to

$$(\hat{w}_j^{k+1}, \bar{w}_j^{k+1}) \leftarrow \underset{(\hat{w}_j, \bar{w}_j)}{arg\,min} \left\{ -\gamma_j^k \cdot \hat{w}_j + \frac{\rho}{2} \|z_j^k - \hat{w}_j\|^2 : \; G(\hat{w}_j, \bar{w}_j) \leq 0, \; \bar{w}_j \in \bar{\mathcal{W}}_j \right\} \tag{3.36}$$

for every $j = 1, \ldots, J$.

It is possible to show that when the functions F, G_j are convex (possibly nonsmooth) and if strong duality holds then the algorithm converges to a global optimum (Boyd et al., 2011). Even if F is nonconvex but smooth, it can still converge to stationary solutions (Hong et al., 2016). When the constraints $G(\hat{w}_j, \bar{w}_j) \leq 0$ define nonconvex regions then one can use the algorithm as a heuristic. Takapoui et al. (2017) show that for problems with quadratic objectives and nonconvex separable constraints, the algorithm can rapidly converge to approximated solutions, and in many cases to the global optimum.

A closely related method, extensively used in large-scale stochastic optimization, is Progressive Hedging (PH). As the alternating direction method of multipliers, this algorithm—introduced by Rockafellar and Wets (1991)—is a specialization of the Proximal Point algorithm (Rockafellar and Wets, 1991). While PH converges to stationary points for a large class of stochastic optimization problems, for models involving discrete variables the method becomes a heuristic. In practice, it has proven to be a very effective technique to tackle large-scale, stochastic mixed integer programs (see, e.g., Løkketangen and Woodruff 1996; Listes and Dekker 2005), and more recently to solve stochastic MPECs (see, e.g., Fan and Liu 2010; Guo and Fan 2016).

We propose using using ADMM as a heuristic for the particular stochastic MPEC that we introduce in this paper. The experience of Fan and Liu (2010), Guo and Fan (2016) and Takapoui et al. (2017) suggests that this is a suitable approach, and Subsection 3.4 provides additional evidence.

Implementing ADMM

Before showing how we implement the alternating direction method of multipliers, we make some clarifications that ease the understanding of our approach. Henceforth we treat a random vector $\boldsymbol{\xi}$ also as a block vector, i.e., $\boldsymbol{\xi} = (\xi_1^\top, \ldots, \xi_{|\Omega|}^\top)^\top$ is an alternative representation. If $\boldsymbol{\xi}$ is a block vector of random vectors, that is $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top)^\top$, another way of expressing this object is $\boldsymbol{\xi} = (\xi_{11}^\top, \ldots, \xi_{|\Omega|1}^\top, \ldots, \xi_{1n}^\top, \ldots, \xi_{|\Omega|n}^\top)^\top$.

We start our description showing that the problem RMPEC is an instance of (3.29). To see this, define $z \coloneqq (\alpha^\top, \boldsymbol{d}^\top, \boldsymbol{p}^\top)^\top$, the constraint set $Z \coloneqq \mathcal{A} \times \mathcal{D} \times \mathcal{P}$, where $\mathcal{D} = \mathbb{R}^{|\Omega|\kappa_1^1} \times \cdots \times \mathbb{R}^{|\Omega|\kappa_n^1}$ and κ_h^1 is the number of columns of A_h. In addition, define the functions

$$F(z) \coloneqq C\left(\sum_{h=1}^{n} \alpha_h \Psi_h \boldsymbol{d}^h\right) - \sum_{h=1}^{n} \alpha_h E\left[U(\boldsymbol{d}^h; \boldsymbol{\theta}^h)\right], \tag{3.37}$$

and

$$G_{\omega h}(z_{\omega h}, \bar{w}_{\omega h}) \coloneqq \begin{bmatrix} \nabla U(d_\omega^h; \theta_\omega^h) - \Lambda^h p_\omega^h - (A^h)^\top \nu_\omega^h \\ -\nabla U(d_\omega^h; \theta_\omega^h) + \Lambda^h p_\omega^h + (A^h)^\top \nu_\omega^h \\ (b^h - A^h d_\omega^h)^\top \nu_\omega^h \\ -(b^h - A^h d_\omega^h)^\top \nu_\omega^h \\ -b^h + A^h d_\omega^h \end{bmatrix} \quad \forall \omega \in \Omega, \;\; h = 1, \ldots, n, \tag{3.38}$$

where $z_{\omega h} := (d^h_\omega, p^h_\omega)$, $\bar{w}_{\omega h} := \nu^h_\omega$, and $\bar{\mathcal{W}}_{\omega h} = \mathbb{R}^{\kappa^2_h}_+$, with κ^2_h the number of rows of A_h.

In order to decouple the problem, we duplicate the vectors $z_{\omega h}$ introducing the variables $\hat{w}_{\omega h} = (\hat{d}_{\omega h}, \hat{p}_{\omega h})$ and the coupling constraints

$$\alpha_h(d^h_\omega - \hat{d}^h_\omega) = 0 \;\; \forall \omega \in \Omega, \;\; h = 1, \ldots, n \;\; \text{and} \tag{3.39}$$

$$p^h_\omega - \hat{p}^h_\omega = 0 \;\; \forall \omega \in \Omega, \;\; h = 1. \ldots, n, \tag{3.40}$$

Because we will use it later, we now define $Q(w, z)$ as a block vector function that has in each block (ω, h) the left hand side of (3.39) and (3.40). In addition, we call SMPEC the stochastic MPEC that results from replacing (3.32) with (3.39) and (3.40) in (3.30)–(3.33). Note that, while (3.32) defines a set of linear constraints, those in (3.39) are nonlinear. This does not alter the structure of Algorithm 2, however, it does changes the update of s^{k+1}. We will describe this update in detail later in this subsection. In what follows, we concentrate in the optimization steps.

The Augmented Lagrangean of the reformulated problem is

$$\begin{aligned} L_\rho(\boldsymbol{w}, z; \boldsymbol{\gamma}) = &- \sum_{h=1}^{n} \alpha_h E \left[U(\boldsymbol{d}^h; \boldsymbol{\theta}^h) - (\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h)^\top \boldsymbol{\gamma}^h_d - \frac{\rho}{2} \alpha_h \|\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h\|^2 \right] \\ &+ \sum_{h=1}^{n} E \left[(\boldsymbol{p}^h - \hat{\boldsymbol{p}}^h)^\top \boldsymbol{\gamma}^h_p + \frac{\rho}{2} \|\boldsymbol{p}^h - \hat{\boldsymbol{p}}^h\|^2 \right] + C \left(\sum_{h=1}^{n} \alpha_h \Psi_h \boldsymbol{d}^h \right), \end{aligned} \tag{3.41}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^h_d, \boldsymbol{\gamma}^h_p)$ corresponds to the block vector of Lagrange multipliers of the constraints (3.39) and (3.40), and $\boldsymbol{w} := (\hat{\boldsymbol{d}}^\top, \hat{\boldsymbol{p}}^\top, \boldsymbol{\nu}^\top)^\top$.

The $\boldsymbol{w}$-minimization step

In order to update w^{k+1}, we solve for every $\omega \in \Omega$ and $h \in \{1, \cdots, H\}$ the MPEC that follows

$$\min_{(\hat{d}_{\omega h}, \hat{p}_{\omega h}, \nu_{\omega h})} \alpha_h \left[(d^h_\omega - \hat{d}^h_\omega)^\top \gamma^h_{\omega d} + \frac{\rho}{2} \alpha_h \|d^h_\omega - \hat{d}^h_\omega\|^2 \right] + (p^h_\omega - \hat{p}^h_\omega)^\top \gamma^h_{\omega p} + \frac{\rho}{2} \|p^h_\omega - \hat{p}^h_\omega\|^2 \tag{3.42}$$

$$\text{subject to (3.27) and (3.28),} \tag{3.43}$$

and assign its solution to $(\hat{d}^{k+1}_{\omega h}, \hat{p}^{k+1}_{\omega h}, \nu^{k+1}_{\omega h})$.

There are various approaches to solve this optimization problem. We refer the reader to Giallombardo and Ralph (2008) for a comprehensive review. Here we describe the approach we take, which casts the MPEC as a Mixed-Integer Quadratic Program (MIQP). The technique, proposed by Fortuny-Amat and McCarl (1981), introduces integer variables and new constraints to reformulate the complementarity conditions. Letting $\sigma^h_\omega \in \{0, 1\}^{\kappa^2_h}$, $\bar{M}^h_\omega, \tilde{M}^h_\omega > 0$ scalars, and e a vector of ones of dimension κ^2_h, in our setting the procedure involves replacing (3.28) with the following constraints

$$\bar{M}^h_\omega \sigma^h_\omega \geq b^h - A^h d^h_\omega, \tag{3.44}$$

$$\tilde{M}^h_\omega (e - \sigma^h_\omega) \geq \nu^h_\omega, \tag{3.45}$$

$$b^h - A^h d^h_\omega \geq 0, \tag{3.46}$$

$$\nu^h_\omega \geq 0, \tag{3.47}$$

$$\sigma^h_\omega \in \{0,1\}^{\kappa_h^2}. \tag{3.48}$$

Researchers can solve the reformulated problem (3.42), (3.27), (3.44)–(3.48), which we refer to as w-MPEC, using standard MIQP techniques.

In principle, the approach of Fortuny-Amat and McCarl (1981) can be inaccurate and inefficient. Inaccurate because if $\bar{M}^h_\omega, \tilde{M}^h_\omega$ are too small, $b^h - A^h d^h_\omega$ or ν^h_ω may be far from the optimum of (3.42)–(3.43). On the other hand, increasing these scalars too much can lead to inefficiencies because the problem could be ill-conditioned. Besides, even if they are selected properly, since MIQP is in NP-complete (Pia et al., 2017), it can take a long time for a solver to reach a solution of acceptable quality. In practice, for the class of problems that we focus on, non of these concerns posed significant difficulties to use this approach. Based on the physical nature of the constraints of the customer problem (3.46), one can find adequate values for $\bar{M}^h_\omega$ and, with some additional work, for $\tilde{M}^h_\omega$; and while the complexity of the problem translated into slow performance, we were able to overcome this issue by providing the MIQP solver with good starting points.

A good starting point decreases the solution time of w-MPEC by pruning several sub-optimal branches of the Branch and Bound tree, and by providing an adequate range for $\tilde{M}^h_\omega$. The procedure we used to construct a good starting point exploits that, in general, solutions across iterations do not differ as much, and that depending on the magnitude of α_h the first two terms of (3.42) might be irrelevant. Then, one way of constructing a starting point is using the price of the previous iteration and solve the customer problem to find the associated demand and dual variables. In addition, assuming the first two terms of (3.42) are irrelevant (because α_h is relatively small) then, one can construct another starting point finding $(\hat{p}^h_\omega)^* = \mathrm{a}rg\,min_{\hat{p}^h_\omega} \left\{(p^h_\omega - \hat{p}^h_\omega)^\top \gamma^h_{\omega p} + \frac{\rho}{2}\|p^h_\omega - \hat{p}^h_\omega\|^2\right\}$ and solving the customer problem for this price. A third technique is to simply use $\hat{p}^h_\omega$ as input to the customer problem. Then, the starting point is the one with the smallest objective value for w-MPEC.

The z-minimization step

Note that the z-minimization problem can be separated into two subproblems. One that finds the optimal $(\alpha, \boldsymbol{d})$ and another that optimizes $\boldsymbol{p}$. This is possible because there are no constraints coupling these block of variables. The problem that optimizes $\boldsymbol{p}$ is

$$\min_{\boldsymbol{p}} \left\{ \sum_{h=1}^{n} E\left[(\boldsymbol{p}^h - \hat{\boldsymbol{p}}^h)^\top \boldsymbol{\gamma}^h_p + \frac{\rho}{2}\|\boldsymbol{p}^h - \hat{\boldsymbol{p}}^h\|^2\right] : \boldsymbol{p} \in \mathcal{P} \right\}, \tag{3.49}$$

which depending on the structure of $\mathcal{P}$ (a convex set), can be solved either analytically or using a standard convex optimization solver. To find the optimal $(\alpha, \boldsymbol{d})$, we solve

$$\min_{(\alpha,\boldsymbol{d})} \left\{ C\left(\sum_{h=1}^{n} \alpha_h \Psi_h \boldsymbol{d}^h\right) - \sum_{h=1}^{n} \alpha_h E\left[U(\boldsymbol{d}^h;\boldsymbol{\theta}^h) - (\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h)^\top \boldsymbol{\gamma}_d^h - \frac{\rho}{2}\alpha_h \|\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h\|^2\right] : \ \alpha \in \mathcal{A} \right\}. \tag{3.50}$$

We could use a standard nonlinear solver to handle this problem. However, specially for problems of large size, this approach is less attractive than using specialized convex optimization algorithms. We are able to take this approach by casting the problem (3.50) as a conic program. In order to do so, we first rewrite the term inside the expectation in the objective function as follows

$$U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h) + (\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h)^\top(\nabla U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h) - \boldsymbol{\gamma}_d^h) + \frac{1}{2}(\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h)^\top \left[\nabla^2 U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h) - \rho\alpha_h I_h\right](\boldsymbol{d}^h - \hat{\boldsymbol{d}}^h), \tag{3.51}$$

where we replaced $U(\boldsymbol{d}^h;\boldsymbol{\theta}^h)$ by its taylor expansion about $\hat{\boldsymbol{d}}^h$, and I_h is an identity of dimension κ_h^1. Next, we introduce variables $\tilde{\boldsymbol{d}}^h = \alpha_h \boldsymbol{d}^h$, $\boldsymbol{v}_0^h$, $\boldsymbol{v}_1^h$, and write the following reformulation of the problem

$$\min_{(\alpha,\tilde{\boldsymbol{d}})} C\left(\sum_{h=1}^{n} \Psi_h \tilde{\boldsymbol{d}}^h\right) - \sum_{h=1}^{n} E\left[\alpha_h U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h) + (\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h)^\top(\nabla U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h) - \boldsymbol{\gamma}_d^h) - \boldsymbol{v}_0^h - \boldsymbol{v}_1^h\right] \tag{3.52}$$

subject to

$$2\alpha_h \boldsymbol{v}_0^h \geq (\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h)^\top \left[-\nabla^2 U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h)\right](\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h), \ \forall h = 1,\ldots,n \tag{3.53}$$

$$2\boldsymbol{v}_1^h \geq \rho(\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h)^\top(\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h), \ \forall h = 1,\ldots,n \tag{3.54}$$

$$\alpha \in \mathcal{A}. \tag{3.55}$$

To obtain a conic program, we introduce the variables $\boldsymbol{\phi}_0^h$, $\boldsymbol{\phi}_1^h$, $\boldsymbol{v}_2^h$, replace (3.53) with the conditions

$$2\alpha_h \boldsymbol{v}_0^h \geq (\boldsymbol{\phi}_0^h)^\top \boldsymbol{\phi}_0^h, \ \forall h = 1,\ldots,n \tag{3.56}$$

$$\boldsymbol{\phi}_0^h = \left[-\nabla^2 U(\hat{\boldsymbol{d}}^h;\boldsymbol{\theta}^h)\right]^{\frac{1}{2}}(\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h), \ \forall h = 1,\ldots,n, \tag{3.57}$$

and (3.54) with

$$2\boldsymbol{v}_1^h \boldsymbol{v}_2^h \geq (\boldsymbol{\phi}_1^h)^\top \boldsymbol{\phi}_1^h, \ \forall h = 1,\ldots,n \tag{3.58}$$

$$\boldsymbol{\phi}_1^h = \rho^{\frac{1}{2}}(\tilde{\boldsymbol{d}}^h - \alpha_h\hat{\boldsymbol{d}}^h), \ \forall h = 1,\ldots,n, \tag{3.59}$$

$$\boldsymbol{v}_2^h = 1, \ \forall h = 1,\ldots,n. \tag{3.60}$$

While constraints (3.56) and (3.58) define two rotated quadratic cones, the others are simply linear constraints. The problem that (3.52), (3.56)–(3.60) and (3.55) define is a conic program. Researchers can solve it using an off-the-shelf conic optimization package.

Updates

After the two optimization steps Algorithm 2 updates the dual variable γ^{k+1}, and residuals r^{k+1} and s^{k+1}. We now show how we adapt these steps for SMPEC. The dual variable and (primal) residual r^{k+1} update in a similar fashion. That is,

$$\gamma^{k+1} \leftarrow \gamma^{k+1} + \rho Q(w^{k+1}, z^{k+1}), \tag{3.61}$$

$$r^{k+1} \leftarrow Q(w^{k+1}, z^{k+1}). \tag{3.62}$$

However, the update of the (dual) residual s^{k+1} is different. In order to derive an update rule for this vector, we follow a line of reasoning similar to that used in the standard ADMM algorithm. In this technique, if the norms of r^{k+1} and s^{k+1} are small then, provided that the problem is convex, the algorithm converged to an optimal solution (Boyd et al., 2011). We now discuss why this is the case considering (3.30)–(3.33), and then show how we can adapt this reasoning to construct a residual update for SMPEC.

Let $\mathcal{X}$ be a convex set. We denote $\mathbb{I}_{\mathcal{X}}(x)$ the indicator function, which is 0 when $x \in \mathcal{X}$ and ∞ otherwise. In addition, define the set $\mathcal{W} \coloneqq \big\{(\hat{w}, \bar{w}) : \; G_j(\hat{w}_j, \bar{w}_j) \leq 0, \; \forall j, \; \bar{w} \in \bar{\mathcal{W}}\big\}$. The Lagrangian of (3.30)–(3.33) is

$$L(w, z; \mu, \gamma) = F(z) + \gamma^\top (B_1 w + B_2 z) + \mathbb{I}_{\mathcal{W}}(w) + \mathbb{I}_Z(z). \tag{3.63}$$

Assuming that the Slater constraint qualification holds then, the necessary and sufficient condition for (w^*, z^*) to be optimal is $0 \in \partial_{(z,w)} L(w^*, z^*; \mu, \gamma)$, where $\partial_{(z,w)}$ denotes the subdifferential with respect to (z, w) (Schirotzek, 2007, Theorem 5.2.1). Using subdifferential calculus rules (e.g., Rockafellar and Wets 2009, ch. 10), we can write this condition as follows

$$0 \in B_1^\top \gamma + N_{\mathcal{W}}(w), \tag{3.64}$$

$$0 \in \nabla F(z) + B_2^\top \gamma + N_Z(z), \tag{3.65}$$

with $N_{\mathcal{X}}(x)$ denoting the normal cone of $\mathcal{X}$ at x.

On the other hand, we have that the *z-minimization* step of Algorithm 2 is such that z^{k+1} satisfies

$$\begin{aligned} 0 &\in \partial_\omega L_\rho(w^{k+1}, z^{k+1}; \gamma^k) + N_Z(z^{k+1}) && (3.66)\\ &= \nabla F(z^{k+1}) + B_2^\top \gamma^k + \rho B_2^\top r^{k+1} + N_Z(z^{k+1}) && (3.67)\\ &= \nabla F(z^{k+1}) + B_2^\top \gamma^{k+1} + N_Z(z^{k+1}), && (3.68) \end{aligned}$$

where the last equality follows from the update rule for γ. In other words, the points that Algorithm 2 generates always satisfy (3.65). The same does not hold for condition (3.64). Indeed, w^{k+1} satisfies

$$\begin{aligned} 0 &\in \partial_\omega L_\rho(w^{k+1}, z^k; \gamma^k) + N_{\mathcal{W}}(w^{k+1}) && (3.69)\\ &= B_1^\top \gamma^k + \rho B_1^\top (B_1 w^{k+1} + B_2 z^k) + N_{\mathcal{W}}(w^{k+1}) && (3.70)\\ &= B_1^\top \gamma^k + \rho B_1^\top r^{k+1} + N_{\mathcal{W}}(w^{k+1}) + \rho B_1^\top B_2(z^k - z^{k+1}), && (3.71)\\ &= B_2^\top \gamma^{k+1} + N_{\mathcal{W}}(w^{k+1}) + s^{k+1}, && (3.72) \end{aligned}$$

which differs from (3.64). However, if s^{k+1} vanishes at some iteration $k+1$ then, (3.72) becomes (3.64). If in addition, $r^{k+1} = 0$ then (w^{k+1}, z^{k+1}) is optimal for (3.30)–(3.33).

We can follow a similar strategy to define the update of s^k in the context of SMPEC. However, we need to address two properties that distinguish this problem from (3.30)–(3.33). One is that in our setting the coupling constraints (3.39) and (3.40) are nonlinear, the other that $G_{\omega h}(\hat{w}_{\omega h}, \bar{w}_{\omega h})$ are nonconvex functions. Because of these, we cannot use techniques from convex analysis in order to derive an update rule for s^k. However, if F, $G_{\omega h}$ and Q were locally Lipchitz-continuous,[3] and the sets Z and $\mathcal{W}$ were closed, we could leverage classic results from nonsmooth analysis. It easy to see that Z and $\mathcal{W}$ are closed sets. We now show that F, $G_{\omega h}$ and Q are locally Lipchitz-continuous functions.

