ON PRICE RESPONSIVE CONSUMER BEHAVIOR IN ELECTRICITY MARKETS: TO MACHINA ECONOMICUS FROM HOMO AGENS

A Dissertation

by

JAEYONG AN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTORATE OF PHILOSOPHY

Chair of Committee,	Panganamala R. Kumar
Committee Members,	Le Xie
	I-Hong Hou
	Suman Chakravorty
Head of Department,	Miroslav Begovic

December 2019

Major Subject: Electrical and Computer Engineering

ABSTRACT

The electricity power market is well known for its highly volatile nature due to its innate variability characteristic of demand and the absence of practical bulk storage at reasonable cost. Any discordance between rapid fluctuation in wholesale prices and near flat retail prices not only incurs economic inefficiency in terms of social welfare, but also creates price-inelastic wholesale demand which severely exacerbates the volatility of wholesale electricity prices. While the market has a fundamental dynamic nature, the behavioral aspect of power consumption in response to price changes is not well understood. This necessitate to develop a empirical modeling methodology of demand which can potentially provide practical insights into demand response.

In the former part of this work, we focus on dynamic aspect of demand response in Chapter 2. We first show that (i) demand is well responsive to outlier high price surges, and (ii) demand response can incur a certain amount of delay. Examining further data, it appears that demand is responsive to anticipated prices. This is in conformity with our previous observations on the inertia of demand, and testing the hypothesis that demand actually responds to anticipated prices rather than actual real time prices is an important next step. While it is impractical to obtain a particular individual's own price prediction, We propose to test the hypothesis with day-ahead electricity prices (DAP). In addition, as an initial step toward the derivation of a quantitative model of electricity load and price, we propose a model of "appliance" usage as a representative basic component of electricity load.

In the latter part of this work, we investigate more fundamental aspect of data-centric modeling in Chapter 3. First, we show the limitation of pure data-centric modeling strategy by proving that having a perfect knowledge on the joint distribution on price and load does not identify the load behavior in response to price. As it turns out that the causal structure of the variables of interest is the central matter that determines load behavior identifiability, we derive a minimal identifiable causal structure of demand response from the preexisting economic theories. Based on the discovered causal structure, we propose a minimal Bayesian model representation called

"stochastic neuron" which connects machine learning technique to demand response modeling. We show that a stochastic neuron is an explainable tool as expressive as an ordinary neural network, and well extends the arguments from "appliance" usage model.

DEDICATION

To Suhyun, Claire, and my mother and father.

ACKNOWLEDGMENTS

I would like to thank my supervisor Panganamala R. Kumar for his insightful comments and discussions throughout writing and composing this material, and carefully reviewing this thesis. I owe special thanks to him for his support and encouragement throughout my graduate student years in Texas A&M. In addition, I would like to thank him for all that he has taught me over the years. I would further like to thank Le Xie for his invaluable comments and discussions. I thank the other two members of my dissertation committee I-Hong Hou and Suman Chakravorty for inspiring discussions and suggestions. I appreciate Steve Puller for sharing the dataset for this study.

Above all, I thank my wife Suhyun for her trust in me and her selfless support. Without her devotion, it would be beyond the bounds of possibility to complete my PhD study. I thank my daughter Claire for her enormous patience. As a daughter of a busy and inattentive grad school dad, she was a truly empathetic daughter who did her best for supporting her dad. I cannot thank my parents enough for their endless support, and I would further like to thank my parents-in-law for their care and concern for me. With love, this dissertation is dedicated to them.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

Part 1, faculty committee recognition

This work was supervised by a dissertation committee Professor Panganamala R. Kumar of the Department of Electrical and Computer Engineering.

Part 2, student/advisor contributions

All work for the dissertation was completed by the student, under the advisement of Professor Panganamala R. Kumar and Le Xie of the Department of Electrical and Computer Engineering. The dataset used for this dissertation was provided by Professor Steve Puller of the Department of Economics.

Funding Sources

Graduate study was supported by NSF under Contract Nos. CCF-0939370, ECCS-1546682, IIS-1636772, ECCS-1547075, and CNS-1239116.

NOMENCLATURE

□	The End of the Proof (*Quod Erat Demonstrandum*)
ACE	Average Causal Effect
ACF	Autocorrelation Function
AIC	Akaike's Information Criterion
AMI	Advanced Metering Infrastructure
ANOVA	analysis of variance
AR	Autoregressive
ARX	Autoregressive Exogenous
AS	Ancillary Service
AutoML	Automated Machine Learning
BIC	Bayesian Information Criterion
BFN	Bayesian Feedforward Network
C/I	Commercial or Industrial
CR	Competitive Retailers
CRR	Congestion Revenue Right
CSC	Commercially Significant Constraint
DAG	Directed Acyclic Graph
DAM	Day-Ahead Market
DAP	Day-Ahead Price
DR	Demand Response
ELBO	Evidence Lower Bound
ERCOT	Electric Reliability Council of Texas

FFN	feedforward network
GMM	Gaussian Mixture Model
IQR	Interquartile Range
ISO	Independent System Operator
LMP	Locational Marginal Price
LS	Least Squares
MAP	Maximum A Posteriori
MDL	Minimum Description Length
MDP	Markov Decision Process
MML	Minimum Message Length
MRAP	Most Rational Account Principle
MLE	Maximum Likelihood Estimate
MSE	Mean Squared Error
NAS	Network Architecture Search
NBFN	Neurosymmetric Bayesian Feedforward Neural Network
NS	Non-spinning Service
OCP	Optimal Choice Problem
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
POMDP	Partially Observable Markov Decision Process
PUCT	Public Utility Commission of Texas
RD	Down Regulating Reserve Service
RR	Responsive Reserve Service
RTP	Real Time Prices
RTRP	Real-Time Retail Pricing

RU	Up Regulating Reserve Service
RUC	Reliability Unit Commitment
SE	Standard Error
SEM	Structural Equation Modeling
SN	Stochastic Neuron
TCR	Transmission Congestion Right
TF	Transfer Function

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Electricity is an example of a real-time commodity, as its production and consumption occur simultaneously. Such simultaneity originates from the economic infeasibility of storing electricity. The lack of efficient storage stresses system operators to dispatch and adjust generation to match the load and generation at every moment. Without such real-time balancing, power systems would suffer instability. On the other hand, a market is a system to lead supply and demand into a balanced status of an equilibrium in an efficient fashion. Therefore, real-time operation and the associated markets are both crucial elements in the electricity industry.

Balancing generation and load is a complicated task due to various generation characteristics. Load volatility can be caused by weather condition changes, while generation and transmission volatility can be caused by generator tripping or transmission congestion. These factors can result in abrupt and drastic changes in electricity price. Singularly, the electricity power market is a market characterized by innate volatility of demand as well as the absence of inexpensive bulk storage. In addition, it is expected that the volatility of supply will be more severe if renewable energy becomes a major part of the energy source mix.