First, note that convex or smooth functions are locally Lipchitz-continuous (Clarke, 1990, p. 228), and that the summation and composition preserve this property. The function $F(z)$, defined in (3.37), is locally Lipchitz-continuous because it is the summation of two locally Lipchitz-continuous functions. The first term in the right hand side of (3.37) has this property since it is the composition of a convex with a smooth function, both locally Lipchitz-continuous; the second term is a smooth function of $(\alpha, \boldsymbol{d})$, and thus locally Lipchitz-continuous. In addition, $G_{\omega h}(\hat{w}_{\omega h}, \bar{w}_{\omega h})$ are smooth vector functions since all their components are smooth, which is also the case for $Q(w, z)$.

To write in a simple manner necessary conditions for optimality, we first introduce the block diagonal matrices $\amalg_d$ and $\amalg_p$ such that the block (ω, h) is equal to $q_\omega I_{\kappa_h^1}$ for the former and $q_\omega I_T$ for the latter, with q_ω the probability that ω occurs. In addition, we define the two-block diagonal matrix $\amalg$, having in its first block $\amalg_d$ and $\amalg_p$ in the second. It is direct to verify that for any pair of vectors, $(\boldsymbol{d}', \boldsymbol{p}')$ and $(\boldsymbol{d}'', \boldsymbol{p}'')$, with the dimensions of $(\boldsymbol{d}, \boldsymbol{p})$, we have

$$(\boldsymbol{d}', \boldsymbol{p}')^\top \amalg (\boldsymbol{d}'', \boldsymbol{p}'') = E\left[\sum_{h=1}^{n} (\boldsymbol{d}'_h)^\top \boldsymbol{d}''_h + (\boldsymbol{p}'_h)^\top \boldsymbol{p}''_h\right] \tag{3.73}$$

We now define the following Lagrangian for the problem SMPEC

$$L(w, z; \mu, \gamma, \eta) = \mu F(z) + \gamma^\top \amalg Q(w, z) + \eta \|(\mu_1, \mu_2)\| d_{\mathcal{W}\times Z}(w, z), \tag{3.74}$$

where η is some positive scalar and $d_{\mathcal{X}}(x) = \inf\{\|x - x'\| : \; x' \in \mathcal{X}\}$. Then, in virtue of Theorem 6.1.1 in Clarke (1990), also known as the Lagrange multiplier rule, we have that if (w^*, z^*) solves globally or locally SMPEC

$$\exists \mu \geq 0 \text{ and } \gamma \text{ not all zero, such that } 0 \in \partial_{(w,z)} L(w^*, z^*; \mu, \gamma, \eta), \tag{3.75}$$

for every $\eta \geq \hat{\eta}$, with $\hat{\eta}$ the Lipchitz constant of the function $[F, Q]$ in a neighborhood of (w^*, z^*), and $\partial_{(w,z)}$ denoting the Clarke subdifferential (see Definition 7.3.4 in Schirotzek 2007). Using calculus rules for this type of subdifferentials (see, e.g., Clarke 1990, ch. 2), we

[3] For a definition of local Lipchitz-continuity, we refer the reader to (Clarke, 1990, p. 25).

can rewrite the expression $0 \in \partial_{(w,z)} L(w, z; \mu, \gamma, k)$ as follows

$$0 \in \mu \nabla F(z^*) + J_z Q(w^*, z^*)^\top \Pi \gamma + N_Z(z^*) \tag{3.76}$$

$$0 \in J_w Q(w^*, z^*)^\top \Pi \gamma + N_{\mathcal{W}}(w^*), \tag{3.77}$$

with J the standard Jacobian, and $N_{\mathcal{X}}(x)$ denoting the normal cone of $\mathcal{X}$ at x, defined in Clarke (1990, p. 51).

As in the convex case, we now analyze what first order conditions w^{k+1} and z^{k+1} satisfy. Before, we note that one can write the augmented Lagrangean of SMPEC, defined in (3.41), as follows

$$L_\rho(w, z; \gamma) = F(z) + \gamma^\top \Pi Q(w, z) + \frac{\rho}{2} Q(w, z)^\top \Pi Q(w, z). \tag{3.78}$$

As we showed before, the problem of the *z-minimization* step is convex. Thus, we have z^{k+1} satisfies

$$0 \in \nabla F(z^{k+1}) + J_z Q(w^{k+1}, z^{k+1})^\top \Pi \gamma^k + \rho J_z Q(w^{k+1}, z^{k+1})^\top \Pi Q(w^{k+1}, z^{k+1}) + N_Z(z^{k+1}) \tag{3.79}$$

$$= \nabla F(z^{k+1}) + J_z Q(w^{k+1}, z^{k+1})^\top \Pi \gamma^{k+1} + N_Z(z^{k+1}), \tag{3.80}$$

which is exactly (3.76), with $\mu = 1$. Before, analyzing the case of w^{k+1}, we introduce the following notation $Q_{k,k} \coloneqq Q(w^k, z^k)$. In this case, using the Lagrange multiplier rule, we have that w^{k+1} satisfies

$$0 \in J_w Q_{k+1,k}^\top \Pi \gamma^k + \rho J_w Q_{k+1,k}^\top \Pi Q_{k+1,k} + N_{\mathcal{W}}(w^{k+1}) \tag{3.81}$$

$$= J_z Q_{k+1,k+1}^\top \Pi \gamma^{k+1} + N_W(w^{k+1}) + s^{k+1}, \tag{3.82}$$

with

$$s^{k+1} = J_w Q_{k+1,k}^\top \Pi [\gamma^k + \rho Q_{k+1,k}] - J_w Q_{k+1,k+1}^\top \Pi \gamma^{k+1}. \tag{3.83}$$

In (3.81) we use that the Linear Independence constraint qualification for MPECs (MPEC LICQ) holds for the problem of the *w-minimization* step. In virtue of Theorem 2 in Scheel and Stefan (2000), and the fact that MPEC LICQ implies the MPEC Mangasarian-Fromovitz constraint qualification,[4] we have that the multiplier of the objective function of the *w-minimization* problem is equal to 1.

The definition of dual residual in (3.83) is such that that our variation of the ADMM algorithm applied to the SMPEC stops at stationary point for this problem. While we do not prove convergence, Subsection 3.4 shows that the algorithm converges in most instances.

To gain a some intuition on when this residual becomes zero, we consider the case when α does not change between consecutive iterations. Before doing so, we note that

$$Q(w, z) = H(\alpha) \begin{bmatrix} \boldsymbol{d} - \hat{\boldsymbol{d}} \\ \boldsymbol{p} - \hat{\boldsymbol{p}} \end{bmatrix} = H(\alpha)\Delta, \tag{3.84}$$

[4] See Ye (2005) for various definitions of constraint qualifications, including MPEC LICQ and MPEC Magasarian–Fromovitz CQ.

where $H(\alpha)$ is a block diagonal matrix such that the block corresponding to $\boldsymbol{d}_h$ is equal to $\alpha_h I_{k_h^1}$, and that corresponding to $\boldsymbol{p}$ is an identity matrix. In addition, we observe that (3.83) can be expressed as follows

$$\begin{aligned} s^{k+1} &= J_w Q_{k+1,k}^\top \amalg[\gamma^{k+1} + \rho(Q_{k+1,k} - Q_{k+1,k+1})] - J_w Q_{k+1,k+1}^\top \amalg \gamma^{k+1} && (3.85) \\ &= [J_w Q_{k+1,k} - J_w Q_{k+1,k+1}]^\top \amalg \gamma^{k+1} + \rho J_w Q_{k+1,k}^\top \amalg [Q_{k+1,k} - Q_{k+1,k+1}], && (3.86) \end{aligned}$$

which in view of (3.84) is equal to

$$H(\alpha^{k+1} - \alpha^k)\amalg\gamma^{k+1} - \rho H(\alpha^k)\amalg[H(\alpha^k)\Delta_{k+1,k} - H(\alpha^{k+1})\Delta_{k+1,k+1}]. \qquad (3.87)$$

If α does not vary between k and $k+1$

$$\begin{aligned} s^{k+1} &= -\rho H(\alpha^k)\amalg H(\alpha^k)\begin{bmatrix}\boldsymbol{d}^k - \boldsymbol{d}^{k+1} \\ \boldsymbol{p}^k - \boldsymbol{p}^{k+1}\end{bmatrix} && (3.88) \\ &= -\rho H(\alpha^k)\amalg\begin{bmatrix}\tilde{\boldsymbol{d}}^k - \tilde{\boldsymbol{d}}^{k+1} \\ \boldsymbol{p}^k - \boldsymbol{p}^{k+1}\end{bmatrix} && (3.89) \\ &= -\rho E\left[\sum_{h=1}^{n} \alpha_h^k(\tilde{\boldsymbol{d}}_h^k - \tilde{\boldsymbol{d}}_h^{k+1}) + (\boldsymbol{p}_h^k - \boldsymbol{p}_h^{k+1})\right], && (3.90) \end{aligned}$$

a residual akin to that of Algorithm 2.

Testing the performance of the approach

We now focus on the computational performance of the method. In order to explore this, we work with moderately sized instances which we describe in detail in the next section. Here we concentrate on the convergence properties and the resolution times of the algorithm we propose for various sizes and parameter configurations. We change the size of the instances by varying the number of scenarios (or states of nature); the number of variables and complementarity constraints are affine transformations of this parameter. We consider 5 instance sizes, corresponding to 2, 3, 10, 25 and 50 scenarios, 2462, 3686, 12 254, 30 614 and 61 214 variables, or 1200, 1800, 6000, 15 000 and 30 000 complementarity constraints. For a given instance size, we run the algorithm for 40 different parameter configurations which corresponds to combinations of different rate structures, network constraints, and costs for the intermittent and distributed technologies.

We ran our experiment in an ASUS M51AC PC, with a 4 core processor Intel i7-4770 of 3.40 GHz, 16 GB of RAM, and a 64-bit Windows operating system. Of a total of 200 instances, we considered convergent the instances that meet the stopping criterion before 400 iterations. These corresponded to 188 or 94% of the total. The results we report in this subsection are those associated to the convergent instances. For convergence, we required the primal residual to be lower than or equal to 0.25% and a dual residual being lower than or equal to 2.5%. Intuitively, this criterion places more emphasis on the feasibility than the stationarity of the solution.

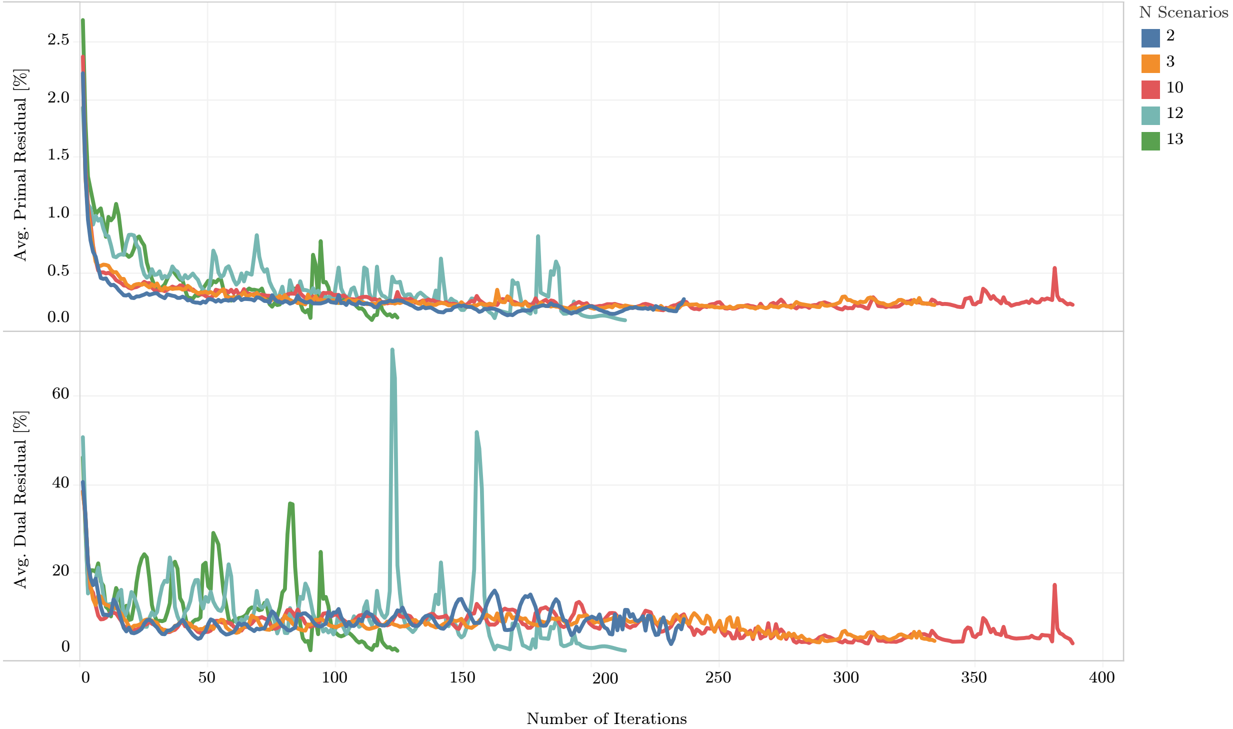

Figure 3.1: Primal and dual residual evolution across iterations.

Before discussing the results, we note that in order to have a benchmark we attempted to run this experiment with an alternative algorithm. Specifically, we tried solving RMPEC with Knitro, a nonlinear optimization package which has specialized routines to solve MPECs (see Byrd et al. 2006). We could not obtain results within 24 hours, even for the smallest instances we considered in this test. In contrast, with the same machine and using the technique we present in this paper, we were able to obtain results within a few hours for the majority of the instances.

Figure 3.1 shows the evolution of the average primal and dual residuals across iterations, for the instance sizes we tested. For most instances, after a few tens of iterations the algorithm finds a solution of reasonable quality, with small primal and dual residuals. The iterations that follow, while improving the solution quality, do this in a relatively slow fashion when compared with the improvement achieved in the first 50 iterations. Interestingly, this convergence behavior seems not to depend on the size, at least for the instances with 10 or more scenarios. We see a different behavior for the instances with 2 and 3 scenarios. While all instances of these sizes converged, the convergence metrics oscillate much more across iterations than the other instances. We only observe this oscillating behavior in the resolution of instances where the transmission is constrained.

In order to have an idea on the potential scalability of the algorithm, we plot in the Panel A of Figure 3.2 the resolution and average iteration times, and number of iterations for different instances sizes. We note that while the graph shows that there is an increase in total resolution times, the median is always below 2000 seconds. In fact, for more than 75% of the instances that converged, the algorithm found a stationary point in less than one hour. This result stands in sharp contrast to the more than 24 hours it took to the off–the–shelf solver to find a solution for instances with just 2 scenarios.

In addition, Panel A shows that the main driver of the rise in resolution times as the size of the instances increases is the average iteration time. As the number of states of nature rises, the distribution of the average time per iteration shifts upwards. This does not happen for the number of iterations. While its variance increases for instances with larger sizes, its median is always below 75. This suggests that if the time per iteration does not grows exponentially with the size of the problem, then it is unlikely that the total resolution time will grow at an exponential rate.

The plots in Panel B provide additional evidence supporting that this is the case. These plots graph the logarithm of the number scenarios versus the logarithm of the average iteration times and the number of iterations. The upper plot shows that a line of slope 0.94 and intercept −0.3 fits well the data. That is, for every additional scenario the average time per iteration increases on average by about a half of a second. On the other hand, the number of iterations increases in proportion to the square root of the number of scenarios.

These results highlight the potential of the algorithm to scale. The increase in the average time per iteration can be handled with more computing nodes, keeping approximately constant the duration of the w-minimization step. While the resolution time of the z-minimization step increases, it does so at a slower rate than linear.

3.5 A Simple Application

We conduct an analysis of rate structures using the method we develop in this paper. Our objective here is to show with an exercise the type of analysis a researcher can conduct using the technique. To keep the analysis simple, we perform this exercise with small sized instances, all with 3 representative scenarios. We explain bellow how we select this scenarios.

Designing the analysis

In this analysis, we compare rate structures that have been proposed as possible alternatives for future electricity systems. We explore five tariffs, ranging from the simplest one, a flat-rate (FR), to the most complex, a real-time price (RTP). In addition, we study how the results change when changing the parameters of the transmission system, and the economics of the renewable generating technologies and distributed energy resources. Figure 3.3 describes the structure of the analysis.

As Figure 3.4 depicts, the network we consider has five buses, two load and three generation buses. In the base case scenario all branches have unlimited capacity; we change only the capacity of branch 1–4 to introduce congestion. In this network, there are five different

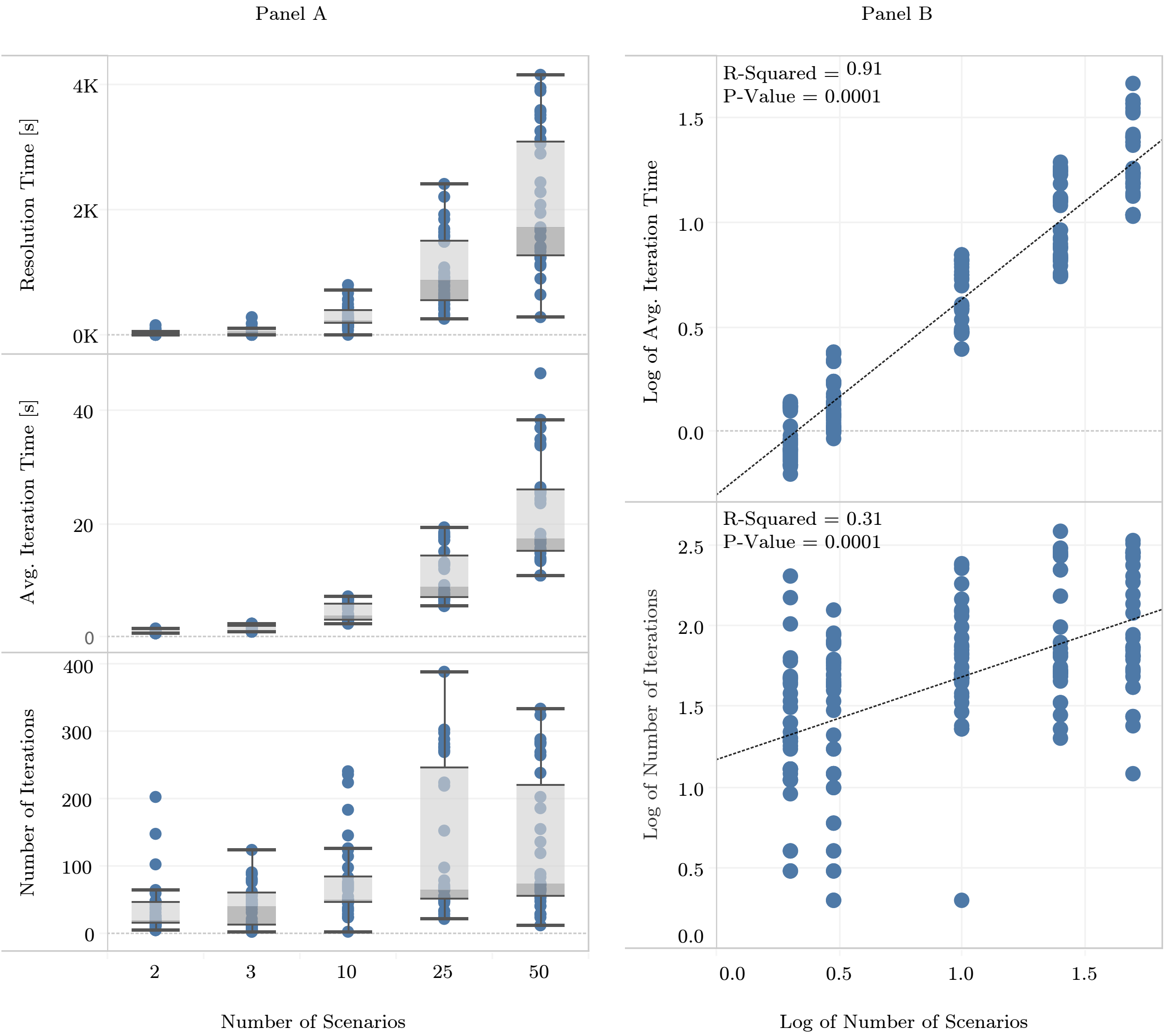

Figure 3.2: Iteration metrics versus instances sizes.

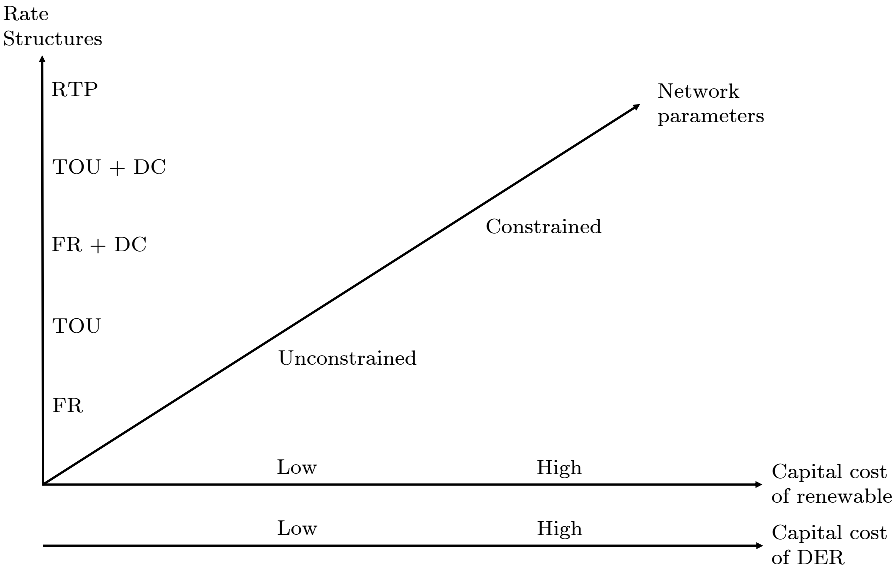

Figure 3.3: Structure of analysis. An instance corresponds to the combination of a rate structure, network parameter and the renewable and DER costs. There are 30 instances in total.

technologies distributed as depicted in Figure 3.4. The economic parameters of the generation technologies are standard for capacity expansion studies (see, for instance, De Jonghe et al. 2012). Table 3.1 summarizes these parameters.

Table 3.1: Economic parameters of supply-side technologies

		Base-load	Mid-merit	Peak	High-peak	Wind
Capital cost	k$/MW-yr	207	85	27	16	225
Fixed O&M		69	21	16	11	40
Total fixed		227	106	43	27	265
Fuel	$/MWh	11	27	43	66	0
Variable O&M		5	11	11	11	0
Total variable		16	38	54	77	0

Besides the network includes two load buses—1 and 2 in Figure 3.4, where a set of representative household are located. These households can differ in terms of their appliances and the DER they can adopt (for instance, because of differences in income). For this exercise, we consider two representative households per location. One can adopt DER, the other cannot. In addition, households located at bus 1 do not have air conditioning while those at bus 2 do. Figure 3.5 summarizes this configuration. As in Example 3.3.2, we model the consumption preferences for each household, assuming the utility function is additively

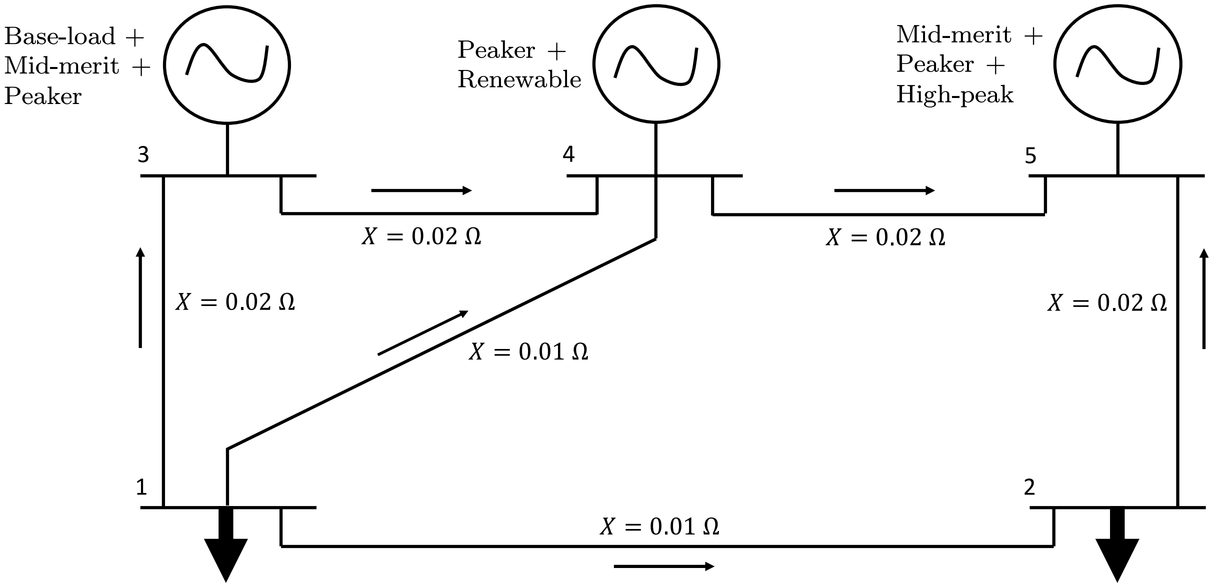

Figure 3.4: Network model. The letter X denotes the reactance of the branch, and the arrow the default direction of the flows. Buses 1 and 2 correspond to loads, and 3, 4 and 5 to generating technologies.

separable. For the purposes of this exercise, we refer to the demand of all other appliances as baseline.

Another relevant input are the time series we use in this exercise. We model weather and consumption patterns with six time series, including temperature profiles and rooftop solar availability factors at each bus. We also include the baseline consumption of the households—which we assume is the same for both buses, and the availability factors of the wind power generator. All time series correspond to one year of hourly data, and we define a scenario as one day, or a 24 hour period. We sample from the data three representative scenarios using the imortance sampling technique described in Papavasiliou and Oren (2013), which minimizes the distortions introduced by the sampling. Figure 3.6 depicts the time series we use and the scenarios we selected.

To the best of our knowledge, an analysis of these characteristics has not been undertaken before. The paper of Gallant and Koenker (1984) presents the closest attempt. The authors develop a model that allows them to estimate the parameters of a demand system with own and cross-price elasticities, and conduct a welfare comparison of time-varying rates and combinations with demand charges. The key difference between the analysis of Gallant and Koenker (1984) and the present is the representation of the supply side. These authors consider a simple supply cost function, which does not permit including renewable generation nor it allows modeling a network, or distributed energy resources. As we see below, different configurations of theses factors can drastically impact the results.

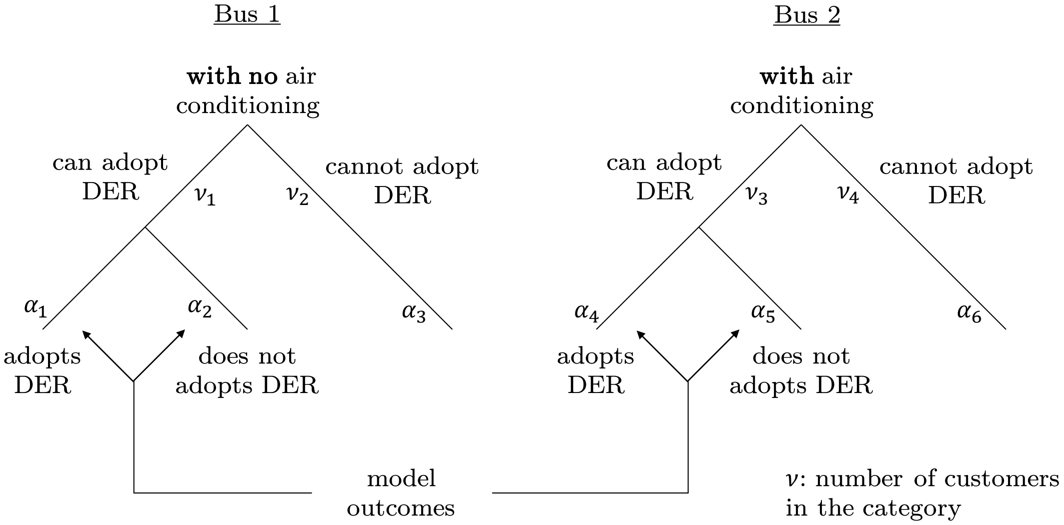

Figure 3.5: Household configuration per bus. There are in total 6 different households types. The final number per type, α, is an outcome of the model, and is constrained by the number of representative households, ν.

Results

Welfare Analysis

We first analyze welfare differences with respect to the theoretically inferior rate, the flat rate. Table 3.2 shows these differences for all the parameter combinations and rates. We see that across all these configurations RTP is the structure that improves welfare the most; on the opposite extreme is the flat rate supplemented with a demand charge. The other two tariffs are close in the middle. On average, across all parameters configurations, the time of use and the combination of this structure with a demand charge obtain close to 90% of the welfare gains that RTP achieves, while the other structure only approximately 50%. These results suggests that while adding a demand charge to a flat rate structure can improve welfare (on average 1.93%), adding a demand charge does not have an effect when complementing a TOU structure. In addition, we see that the TOU outperforms the FR + DC structure, producing almost the same welfare improving effects as the RTP. We observe, however, that the relatively small differences between TOU and RTP are due to our definition of time-windows for the former tariff. This structure has 24 time-windows, 1 per hour; the main difference with respect to the RTP rate is that in the case of the TOU the price for a given hour cannot vary across scenarios. As the number of representative scenarios increases, we should observe a widening in the welfare gains that these two structures can achieve.

Table 3.2 shows, in addition, the affect of the other factors we explore in this exercise in the rate comparison. The network status has the most pronounced effect. When the

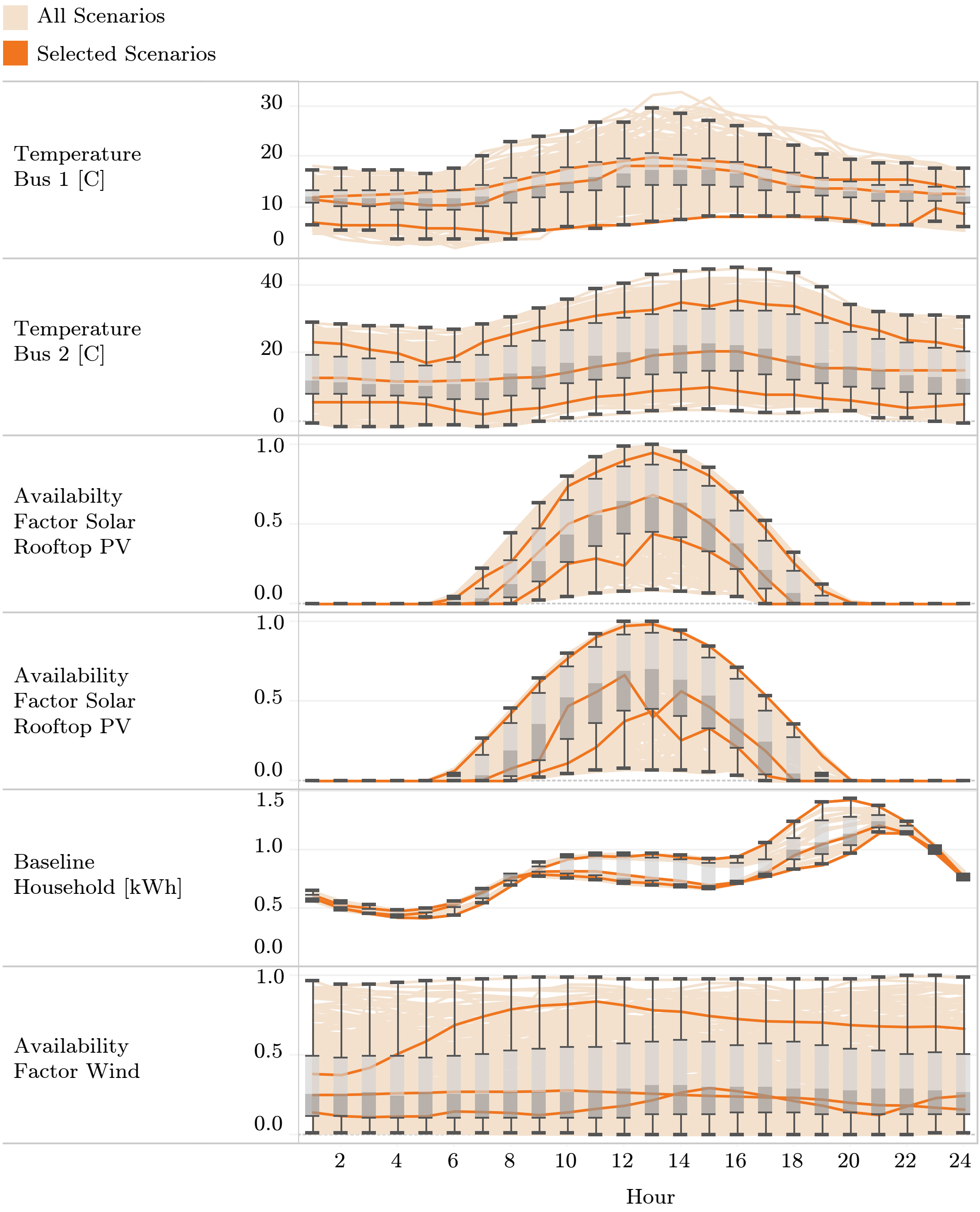

Figure 3.6: Time Series.

Table 3.2: Welfare gains with respect to FR in percentages

			Tariff			
Network Status[1]	Cost DER	Cost Renewable	RTP	TOU	TOU + DC	FR + DC
const.	low	low	8.43	8.23	8.22	4.39
		high	8.01	7.81	7.71	4.70
	high	low	7.52	6.76	6.18	3.07
		high	7.52	6.78	6.49	3.08
unconst.	low	low	0.63	0.47	0.40	0.03
		high	0.63	0.41	0.40	-0.05
	high	low	0.49	0.32	0.30	0.11
		high	0.48	0.32	0.31	0.12

[1] const. = constrained, unconst. = unconstrained.

network is constrained the welfare gains of switching from FR to any other rate are much higher, on average 31 times the gains that occur when the network is unconstrained. A model that does not account for a network implicitly assumes that there are no network constraints. This simple example shows that this assumption can have meaningful impacts on the rate comparison. A policy maker could conclude that given that the economic benefits of switching to a more complex rate are small, it is better to leave households enrolled in a FR tariff. The same conclusion would be hardly sustainable in the presence of the results of the constrained case.

The other factor that has an effect—though smaller—is the cost of the DER. When this technology is less expensive, the gain of switching from FR to a more complex rate is greater than when it has a higher cost. This happens because when the DER is less expensive more households adopt this technology, which increases the value of a flexible demand side. Households enrolled in more complex rates alter their loads to consume electricity when it is less expensive, which is when rooftop PV systems produce it. This in turn increases the utilization of the overall fleet of (supply and demand side) generating technologies, allowing households to get more energy per unit of capacity installed. Table 3.3 shows an increase in adoption and capacity utilization when the cost of the DER is low with respect to the case when it is high.

This table also highlights the potential complementarity between DERs and time-varying rates. In this application, when the cost of the DER is low, not only dynamic rates become more valuable. The adoption of rooftop solar PV increases as rates become more dynamic. This indicates an increase in the relative value of the DER with respect to other generation alternatives. That is, time-varying rates may improve the economics of DERs.

Table 3.3: Rooftop solar PV adoption and generation fleet utilization

			Tariff				
Network Status[1]	Cost DER	Cost Renewable	RTP	TOU	TOU + DC	FR + DC	FR
			population with rooftop solar PV [%]				
const.	low	low	24	23	23	19	16
		high	24	23	22	20	16
	high	low	7	11	15	16	7
		high	7	10	12	15	7
unconst.	low	low	24	23	23	22	21
		high	24	23	22	22	21
	high	low	0	0	0	0	0
		high	0	0	0	0	0
			capacity utilization [%]				
const.	low	low	59	56	56	49	44
		high	59	56	54	51	45
	high	low	55	54	54	47	41
		high	56	52	52	47	41
unconst.	low	low	59	56	55	47	46
		high	59	56	55	45	46
	high	low	49	48	48	43	41
		high	49	48	48	43	41

[1] const. = constrained, unconst. = unconstrained.

Distributional Analysis

We explore now the results from a distributional perspective. As Figure 3.5 shows, a group of households cannot adopt the DER. This element of our setting seeks to reflect that in a population there will be some households which cannot afford adopting a distributed energy resource, such as a rooftop PV system. Now we ask the question of how the switching from a simple flat rate to a more complex structure benefits customers with different financial means—in our setting those who can and cannot adopt DERs.

Table 3.4 shows the surplus increase as a result of the switch in rate structures, for the case where the cost of the DER is low. When the switch improves welfare the most, which is when the network is constrained, the disparities are more significant. In one of the buses, for those household that can afford the DER, the increase in surplus is approximately 1.5 times that of the households that cannot afford a PV system. It is interesting to notice that this is

Table 3.4: Net surplus increase per household with respect to FR [%]

		Household Type		Tariff			
Network Status[2]	Cost Renewable	Bus	Can Adopt DER	RTP	TOU	TOU + DC	FR + DC
const.	low	1	yes	7.28	7.26	7.24	2.48
			no	7.29	7.26	7.25	2.44
		2	yes	10.98	10.72	10.46	4.53
			no	6.60	6.55	6.80	3.48
	high	1	yes	6.60	6.55	6.80	3.48
			no	6.60	6.55	6.51	3.48
		2	yes	10.46	10.06	9.81	5.92
			no	6.94	6.75	6.50	5.28
unconst.	low	1	yes	0.28	0.20	0.21	(0.00)
			no	0.28	0.20	0.21	(0.00)
		2	yes	1.47	1.50	1.32	(0.02)
			no	(0.06)	(0.72)	(0.74)	0.20
	high	1	yes	0.22	0.22	0.24	(0.05)
			no	0.22	0.21	0.21	(0.05)
		2	yes	1.59	1.58	1.28	(0.19)
			no	(0.14)	(1.15)	(0.74)	0.19

[1] A 10 % increase is between \$100 and \$250 per year.

[2] const. = constrained, unconst. = unconstrained.

not always the case, as there no differences in the surplus increase between the households located in bus 1. When weather patterns and network conditions are favorable, wealthier households can take more advantage of sophisticated rates by adopting distributed energy resources. Additionally, the model indicates that the flexibility of the structures across time and states of nature makes this phenomenon more pronounced. In fact, the switch to RTP translates into surplus increases for households that can adopt the DER being 1.6 times that of households with less financial means. On the other hand, when the switch is to the FR + DC structure, the surplus increase of wealthier households is only 1.2 times that of those that cannot afford the DERs.

In Table 3.4 we can also see that in the constrained case the switch to a more complex rate is always Pareto improving. All type of households are better off after the switch. This does not happen when the network is unconstrained. In most cases all households increase or maintain their net surpluses. However, when the cost of the renewable generating technology is high and the switch is to a time of use rate, households that cannot adopt DERs and are located at Bus 2 are worse off after switching. While the absolute harm is small, approximately \$31 per year, since it happens to household with comparatively less financial resources, it potentially has more negative consequences than if it were borne by those that can afford adopting DERs. In addition, we see that in the unconstrained case time-varying rates continue to accentuate disparities. While the switch to a FR + DC rate practically has no effect on consumer surplus, it does has an affect when it is to a time-varying rate. As in the constrained case, households that can afford rooftop PV systems take advantage of this technology to increase their surpluses under the time-varying rate.

3.6 Conclusions

This study develops a technique to compare rate structures. It allows researchers to transparently model a wide range of tariffs, distributed energy resources and supply side configurations. The specific values of the tariffs, as well as the demand and supply side consumption, production and investment decisions result from solving the problem of a Ramsey-Boiteux planner. This is the problem of a regulator that anticipates the impacts of it choice of rates on demand and supply side short- and long-run decisions.

We cast this Bilevel Model as a Mathematical Program with Equilibrium Constrains, and solve it using a variant of the alternating direction method of multipliers. This variant handles the complementary constraints in a distributed manner. It solves the MPEC iteratively, at each iteration solving independently several small MPECs and one nonlinear program. Using the technique of Fortuny-Amat and McCarl (1981) we cast the small MPECs as MIPs and develop heuristics to find reasonable starting points; we handle the nonlinear program via a conic reformulation. A computational exercise demonstrates that our solution approach has several desirable properties for practical applications. First, it vastly outperforms Knitro, a popular commercial solver for MPECs. Second, the method shows good convergence behavior, reaching solutions of reasonable quality after a few tens of iterations. Third, the algorithm scales well with size as part of the resolution of the problem can be distributed to various computing nodes and the solution time of the centralized step increases at a rate lower than linear.

With a numerical analysis, we demonstrate the importance of the modeling flexibility of the present technique to compare rate structures. In contrast to previous approaches, our method allows capturing complex supply side configurations. In particular, researchers can model a network. Our exercise shows that abstracting away this element can have meaningful impacts on the analysis of rate structures. In the example, the presence of network constraints translates into significantly higher welfare gains when switching from a flat rate to more sophisticated structures. Omitting this element underestimates the benefits by about 3000 times. In addition, the analysis highlights the value of being able to model DERs. It shows how the economics of rooftop PV systems impact the rate comparison, and that time-varying rates and inexpensive DERs could complement each other; the former can improve the relative value of the latter while DERs being inexpensive increases the gains of switching from a time invariant rate. Finally, the distributional analysis portrays how our method permits regulators and policy makers to study impacts of a rate update on a heterogeneous population. While a switch in rates could have a positive impact on the aggregate of households, it could benefit some more than others. It could even harm some customers, which can be particularly problematic if those harmed were low income households. A technique such as the one introduced in this paper permits to anticipate these impacts, letting regulators and policy makers to decide among rates structures with considerably more information that what would be available with alternative techniques.