While extreme price fluctuation is a widely observed phenomenon in the restructured electricity wholesale competitive markets, end-customers in most electricity markets do not face frequent price changes. Wholesale electricity prices vary from hour to hour, but retail prices do not change for months in most electricity markets. Such discordance between rapid fluctuation in wholesale prices and near flat retail price not only incurs economic inefficiency in terms of social welfare, but also creates price-inelastic wholesale demand which severely exacerbates volatility of wholesale electricity prices. In practice, wholesale electricity prices sometimes vary by an order of magnitude in real-time. Moreover, a combination of inelastic demand and the inherent real-time nature of the market makes electricity markets vulnerable to the exercise of market power [14].

Smart grid refers to a flexible and cost effective power delivery network transferring power between a diverse set of energy suppliers and informed power consumers. Among many smart

grid investments, a major one around the world is the massive deployment of advanced metering infrastructure. The payback from this major investment in data infrastructure is anticipated to be (a) enhanced flexibility from demand response participation for smart aggregators; and (b) improved real-time situational awareness for grid operators. While streaming data in the smart grid provides unprecedented opportunities to transform grid operation, it appears that most prior research in this area falls into two categories: (i) Data-driven static analytics tailored for power system domain applications, which do not capture the underlying coupling between the data and the dynamics in complex human-physical power systems [24] [25]; and (ii) Model-based system theoretical studies which are difficult to scale up to permit real-world testing [26] [27].

Among the smart grid technologies, *demand response* (DR) provides a key mechanism to extract flexibility from informed consumers. It is supposed that introducing DR will provide a wide range of benefits, especially with respect to system operability and market efficiency. In terms of market efficiency, one main objective of DR is to manage consumption of power in response to the supply condition. Concerning that, traditional economic theory provides a simple and insightful static model of optimal pricing for a given supply and demand condition through the concept of demand elasticity. There is extensive literature on DR design based on this traditional model [31].

However, because of its static setting, the traditional model has an innate limitation in explaining the time-varying characteristic that a market typically exhibits. The time-evolving behavior pattern of an economic agent is commonly observed in many economic activities; in fact, the evolutionary process of economical behavior is an inevitable consequence of the inherent inertia involved. Specifically, a sudden change of price or demand may impact supply, but most production processes necessarily require a certain amount of time to alter their production speed, and demand also is slow to change its consumption pattern in response to a price change. Such delays in making decisions and acting in response to the price changes necessarily results in a dynamic system.

From the perspective of dynamic systems, the market clearing price in a static traditional model can be interpreted as the stable price after market equilibrium. In this regard, the instantaneous

fluctuation of price caused by the change of supply conditions may not impact real time demand instantaneously, but rather impacts it slowly due to the composition of the inertia of demand and the effects of past price changes. However, such a mechanism may not be clearly captured by the traditional model. While the traditional model provides an eventual asymptote after which the market stabilizes into an equilibrium state, there is specific interest in understanding the transient period resulting from the dynamical interactions between price and demand. The market mechanism based on the traditional model has a fundamental limitation when explaining demand in a market with high volatility and rapid price variation. For this reason, it is necessary to seek and develop an alternative model of demand elasticity for understanding the dynamic characteristic of DR. This is the primary goal of the Chapter 2.

On the other hand, we are concerned with the fundamental limitation of a naive data-driven approaches and applications in the latter part of this work. As it is natural that data-driven approaches for power economics domain applications gaining greater popularity along with powerful analytic tools in the big data era, there is an increasing belief among scientists that the data will do all of the speaking, and that theory will become obsolete [2]. However, we discover that it is necessary to posit a certain untested presumption to model demand response as a decision maker's behavior in general in chapter 3. In this chapter, we study the identifiability of the generic data-centric modeling of a price responsive demand from given data by addressing two core problems in modeling:

1. How to establish an identifiable demand response model $y = g(x)$ from given a joint distribution $P(X, Y)$ in the presence of unknown confounders, where x denotes price and y denotes demand.

2. How to find an irreducible representation of the acquired posterior $P(\mathbf{w})$ of the parameters with the model $g(\cdot)$, from data.

The main contributions made in chater 3 are:

1. To present a holistic methodology for modeling and analysis of price-responsive electricity demand.

2. To construct necessary postulates taking account of domain specific characteristics of price-responsive electricity demand that intrinsically limit consistent modeling.

3. To propose a new model representation that effectively displays the causal mapping between stochastic input factors and the corresponding output features.

Overall, this work addresses a complete methodology to incorporate what is now called "machine learning techniques" for the coherent modeling of an economic decision maker from non-experimental records of her response. It is our hope that this work provides a standardized approach for addressing such problems of obtaining models from data.

2. DYNAMIC MODELING OF PRICE RESPONSIVE LOADS[1]

From the viewpoint of demand response as the output of a dynamical system, this chapter shows that (i) demand is well-responsive to high price events, and (ii) demand response incurs a certain period of delay in its manifestation. As an initial step towards the derivation of a quantitative model of electricity load and price, we propose a model of “appliance" usage as a representative basic component of electricity loads. We propose to further develop a methodology for the identification from empirical data of the demand response system as an aggregated form of the proposed appliance usage model, expecting that this can lead to a greater analytic understanding of the economic efficiency of electricity markets in terms of their volatility.

This chapter is structured as follows. In the first half of Chapter 2.1, the market structure of the Electric Reliability Council of Texas (ERCOT) after its restructuring is introduced as background. The latter half of the chapter covers previous efforts by economists to address the high volatility found in electricity markets. The first half of Chapter 2.2 describes our preliminary work on the dynamic characteristic of loads in response to real time prices (RTP) [1] as well as day-ahead prices (DAP). In the latter part of the chapter, a model is proposed to explain the characteristics of load based on the empirical study. Chapter 2.3 contains concluding remarks along with a description of potential future works.

2.1 Background and Related Works

A wave of electricity restructuring from the early 2000s aimed at introducing competition ended the era of vertically integrated monopolies in many states including Texas. The major purpose of such restructuring was to improve efficiency and lower consumer costs through the incentives provided by competition. In the first half of this chapter, the market structure of the Electric Reliability Council of Texas (ERCOT) after its restructuring is introduced.

[1] © 2015 IEEE. Reprinted, with permission, from J. An, P. R. Kumar, and L. Xie, On transfer function modeling of price responsive demand: an empirical study, Proc. of *IEEE Power & Energy Society General Meeting (PESGM 2015)*, July 2015 [1].

While the primary goal of electricity markets is to maintain system reliability at least cost in a stable manner, the intense real-time balancing requirement of electricity systems combined with inelastic demand keep markets unstable, manifested as volatile prices. The latter part of this chapter covers previous efforts by economists to address the high volatility found in electricity markets.

2.1.1 Electric Reliability Council of Texas (ERCOT) Electricity Market

The objective of this section is to provide background concerning the empirical data I use in this study. The Electric Reliability Council of Texas (ERCOT) is an Independent System Operator (ISO) serving about 85% of the electricity load in Texas under regulation by Public Utility Commission of Texas (PUCT). The major responsibilities assigned to ERCOT by the restructuring in 1999 are as follows [4] [5]:

- Maintain system reliability in terms of planning and operations,
- Ensure open access to transmission system,
- Facilitate retail switching process for customer choice, [2] and
- Wholesale market settlement for electricity production and delivery.

It should be emphasized that ERCOT implemented the current ERCOT Nodal Market in December 2010. The current ERCOT electricity market is composed of the following markets [4]:

- Day-Ahead Market (DAM), a day-ahead forward energy market,
- Real-time Energy Market with locational marginal prices (LMPs),
- Congestion Revenue Right (CRR) Markets, the markets for CRR trade, [3] and
- Reliability Unit Commitment (RUC) market for ancillary service (AS).

[2] ERCOT maintains a registration system about the association between every customer and a retailer to properly share the meter consumption data between retailers and transmission providers.

[3] CRR is a financial instrument that entitles the CRR owner to be charged or receive compensation when the ERCOT transmission grid is congested.

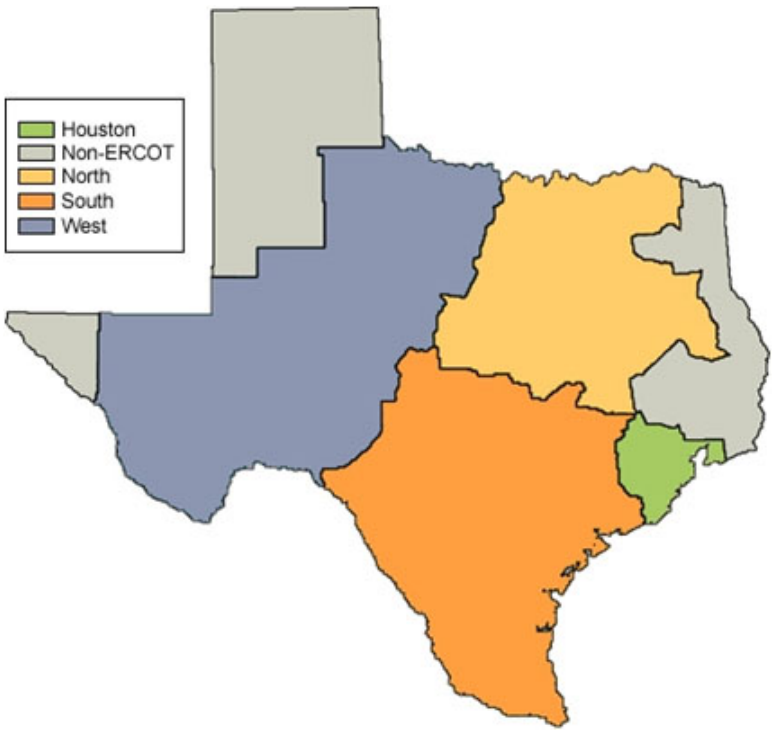


Figure 2.1: The boundaries of four zones in ERCOT in 2008 [6].

In this section, I describe the ERCOT market structure called the Zonal Market previously existing in 2008, rather than the currently existing Nodal Market structure, since the data used throughout this document is the price and consumption history of a commercial or industrial (C/I) customer in the ERCOT area from 2008. The boundaries of four zones are depicted in Figure 2.1.

The ERCOT electiricity market in 2008 was composed of following markets [4]:

- Real-time Energy Market with zonal prices,
- Ancillary Service (AS) Market running one day-ahead,
- Transmission Congestion Right (TCR) Market on the financial instruments associated with zonal congestion.

In this dissertation, I use a variety of customers from the ERCOT area in 2008. In this document, I use the data of the year 2008 from an anonymous commercial/industrial (C/I) customer in the Houston zone.

2.1.1.1 Real Time Energy Wholesale Market

The energy transaction wholesale market in ERCOT is the real-time market in which the price varies by every 15 minutes in each zone. The market is a bid-based, least-cost auction in which clearing prices are determined by security constrained unit commitment dispatch [8]. The zonal model is based on the transfer capability of the 345kV transmission system. These major transmission paths form the basis of the commercially significant constraints (CSCs). [7]

2.1.1.2 Competitive Retail Market

The ERCOT electricity retail market is a competitive market based on bilateral transactions between retail customers and the competitive retailers (CRs) [7]. CRs purchase power in the wholesale market to provide it to retail customers typically at fixed rates. Due to this structure, most retail customers are unexposed to real time electricity prices.

2.1.1.3 Ancillary Service (AS) Market

2.1.1.3.1 Ancillary service (AS) An Ancillary Service (AS) is an obligation between the ISO and a generator that the generator must provide a variety of services to maintain system reliability upon request of ISO.

There are four types of ancillary service markets in ERCOT. [5]

- Non-spinning Service (NS):

  NS is the reserves maintained by ERCOT, that are deployed for the operating hour in response to loss-of-Resource contingencies on the ERCOT System. Generator should provide off-line generation resource capacity, or reserved capacity from on-line generation resources, which are capable of being ramped to a specified output level within 30 minutes, or loads acting as a resource that are capable of being interrupted within 30 minutes and that are capable of running (or being interrupted) at a specified output level for at least one hour upon ISO's request.

- Responsive Reserve Service (RR):

RR is the operating reserves ERCOT maintains to restore the frequency of the ERCOT System within the first few minutes of an event that causes a significant deviation from the standard frequency. In addition, RR also provides reserved resources that are deployed for the operating hour in response to loss-of-Resource contingencies on the ERCOT System. Generators should provide the capacity from unloaded generation resources that are on line, with the resources controlled by high-set under-frequency relays [4] or from Direct Current (DC) tie-line [5] response. The capacity from unloaded generation resources or a DC Tie response should be deployable within 15 seconds.

- Up Regulating Reserve Service (RU):

  RU is deployed in response to a decrease in ERCOT System frequency to maintain the target ERCOT System frequency.

- Down Regulating Reserve Service (RD):

  RD is deployed in response to an increase in ERCOT System frequency to maintain the target ERCOT System frequency.

Unlike the period since December 2010, in 2008 the AS market was the only day-ahead market in ERCOT. Since the AS market concerns the obligation between the ISO and generator companies, and the main purpose of AS is to maintain system reliability, no transactions of energy can be made by any customers. For this reason, there is no economic reason for a customer to show its interest and respond to AS prices. However, my hypothesis throughout this study is that a customer may well respond to AS prices if the AS price is a reasonable predictor of RTP. According to my hypotheses, a customer with a larger demand inertia tends to be more responsive to DAP than RTP.

[4] An underfrequency relay is a device that functions to protect the load when the event comprised of system frequency decreasing below preset limits is detected. A high-set instantaneous overcurrent setting is intended to operate for close faults with high short circuit current. The setting applied is usually higher than the maximum short circuit current beyond the downstream devices (breaker or fuse). The purpose of such a setting is to prevent unselective tripping of the feeder breaker for faults on taps, which are normally cleared by the tap fuse.

[5] A DC tie-line is a transmission line between neighboring interconnections.

2.1.1.3.2 Operation of Day-Ahead Market (DAM) The DAM transpires from 6:00 AM to 6:00 PM on the day prior to the operating day. Generators may submit balanced schedules and ancillary services bids for each one hour period based on operational forecasts. After the AS for each of the four markets clears, ERCOT publishes the results.