Chapter 4

Utility Pricing in the Prosumer Era: An Analysis of Residential Electricity Pricing in California

4.1 Introduction

The increasing penetration of *distributed energy resources* (DER) including advanced metering infrastructure (AMI), energy management systems (HEMS), solar photovoltaic and battery storage systems are enabling residential consumers, or "prosumers", to interact bidirectionally with electricity systems. Customers can not only purchase electricity from the grid but also provide the system with energy and reliability services (Schleicher-Tappeser, 2012). In regulated retail sectors, residential rate structures greatly influence this interaction. Rates can impact adoption decisions, by changing the economic value of the technology, and can influence the usage of this resource. Rates may also increase distributional disparities, creating cross subsidies between those who can and cannot adopt the distributed technology (Eid et al., 2014). Despite this crucial role, no consensus exists with respect to the ideal rate structure. Factors explaining this reality include limitations of the theory and the focus of the empirical work. While the theory asserts that real-time pricing (RTP) is optimal from an economic efficiency perspective (Joskow and Tirole, 2006), regulators must balance efficiency with other public goals, such as rate simplicity, equity or meeting environmental directives (Stanton, 2015). Absent a comprehensive quantification of the impacts associated with implementing RTP, it is difficult for regulators to judge the value of this alternative, especially when it may compromise other regulatory goals. The empirical literature has tried to quantify impacts, however, the scope has been somewhat limited. Researchers have focused on estimating price responsiveness to quantify changes in efficiency in the short-run.[1] Other relevant metrics such as long-run welfare effects, equity or environmental implications

[1] Examples include the work of Allcott (2011), Caves et al. (1984a), Faruqui and Sergici (2010), Faruqui and Sergici (2011) and Herter (2007).

have been explored either in isolation or with stylized analyses, but never with an applied approach, within a unified framework.[2]

This work contributes to rate regulation policy with an applied study in the context of California's residential electricity sector. Our focus is the comparison of the long-run welfare effects as well as the equity and environmental implications of a set rate structures. The analysis considers the hypothetical scenario in which HEMS is widely adopted. Under this circumstance the household responses to price signals are likely to be fully rational, driven by an algorithm optimizing the consumption of the appliances (Beaudin and Zareipour, 2015).

A second contribution is the development of a framework that permits comparing rate structures along each of the metrics we consider in our analysis. We develop a model of optimal pricing that accommodates a wide variety of tariffs. Our framework builds upon *peak-load pricing*[3], and borrows some elements from the literature of *generating capacity expansion*.[4] We embed a detailed model of household behavior in this setting, expanding the basic model of peak-load pricing to include heterogeneous households and the adoption of DERs. Our approach allows us to capture a wide variety of temporal and spatial demand substitution patterns, without needing to use a large number of estimates.

Our analysis is particularly relevant to the current regulatory situation in California, in which the default increasing block pricing program is gradually being retired and time-of-use becomes the default rate. We compare this prospective structure with two variants of this program, a TOU combined with demand charges (TOU&DC) and a TOU combined with a critical peak pricing program (TOU&CPP). We also include in the comparison the cases in which households enroll in a flat rate (FR) and in a real-time pricing program.

Considering all households under the FR rate as a reference case, our analysis shows that implementing any other pricing alternative produces gains for the average household that are mild at best. When implementing time-of-use pricing, or any of its variants, the average gain is not greater than 1.2 dollars per month. Even though real-time pricing performs better than the other rate structures, the improvement over the FR tariff seems mild. The average household increases its net surplus by 2 dollars per month under this rate scenario.

In addition, our analysis shows that the net surplus gain varies considerably across households. Factors such as the presence of an air conditioning system or the temperature outside the dwelling are major drivers of this variation. For all rates, households with air conditioners experience higher average gains than household without these appliances. However, the gains in net surplus vary more across the former group of households. The exterior temperature profile is a key factor. Its relationship with the net surplus gains, however, is not simple, and depends on the specific rate. For instance, under the RTP case customers experience greater gains in areas with higher average temperature. But this statistic has no correlation with gains under the TOU&DC program.

These two results combined suggest that defaulting all residential customers into a time-

[2] See, for instance, Borenstein (2006), Borenstein (2013), Crew et al. (1995) and Joskow and Tirole (2006).

[3] For a comprehensive survey of the literature we refer the reader to Crew et al. (1995).

[4] Sauma and Oren (2006) provide a good example of these models.

of-use rate structure, which is the current path California is following for the residential sector, may not be the best strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

4.2 California Electricity Sector and the Emergence of Prosumers

An overview of the sector

The California electricity sector serves approximately 30 million people across the state. With 59 GW of power plant capacity, the sector delivers near 309 TWh of electricity annually. Its market size is close to $8 billion per year and its transmission system, spanning $25,627$ circuit-miles, is part of the Western Interconnection. In terms of market regulation and oversight, there are three institutions involved, each with different roles. The first, the Federal Energy Regulatory Commission (FERC), has jurisdiction over the interstate transmission of electricity. Its responsibilities include the oversight of important merger and acquisitions, reviewing applications for transmission projects, as well as licensing and inspecting private, municipal and state hydroelectric projects. The commission also sets mandatory reliability standards and monitors energy markets across the US. The other two regulatory agencies have jurisdiction in the sate of California. One is the California Energy Commission which is the primary energy policy and planning agency of the state. The other is the California Public Utilities Commission (CPUC). Its main role is regulating the three investor-owned electric utilities of California, including Pacific Gas and Electric Company (PG&E), Southern California Edison (SCE) and San Diego Gas and Electric Company (SDG&E), which collectively serve two thirds of the electricity demand throughout California. Among other functions, the CPUC sets and approves the retail rates, is responsible for ensuring that utilities meet state environmental policies and ensures electricity safety at the distribution level.

In terms of system and market operations, the California's Independent System Operator (CAISO) is responsible for maintaining a reliable transmission of power as well as the comprehensive long-term planning of grid infrastructure. This entity also coordinates forward and spot markets for energy and ancillary services. In addition, CAISO complies with the reliability standards set by the North American Electric Reliability Corporation (NERC) and the Western Electricity Coordinating Council (WECC). While the former is a non-profit organization developing and enforcing reliability standards for the continental United States, Canada and Baja California, Mexico, the latter is a regional entity promoting bulk electric system reliability in the Western Interconnection.

Residential rates in California

The distinctive characteristic of residential electricity rates in California is its increasing block structure. They have had this form since 1976, when the Miller-Warren Energy Lifeline Act was enacted. This legislation sought to provide California's residential customers with a minimum necessary quantity of gas and electricity at a fair price, and also to encourage conservation. The legislation set a precedent, providing a conceptual justification for implementing increasing block rates. Since 1976 rates did not change meaningfully until California's electricity crisis.

Beginning in the summer of 2000, tight supply margins, weak federal oversight, lack of an elastic demand and flaws in the market design yielded a period of highly volatile electricity prices, known as California's Electricity Crisis. As a result of the crisis the sector underwent a period of drastic reforms. At the retail level, a first response to the high wholesale prices was to lift the retail price cap. This triggered notorious increases in electricity bills, which were then mitigated by freezing the charges of the lower two tiers. The result of this legislation was the replacement of a two tier system by a five tier structure, with prices of Tiers 3 to 5 considerably higher than those of the remaining lower tiers. From 2000 to 2009 differences among tiers increased. However, the enactment of SB695 in 2009 began to allow limited annual increases for Tiers 1 and 2.

At the time of this writing, decision $D.15-07-001$ is the main piece of regulation laying the path for the future of residential electricity rates in California. Key elements of this regulation include the promotion of the consolidation of the tiers and the development of rates that reflect better cost causation. In particular, the decision approves transitioning all residential customers to a default time-of-use tariff by 2019.

The emergence of prosumers

There are two main forces pushing the emergence of prosumers in California: Environmental policy directives and distributed technologies reaching maturity. The relevant environmental policy is the renewable portfolio standard which established, among others, targets for distributed generation. The policy has triggered the development of an array of incentives for generation at the customer's premises which has caused the massive deployment of distributed energy resources, such as solar photovoltaic panels. As for technological evolution, California has taken major steps towards modernizing its distribution grid, having the largest installation of AMI in the US. This technology constitutes a vital element for implementing time-varying-rates. By enabling two way communication between customer and utility on time intervals of an hour or less, AMI allows utilities measuring consumption on an hourly basis as well as sending price signals on a consistent time scale.

4.3 A Modeling Framework to Compare Rate Structures

Utility pricing: An overview of the theory

An important function of regulators is determining the rates that a regulated utility can charge for the provision of its services. This process, also known as *rate regulation*, includes the determination of the rate *rate level* and the *design of rate structures* (Phillips Jr, 1993, pp. 176 - 180). Establishing the rate level entails specifying the total compensation that the utility, or load serving entity (LSE), receives for its services. The design of rate structures, which is the focus of this work, defines how the LSE collects its compensation. Designing rate structures is far from trivial. So it is not surprising that a myriad of methodologies have been suggested and adopted in different jurisdictions. Brown and Sibley (1986, pp. 44 - 60) divide the approaches according to how they assign *common* or *non-attributable* costs across different services and consumer classes. There are two broad categories: the *cost-based pricing* and pricing based on the concept of *marginal cost*. The first group of approaches allocate costs based on criteria other than efficiency. One example is the *fully distributed costs method*, in which common costs are assigned according to the relative shares of magnitudes that can be attributed to a service or group of customers, such as peak-demand, output or revenue.[5] On the other hand, in pricing based on the concept of marginal cost efficiency has a prominent position. How common costs are attributed to the different groups of customers is a byproduct of a welfare maximization process. Our framework falls into the second category of approaches. The model that this work introduces produces a set of rates and allocations of costs that emerge from the welfare maximization of the system under study. This is the approach to rate design that the literature of peak-load pricing studies.

Peak-load pricing develops a normative theory of efficient or welfare maximizing pricing for industries with limited storage capability and time-varying demand. The modern version of the theory originates with the contributions of Boiteux (1960) and Steiner (1957), and intended to provide guidelines in the context of price regulation of natural monopolies, such as vertically integrated electric utilities (Crew et al., 1995). The basic model considers the problem of a social planner choosing prices that maximize welfare, i.e., the surplus of customers and the public utility's profits. Prices coordinate production and consumption decisions over a time horizon. The monopolist invests in production capacity at the beginning of the horizon and prices are such that the utilization and the level of the installed capacity are optimal (Drèze, 1964). Studies including Carlton (1977), Chao (1983), Crew and Kleindorfer (1976) and Panzar (1976) further refine the model to include a stochastic demand, supply-side uncertainties and multiple technologies.

Borenstein and Holland (2005) and Joskow and Tirole (2006), in examining the merits of

[5] Other approaches seek to minimize cross subsidies across services and consumer classes and others build a set of axioms and derive rate structures consistent with them. For more details on the subject of cost-based pricing, see (Brown and Sibley, 1986, pp. 44 - 60).

retail competition in the electricity industry, show that the theory also applies to restructured electricity sectors. In these models, a competitive wholesale market replaces the production side of the vertically integrated utility. At the retail level, both papers distinguish the cases of a regulated distribution company and competitive retailers. While Borenstein and Holland (2005) consider a setting with linear or uniform prices, Joskow and Tirole (2006) contemplate the case of a two-part, non-linear price. Further, Borenstein and Holland (2005) explore the long-run effects of different pricing policies, analyzing the equilibria that emerge at the wholesale and retail levels. More recently, Zöttl (2010) investigates a setting in which there is imperfect competition at the wholesale level, and Chao (2011) updates earlier work exploring the interaction of different pricing policies and renewable technologies.

Our model departs from previous work first by generalizing the type of rate structures present in peak-load pricing. In addition, we introduce a mechanism linking pricing and technology adoption decisions. Finally, our model accommodates household heterogeneity beyond a scale factor.

The regulator's problem

The regulator's problem combines elements of the peak-load pricing and capacity expansion literature. The key element from capacity expansion not present in peak-load pricing is a transmission network. For simplicity, we do not detail this element of the model in this section. We refer the interested reader to the appendix B.4. As in peak-load pricing, our model falls into the broad category of two-stage stochastic optimization models. Agents in these models make long-run decisions at the beginning of the horizon before uncertainty is realized and define state contingent strategies for the short-run stage. These are static models that can describe systems in steady state. Our framework, therefore, is not suitable for studying system dynamics. In terms of the institutional setting, at the retail level we consider a distribution utility as the load serving entity. In general, however, one can consider settings within two polar cases. While the utility could be fully integrated with the supply side in one case, in the opposite, it could be just a distribution company. Under the assumption of perfect competition at the wholesale level, both cases are equivalent (Joskow and Tirole, 2007), however.

Let ω index a finite and countable set Ω of states of nature, π the corresponding probability vector, $E\left[\cdot\right]$ the expectation operator, and a time horizon of $t \in T$ time steps. Given a random vector of consumption d, the household pays $l + \eta(d, p)$ to the utility, where l is a fixed charge, $p \in \mathcal{P}$ a vector of rate parameters and $\eta(\cdot)$ a fee contingent on consumption and the rate parameters. We call the triple $(l, \eta(\cdot), p)$ a rate structure. Our setting is similar to the one in (Joskow and Tirole, 2006) insofar we focus on two-part structures with a state and time contingent demands and prices. However, we generalize this model to accommodate more complex rate structures. In our setting the vectors d do not only have one component for every time and state of nature but also may include other relevant metrics associated to the demand profile, such as peak or total consumption across the time horizon. Similarly, price parameters may include charges for peak or total demand. Specifically, we focus on the case in which $\eta(\cdot)$ is bilinear on the demand vector and price parameters. The following assumption formalizes this specification.

Assumption 4.3.1 *The demand contingent charge $\eta(d,p) = d^\top Mp + Ind\left\{\tilde{b} - \tilde{A}d\right\}$, where $Ind\{x\}$ is 0 if $x \geq 0$, and ∞ otherwise.*

As subsection 4.3 shows, this specification is fairly general, allowing researchers to model a wide range of the rate structures used in practice.

As in peak-load pricing, we consider homogeneous households, with a mapping $D_\omega(p) : \mathcal{P} \rightarrow \mathbb{R}^{|T|}$ and a real valued function $U_\omega(d_\omega)$ representing their demand and gross surplus metric, respectively. Given a set of wholesale prices $\{\lambda_\omega\}$, the planner problem optimizes the household net surplus, $E\left[U_\omega\left(D_\omega(p)\right) - \eta\left(D_\omega(p), p\right)\right] - l$, and guarantees that the utility meets its revenue requirement, $E\left[\eta\left(D_\omega(p), p\right) - D_\omega(p)^\top \lambda_\omega\right] + l - \Pi$, with Π an exogenous fixed cost. As (Joskow and Tirole, 2006) shows, this amounts to find (l^*, p^*) such that

$$(d^*, p^*) \in \underset{(d,p)}{arg\,max} \left\{E\left[U_\omega(d_\omega) - d_\omega^\top \lambda_\omega\right] : \, d_\omega = D_\omega(p) \; \forall \omega \in \Omega\right\}, \tag{4.1}$$

$$l^* = E\left[\lambda_\omega^\top d_\omega^*\right] - \eta(d^*, p^*) + \Pi. \tag{4.2}$$

A competitive wholesale electricity market

The wholesale market representation in this model is a variant of the supply-side model studied in the peak-load pricing and capacity expansion literature. More specifically, we follow closely the representation in (Chao, 2011). In this model, infinitesimal competitive firms interact in a spot market for electricity. Each decides on their long-run installed capacity and short-run generation profiles. We denote the total installed capacity of technology $k \in K$ as x_k and its cost of carrying capacity as $\tilde{r}_k$. The aggregated production profile of this technology in state of nature ω is $y_{\omega k} \in \mathbb{R}_+^T$, and variable costs per unit of power production is $c_{\omega k} \in \mathbb{R}_+^T$. We capture variability in a technology's availability – e.g. due to outages – with an availability factor per technology contingent on the states of nature, $\rho_{\omega k} \in R^T$. In a perfectly competitive market firms are price takers, thus, production and capacity for technology k are the solution of the problem

$$\max_{(y_{\omega k}, x_k)} \left\{E\left[(\lambda_\omega - c_{\omega k})^\top y_{\omega k}\right] - x_k \tilde{r}_k \colon 0 \leq y_{\omega k} \leq x_k \rho_{\omega k}\right\}. \tag{4.3}$$

The market equilibrium is a tuple $(d^*, p^*, y^*, x^*, \lambda^*)$ such that (d^*, p^*) solves the regulator's problem at λ^*, (y^*, x^*) solves the problem of the producer at that price, and supply equals demand. It is easy to verify that the market equilibrium is the solution of

$$\max_{(d,p,x,y)} E\left[U_\omega(d_\omega) - \sum_{k\in K} y_{k\omega}^\top c_{k\omega}\right] - x^\top \tilde{r} \tag{4.4}$$

$$\text{subject to}$$

$$d_\omega = \sum_{k\in K} y_{\omega k} \colon \lambda_\omega, \tag{4.5}$$

$$0 \leq y_{k\omega t} \leq x_k \rho_{\omega k}, \tag{4.6}$$

$$p \in \mathcal{P}, \tag{4.7}$$

$$d_\omega = D_\omega(p). \tag{4.8}$$

The household behavior

Except for our specification of the demand contingent fee, $\eta(\cdot)$, (4.4) – (4.7) is the classic peak-load pricing problem. Researchers can use the model to analyze theoretically and numerically implications of different constraint sets for the vector of retail prices. A key assumption that facilitates the study of these models is a demand system with analytic expression. Our framework drops this assumption because our specification of $\eta(\cdot)$ implies, in general, demands with no analytic definition. Consistently, our model updates the peak-load pricing problem replacing (4.8) with the following condition,

$$d \in \mathrm{ar}g\,\underset{d}{max} \left\{E\left[U_\omega\left(d_\omega\right) - d_\omega^\top M_\omega p_\omega\right] \colon \bar{b}_\omega - \bar{A}_\omega d_\omega \geq 0,\ \forall \omega \in \Omega\right\}, \tag{4.9}$$

where $(\bar{b}, \bar{A})$ contain the parameters of the rate structure, $(\tilde{b}, \tilde{A})$, and possibly others. Henceforth we refer to (4.9) as the *household problem*, and to (4.4) – (4.7), (4.9) as the *pricing problem.*

Illustrative examples

Our specification of the household demand allows to model the influence on demand of several rate structures and, also, represent demand-side technologies of interest. Here we show how to implement the models that we use in our analysis. The appendix B.2 provides additional examples. Some notation will prove useful. The matrix I_m corresponds to the identity of m by m. The vectors e_m and z_m are, correspondingly, vectors of ones and zeros of m dimension.

Modeling rate structures. Our analysis compares *time-varying pricing* (TVP) and a TVP combined with a *demand charge* (DC). A time varying pricing is the simplest type of rate to model. Set $M_\omega = I_{|T|}$, let the vectors $p_\omega \in \mathbb{R}_+^T$ and $d_\omega \in \mathbb{R}^T$, and define define $\mathcal{P}$ as follows,

$$\mathcal{P} := \begin{cases} \Big\{p \in \mathbb{R}_+^{|T|\times|\Omega|} : p_{\omega t} = p_{\omega' t'} \,\forall(\omega,t),(\omega',t')\Big\} & \text{for FR,} \\ \Big\{p \in \mathbb{R}_+^{|T|\times|\Omega|} : p_{\omega t} = p_{\omega' t'} \,\forall(\omega,t),(\omega',t') \in TW(\omega,t)\Big\} & \text{for TOU,} \\ \Big\{p \in \mathbb{R}_+^{|T|\times|\Omega|}\Big\} & \text{for RTP,} \end{cases} \tag{4.10}$$

where $TW(\omega, t)$ is the set of time windows (ω', t') in the same time window as (ω, t).