2.1.2 Literature Review

In support of the recent extensive deployment of advanced metering infrastructure, there has been a consensus on the potential benefits of *real-time retail pricing* (RTRP) among economists and some policy makers [10] [11] [15] [16] [19] [23]. As RTRP is one of the most important issues in the power industry, there has been an abundant literature supporting the economic benefits realizable from RTRP. The first potential benefit mostly discussed in the previous literature is the allocative efficiency improvement resulting from resolving the market inefficiency caused by (near) constant retail electricity prices justified both via an econometric approach [9] and by theoretical analysis [12] [16] [18] [22] [23]. The second benefit studied is the increased robustness of market with RTRP resulting from the exercise of market power [6] [21] [13], [14] [17]. The last benefit considered is that the mitigation of demand volatility induced by real-time price signals will also relieve the need for a huge reserve capacity which incurs a large portion of the social costs [10] [20] [22]. However, all the potential economic benefits of RTRP substantially depend on how much demand is responsive to price, i.e., the price elasticity of demand [20] [22].

The allocative efficiency improvement of RTRP is well analyzed in the literature [12] [18] [22] [23], as depicted in Figure 2.2. Since the demand curve has a time variant property, it is not likely to happen that the fixed rate meets C or C', which are the optimal market clearing prices in terms of social welfare maximization. Thus, the shaded triangles ΔABC and $\Delta A'B'C'$ are the deadweight loss, the economic inefficiency caused by fixed rate the P_0.

[6] Market power is the ability of a firm to profitably raise the market price of a good or service over marginal cost mostly based on its own market share. The exercise of market power is an attempt by a firm to manipulate market price utilizing its market power. A critical condition for attaining the efficiency of a free competitive market is that every market participant is a price-taker. In that sense, a greater market power of a firm may bring a greater risk of market failure. If demands are inelastic, withholding a small portion of supply by a firm may drive the market price higher.

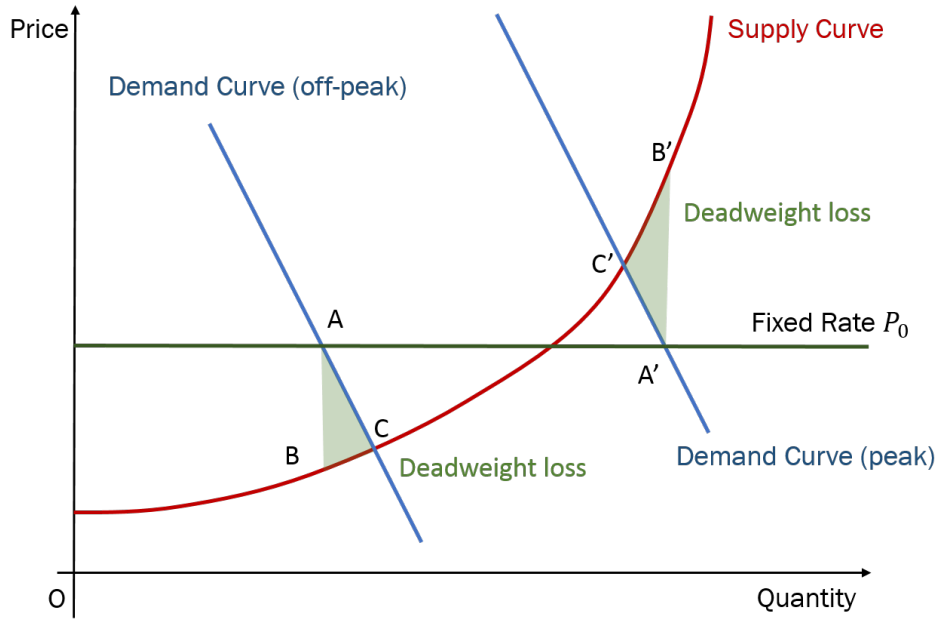


Figure 2.2: The economic inefficiency resulting from a fixed retail electricity tariff.

Due to the instantaneousness of electricity, it is reasonable to assume that the electricity for each time slot is a distinct commodity. For instantaneous and accurate market convergence to the optimal market clearing prices (C or C' in Figure 2.2), RTRP advocates argue that the ultimate real-time retail price is the optimal pricing policy [18] in terms of economic efficiency.

Although the analysis shown in Figure 2.2 seems reasonable, it requires several assumptions to be justified: (1) *Demand converges to C or C' almost immediately, in at most one time slot as determined by the market rules,* and (2) *Utility from the consumption in each time slot is attained in that same time slot.* However, both assumptions are not likely to be true for every case unless the market is slow-paced. However, the electricity consumption for running a laundry machine (say) does not provide utility until the laundering is complete, which may take more than one period. Another fundamental limitation in the demand-supply curve model is that it is difficult to obtain any insight concerning the dynamic structure from demand curve, which makes it difficult to estimate and predict demand from this model.

2.2 Preliminary Results and Discussion

In this chapter, my preliminary work on dynamic behavior of loads in response to RTP [1] as well as DAP described in the first half. In the latter half, a model is proposed to explain the characteristics of loads based on the observations made in my preliminary work.

2.2.1 Preliminary Results

In prior work [1] we have conducted an empirical study of data from two sources. In our first study, all customers are exposed to real-time prices, and we study their empirical response to such real-time prices. In the second study, we study their responses to day ahead prices as well as subsequent reactions to real-time prices as they later manifest themselves.

2.2.1.1 Empirical Transfer Function Modeling of Price Responsive Demand

We first describe our prior work [1] on the problem of determining a dynamic model of demand response to RTP from empirical data.

2.2.1.1.1 Introduction The operation of power systems has traditionally adopted the philosophy of controlling generation to balance the stochastic demand. As a result, the dynamic modeling and control of power systems has primarily focused on generator side. Governor-turbine-generator (GTG) modules from various fossil fuels have been modeled from first principles, resulting in a mature modeling taxonomy with well engrained notions such as droop characteristics and ramp rate [28]. More recently, there has also been work on modeling renewable energy sources such as wind farms as stochastic dynamic systems controlled by doubly-fed induction generators [29]. During the era when the prevailing paradigm was that supply follows demand, this modeling of supply side was enough to develop a coherent resource allocation framework for power systems. However, with the advent of demand response where demand too can be viewed as a *controllable* entity, it has become imperative to symmetrically develop models for analyzing demand response too as a dynamical system with well defined inputs and outputs. The goal of our first prior work in Section 2.2.1.1 is to develop just such a dynamic system viewpoint for the demand side.