Adding a demand charge to any of these structures requires redefining $d := [\bar{d}, \hat{d}]$ and $p := [\bar{p}, \hat{p}]$, where $\bar{d}, \bar{p} \in \mathbb{R}_+^{|T|\times|\Omega|}$, and $\hat{d}_\omega, \hat{p}_\omega$ correspond to the maximum consumption and demand charges under ω, respectively. The matrix M_ω is now equal to $I_{|T|+1}$, and the analyst may add additional conditions to the set $\mathcal{P}$ to model demand charges constant across some scenarios. A final element of this structure is the constraint linking the hourly consumption profile $\bar{d}$ and the maximum consumption $\hat{d}$, which we model via the following definitions

$$\tilde{b} := z_{|\Omega|\cdot|T|}, \; \tilde{A} := \begin{bmatrix} I_{|\Omega|\times|T|} & -I_{|\Omega|} \otimes e_{|T|} \end{bmatrix}. \tag{4.11}$$

Household as composite of devices. Following the approach of Reiss and White (2005), we consider that households are composite of devices and assume their utility functions are additively separable. For reasons we explain later, we consider in our analysis two devices, a central air conditioning unit and a rooftop solar panel, and the household baseline. Central air conditioning falls in the more general category of *thermostatically controlled loads* (TCL's), whose behavior follows the laws of thermodynamics. Mathieu et al. (2015) presents a model describing the behavior of these appliances, which links the household's inside temperature, θ, with the outdoor temperature $\tilde{\theta}$, the thermal characteristics of the dwelling, ξ, and the electricity consumption of this appliance, d. It is possible to show that the inside temperature profile has the following form

$$\theta(d; \xi, \tilde{\theta}) = \Theta_1(\xi)d + \theta_2(\xi, \tilde{\theta}), \tag{4.12}$$

where Θ_1 and θ_2 are a matrix and a vector, functions of the thermal parameters and temperature outside the dwelling. Appendix B.3 shows the full derivation of this relationship. Here we close the model of the TCL behavior introducing a mechanism capturing household's preferences for thermal comfort. The simplest approach involves a penalty for deviating from an ideal inside temperature, $\hat{\theta}$. Equation (4.13) shows a utility function consistent with this approach, which we use in our analysis.

$$U(d) = -\beta\|\theta(d; \xi, \tilde{\theta}) - \hat{\theta}\|^2 \tag{4.13}$$

For modeling the household baseline, we assume a linear demand system and compute the associated utility function using a standard procedure. For the rooftop solar panel we add to the household problem a constraint limiting its hourly production given the hourly availability of the solar resource. A final element of the household model links the demands of each device with the net demand of the customer,

$$\bar{d} = d_{baline} + d_{ac} - d_{solar}. \tag{4.14}$$

Household heterogeneity and DER adoption

The pricing problem contemplates one representative customer and there is no mechanism modeling customer adoption. In our analysis, however, we consider heterogeneous customers and analyze impacts of pricing on adoption. We incorporate these two elements using the framework that Castro and Callaway (2016) develop. The paper distinguishes different customer types $i \in I$,[6] each of which decides a set of technologies to adopt $j \in J$. [7] Calling the combination $h := (i, j)$ a *segment* and defining α_h and r_h, respectively, as the number of households and cost associated to a segment,[8] the paper shows how modifying the pricing problem permits modeling adoption decisions. Specifically, equation (4.4) becomes

$$\max_{(\alpha,d,p,x,y)} E\left[\sum_h \alpha_h \left[U_{h\omega}(d_{h\omega}) - r_h\right] - \sum_{k\in K} y_{k\omega}^\top c_{k\omega}\right] - x^\top \tilde{r} \tag{4.15}$$

and (4.5) updates to

$$\sum_h \alpha_h d_{h\omega} = \sum_{k\in K} y_{\omega k} : \lambda_\omega. \tag{4.16}$$

A final element to include for modeling adoption is the feasible region for α, which ensures that the number of households per segment is consistent with the number of households per customer type.

Comparing rate structures

Subsections 4.3 to 4.3 develop an analytic tool to compare rate structures. Researchers can explore the effects of different structures on welfare and other metrics by changing the specification of the consumption contingent fee, $\eta(\cdot)$, solving the pricing problem and comparing the metrics of interest. While the method is not intended to predict what would happen were the tariff under analysis in place, it provides a consistent assessment of the potential differences between them.

Solving the pricing problem is not straightforward. The problem falls into the broad category of *Bilevel Problems*, in which a leader – in our setting, the regulator – indirectly controls the actions of the follower – the household – changing one or more parameters of her problem. Bilevel problems are hard to solve in general, and state of the art solvers can only handle problems of moderate size. For large instances, researchers have to devise specialized algorithms. Given the size of the instance we explore in this study, we had to develop a specialized algorithm as well. However, the development of the algorithm is beyond the scope of this work. It constitutes a completely separate research effort, which Castro et al. (2016) describes in detail. The basic idea is to decompose the pricing problem

[6] In order to fix ideas consider the following two examples: $I = \{\text{with central AC}, \text{without central AC}\}$ or $I = \{\text{live in hot weather}, \text{live in cold weather}\}$.

[7] An instance of this set is $J = \{\{\text{solar PV, battery storage}\}, \{\text{solar PV}\}, \{\text{battery storage}\}\}$.

[8] The cost associated to a segment (i, j) is the annualized capital cost of the set of technologies j.

into one problem per household and state of nature, and one problem that coordinates the demands of the households. The algorithm iterates solving all problems at each repetition and stops when consecutive solutions do not change. The key aspect of the algorithm is its distributed nature which, by enabling its implementation in cluster computing facilities, makes our modeling framework practical.

4.4 Modeling California's Electricity Sector

We construct our model of the California electricity sector supplementing the network model that Price and Goodin (2011) developed for market analysis. The model consists of a network with 240 nodes, or buses, which corresponds to a topological reduction of the transmission system encompassing the Western Interconnection.[9] This reduced system also provides generation technologies and demands at each node, and the physical characteristics of the network, including transmission constraints. The generating power plants correspond to aggregations per type of fuel. Non-dispatchable generating technologies[10] such as solar or wind generation, and reservoirs such a geothermal or hydro power plants come with a year of hourly energy production. Fossil fuel technologies, on the other hand, only include physical and short-run economic parameters, such as heat rates and fuel costs. As for the demands, the network model includes a year of hourly energy consumption at nodes of the network corresponding to demand centers.

Before describing how we complete this data set, we make two clarifications. At first sight, the network model has more information than this analysis requires, because the Western Interconnection includes more states than just California. Using the full interconnection model, however, allows us to produce realistic import and export flows and provides the opportunity to study the impacts of residential pricing policy outside California. Because the main computational difficulties emerge from our detailed modeling of the demand, and because we do not model in detail demand at nodes outside California, the full network model did not increase significantly the computational complexity of our analysis.

A second clarification relates to our treatment of the Western Interconnection as an integrated market. That is, in our model the physics of the transmission lines is the unique factor limiting the flow of electricity through the interconnection. In practice, the administration of this system is divided among 38 *balancing authorities*, each controlling one portion of the network, and whose central role is to guarantee the reliable operation of their respective sub regions. This adds additional limitations to the flows of electricity which our model does not capture. At the time of this writing, however, California is leading the efforts to assess the impacts of a multistate regional market for the Western Interconnection.[11] Thus, an integrated market is plausible for the future of the interconnection.

[9] We refer the reader to the Western Electric Coordinating Council website for detailed information on the interconnection.

[10] Plants with outputs that are determined to great extent by exogenous factors such as weather conditions.

[11] See Brattle et al. (2016) for further detail.

Generating technologies

Because we are interested in studying long-run impacts, we replace the cost functions in the network model with the functions we described in subsection 4.3. A fixed and variable cost implement these functions. The fixed cost includes the annuity associated with developing and installing the generating technology and the fixed O&M costs. The variable cost, on the other hand, encompasses fuel and variable O&M costs.

In addition to these economic parameters, technologies have associated emissions and availability factors. While the former captures the fact that different fuels have different GHG emissions, the latter reflects the fact that power plants experience unplanned outages. Emissions factors as well as the economic parameters of the generating technologies come from EIA (2016), and table 4.1 summarizes the specific values we use in this study. As for availability factors, we use the magnitudes that NERC makes publicly available through its Generating Availability Data System (GADS).

Table 4.1: Generating technologies costs and GHG emissions

technology	variable cost	fixed cost	emissions factor
	[2015 $/MWh]	[2015 k$/MW-year]	[tCO2eq/MWh]
nuclear	11	448	-
biomass	35	424	1.93
coal	32	361	0.65
geothermal	-	341	-
solar	-	192	-
wind	-	184	-
gas adv CC	39	84	0.33
gas conv CC	42	77	0.35
gas adv CT	63	52	0.52

We include in the model the existing plants, by technology, for each node.[12] As mentioned, investment decisions to expand this fleet are endogenous to the model. We treat hydro power generation as exogenous because a correct treatment of this technology, which involves the stochastic dynamic optimization of reservoirs, is beyond the scope of our model. Even though solar and wind are non-dispatchable generators, we use the time series in the data set only to compute hourly availability factors.[13] The actual hourly production for these technologies is ultimately determined by their installed capacity, an outcome of our model.

[12] Appendix B.1 shows the geographic distribution.

[13] Hourly availability factors are the ratio between the hourly production and the nameplate capacity of the technology in the data set.

Developing a model of California's residential demand

We construct a model of the residential sector of California calibrating our model of household behavior at each node of the network. Households can consume and produce electricity on an hourly basis, and impact the system via their net demands. In terms of consumption, we consider two major categories of household end uses: cooling and non-cooling. We take this approach for two reasons. In California central air conditioning is a major source of electricity consumption, and approximately one out of every two households has this type of appliance (Palmgren et al., 2010). In addition, studies at the appliance level report air conditioning to be a major source of demand responsiveness (Mathieu et al., 2015; Reiss and White, 2005).

The second reason relates to the hourly demand data available for this study. In our framework the baseline of a household is the fraction of its demand not modeled as any particular device. In order to calibrate this function one needs an intercept, i.e., a time series of electricity consumption and the prices in effect when this happened. For the demand part of the intercept we use the load shapes developed by Itron Inc., described in Wei et al. (2013). This data set disaggregates residential consumption into space conditioning and other loads.

Utility functions summarize household preferences for each end use. We calibrate them using the appliance level elasticities and marginal effects estimates that Reiss and White (2005) report. We model the baseline consumption as a linear demand system and use an elasticity of -0.08, corresponding to the estimate for households with no space conditioning. For the price intercept, we follow the procedure that Borenstein and Holland (2005) describe, assuming the rate structure of the intercept to be a flat rate.

The model for cooling corresponds to the TCL model developed in subsection 4.3. This has three groups of parameters which can be categorized as technical, behavioral and weather related. Technical parameters include the thermal resistance and capacitance of the household, and the efficiency of the air conditioner. Mathieu et al. (2015) provide ranges for these parameters in California, and we use the midpoint of those ranges in their study. The behavioral parameters are the ideal interior temperature, which we set to 22°C (or 72°F), and the discomfort penalty, β. Using our model of the TCL, we link this latter parameter with the estimate of the marginal effect for central air conditioning of Reiss and White (2005). The expression linking the two magnitudes is

$$\beta = -\frac{e_{|T|}^{\top}\Theta_1(\xi)^{\top}\Theta_1(\xi)e_{|T|}}{\frac{d}{dp}E[e_{|T|}^{\top}d_{\omega}]}. \tag{4.17}$$

The weather related parameter corresponds to the outside temperature. The National Renewable Energy Laboratory (NREL) develops the Typical Meteorological Year (TMY) data set for modeling energy conversion systems.[14] This data set contains 12 months of hourly data at selected locations across the US. The data for each month typifies conditions for the location over a longer period of time, such as 30 years. There are 73 locations

[14] For a detailed description of the data we refer the reader to Wilcox and Marion (2008).

corresponding to California. Based on distance we assign each of these locations to one of the buses with demand within the California portion of the network model. The TMY data set also comes with hourly values for solar radiation. We use them for our model of rooftop solar panels.

Another input for the analysis is the count of households types per bus. We distinguish four types of households in our analysis, corresponding to the combinations of tenure status and the presence of central air conditioning. While section 4.5 discusses this categorization further, here we focus on the calibration of the household counts. The sources of data for this task are the 2010 census and the Residential Appliance Saturation Survey 2010 (RASS), described by Palmgren et al. (2010). The General Housing Characteristic data set of the 2010 census contains counts of occupied households by tenure. Based on distance we assign census counts at the tract level to each bus in order to calibrate the total number of renter and owner occupied households. Similarly, we use the distance between the zip code centroid and the bus to assign the RASS responses to each bus. The survey, besides recording appliance ownership per survey participant, includes their tenure status. We assume that the fraction of household under each of the household types is that of the RASS survey and we multiply this fraction by the census counts per bus to estimate the final number of households under each of the categories that we analyze.

A final piece of the demand side is that corresponding to commercial and industrial customers. The test system comes with total load profiles per bus. The commercial and industrial load at each node of the network model is the difference between the total and aggregated residential demand at each node. We assume the latter quantity to be equal to the baseline profiles multiplied by the households counts.

4.5 An Analysis of Residential Rate Structures in California

This analysis explores efficiency, as well as the distributional and environmental impacts of residential electricity rate structures. We use our model to quantify these metrics for five different tariffs that constitute plausible future residential rates in the Californian electricity sector.

Because our analysis focuses on long-run impacts, ideally one would have to compare net surplus distributions with respect to wealth levels under the different pricing regimes in our study. This approach is impractical, however, for at least two reasons. First, to the best of our knowledge information on households wealth for California is not publicly available. Thus, having this input would require an indirect calculation, which is an effort beyond the scope of this research. A second reason is that our model does not directly account for household wealth. There is no explicit mechanism linking this metric with either short- or long-run decisions. We assume, alternatively, that the level of wealth of a household translates into differences in its technology options. Wealthier customers have access to a wider variety of technologies.

Consistently, we split the population according to whether they can or cannot adopt distributed energy resources. The splitting criteria is household tenure status. That is, we assume homeowners have enough resources to purchase DERs while renters do not. Even though this criteria reflects the reality in California in the past decade (Borenstein, 2015), with better financing alternatives and new business models such as community solar[15] our assumption may not be adequate for future analyses.

An additional clarification relates to the specific rate structures we use in our analysis. An important element determining the final definition of the time-of-use schedules are the time windows associated with the different charges. Commonly, these rate structures distinguish valley and peak periods and also seasons in order to set the volumetric charges. The traditional approach is to consider existing time windows as inputs. In this analysis we take a different path which avoids two difficulties involved with the traditional approach. One is that existing time windows are likely inadequate for future system conditions. In California only a small fraction of households have been enrolled in TOU's programs. As TOU becomes the default rate for residential customers and the *net load*[16] shape changes due to the increasing penetration of renewable generation, the existing time windows will likely be obsolete. Furthermore, a TOU with a demand charge rate has not yet been implemented at the residential level in California. A second difficulty is the sub-optimality of setting time windows exogenously. This makes the comparison among rates inconsistent because our framework computes optimal retail prices for the RTP and FR programs.

In order to avoid these shortcomings, we consider the most flexible type of TOU possible. That is, tariffs in which the energy charge can vary hourly and across seasons but not for days occurring in the same season. Similarly, we assume a demand charge that changes across seasons. Our preliminary analysis indicates that four to five time-windows, depending on the season, can approximate with no meaningful efficiency loss the hourly windows. The results we report in this analysis, however, correspond to the hourly energy charges.

A final clarification relates to the DERs we include in our exercise. Besides considering all residential customers having AMI and HEMS, originally, homeowners were able to adopt either rooftop solar PV systems or battery storage units. Preliminary results indicated that the latter two DERs were not cost-effective alternatives. No customer under any of the rates we study, nor under current or projected economic parameters for these technologies, adopted these DERs. In the case of the solar PV systems, this result indicates that the factors driving the current adoption levels of this DER are policies specially designed to promote this technology. These include the California Solar Initiative (CSI), federal subsidies and the increasing block rate structure for residential customers. Because the CSI is not effective anymore, and the future of federal subsidies is uncertain, we do not include these policies in our analysis. On the other hand, even though increasing block structures are being phased

[15] We refer the reader to Huijben and Verbong (2013) for further discussion.

[16] Net load is the net of the aggregated demand, or system load, and the non-dispatchable generating technologies.

out, a surcharge for high monthly consumption will remain. This makes the study of this structure relevant. Future versions of this analysis will include this rate.

In the case of battery storage systems, our preliminary results do not make a case against this technology. It simply reflects the fact that our model only accounts for the energy arbitrage value that a battery storage unit can create. In addition, however, this technology can provide ancillary services for the distribution grid and serve as a mean of transportation when being part of an electric vehicle. Because our model does not capture any of these value streams, we do not include this technology in the final analysis.

Aggregated efficiency gains

Table 4.2 shows efficiency gains of four tariffs with respect to the base case scenario: the flat rate structure. At the level of the Western Interconnection, the RTP rate achieves greater efficiency gains followed by the time-of-use combined with a critical peak pricing program. The TOU combined with a demand charge produces similar gains but the time-of-use program alone only increases the net benefit by one half of the value when combined with another program. All programs reduce the aggregated benefit - or gross surplus - of the residential sector. However, the reductions in costs more than compensate the reductions in gross surplus.

In terms of efficiency increases, the same ranking does not hold when focusing on the residential sector in California. The main difference is that the TOU&DC rate structure increases the net benefit the least. This is the result of differences in bill reductions for customers inside and outside California. The presence of a transmission network explains this outcome. In our framework, the bill of a customer is equal to the multiplication of the *locational marginal prices* (LMPs)[17] by her consumption profile. The topology of the network significantly influences the magnitude of the LMPs at different nodes and, thus, the household bill at different locations. Differences in LMPs then explain differences in the distribution of bill reductions, in and outside California. In the case of the TOU&CPP rate, customers outside California capture an important fraction of the cost reductions with respect to the FR case. In all the other cases, on the other hand, the Californian residential sector captures most of the reductions in costs.

The average efficiency gains per household are mild at best, being not greater than 2 dollars per month. Importantly, in all cases, with the exception of the RTP program, the gains appear insufficient to justify the implementation of time-varying rate structures. Implementing any time-varying rate requires the deployment of AMI. Estimates of the cost of this infrastructure vary. However, one can construct a reasonable range using the documentation of pilot projects conducted under the American Recovery and Reinvestment Act of 2009 in DOE (2012b,d). Considering the cost of AMI, the net of the average expenditure per household on advanced metering infrastructure and the operational savings, plus the cost

[17] In many jurisdiction, in particular in California, the wholesale electricity prices differ at different nodes of the transmission network, reflecting network congestion and transmission losses. These nodal prices are called locational marginal prices.

Table 4.2: Benefits and costs: Changes with respect to flat rate structure

Level	Tariff	Net benefit	Benefit	Cost	Net benefit as a percentage of the cost
		[millions $/year]			[%]
Western Interconnection	RTP	340	-100	-440	1.50
	TOU & CPP	155	-38	-193	0.68
	TOU & DC	137	-22	-159	0.61
	TOU	77	-22	-99	0.34
California's residential sector	RTP	274	-100	-374	5.18
	TOU & CPP	46	-38	-84	0.87
	TOU & DC	176	-22	-198	3.28
	TOU	82	-22	-104	1.53
		[$/year]			
Average per household in California	RTP	22	-8	-30	5.18
	TOU & CPP	4	-3	-7	0.87
	TOU & DC	14	-2	-16	3.28
	TOU	6	-2	-8	1.53

of a standard meter, a reasonable approximation of this cost lays between 1 and 2.5 dollars per month. The lower bound at least doubles the gains of TOU and TOU&CPP, warning against the deployment of AMI if these tariffs are the pricing alternatives. Even tough in California AMI is already deployed, unless the cost of AMI decreases, in the long run the state might do as well with simpler rates and a simpler infrastructure.

Implications for different households

Figure 4.1 shows the distribution of households across net surplus gains with respect to the flat rate tariff. The figure has four panels, one per each type of household we distinguish in this analysis. In terms of the average gain, the ranking that we observe at the household level in Table 4.2 also holds when disaggregating per type of household. While RTP remains the most beneficial rate structure, the combination of a TOU and critical peak pricing program is the least favorable. Even some household types would be better off with a simple flat rate tariff than with the TOU&DC structure.

For all rates, the average net surplus gain is different for different households. Those with central air conditioning have a greater average surplus gain when compared to households without this appliance. This difference translates in turn into homeowners having a greater average surplus gain than renters. This happens simply because the proportion of homeowners with central AC is greater than 50%, and the opposite is true for renters. If one consider home-ownership as a proxy for wealth, wealthier customers benefit more from the rate structures we explored in this analysis.

Customers with no AC systems experience small average net surplus gains and this metric has small variance across households. The small increase in net surplus is driven by the elasticity we assume for the baseline consumption. The demands of households with no AC system are inelastic. This small elasticity helps to explain the small variance as well. In addition, other two elements influence the variance. One is the fact that we use one baseline profile for all households. Another is that the LMPs show small variation in the demand nodes of the California portion of the network.

The net surplus gains vary more across households with central AC systems. The variance is driven by differences in the temperature profiles at the different locations. Interestingly, the order in terms of net surplus gains induced by the different temperature profiles is different for each rate we explore. For instance, in the case of real-time pricing locations with higher within-day temperature variance and higher average temperature tend to have greater net surplus gains. One observes a similar pattern when households are enrolled in the TOU&CPP and TOU programs. However, this correlation disappears when the time-of-use rate is combined with demand charges. In this latter case, households located in places where the between-day variance of temperatures is lower tend to benefit the most.

The variance in net surplus for household with AC systems suggest targeting as a strategy for implementing time-varying rates. In particular, the TOU&DC and the RTP program appear to be the most attractive alternatives. However, the non-trivial relationship between surplus gains and temperature profiles suggests that regulators should analyze carefully where to implement these structures.