The central contribution of this work is to exhibit from analysis of anonymous commercial/industrial

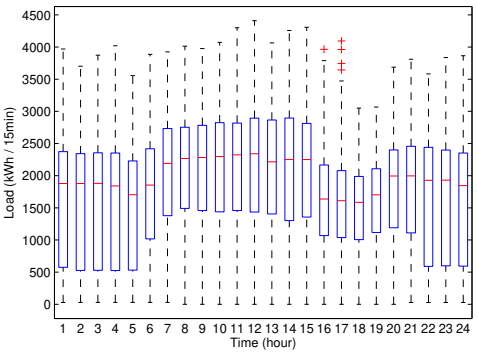
(C/I) data[7] that it is indeed possible to model demand response to prices as such a dynamical input-output system. Our contributions based purely on analysis of empirical data are twofold:

1. The response to large prices (over 95%-quantile: \$144.42) can be modeled as a *Hammerstein system*, i.e., a static nonlinearity followed by a linear transfer function [30]. Such large prices rarely persist for longer than a quarter-hour duration, and so the demand response can be viewed as a response to a price spike of a specific amplitude. We show that the resulting demand response indeed appears to be an impulse response of a nonlinear transformation of the initiating large price. The nonlinear transformation captures the fact that the demand reduction is not proportionate or linear in the price swing initiating the demand response, i.e., a $100\times$ price increase does not result in a reduction that is five times the response to a $20\times$ price increase. After accounting for this nonlinear transformation, which is typically concave since the response is sublinear, the response exhibits a reduction after a delay of about 0.75-2.5 hours, before subsequently reverting back to normal levels. Fig. 2.5 shows a typical demand response gleaned from the nine months data (Jan.1 - Sep. 30, 2008) available to and analyzed by us.

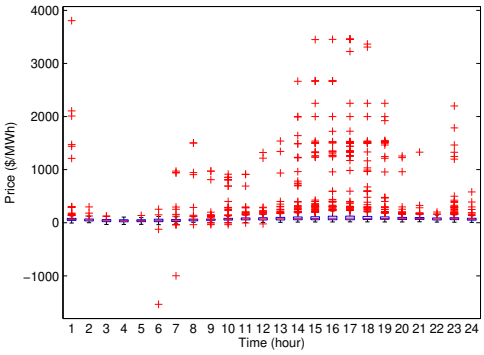
2. The response to moderate prices (up to \$144.42) can be modeled as a linear stochastic system, specifically as an *autoregressive exogenous* (ARX) system, i.e., an autoregressive (AR) system with exogenous input and white noise.

2.2.1.1.2 Preliminary Data Analysis In Fig. 2.3, the C/I load and prices from *Electric Reliability Council of Texas* (ERCOT) measured at intervals of 15 minutes from Jan. 1, 2008 to Sep. 30, 2008 is plotted with respect to time. Figs. 2.3(a) and 2.3(b) are presented as boxplots. A boxplot is a graphical approach of depicting groups of data through their quartiles. While the bottom and top of the box are the first and third quartiles, the length of each whisker is equal to $1.5\times$ interquartile range (IQR), i.e., the height of the box. The first point which can be easily observed here is that the plot on price (Fig. 2.3(b)) shows many outliers while the plot on load rarely

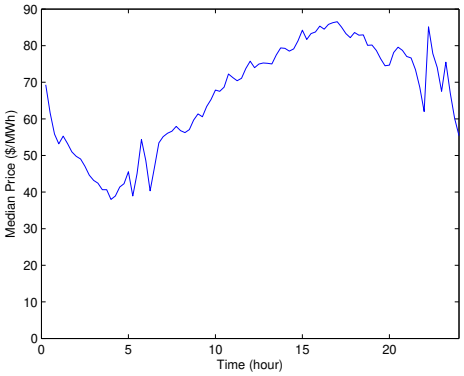
[7] Anonymous even to us.



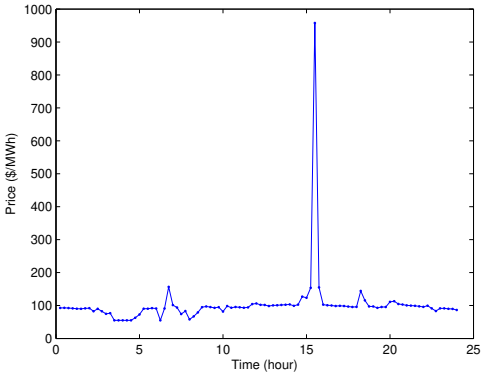
(a) The boxplot of hourly load.



(b) The boxplot of hourly prices.



(c) The median price by time of day (at 15-minute intervals).



(d) The price time series on July 9.

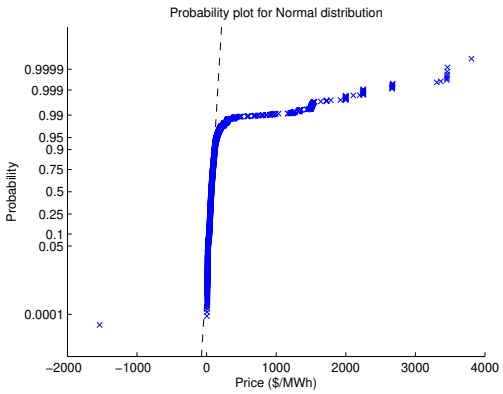
Figure 2.3: Figs. 2.3(a) and 2.3(b) show the hourly plots of a C/I load and prices from ERCOT, based on 15-minute measurements from Jan. 1, 2008 to Sep. 30, 2008. Fig. 2.3(c) shows the median price over these nine months by time of day. Fig. 2.3(d) shows a particular sample of the price series on July 9, 2008 [1]. (© 2015 IEEE)

Table 2.1: Statistics of price (P) and load (Q) [1]. (© 2015 IEEE)

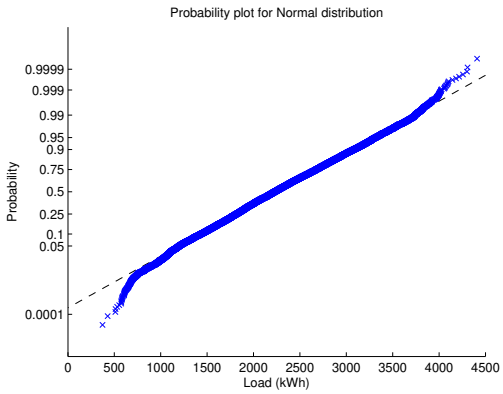
	Kurtosis	**Skewness**	**Mean**	**Std. Deviation**
P	149.0002	10.9133	67.9700	173.2434
Q	2.7712	0.1069	2246	631.0869

has them. This shows the “spiky" nature of price series, an abrupt and irregular sudden extreme price change for a very short term of 15-30 minutes duration (Fig. 2.3(d)). This makes the price highly non-normally distributed with heavy tail. Such spiky nature of prices can be explained by either the high marginal cost of production by the generators with the ability to respond rapidly to meet peak demand (e.g., gas or oil fired plants), or the bidding or withholding strategy of utility companies to maximize their profit. The other notable feature we see in Fig. 2.3 is that the load time-series exhibits a depressed demand in the afternoons, over time intervals overlapping fairly well with the time intervals which show frequent large outliers in the price time series. We infer that the depressed demand is developed as a consequence of demand response, and that the demand response is highly related to the outliers of price, because the depression can be hardly explained by the plot of the median of prices (Fig. 2.3(c)).

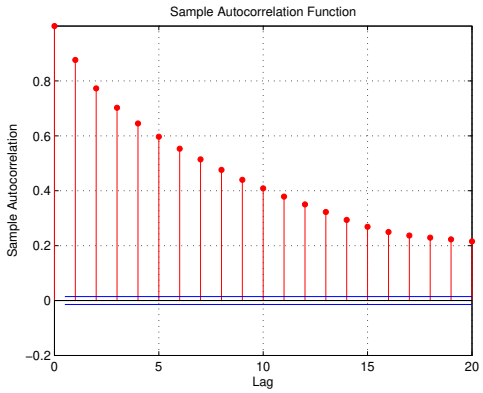
In Fig. 2.4, the statistics of the prices (P) and C/I load (Q) on workdays are depicted. From Fig. 2.4(b), the empirical probability plot of load versus the normal distribution shown by the diagonal dashed line, we can see that the empirical distribution of the load is fairly close to the normal distribution. For further validation, we can also check an estimate of the kurtosis, μ_4/σ^4, where μ_n is the nth moment about the mean and σ is the standard deviation. It is 2.77, which is close to the value 3.0 for the normal distribution. Also its skewness, μ_3/σ^3, is 0.11, which is close to the value 0 for the normal distribution. (Table 2.1). Therefore, we can conclude that the distribution of the load is very close to a normal distribution. We also see that the load shows a highly correlated structure with the past load in Fig. 2.4(c), the plot of autocorrelation (ACF) of the load, while the partial autocorrelation (PACF) of the load (Fig. 2.4(d)) decays rapidly not exceeding lag of five quarter hours (75 minutes). Taking these facts into account, it is highly likely that a simple



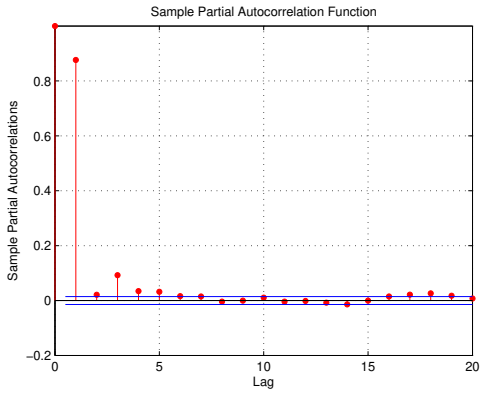
(a) The cumulative probability distribution of price (P) versus that of the normal distribution.



(b) The empirical cumulative probability distribution of the demand (Q) and the comparison with the normal distribution.



(c) The autocorrelation function (ACF) of Q. (One discrete unit of time = 15 mins)



(d) The partial autocorrelation function (PACF) of Q (The autocorrelation of each lag k after the dependence on lags $1, 2, \ldots, k-1$ is removed).

Figure 2.4: The statistics of Price (P) from ERCOT and the C/I load (Q) on workdays (i.e., weekends removed) based on 15-minute measurements from Jan. 1, 2008 to Sep. 30, 2008 [1]. (© 2015 IEEE)

autoregressive (AR) model of order 3 or 5 would sufficiently well describe the load process.

On the other hand, the first characteristic of price we can observe in Fig. 2.4(a) is that the distribution of prices is highly non-normal. At low to moderate prices, the cumulative distributions matches with the diagonal dashed line, suggesting closeness to the normal distribution. However, the top 5% of the prices severely deviate from the line, reflecting the spiky nature of electricity prices. Such a long-tail property yields huge kurtosis (149.0002) and skewness (10.9133) as presented in Table 2.1.

From the above, it is clear that it is not feasible to get a linear relationship between load and price over all values of P and Q. Hence, we conclude that it is not possible to obtain one universal linear dynamic system model between price and demand. As an alternative, it is natural to continue the analysis by assuming that there are *two transfer functions* (TFs), one for *moderate prices* which is a linear model, and one for *high prices* where there are large values. The deviation from normality of the top 5% in Fig. 2.4(a) gives a reasonably good demarcation between moderate prices and peak or high prices.

2.2.1.1.3 Estimation of Dynamic model on Load and Price From the aforementioned preliminary data analysis in previous section, we conjecture that there exist two qualitatively distinct regimes, a low to moderate price regime, and a high price regime. In the former we consider a linear transfer function between price and load with additional noise to account for uncertainty, i.e., an ARX model driven by white noise. In the high price regime we consider a concave transformation of peak prices to account for non-normality of the process. In this section, we further address this problem of identifying the dynamic model of DR.

2.2.1.1.4 Methodology We briefly discuss the estimation and validation methodology to estimate the dynamic model of DR. For the basic dynamic model of DR, we consider an ARX model driven by white noise, one of the simplest but most useful models for forecasting and control. For estimation, we consider the least squares (LS) method for estimating the unknown parameters in a linear regression model [32], [33]. To verify the existence of DR and the significance of the results of estimated parameters, we use the analysis of variance (ANOVA) method [34]. For examining

the minimum net contribution of price information to reduction of error in load estimation, we consider a two-step estimation procedure. To achieve parsimony of the model, we cross-validate the model by a random division of each complete data set under two separate conditions (i.e., moderate prices and high prices) into two sets of the same size, namely, a training set and a test set. We estimate the model from the training set and evaluate it on the test set.

2.2.1.1.5 Autoregressive Exogenous (ARX) model Denote by $\{P(t)\}_{t=1}^{N}$ and $\{Q(t)\}_{t=1}^{N}$ the time series of prices and loads, each consisting of N observations. If we denote by z^{-1} the backshift operator $z^{-1}X(t) \coloneqq X(t-1)$, the ARX model can be described as follows:

$$\alpha(z^{-1})Q(t) = \beta(z^{-1})P(t) + \epsilon_t, \tag{2.1}$$

where vectors $\alpha \coloneqq [1 \ -\alpha_1 \ -\alpha_2 \ ... \ -\alpha_m]'$ and $\beta \coloneqq [\beta_1 \ \beta_2 \ ... \ \beta_n]'$ are unknown parameters to be estimated, $\alpha(z^{-1}) \coloneqq \alpha' \cdot [z^{-i}]_{i=1}^{m}$ and $\beta(z^{-1}) \coloneqq \beta' \cdot [z^{-i}]_{i=1}^{n}$ are the characteristic and numerator polynomial of TF respectively, and ϵ_t is an error which is an independent and identically distributed (i.i.d.) noise process with $E\epsilon_t = 0$ and VAR$\epsilon_t = \sigma^2$.

2.2.1.1.6 Two-step Estimation Our primary objective in this work is to show the existence of DR and understand it from a dynamic system perspective. We employ the following two-step estimation procedure to examine the net contribution of price information to reduction of error in load estimates.

1. First estimate the regression parameters $\hat{\alpha}$, and obtain $Q_{res}(t) \coloneqq (1 - \sum_{i=1}^{m} \hat{\alpha}_i z^{-i})Q(t)$.
2. Estimate $\hat{\beta}$ using the equation $Q_{res}(t) = (\sum_{i=1}^{n} \beta_i z^{-i})P(t) + \epsilon_t$.

Then, the overall estimated ARX model is the following:

$$Q(t) = (\sum_{i=1}^{m} \hat{\alpha}_i z^{-i})Q(t) + (\sum_{i=1}^{n} \hat{\beta}_i z^{-i})P(t) + \epsilon_t, \tag{2.2}$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the LS estimators of α_i and β_i.