On carbon emissions

A final element we explore in this analysis is how the different rate structures impact carbon emissions. A first observation is that not all technologies we consider are economical. Neither coal, nor biomass or nuclear are profitable. Perhaps one could have anticipated this outcome in light of the figures in table 4.1, which shows that geothermal, solar and wind dominate nuclear, biomass and coal. This is not a fair comparison, however, because these resources are of a different nature. While geothermal power plants have important geographic limitations, wind and solar are intermittent resources. Thus, one cannot discard a priori technologies with dominated economic characteristics.

A second observation is that some technologies do not change their total production profile or capacities across the rate scenarios. Consistently, those technologies do not alter their carbon emissions. This technologies include hydro and wind generating power plants. We expected hydro power generation to be invariant because it was exogenous in this analysis. The invariance of wind generation, on the other hand, is a outcome of the model.

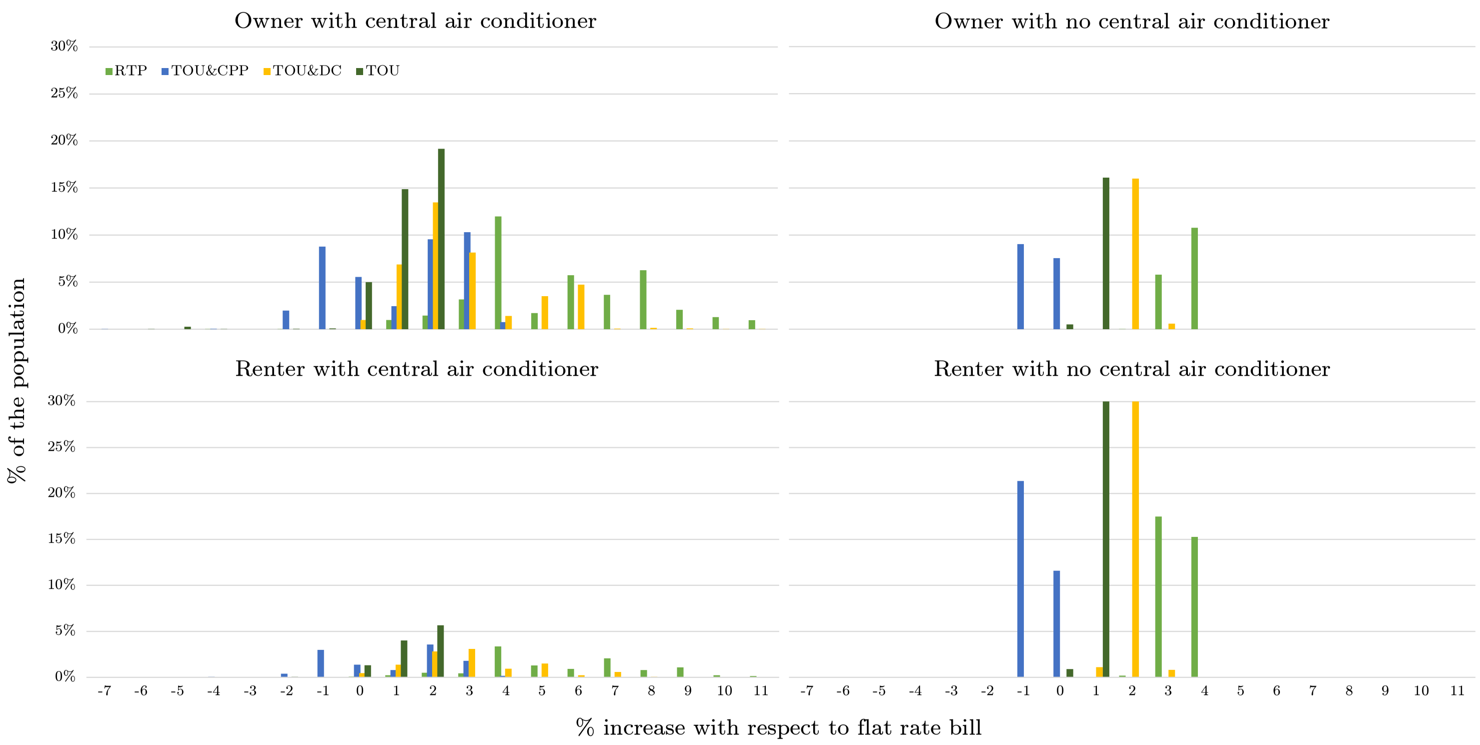

Figure 4.1: Distribution of households across net surplus gains

Table 4.3: Capacity, production and emissions changes with respect to FR scenario

Metric	Tariff	Gas adv. CC	Gas adv. CT	Gas conv. CC	Solar	Total	Change relative to FR total
				[MW-year]			%
Capacity	RTP	315	(8,561)	(177)	439	(7,984)	(6.09)
	TOU&CPP	(715)	(3,609)	(754)	2,449	(2,629)	(2.01)
	TOU&DC	(598)	(3,043)	(142)	1,463	(2,320)	(1.77)
	TOU	309	(1,600)	(642)	15	(1,918)	(1.46)
				[GWh/year]			%
Production	RTP	2,197	(2,900)	156	1,620	1,074	0.17
	TOU&CPP	(6,932)	(610)	(1,177)	9,268	549	0.09
	TOU&DC	(4,356)	(793)	(33)	5,536	352	0.06
	TOU	1,693	(216)	(1,135)	55	398	0.06
				[kt of CO2eq/year]			%
Emissions	RTP	725	(1,508)	55	-	(728)	(0.62)
	TOU&CPP	(2,288)	(317)	(412)	-	(3,017)	(2.56)
	TOU&DC	(1,438)	(412)	(12)	-	(1,861)	(1.58)
	TOU	559	(112)	(397)	-	49	0.04

Table 4.3 shows changes in capacity production and emissions with respect to the reference case. In addition, the table shows total change as a percentage of the Western Interconnection total for the FR scenario. We do not include unprofitable technologies nor technologies that do not vary across rate scenarios.

Agreeing with the basic insight of peak-load pricing, the total installed capacity decreases the most under real-time pricing, followed by TOU&CPP, TOU&DC and TOU. The order in terms of total installed capacity is not the same for total production. Indeed, the figures in Table 4.3 show that the order is somewhat reversed, with RTP increasing production the most. These changes, however, are a minor fraction of the total production of the Western Interconnection.

Even though production always changes positively, emissions do not. This is true for all rate scenarios with the exception of the TOU case. What happens is that solar production increases considerably in all but the TOU scenario. This suggests long-run complementarities between the demand responsiveness of the residential sector in California and solar generating plants in the Western Interconnection. Interestingly, RTP is not the rate that increases this complementarity the most. The combination between a TOU and a CPP program is the rate alternative that increases solar production and reduces emissions more notoriously.

4.6 Conclusions

We conduct an analysis of rate design in California's residential electricity sector. Beyond the applied insights, we contribute with a modeling framework to evaluate rate structures. The framework gives an important step towards bridging top down models of pricing and investment with bottom up models of household behavior. Building upon the theory of peak-load pricing, we illustrate how to modify the basic model to accommodate household heterogeneity, as well as the adoption of distributed energy resources and more general types of rate structures.

Our analysis seeks to quantify efficiency, distributional and environmental implications of rate structures that are plausible alternatives for California's future residential sector. The analysis shows that the average gains of implementing time-varying rates with respect to a simple flat rate program are rather mild, even in the real-time pricing scenario. Our results also show that factors such as the presence of an air conditioning system and the exterior temperature profile can have a meaningful impact on the surplus gains that different rates generate on households. These two results combined suggest that defaulting all residential customers into a time-of-use rate structure, which is the current path California is following for the residential sector, may not be an ideal strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

Chapter 5

Conclusion

This dissertation addresses the problem of how to price the services of a utility in regulated sectors. It is motivated by the current developments in the electricity industry, where the nature of the supply and demand side is changing. On one hand, intermittent technologies such wind or solar generation are becoming a salient feature of current generation mixes. At the demand side, the penetration of distributed energy technologies gains momentum, being a reality in some systems such as that of California or Hawaii. These technological changes create challenges and opportunities for the sector. Sound regulatory policy, in general, and judicious rate design, in particular, can help to unleash the potential of these technologies, while addressing some of their inherent challenges. Carefully constructed rate structures could send proper signals for energy consumption, and for adjusting the stock supply and demand technologies at the right times and locations. While these are desirable outcomes from an economic point of view, in addition, regulators must take into account other consequences such as how simple rates are, what are the distributional or environmental implications of a rate redesign, and what could be good strategies to roll-out new tariff structures. While a rate update could produce gains in terms of efficiency, it could score poorly in the other metrics, making it less attractive or even infeasible from a practical perspective. Having analytic tools that permit regulators and policy makers to anticipate implications, beyond just welfare changes, becomes increasingly relevant given the complexities unfolding in the utility sector.

This dissertation develops an analytic technique to systematically compare a wide variety of rate structures, including time varying rates, rates with demand charges, and block rates. With the aim of contributing to the ongoing debate on the future of electricity pricing in California, it uses the technique in an applied analysis of tariffs structures in this sector. The method is based upon a model which is a variant of the Ramsey-Boiteaux problem: the problem faced by a regulator that endogenizes the impacts of his choice of rate structures in the long-run equilibrium of the system. Our variant of the problem advances previous work by considerably increasing the realism of the regulator's problem. With our setting, researchers can explore in their analysis the impacts of transmission networks, intermittent generation technologies, and distributed energy resources, all elements missing in previous models. In addition, a key building block of our approach is a bottom up representation of consumption behavior. Households are a composite of devices, face income and technological constraints and have preferences for electricity consumption. As result, demand heterogeneity in our setting is driven by fundamentals, such weather patterns or the appliance stock of the households, among others. We exploit the richness of this model, and the availability of detailed data on appliance ownership, weather patterns and consumption behavior, in our analysis of California's future rates. We supplement this demand data with a realistic model of the western interconnection, including a simplified representation of the network's topology and a comprehensive set technical parameters for the power production technologies. A distinctive element of our analysis is the variety of perspectives it offers. We are able to asses the impacts of diverse rate updates from an efficiency point of view, while at the same time we are able to grasp the potential distributional and environmental consequences.

5.1 A Method to Compare Rate Structures

The first two chapters focus on the methodological effort. This encompasses developing sound models, a technique to solve the underlying mathematical programs efficiently, and applied exercises portraying the type of analyses a researcher could conduct with the aid of our methodologies.

In the first chapter, we introduce our first method for evaluating tariffs based on mathematical programming. In contrast to previous approaches, this technique allows consistently comparing portfolios of rates while capturing complexities emerging in modern electricity sectors. Welfare analyses conducted with the method can account for interactions between intermittent renewable generation, distributed energy resources and tariff structures.

We explore the theoretical and practical implications of the model that underlies the technique. At a theoretical level, our analysis suggests that our model compares rates which are not only socially optimal but could also be implemented. A regulator could induce the welfare maximizing configuration of the demand by properly updating the portfolio of tariffs. In the context of our model, the proper update translates into recomputing volumetric and fixed charges every time the system conditions change meaningfully as a result of a change in demand behavior. If a regulator follows this pricing rule, he will be providing incentives consistent with a welfare maximizing long-run demand configuration.

We also derive conclusions that help practical purposes. These allow us to exploit the structure of the model to construct a simple algorithm which finds globally optimal solutions of the underlying nonlinear optimization problem. A computational experiment suggests that the specialized procedure can outperform standard nonlinear programming techniques. The speed dominance is specially stark for large sized instances.

To illustrate the practical relevance of the rate analysis method, we compare portfolios of tariffs with data from two electricity systems: California and Denmark. Although portfolios with sophisticated rates create value in both systems, these improvements differ enough to advise very different portfolios. While in Denmark RTP appears unattractive, it at least deserves further revision in California. This conclusion is beyond the reach of previous techniques to analyze rates, since the key factor driving these different outcomes is the relationship between load and renewable production. The correlation between these two time series is greater in Denmark than in California. As a result, modifying load via time-variant pricing is more valuable in the latter as it is more needed to induce cheap electricity consumption. Models to compare rates that neglect the supply side, or over simplified it, cannot capture these types of effects, and thus could produce misleading conclusions.

5.2 Improving the Applicability of the Method

While useful, the model that underlies the method that we develop in the first chapter has two important limitations. One is the lack of transparency of the parameters that determine demand substitution patterns, and demand heterogeneity; the other is the narrow range of rate structures that could be studied with the technique. All models based upon the theory of Peak-Load Pricing inherit these limitations. The theory assumes that the demand, the solution of the consumer utility maximization problem, is a singleton. Thus, the models based upon this theory take this consumer demand as a primitive, instead of considering the consumer problem as the fundamental building block. While simplifying mathematical analysis, this assumption considerably restricts modeling flexibility. Important determinants of household electricity consumption behavior such as weather patterns or appliance ownership cannot be transparently included in models where a representative demand function is the starting point. In addition, while a researcher could represent a wide variety of time-varying rate structures, more complex designs such a combining a time-varying rate with a demand charge or a block rate cannot be captured. Modeling these rates requires to modify the consumer problem, breaking the basic assumptions of Peak-Load Pricing.

In the second chapter, we develop a model that tackles both limitations by considering as basic building block the consumer utility maximization problem. As a result, the technique allows researchers to consistently benchmark a large class of tariffs. It also makes possible capturing an unprecedented level of detail, making direct and transparent the inclusion of important technological aspects emerging in the utility industry. Our model finds the Ramsey-Boiteux prices while simultaneously determining production schedules, supply and demand side investments and consumption allocations.

This significant increase in flexibility comes at a cost, however. Finding a solution for this variant of the Ramsey-Boiteux problem is much more difficult. The resulting underlying mathematical program falls within the class of Bilevel Problems, which are hard to solve. Thus we dedicate an important part of the second chapter to developing a solution technique. We cast this Bilevel Model as a Mathematical Program with Equilibrium Constrains, and solve it using a variant of the Alternating Direction Method of Multipliers. This variant handles the complementary constraints in a distributed manner. It solves the MPEC iteratively, at each iteration solving independently several small MPECs and one nonlinear program. Using the technique of Fortuny-Amat and McCarl (1981), we cast the small MPECs as MIPs and develop heuristics to find reasonable starting points; we handle the nonlinear program via a conic reformulation.

A computational experiment demonstrates that our solution approach has several desirable properties for practical applications. First, it vastly outperforms Knitro, a popular commercial solver for MPECs. Second, the method shows good convergence behavior, reaching solutions of reasonable quality after a few tens of iterations. Third, the algorithm scales well with size. Part of the resolution of the problem can be distributed to various computing nodes and the solution time of the centralized step increases at a rate lower than linear.

A numerical analysis of tariff structures demonstrates the importance of the increased modeling flexibility of this method. The analysis highlights the value of being able to model DERs. It shows how the economics of rooftop PV systems impact the rate comparison, and that time-varying rates and inexpensive DERs could complement each other; the former can improve the relative value of the latter while DERs being inexpensive increases the gains of switching from a time invariant rate. In addition, the exercise shows that ignoring transmission constraints can have meaningful impacts on the analysis of rate structures. In the exercise, the presence of network constraints translates into significantly higher welfare gains when switching from a flat rate to more sophisticated structures. Omitting this element underestimates the benefits by about 3000 times. Finally, we are able to conduct a distributional analysis, which portrays how our method permits regulators and policy makers to study impacts of a rate update on a heterogeneous population. While a switch in rates could have a positive impact on the aggregate of households, it could benefit some more than others. It could even harm some customers, which can be particularly problematic if those harmed were low income households. A technique such as the one introduced in this chapter permits to anticipate these impacts, letting regulators to decide among rate structures with considerably more information that what would be available with alternative techniques.

5.3 An Applied Analysis in California

After having developed the most advanced method for analyzing rates, we use it in an applied study. The estate of California is transitioning its residential customers from an increasing block rate structure to a time-of-use program. Since 1976, when the Miller-Warren Energy Lifeline Act was enacted, California's residential consumers have had an increasing block rate as a default tariff. In the post crisis period, between 2000 and 2009, differences between tiers became more prominent (one of the crises consequences), and have been decreasing—some tiers even being consolidated—thereafter. In July of 2015, decision $D.15-07-001$ set a schedule and a series of steps to transition all residential customers to a default time-of-use rate by 2019.

In the third chapter, we embark in applied exploration of the long-run effects of this rate redesign, and analyze alternatives. In concrete, we compare the prospective structure with two variants of this program, a TOU combined with demand charges and a TOU combined with a critical peak pricing program. We also include in the comparison the cases in which households enroll in a flat rate and in a real-time pricing program. In our analysis, we focus on a scenario in which AMI and HEMS are widely adopted. Our approach allows us to capture a wide variety of temporal and spatial demand substitution patterns without the need of estimating a large number of parameters. We calibrate the model using data of appliance ownership, census household counts, weather patterns, and a model of California's electricity network. The analysis shows that the average gains of implementing time-varying rates with respect to a simple flat rate program are rather mild, not greater than 2 dollars per month, even in the real-time pricing scenario. Our results also show that factors such as the presence of an air conditioning system and the exterior temperature profile can have a meaningful impact on the surplus gains that different rates generate on households. These two results combined suggest that defaulting all residential customers into a time-of-use rate structure, which is the current path California is following for the residential sector, may not be an optimal strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

Bibliography

Acton, J. P. and Bridger M., M. (1983). Welfare Analysis of Electricity Rate Changes.

Allcott, H. (2011). Rethinking real-time electricity pricing. *Resource and Energy Economics*, 33(4):820–842.

Beaudin, M. and Zareipour, H. (2015). Home energy management systems: A review of modelling and complexity. *Renewable and Sustainable Energy Reviews*, 45:318 – 335.

Boiteux, M. (1960). Peak-load pricing. *The Journal of Business*, 33(2):157–179.

Bonbright, J. C. (1988). *Principles of public utility rates.* Public Utilities Reports, Incorporated, 1988, 2nd edition.

Borenstein, S. (2005). The Long-Run Efficiency of Real-Time Electricity Pricing. *The Energy Journal*, 26(3):93–116.

Borenstein, S. (2006). Customer risk from real-time retail electricity pricing: Bill volatility and hedgability. Technical report, National Bureau of Economic Research.

Borenstein, S. (2013). Effective and Equitable Adoption of Opt-In Residential Dynamic Electricity Pricing. *Review of Industrial Organization*, 42(2):127–160.

Borenstein, S. (2015). The Private Net Benefits of Residential Solar PV: The Role of Electricity Tariffs, Tax Incentives and Rebates. Working Paper 21342, National Bureau of Economic Research.

Borenstein, S. and Holland, S. (2005). On the efficiency of competitive electricity markets with time-invariant retail prices. *The RAND Journal of Economics*, 36(3):pp. 469–493.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2009). Chapter 11: Interior-point methods. In *Convex optimization.* Cambridge University Press, illustrated, reprint edition.

Brattle, E3, BEAR, and Aspen (2016). The impacts of a Regional ISO-Operated Power Market in California. Technical report, California Independent System Operator.

Brown, S. J. and Sibley, D. S. (1986). *The theory of public utility pricing.* Cambridge University Press.

Byrd, R. H., Nocedal, J., and Waltz, R. A. (2006). Knitro: An Integrated Package for Nonlinear Optimization. In Pillo, G. D. and Roma, M., editors, *Large-Scale Nonlinear Optimization*, number 83 in Nonconvex Optimization and Its Applications, pages 35–59. Springer US. DOI: 10.1007/0-387-30065-1_4.

Carlton, D. W. (1977). Peak load pricing with stochastic demand. *The American Economic Review*, 67(5):pp. 1006–1010.

Castro, F. A. and Callaway, D. S. (2016). Optimal rate design in modern electricity sectors. Manuscript submitted for publication.

Castro, F. A., Lara, J., and Callaway, D. S. (2016). A mathematical programming approach to utility pricing. Manuscript in preparation.

Caves, D. W., Christensen, L. R., and Herriges, J. A. (1984a). Consistency of residential customer response in time-of-use electricity pricing experiments. *Journal of Econometrics*, 26(1):179–203.

Caves, D. W., Christensen, L. R., Schoech, P. E., and Hendricks, W. (1984b). A comparison of different methodologies in a case study of residential time-of-use electricity pricing: Costbenefit analysis. *Journal of Econometrics*, 26(12):17 – 34.

Chao, H.-p. (1983). Peak load pricing and capacity planning with demand and supply uncertainty. *The Bell Journal of Economics*, 14(1):pp. 179–190.

Chao, H.-p. (2011). Efficient pricing and investment in electricity markets with intermittent resources. *Energy Policy*, 39(7):3945 – 3953. Special Section: Renewable energy policy and development.

Clarke, F. (1990). *Optimization and Nonsmooth Analysis.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611971309 DOI: 10.1137/1.9781611971309.

Crew, M. A., Fernando, C. S., and Kleindorfer, P. R. (1995). The theory of peak-load pricing: A survey. *Journal of Regulatory Economics*, 8(3):215–248.