Table 2.2: Estimated AR Model of $Q(t)$ in the moderate price regime [1]. (© 2015 IEEE)

$Q(t) = \alpha_1 Q(t-1) + \alpha_3 Q(t-3) + \alpha_5 Q(t-5) + \alpha_0 + Q_{res}(t)$

Coeff.	**Estimate**	**SE**	**tStat**	**pValue**
α_0	238.07	13.989	17.018	8.883×10^{-64}
α_1	0.81268	0.0085477	95.075	0
α_3	0.046086	0.010267	4.4886	7.2744×10^{-6}
α_5	0.036614	0.0085466	4.284	1.8579×10^{-5}
$\sqrt{\text{MSE}}$: 301				R^2: 0.775
F-statistic vs. constant model: 8.81×10^3				p-value = 0

2.2.1.1.7 Demand Response to Moderate Price We now present an ARX model for DR in the moderate price regime, the prices below the 95%-quantile. Tables 2.4 shows the overall estimation results of the ARX model. The estimated TF of the model is:

$$TF_{\text{Low}} = \frac{-0.8555z^{-1} + 0.5273z^{-2}}{1 - 0.8127z^{-1} - 0.0461z^{-3} - 0.0366z^{-5}}. \tag{2.3}$$

This model explains 77.6% of the variance that $Q(t)$ initially possesses. Tables 2.2 and 2.3 present the results of the analysis for each of the two steps of estimation. The *Estimate* column shows the estimated coefficient value, the *SE* refer to the standard error of the estimate, the *tStat* indicates the t-statistic for a hypothesis test that the coefficient is zero, and the *pValue* is the p-value for the t-statistic.

What we see here is that though price has sufficient statistical significance due to its low p-value (0.0147), its innovative contribution to the load forecast is relatively small (less than 0.1%), and most of the change in $Q(t)$ can be explained by the past of the load itself (AR(5) model). This suggests that a moderate price has very little impact in eliciting demand response, which is also consistent with our observation in the preliminary analysis shown in preliminary data analysis part.

2.2.1.1.8 Demand Response to High Price We now present an ARX model for the high price regime, where the prices are over the 95%-quantile (144.4187 $/MWh). A sample time-series of a

Table 2.3: Estimated Linear Model of $Q_{res}(t)$ in the moderate price regime [1]. (© 2015 IEEE)

$$Q_{res}(t) = \beta_1 P(t-1) + \beta_2 P(t-2) + \beta_0 + \epsilon_t$$

Coeff.	**Estimate**	**SE**	**tStat**	**pValue**
β_0	22.506	10.054	2.2385	0.025218
β_1	-0.8555	0.42677	-2.0046	0.045043
β_2	0.5273	0.43006	1.2261	0.2202
$\sqrt{\text{MSE}}$: 301			R^2: 0.00084	
F-statistic vs. constant model: 4.22			p-value = 0.0147	

Table 2.4: The ARX Model on $Q(t)$ in the moderate price regime [1]. (© 2015 IEEE)

$$(1 - \alpha_1 z^{-1} - \alpha_3 z^{-3} - \alpha_5 z^{-5})Q(t) = (\beta_1 z^{-1} + \beta_2 z^{-2})P(t) + \epsilon_t + \epsilon_0$$

Coeff.	**Estimate**	**Coeff.**	**Estimate**
α_1	0.81268	β_1	-0.8555
α_3	0.046086	β_2	0.5273
α_5	0.036614	ϵ_0	260.126
$\sqrt{\text{MSE}}$: 301		R^2: 0.776	

Table 2.5: ANOVA Results for Fig. 2.5(b) [1]. (© 2015 IEEE)

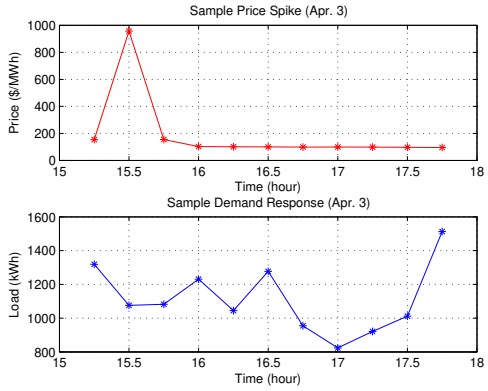
Source	SS	DF	MS	F	p-value
Groups	1.21×10^7	10	1.21×10^6	3.21	3.86×10^{-4}
Error	3.89×10^9	10351	3.76×10^5		
Total	3.90×10^9	10361			

SS: Sum of squares; DF: Degree of freedom of error;
MS: Mean square; F: F-statistic.

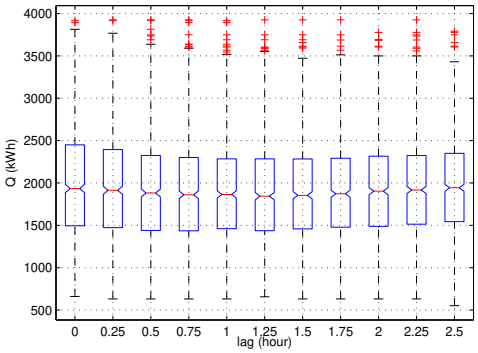
typical load evolution after a high price spike is shown in Fig. 2.5(a). Here, we observe a huge drop of the load after a one and half hour lag. Fig. 2.5(b) shows that such a load drop phenomenon is not an isolated event; we generally see such a general load drop and recovery pattern over two and half hours after price surges. The ANOVA result in Fig. 2.5(b) in Table 2.5 supports our observation that there exists a significant load drop 0.5-1.5 hours after a price surge due to its extremely low p-value (3.86×10^{-4}). This is sufficiently low to reject the null hypothesis of a constant model for the load.

In addition, we also observe that the height of price surge is correlated to the depth of load drop from Fig. 2.5(c) and 2.5(d). Fig. 2.5(c) shows the average curve of the change $\overline{Q(k)}$, at a certain level of price surge P at time t, where $\overline{Q(k)} := \frac{1}{|\mathbf{P}|} \sum_{P(t) \in \mathbf{P}} [Q(t+k) - Q(t)]$ for all $\mathbf{P}$ in a subset of sample prices $\mathbf{P} = \{P(t) : P_{\text{min}} \leq P(t) \leq P_{\text{max}}\}$ for given P_{min} and P_{max}. We see that higher P_{min} and P_{max} result in the greater load drop. Fig. 2.5(d) shows the correlation between the height of the price surge ($\Delta P = P(t) - P(t-1)$) and the load Q, which is most negatively significant after $k = 5$ quarter-hour periods (i.e., one hour and 15 minutes) from a price surge.