Crew, M. A. and Kleindorfer, P. R. (1976). Peak load pricing with a diverse technology. *The Bell Journal of Economics*, 7(1):pp. 207–231.

De Jonghe, C., Hobbs, B., and Belmans, R. (2012). Optimal generation mix with short-term demand response and wind penetration. *Power Systems, IEEE Transactions on*, 27(2):830–839.

Dempe, S. and Dutta, J. (2012). Is bilevel programming a special case of a mathematical program with complementarity constraints? *Mathematical programming*, 131(1-2):37–48.

DOE (2012a). Demand reductions from the application of advance metering infrastructure, pricing programs and customer-based systems. Technical report, U.S. Department of Energy.

DOE (2012b). Demand reductions from the application of advance metering infrastructure, pricing programs and customer-based systems. Technical report, US Department of Energy.

DOE (2012c). Operations and maintenance savings from advance metering infrastructure. Technical report, U.S. Department of Energy.

DOE (2012d). Operations and maintenance savings from advance metering infrastructure. Technical report, US Department of Energy.

Drèze, J. H. (1964). Some postwar contributions of French economists to theory and public policy: With special emphasis on problems of resource allocation. *The American Economic Review*, 54(4):pp. 2–64.

EIA (2013). Updated capital cost estimates for electricity generation plants. Technical report, US Energy Information Administration.

EIA (2016). Annual Energy Outlook 2016, with Projections to 2040. Technical report, US Energy Information Administration.

Eid, C., Reneses Guilln, J., Fras Marn, P., and Hakvoort, R. (2014). The economic effect of electricity net-metering with solar PV: Consequences for network cost recovery, cross subsidies and policy objectives. *Energy Policy*, 75:244–254.

Fan, Y. and Liu, C. (2010). Solving Stochastic Transportation Network Protection Problems Using the Progressive Hedging-based Method. *Networks and Spatial Economics*, 10(2):193–208.

Faruqui, A., Harris, D., and Hledik, R. (2010). Unlocking the 53 billion savings from smart meters in the eu: How increasing the adoption of dynamic tariffs could make or break the EUs smart grid investment. *Energy Policy*, 38(10):6222 – 6231. The socio-economic transition towards a hydrogen economy - findings from European research, with regular papers.

Faruqui, A. and Sergici, S. (2010). Household response to dynamic pricing of electricity: A survey of 15 experiments. *J Regul Econ*, 38(2):193–225.

Faruqui, A. and Sergici, S. (2011). Dynamic pricing of electricity in the mid-Atlantic region: econometric results from the Baltimore gas and electric company experiment. *Journal of Regulatory Economics*, 40(1):82–109.

Faruqui, A. and Sergici, S. (2013). Arcturus: International evidence on dynamic pricing. *The Electricity Journal*, 26(7):55 – 65.

FERC (2006). Assesment of demand response & advanced metering. Staff report Docket AD06-2-000, Federal Energy Regulatory Commission.

Fortuny-Amat, J. and McCarl, B. (1981). A Representation and Economic Interpretation of a Two-Level Programming Problem. *The Journal of the Operational Research Society*, 32(9):783–792.

Frondel, M., Sommer, S., and Vance, C. (2015). The burden of Germanys energy transition: An empirical analysis of distributional effects. *Economic Analysis and Policy*, 45:89–99.

Gallant, A. R. and Koenker, R. W. (1984). Costs and benefits of peak-load pricing of electricity. *Journal of Econometrics*, 26(1):83–113.

Giallombardo, G. and Ralph, D. (2008). Multiplier convergence in trust-region methods with application to convergence of decomposition methods for MPECs. *Mathematical Programming*, 112(2):335–369.

Guo, Z. and Fan, Y. (2016). A Stochastic Multi-agent Optimization Model for Energy Infrastructure Planning under Uncertainty in An Oligopolistic Market. *Networks and Spatial Economics*, pages 1–29.

Herter, K. (2007). Residential implementation of critical-peak pricing of electricity. *Energy Policy*, 35(4):2121–2130.

Hiriart-Urruty, J.-B. and Lemarchal, C. (2013). *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media.

Holland, S. P. and Mansur, E. T. (2008). Is Real-Time Pricing Green? The Environmental Impacts of Electricity Demand Variance. *The Review of Economics and Statistics*, 90(3):550–561.

Hong, M., Luo, Z., and Razaviyayn, M. (2016). Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems. *SIAM Journal on Optimization*, 26(1):337–364.

Howrey, E. P. and Varian, H. R. (1984). Estimating the distributional impact of time-of-day pricing of electricity. *Journal of Econometrics*, 26(1):65–82.

Huijben, J. and Verbong, G. (2013). Breakthrough without subsidies? {PV} business model experiments in the netherlands. *Energy Policy*, 56:362 – 370.

IEA (2013). Secure and efficient electricity supply during the transition to low carbon power system. Technical report, OECD, Paris.

IEA (2016). Re-powering markets: Market design and regulation during the transition to low-carbon power systems. Technical report, OECD/IEA, 9 rue de la Federation, 75739 Paris Cedex 15, France.

Joskow, P. and Tirole, J. (2006). Retail electricity competition. *The RAND Journal of Economics*, 37(4):799–815.

Joskow, P. and Tirole, J. (2007). Reliability and competitive electricity markets. *The Rand Journal of Economics*, 38(1):60–84.

Joskow, P. L. (1976). Contributions to the theory of marginal cost pricing. *The Bell Journal of Economics*, 7(1):pp. 197–206.

Joskow, P. L. (2007). Regulation of natural monopoly. *Handbook of law and economics*, 2:1227–1348.

Joskow, P. L. and Wolfram, C. D. (2012). Dynamic pricing of electricity. *American Economic Review*, 102(3):381–85.

Kahn, A. E. (1988). *The economics of regulation: Principles and institutions*. MIT press.

Kök, A. G., Shang, K., and Ycel, a. (2016). Impact of Electricity Pricing Policies on Renewable Energy Investments and Carbon Emissions. *Management Science*.

Konnov, I. V. (2016). Selective bi-coordinate variations for resource allocation type problems. *Computational Optimization and Applications*, 64(3):821–842.

Lazar, J. and Gonzalez, W. (2015). Smart rate design for a smart future.

Lillard, L. A. and Aigner, D. J. (1984). Time-of-Day Electricity Consumption Response to Temperature and the Ownership of Air Conditioning Appliances. *Journal of Business & Economic Statistics*, 2(1):40–53.

Listes, O. and Dekker, R. (2005). A Scenario AggregationBased Approach for Determining a Robust Airline Fleet Composition for Dynamic Capacity Allocation. *Transportation Science*, 39(3):367–382.

Løkketangen, A. and Woodruff, D. L. (1996). Progressive hedging and tabu search applied to mixed integer (0,1) multistage stochastic programming. *Journal of Heuristics*, 2(2):111–128.

Luderer, B., Minchenko, L., and Satsura, T. (2013). *Multivalued analysis and nonlinear programming problems with perturbations*, volume 66. Springer Science & Business Media.

Luenberger, D. G. and Ye, Y. (2008). Chapter 10 Quasi-Newton methods. In *Linear and nonlinear programming*, volume 116. Springer.

Luo, Z.-Q., Pang, J.-S., and Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge University Press.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory*. Number Book, Whole. Oxford University Press, New York, NY.

Mathieu, J. L., Dyson, M. E., and Callaway, D. S. (2015). Resource and revenue potential of california residential load participation in ancillary services. *Energy Policy*, 80:76 – 87.

Murphy, F. H. and Smeers, Y. (2005). Generation Capacity Expansion in Imperfectly Competitive Restructured Electricity Markets. *Operations research*, 53(4):646–661.

NARUC (2016). Manual on distributed energy resources rate design compensation.

O'Neill, R. P., Sotkiewicz, P. M., Hobbs, B. F., Rothkopf, M. H., and Stewart Jr., W. R. (2005). Efficient market-clearing prices in markets with nonconvexities. *European Journal of Operational Research*, 164(1):269–285.

Palmgren, C., Stevens, N., Goldberg, M., Bames, R., and Rothkin, K. (2010). California Residential Appliance Saturation Survey. Technical Report CEC-200-2010-004, California Energy Commission.

Panzar, J. C. (1976). A neoclassical approach to peak load pricing. *The Bell Journal of Economics*, 7(2):pp. 521–530.

Papavasiliou, A. and Oren, S. S. (2013). Multiarea Stochastic Unit Commitment for High Wind Penetration in a Transmission Constrained Network. *Operations research*, 61(3):578–592.

Parks, R. W. and Weitzel, D. (1984). Measuring the consumer welfare effects of time-differentiated electricity prices. *Journal of Econometrics*, 26(12):35 – 64.

Phillips Jr, C. F. (1993). *The regulation of public utilities*. Public Utilities Reports, Incorporated, 1993, third edition.

Pia, A. D., Dey, S. S., and Molinaro, M. (2017). Mixed-integer quadratic programming is in NP. *Mathematical Programming*, 162(1-2):225–240.

Price, J. E. and Goodin, J. (2011). Reduced network modeling of WECC as a market design prototype. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–6. Conference Proceedings.

RAP (2011). Eletricity regulation in the us: A guide.

Reiss, P. C. and White, M. W. (2005). Household Electricity Demand, Revisited. *The Review of Economic Studies*, 72(3):853–883.

Rockafellar, R. T. and Wets, R. J.-B. (1991). Scenarios and Policy Aggregation in Optimization under Uncertainty. *Mathematics of Operations Research*, 16(1):119–147.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Sauma, E. E. and Oren, S. S. (2006). Proactive planning and valuation of transmission investments in restructured electricity markets. *Journal of Regulatory Economics*, 30(3):261–290.

Scheel, H. and Stefan, S. (2000). Mathematical Programs with Complementarity Constraints: Stationarity, Optimality, and Sensitivity. *Mathematics of Operations Research*, 25(1):1–22.

Schirotzek, W. (2007). *Nonsmooth analysis*. Springer Science & Business Media.

Schleicher-Tappeser, R. (2012). How renewables will change electricity markets in the next five years. *Energy Policy*, 48:64–75.

Schwartz, L., Wei, M., Morrow, W., Deason, J., Schiller, S., Leventis, G., Smith, S., Ling, W., Levin, T., Plotkin, S., Zhou, Y., and Teng, J. (2017). Electricity end uses, energy efficiency, and distributed energy resources baseline. study LBNL-1006983, Lawrence Berkeley National Laboratory, Berkeley, CA, US.

Sioshansi, R. (2012). OR ForumModeling the Impacts of Electricity Tariffs on Plug-In Hybrid Electric Vehicle Charging, Costs, and Emissions. *Operations Research*, 60(3):506–516.

Stanton, T. (2015). Distributed Energy Resources: Status Report on Evaluating Proposals and Practices for Electric Utility Rate Design. Status Report 15-08, National Regulatory Research Institute, 8611 Second Avenue, Suite 2C, Silver Spring, MD 20910.

Steiner, P. O. (1957). Peak loads and efficient pricing. *The Quarterly Journal of Economics*, 71(4):pp. 585–610.

Stoft, S. (2002). *Power System Economics: Designing Markets for Electricity.* Number Book, Whole. IEEE Press, Piscataway, NJ.

Takapoui, R., Moehle, N., Boyd, S., and Bemporad, A. (2017). A simple effective heuristic for embedded mixed-integer quadratic programming. *International Journal of Control*, 0(0):1–11.

Taylor, T. N., Schwarz, P. M., and Cochell, J. E. (2005). 24/7 Hourly Response to Electricity Real-Time Pricing with up to Eight Summers of Experience. *Journal of Regulatory Economics*, 27(3):235–262.

Vitina, A. (2015). Wind energy development in Denmark. In *IEA Wind Task 26 - Wind Technology, Cost, and Performance Trends in Denmark, Germany, Ireland, Norway, the European Union, and the United States: 2007-2012*, volume Chapter 1, pages 16–47. National Renewable Energy Laboratory, Golden, CO, USA, hand, m. m., ed. edition.

von Stackelberg, H. (1954). *The Theory of the Market Economy.* Oxford University Press, Oxford, England, (english translation of marktform und gleichgewicht, springer-verlag, berlin, 1934) edition.

Wei, M., Nelson, J. H., Greenblatt, J. B., Mileva, A., Johnston, J., Ting, M., Christopher Yang, Jones, C., McMahon, J. E., and Kammen, D. M. (2013). Deep carbon reductions in California require electrification and integration across economic sectors. *Environmental Research Letters*, 8(1):014038.

Wilcox, S. and Marion, W. (2008). User manual for TMY3 data sets. Technical Report NREL/TP-581-43156, National Renewable Energy Laboratory.

Willig, R. D. (1976). Consumer's Surplus Without Apology. *The American Economic Review*, 66(4):589–597.

Ye, J. J. (2005). Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *Journal of Mathematical Analysis and Applications*, 307(1):350 – 369.

Zöttl, G. (2010). A framework of peak load pricing with strategic firms. *Operations Research*, 58(6):1637–1649.

Appendix A

Appendix of Chapter 2

A.1 Proofs of Results in Section 2.3

Proof of Lemma 2.3.1

Proof. Willig (1976) shows that when focusing on goods with an associated expenditure relatively small with respect to the customer's budget, such as the consumption of electricity, consumer surplus and compensating variation are equivalent. Thus, in the context of the present study, we can write

$$\Delta Y = CS(l_2, \bar{p}^2) - CS(l_1, \bar{p}^1). \tag{A.1}$$

By definition, the change in producer surplus is

$$\Delta \Pi = \Pi(l_2, \bar{p}^2) - \Pi(l_1, \bar{p}^1). \tag{A.2}$$

$\Pi(l, \bar{p}) = 0$ for the optimal solution of (2.6). Combing this condition with (2.1) and (2.5), and letting $\mathcal{P}_h = \left\{\bar{p}^h\right\}$ for $h \in \{1, 2\}$, we conclude that $v_h = E\left[\bar{S}^I(\bar{p}^h)\right] - C(\bar{D}^I(\bar{p}^h)) - \Pi_0$. Using this fact and adding (A.1) and (A.2), the expression (2.7) follows. □

Proof of Proposition 2.3.1

Proof. First we show that $C(\cdot)$ is a convex function. Consider $\bar{d}^1$, $\bar{d}^2$ non-negative, $\psi \in [0, 1]$ and define $\psi^1 := \psi$, $\psi^2 := 1 - \psi$. We call $\bar{y}$ the block vector that has $\bar{y}_k$ in its k-th block, and $M(\bar{d})$ the problem (2.2)–(2.4) when $\bar{d}$ is the demand parameter. Let $(\bar{y}^j, x^j)$ be the optimal solution of $M(\bar{d}^j)$ for $j \in \{1, 2\}$. Since (2.2)–(2.4) is a convex problem and (2.3) is a linear constraint $(\hat{\bar{y}}, \hat{x}) := \psi^1(\bar{y}^1, x^1) + \psi^2(\bar{y}^2, x^2)$ is feasible for $M(\psi^1\bar{d}^1 + \psi^2\bar{d}^2)$. Thus it holds that

$$C(\psi^1\bar{d}^1 + \psi^2\bar{d}^2) \leq \kappa(\hat{\bar{y}}, \hat{x}), \tag{A.3}$$

where $\kappa(\cdot, \cdot)$ is the objective function of (2.2)–(2.4). Since this is a convex function we can write

$$\kappa(\hat{\bar{y}}, \hat{x}) \leq \psi^1 C\left(\bar{d}^1\right) + \psi^2 C\left(\bar{d}^2\right). \tag{A.4}$$

Now we show the concavity of $g(\cdot)$. Let $\alpha^1, \alpha^2 \in \mathcal{A}$ and $\hat{\alpha} = \psi^1\alpha^1 + \psi^2\alpha^2$. Consider the following point,

$$\hat{\bar{p}}^h := \begin{cases} \frac{\psi^1\alpha^1_h}{\hat{\alpha}_h}\bar{p}^{h1} + \frac{\psi^2\alpha^1_h}{\hat{\alpha}_h}\bar{p}^{h2} & \text{if } \alpha^1_h + \alpha^2_h > 0, \\ \psi^1\bar{p}^{h1} + \psi^2\bar{p}^{h2} & \text{otherwise,} \end{cases} \tag{A.5}$$

where p^j is an optimal solution for the problem P_{α^j}, $j \in \{1, 2\}$. Observe that by the convexity of $\bar{D}^h(\cdot)$ it holds that

$$\underbrace{\sum_{h=1}^{n} \hat{\alpha}_h \bar{D}(\hat{\bar{p}}^h)}_{d(\hat{\alpha})} \leq \psi^1 \underbrace{\sum_{h=1}^{n} \alpha^1_h \bar{D}(\bar{p}^{h1})}_{d(\alpha^1)} + \psi^2 \underbrace{\sum_{h=1}^{n} \alpha^2_h \bar{D}(\bar{p}^{h2})}_{d(\alpha^2)}, \tag{A.6}$$

which implies that the optimal solution of $M(\psi^1 d(\alpha^1) + \psi^2 d(\alpha^2))$ is feasible for $M(d(\hat{\alpha}))$. Then we can write

$$C(d(\hat{\alpha})) \leq C(\psi^1 d(\alpha^1) + \psi^2 d(\alpha^2)) \leq \psi^1 C(d(\alpha^1)) + \psi^2 C(d(\alpha^2)), \tag{A.7}$$

where the last inequality follows from the convexity of $C(\cdot)$. Because $\mathcal{P}$ is convex, $\hat{\bar{p}}$ is feasible for $P_{\hat{\alpha}}$. Thus, it holds that

$$g(\hat{\alpha}) \geq \sum_{h=1}^{n} \hat{\alpha}_h E\left[\bar{S}^h(\hat{\bar{p}}^h) - r_h\right] - C\left(d(\hat{\alpha})\right) \tag{A.8}$$

$$\geq \sum_{j=1}^{2} \psi^j \sum_{h=1}^{n} \hat{\alpha}^j_h E\left[\bar{S}^h(\bar{p}^{hj}) - r_h\right] - \psi^1 C\left(d(\alpha^1)\right) - \psi^2 C\left(d(\alpha^2)\right), \tag{A.9}$$

where (A.9) follows from (A.7) and the fact that $\bar{S}^h(\cdot)$ is concave. □

Proof of Proposition 2.3.2

Proof. The Lagrangian of (2.12)–(2.14) is

$$\begin{aligned} L = &\sum_{h=1}^{n} \alpha_h E\left[\bar{S}^h(\bar{p}^h) - r_h\right] - C\left(\sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right) \\ &+ \eta^\top(\nu - \Gamma\alpha) + \zeta^\top\alpha + \kappa^\top\bar{h}(\bar{p}), \end{aligned} \tag{A.10}$$

where η, ζ, κ are Lagrange multipliers, and $\bar{h}$ is the mapping associated to the constraint set $\mathcal{P}$, such that $\bar{p} \in \mathcal{P} \Leftrightarrow \bar{h}(\bar{p}) \geq 0$. The optimal solution of the Ramsey problem satisfies the following first order necessary condition for α_h

$$E\left[\bar{S}^h(\bar{p}^h) - r_h\right] - \frac{d}{d\alpha_h} C\left(\sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right) - \eta_i\Gamma_{ih} + \zeta_h = 0 \tag{A.11}$$

The second term can be derived using a result from sensitivity analysis for nonlinear programming problems. Define $Q_0(\bar{y}, x)$ as the objective of (2.2)–(2.4), and $\bar{Q}_1(\bar{y}, \bar{d})$ and $\bar{Q}_2(\bar{y}, x)$ as the mapping associated to constraints (2.3) and (2.4), respectively. The Lagrangian of (2.2)–(2.4) is

$$L' = Q_0(\bar{y}, x) + E\left[\bar{Q}_2(\bar{y}, x)^\top \bar{\xi}\right] + E\left[\bar{\lambda}^\top \bar{Q}_1\left(\bar{y}, \sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right)\right] \tag{A.12}$$

We can use the Theorem 6.67 in Luderer et al. (2013) to write

$$\frac{d}{d\alpha_h} C\left(\sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right) = E\left[\bar{D}^h(\bar{p}^h)^\top \nabla_{\bar{d}} L'\right] = E\left[\bar{D}^h(\bar{p}^h)^\top \bar{\lambda}\right]. \tag{A.13}$$

Theorem 6.67 requires $\bar{Q}(\bar{y}, x, \bar{d}) := (\bar{Q}_1, \bar{Q}_2)$ being uniformly bounded and (R)-regular, and the problem (2.2)–(2.4) being convex. In order to obtain $\bar{Q}$ uniformly bounded, one can reformulate (2.2)–(2.4) bounding each component of x by a number large enough so that the solution set does not change. To see the (R)-regularity of $\bar{Q}$, we note that for any $\bar{d}$ there is a point $(\bar{y}, x)$ such that $\bar{Q}(\bar{y}, x, \bar{d}) < 0$. This implies (R)-regularity for convex problems (see Definition 6.9 and Theorem 6.15 in Luderer et al. (2013)). Then, we can write

$$E\left[\bar{S}^h(\bar{p}^h) - \bar{D}^h(\bar{p}^h)^\top \bar{\lambda} - r_h\right] = \eta_i - \zeta_h. \tag{A.14}$$

Given our definition of fixed charges, the left hand side is the consumer surplus of the customers in h, whereas the right hand side is lower or equal to η_i. Thus, as long as there are customer in h, they do not have incentives to switch. □

Proof of Proposition 2.3.3

Proof. For simplicity, we denote the consumer surplus of a customer in rate h (the left hand side of (A.14)) as $cs_h(\alpha)$, when the population distribution is α. Let $\alpha(t)$ be a curve such that

$$\alpha_h(t) := \begin{cases} \alpha'_{h1} - t & \text{if } h = h_1, \\ \alpha'_{h2} + t & \text{if } h = h_2, \\ \alpha'_h & \text{otherwise,} \end{cases} \tag{A.15}$$

for $t \in [0, \alpha'_{h1}]$. A switch from $h1$ to $h2$ is a movement along this curve from t to $t + \Delta t$. Let $f(t) := g(\alpha(t))$, and note that this is a concave function since $g(\cdot)$ is concave and $\alpha(t)$ is linear.

In the steps that follow, we compute the derivative of $f(t)$ applying Theorem 6.67 in Luderer et al. (2013) to the Lagrangian of P_α

$$L = \sum_{h=1}^{n} \alpha_h E\left[\bar{S}^h(\bar{p}^h) - r_h\right] - C\left(\sum_{h=1}^{n} \alpha_h \bar{D}^h(\bar{p}^h)\right) + \kappa^\top \bar{h}(\bar{p}), \tag{A.16}$$

where we define κ and $\bar{h}(\bar{p})$ as in Appendix A.1.

We show first (iii). For Δt small enough we have that the sign of $f(0 + \Delta t) - f(0)$ is equal to the sign of

$$\frac{d}{dt} f(0) = \nabla g(\alpha')^\top \frac{d}{dt}\alpha(0) \tag{A.17}$$

$$= cs_{h2}(\alpha') - cs_{h1}(\alpha') \tag{A.18}$$

$$> 0, \tag{A.19}$$

where (A.18) follows from the fact that $\nabla g(\alpha')_h$ is equal to cs_h (as we showed in the proof of Proposition 2.3.2), and (A.19) holds because otherwise customers in $h1$ would not have incentives to switch.

Now we focus on (ii) and (iv). Note that, in view of (A.18), both claims are equivalent. By the concavity of f, we have

$$\left(\frac{d}{dt} f(0 + \Delta t) - \frac{d}{dt} f(0)\right)(0 + \Delta t - 0) \leq 0, \tag{A.20}$$

from where the result follows.

To show (i) it is enough to observe that ∇g is a continuous mapping, because g is concave and differentiable (Hiriart-Urruty and Lemarchal, 2013, pp. 282–284). Then, for small enough Δt we have

$$cs_{h2}(\alpha(0)) - cs_{h1}(\alpha(0)) > 0 \Rightarrow cs_{h2}(\alpha(\Delta t)) - cs_{h1}(\alpha(0)) > 0. \tag{A.21}$$

□

Proof of Proposition 2.3.4

Before proving the proposition, we introduce notation and make some observations. Define

$$\mathcal{A}_i = \left\{ \alpha \in \mathbb{R}_+^{n_i} : \nu_i \geq \sum_{h \in H_i} \alpha_h \right\}, \tag{A.22}$$

where $H_i = \{h_1^i, \ldots, h_{ni}^i\}$ is the set of rates associated to types i. The set $\mathcal{A}$ is the Cartesian product of these sets. Based on $\mathcal{A}_i$, we define the simplex

$$\mathcal{A}_i^0 = \left\{ v^i \in \mathbb{R}_+^{n_i+1} : \nu_i = \sum_{h \in H_i^0} v_h \right\}, \tag{A.23}$$

were $H_i^0 = H_i \cup \{h_0^i\}$, and $v^i = (v_{h_0^i}, \ldots, v_{h_{ni}^i})$. In addition, define

$$\mathcal{A}^0 \coloneqq \mathcal{A}_1^0 \times \cdots \times \mathcal{A}_{|I|}^0 \tag{A.24}$$

and the problem

$$\min \left\{ f(v) \colon v \in \mathcal{A}^0 \right\}, \tag{A.25}$$

where v is a block vector that has v^i in its i-th component, $f(v) \coloneqq g(v_{-0})$ and v_{-0} results after removing $v_{h_0^i}$ from v, for all $i \in I$.

Note that (A.25) is concave. Moreover, this problem is equivalent to (2.15). The optimal values of both problem coincide, and for $\alpha \in \mathcal{A}$ there is a unique $v \in \mathcal{A}^0$ and vice versa. Further, applying Theorem 6.67 in Luderer et al. (2013) to (A.25), we have that for $h \neq h_0^i$, $\nabla f(v)_h = cs_h$ (the consumer surplus of customers enrolled in h). We define $cs_{h_0^i} \coloneqq 0$ for all $i \in I$. Given this, $\nabla f(v) = cs$.

We also make explicit the way in which customer react to the incentives as well as how the utility updates rates as customers switch. Given the previous definitions, one can interpret an update in α as a switch between two groups, say from $h1$ to $h2$. We distinguish the following three cases:

- If $h1 = h_0^i$, new customers enroll in rate $h2$.
- If $h2 = h_0^i$, existing customers end their utility services.
- Otherwise, existing customers replace $h1$ with $h2$.

We define a mapping $s(v)$ such that if t customers switch the new distribution is $v + t \cdot s(v)$. If customers switch from $h1$ to $h2$ then

$$s(v)_h = \begin{cases} 1 & \text{if } h = h2, \\ -1 & \text{if } h = h1, \\ 0 & \text{otherwise.} \end{cases} \tag{A.26}$$

We consider that customers switch rationally but not strategically. That is, they consider current differences in consumer surpluses among rates but do not internalize the impacts of other customers' decisions on their after-switching surpluses. In terms of timing, rationality implies that the first to switch are those customers with the highest opportunity for increasing their surpluses.

We now make precise what we mean by the utility updating rates relatively fast. Suppose that the first event after the last rate update is a switch from $h1$ to $h2$, i.e., $h1$ and $h2$ produce the lowest and highest surpluses respectively. The maximum number of customers that can switch before the next update is $t(v) = \theta^p \epsilon$, for some $\epsilon > 0$ such that $x_{h1} \geq \epsilon$, $\theta \in (0, 1)$, and p being the smallest number in $\mathbf{Z}_+$ satisfying

$$f(v + \theta^p \epsilon \cdot s(v)) \geq f(v) + \beta \theta^p \epsilon \cdot s(v)^\top \nabla f(v), \tag{A.27}$$

with $\beta \in (0, 1)$. Further, we consider that the utility decreases ϵ as the difference between the maximum and minimum consumer surpluses decreases across rate updates.

Proof. Given our previous considerations, the evolution of v across time follows the iterations of the *Selective Bi-coordinate Method* described in Konnov (2016). The algorithm converges to the optimum for concave problems with compact feasible regions, and with objective functions differentiable, with continous gradient (Konnov, 2016). The problem (A.25) is concave, f is differentiable because P_α meets the conditions of Theorem 6.67 in Luderer et al. (2013), and ∇f is a continuous mapping because f is concave and differentiable (Hiriart-Urruty and Lemarchal, 2013, pp. 282–284). Thus, v converges to the Ramsey optimum. □

We acknowledge that the method described in Konnov (2016) is developed for the case in which I is a singleton. However, the reader can verify that the proof needed to show that the algorithm converges require just minor modifications when I has more than one element.

A.2 Implementation Details

Demand model calibration

We consider customers with linear demands. These functions differ in their price responsiveness but have the same intercept. Let $\bar{d}_0$ and $\bar{p}_0$ be the demand and price intercepts, and $\bar{B}^h \prec 0$ the Jacobian of the demand of customers in h. The statement $\bar{B}^h \prec 0$ indicates that $\bar{B}$ is negative definite. This assumption is implied by the strict concavity of the utility function assumed in Peak-Load Pricing. A system of demand equation that satisfies these conditions is

$$\bar{D}^h(\bar{p}) = \bar{B}^h(\bar{p} - \bar{p}_0) + \bar{d}_0, \tag{A.28}$$

with an associated gross surplus function

$$\bar{S}^h(\bar{p}) = \frac{1}{2}(\bar{p}^\top \bar{B}^h \bar{p} - \bar{p}_0^\top \bar{B}^h \bar{p}_0). \tag{A.29}$$

The latter expression results from assuming that the optimum of the utility maximization problem is interior, a standard assumption in applied microeconomics analysis. This implies that $\nabla\bar{U}(\bar{D}(\bar{p})) = \bar{p}$, which also provides an expression for the Hessian of $\bar{U}(\cdot)$. Normalizing $\bar{U}(0) = 0$, researchers can derive an expression for the Taylor expansion of $\bar{U}(\cdot)$ about $\bar{d}_0$.

It follows that $\bar{d}_0$, $\bar{p}_0$ and $\bar{B}^h$ fully determine the demand and gross surplus functions. The demand intercept $\bar{d}_0$ corresponds to the consumption baseline that we described in the previous subsection. We assume that the price intercept is a flat rate, i.e., $\bar{p}_0 = e \cdot \tau_0$ (with $e \coloneqq (1, \dots, 1)^\top$). Following a procedure similar to De Jonghe et al. (2012), we set τ_0 simply as the average long-run marginal cost of electricity for a system with aggregated demand equal to $\sum_i \nu_i \cdot d_0$ (recall that $\sum_i \nu_i$ is the total number of customers in the population).

To compute $\bar{B}^h$ we use the definition of the price elasticity matrix as follows

$$[E^h]_{tl} \coloneqq \frac{\partial \bar{D}_t^h(\bar{p})}{\partial \bar{p}_l} \cdot \frac{\bar{p}_{l0}}{\bar{d}_{t0}} \Leftrightarrow [\bar{B}^h]_{tl} = [E^h]_{tl} \frac{\bar{d}_{t0}}{\tau_0}, \tag{A.30}$$

where $[E^h]_{tl}$ is the element tl of the price-elasticity matrix of the corresponding segment. To calibrate the elasticity matrix with the information we have, we use the following procedure.

Let ε_o^h and ε_c^h be the pair of own- and cross-price elasticities of customers in h. First, we define $\hat{E}^h$, which is not necessarily symmetric or negative definite,

$$[\hat{E}^h]_{tl} \coloneqq \begin{cases} \varepsilon_o^h & \text{if } t = l, \\ \varepsilon_c^h/10 & \text{if } t \in \{l-5, \dots, l+5\} \setminus \{l\}, \\ 0 & \text{otherwise.} \end{cases} \tag{A.31}$$

To ensure concave utility functions we then compute E^h as the closest elasticity matrix to $\hat{E}^h$ consistent with a negative definite $\bar{B}^h$ via the following conic program:

$$\min_{(\bar{B}^h, E^h)} \left\{ \frac{\|E^h - \hat{E}^h\|}{\|\hat{E}^h\|} : \ [\bar{B}^h]_{tl} = [E^h]_{tl} \frac{\bar{d}_{t0}}{\tau_0} \, \forall tl, \ \bar{B}^h \prec 0 \right\}, \tag{A.32}$$

where $\|\cdot\|$ is the Frobenius norm.

Modeling a Renewable Portfolio Standard

A renewable portfolio standard is a policy in which the regulator mandates the utility to produce a minimum fraction ζ of its total energy from renewable sources. In order to include the effect of this policy in our analysis, we modify the model of the cost function (2.2)–(2.4) by including a new constraint. Let $e \coloneqq [1, \dots, 1]^\top \in \mathbb{R}^T$, the associated constraint is

$$e^\top E \left[\zeta \cdot \sum_{h \in H} \alpha_h \bar{D}^h(\bar{p}^h) - \sum_{k \in K_R} \bar{y}_k \right] \le 0, \tag{A.33}$$

where K_R is the set of qualified renewable technologies.

Appendix B

Appendix of Chapter 4

B.1 Geographic Distribution of Production Technologies

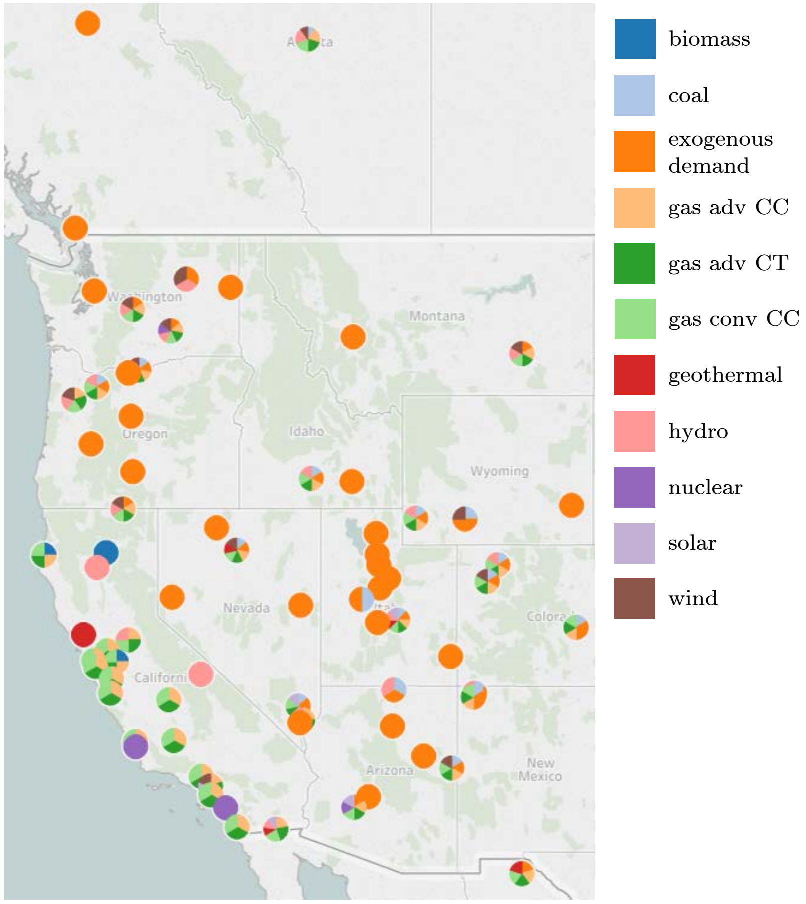

Figure B.1: Geographic distribution of generating technologies

B.2 Examples of Rates and Devices

This appendix provides two additional examples. One describes how to implement an (IBP) structure and the other how to model a battery storage system. We keep the notation in section 4.3 and define Z_m as a square matrix of zeros of m by m.

A model of an IBP structure. Under this rate a customer pays a volumetric charge p_n if its total consumption falls within the tier $n \in \{1, \dots, N\}$, that is, if $\bar{d}^\top e_{|T|} \in [q_{n-1}, q_n]$. To model this let the auxiliary variable $\hat{d}_n$ denote the consumption on tier n, and define

$$\tilde{b} := \begin{bmatrix} 0 \\ q \end{bmatrix}, \ \tilde{A} := \begin{bmatrix} e_{|T|}^\top & -e_N^\top \\ z_N z_{|T|}^\top & I_N \end{bmatrix}, \ M := \begin{bmatrix} z_{|T|} z_N^\top \\ I_N \end{bmatrix}. \tag{B.1}$$

Additionally, define the prices constraint set as follows

$$\mathcal{P} := \left\{ p \in \mathbb{R}_+^N : \, p_n \geq p_{n-1} \, \forall n > 1 \right\}. \tag{B.2}$$

Modeling the dynamics of a battery storage system. A battery of *roundtrip efficiency*[1] ζ changes its state of charge in t, s_t, in response to an energy charge, d_t^1, or discharge, d_t^2, according to the following relationship,

$$s_{t+1} = s_t + d_t^1 \sqrt{\zeta} - d_t^2 \frac{1}{\sqrt{\zeta}}.$$

In addition, the battery has a maximum storage capacity s_{max}, and a minimum state of charge s_{min}. The energy charge and discharge also have boundaries: the maximum charging and discharging rates of the battery, d_{max}.

To include this model in our framework define $\bar{d} := [d^1, d^2]$, and

$$b_{storage} := \begin{bmatrix} (s_{max} - s_1) \cdot e_{|T|} \\ -(s_{min} - s_1) \cdot e_{|T|} \\ d_{max} \cdot e_{|T|} \\ z_{|T|} \\ d_{max} \cdot e_{|T|} \\ z_{|T|} \end{bmatrix}, \ A_{storage} := \begin{bmatrix} B\sqrt{\zeta} & -B\frac{1}{\sqrt{\zeta}} \\ -B\sqrt{\zeta} & B\frac{1}{\sqrt{\zeta}} \\ I_{|T|} & Z_{|T|} \\ -I_{|T|} & Z_{|T|} \\ Z_{|T|} & I_{|T|} \\ Z_{|T|} & -I_{||T|} \end{bmatrix}, \tag{B.3}$$

with B a lower triangular matrix of ones, and s_1 the state of charge at the beginning of the first period.

B.3 Deriving a Linear Model for a TCL

The model of the thermostatically controlled load we use in this paper is similar to the one in (Mathieu et al., 2015). It originates from the thermodynamic identity,

$$C \cdot \dot{\theta}(t) = d(t) \cdot \eta - \frac{\theta(t) - \tilde{\theta}(t)}{R},$$

[1] Round trip efficiency is the maximum output per unit of energy input into the battery storage system.

in which C is the thermal capacity of the interior space, η is the coefficient of performance of central air conditioner and R is thermal resistance of the household. This identity simply sates that the variation in heat in the interior space is equal to the heat extracted by the air conditioner plus the heat that enters the space as a result of a temperature gradient. The solution of this equation is

$$\theta(t) = \theta_0 e^{\frac{-t}{CR}} + e^{\frac{-t}{CR}} \int_0^t e^{\frac{s}{CR}} \kappa(s) ds,$$

where $\kappa(s) := \left[\tilde{\theta}(t) + \eta R d(s))\right] / CR$. Thus, we can write the temperature in $t + \Delta$ as

$$\theta(t+\Delta) = e^{\frac{-\Delta}{CR}} \theta(t) + e^{\frac{-t-\Delta}{CR}} \int_t^{t+\Delta} e^{\frac{s}{CR}} \kappa(s) ds.$$

Assuming that the exterior temperature and the power consumption are constant over the interval $[t, t+\Delta]$ and equal to $\tilde{\theta}(t)$ and $d(t)$, $\kappa(\cdot)$ is also constant. Furthermore, If we normalize Δ to 1, we have that

$$\theta_{t+1} = a\theta_t + (1-a)\tilde{\theta}_t + (1-a)\eta R d_t, \tag{B.4}$$

where $a := e^{\frac{-\Delta}{CR}}$. Equation (B.4) shows a linear relationship between the consumption of the central air conditioner and the interior temperature. Considering as border condition $\theta_{|T|+1} = \theta_1$, which is consistent with the assumption of similar consecutive days, we can express (B.4) in matrix form as in (4.12). The explicit formulae follows

$$\Theta_1(\xi) = \tfrac{(1-a)\eta R}{1-a^{|T|}} \begin{bmatrix} a^{|T|-1} & a^{|T|-2} & \cdots & \cdots & \cdots & a & 1 \\ 1 & a^{|T|-1} & \cdots & \cdots & \cdots & a^2 & a \\ a & 1 & \cdots & \cdots & \cdots & a^3 & a^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ a^{t-2} & a^{t-3} & \cdots & 1 & \cdots & a^t & a^{t-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a^{|T|-2} & a^{|T|-3} & \cdots & \cdots & \cdots & 1 & a^{|T|-1} \end{bmatrix} \text{ and } \theta_2(\xi, \tilde{\theta}) = \tfrac{1}{\eta R} \Theta_1(\xi) \tilde{\theta}.$$

B.4 Modeling a Transmission Network

In order to model a transmission network, we take the standard approach in capacity expansion, e.g. (Sauma and Oren, 2006). In principle, a non-liner system, known as the *power flow equations*, describe how electricity circulates in the transmission network. The non-linearity of this system, however, adds significant complexity to the capacity expansion problem. In order to keep this model tractable researchers approximate the power flow equations. In this paper we use the *dc approximation*, which corresponds to a linearization of the original system. The dc approximation considers a vector w of net imports at each node of the network, a matrix $PTDF$ that maps net imports to flows across each of the network's edges, and a bounds for the flows, (f^+, f^-). There are two conditions associated to the approximation. One expresses constraints on flows,

$$f^+ \geq PTDFw \geq f^-,$$

the other establishes that the summation off all net imports is zero. To incorporate this network model in our framework, modify pricing problem adding these conditions. In addition, modify (4.16) as follows,

$$\sum_h \alpha_h d_{h\omega n} = \sum_{k \in K} y_{\omega k n} + w_{\omega n} \colon \lambda_\omega, \tag{B.5}$$

where n indexes the nodes of the network.