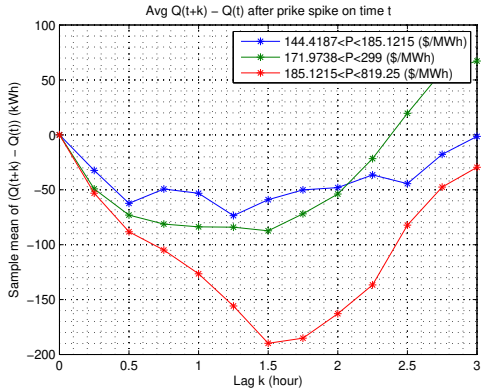
Based on the above observations, we establish a simple dynamic model between the magnitude of the price surge and the load, for high price surges. Taking into account the long-tailed characteristic of prices, we consider a linear model in the convex transformation $\log P(t)$, instead of $P(t)$, for better estimation performance. Moreover, because of the innate time-dependency on DR, we present a TF for a specific time period, from 2:00pm to 2:30pm, in this paper. The estimation results for the ARX model of DR at high price are shown in Tables 2.6, 2.7, and 2.8. The estimated



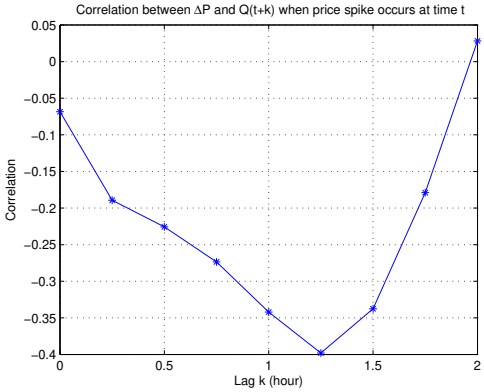
(a) The sample series of load change in response to the price spike at 3:30pm Apr. 3, 2008.



(b) The box plot of Q after a price surge (over 95%-quantile) at lag=0.



(c) The average change in Q after different levels of price surges.



(d) The correlation between ΔP and $Q(t+k)$ after a price surge.

Figure 2.5: The temporal pattern of the change of Q in response to price surge [1]. (© 2015 IEEE)

TF of the ARX model is:

$$TF_{\text{Peak}}^{2:15pm} = \frac{-220.1z^{-4}}{1 - 0.4015z^{-1} + 0.2383z^{-2} - 0.2512z^{-4}}, \tag{2.4}$$

which explains 51.2% of the variance that $Q(t)$ has. The first point we observe here is that the accuracy of the AR model for $Q(t)$ is severely degraded (R^2 = 33.2%) in Table 2.6, compared to the AR model for the moderate price regime (Table 2.2). However, we see that a relatively high portion (27%) of the variance of $Q_{res}(t)$ is explained by the estimated model of $Q_{res}(t)$ shown in Table 2.7, from which we conclude that the innovation from the price information is significant to

Table 2.6: Estimated AR Model for $Q(t)$ in the high price regime [1]. (© 2015 IEEE)

$Q(t) = \alpha_1 Q(t-1) + \alpha_2 Q(t-2) + \alpha_4 Q(t-4) + \alpha_0 + Q_{res}(t)$

Coeff.	**Estimate**	**SE**	**tStat**	**pValue**
α_0	748.26	233.72	3.2015	0.0025097
α_1	0.40153	0.11763	3.4133	0.0013678
α_2	-0.23826	0.1461	-1.6308	0.10992
α_4	0.25124	0.11516	2.1816	0.0344
$\sqrt{\text{MSE}}$: 336			R^2: 0.332	
F-statistic vs. constant model: 7.44			p-value = 0.000377	

Table 2.7: Estimated Linear Model for $Q_{res}(t)$ in the high price regime [1]. (© 2015 IEEE)

$Q_{res}(t) = \beta_4 logP(t-4) + \beta_0 + \epsilon_t$

Coeff.	**Estimate**	**SE**	**tStat**	**pValue**
β_0	1213.4	293.68	4.1316	0.00014688
β_4	-220.1	52.774	-4.1707	0.00012965
$\sqrt{\text{MSE}}$: 281			R^2: 0.27	
F-statistic vs. constant model: 17.4			p-value = 0.00013	

improve R^2 of ARX model up to 51.2% as shown in Table 2.8.

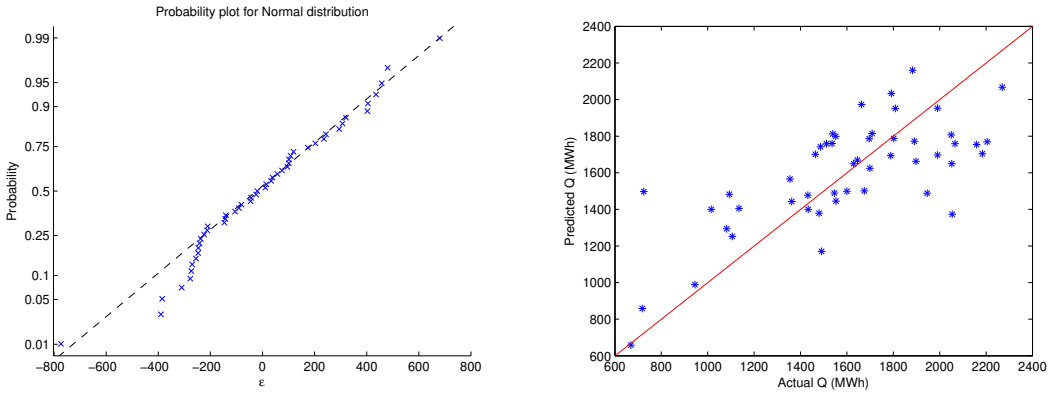
In Fig. 2.6, we check the validity of our model by sample load forecast. Figs. 2.6(a) and 2.6(b) depict the errors in the load forecast at 3:15pm after a price surge at 2:15pm. We see that the forecasted $\widehat{Q}(t)$ and the actual $Q(t)$ at t = 3:15pm are fairly well correlated (correlation ($r_{\hat{Q}Q}$ = 0.7160) in Fig. 2.6(b), and that the errors exhibit normality (Kurtosis = 3.1809) in Fig. 2.6(a).

In Fig. 2.7, we investigate the time dependency of the ARX model for a high price surge. Fig. 2.7(a) shows that the time lag in the TF has some randomness, ranging from 0.75 hours to 2.75 hours. Fig. 2.7(d) suggests that the period of the day in which DR demonstrates statistical significance is from 1:15pm to 2:45pm, with the most significant time slot being from 2:00pm to 2:30pm (Fig. 2.7(c) and 2.7(d)), for which the TF is shown in Equation (2.4).

Table 2.8: The ARX Model for $Q(t)$ in the high price regime [1]. (© 2015 IEEE)

$$(1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \alpha_4 z^{-4})Q(t) = \beta_4 z^{-4} logP(t) + \epsilon_t + \epsilon_0$$

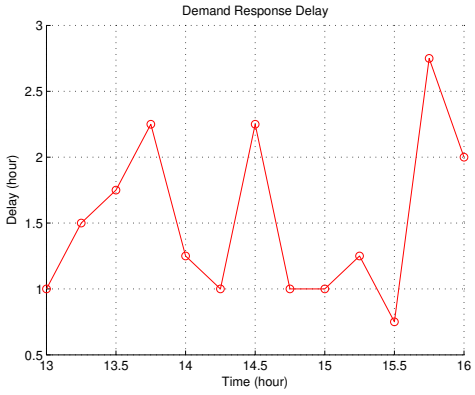
Coeff.	**Estimate**	**Coeff.**	**Estimate**
α_1	0.40153	β_4	-220.1
α_2	-0.23826	ϵ_0	1961.66
α_4	0.25124		
$\sqrt{\text{MSE}}$: 281		R^2: 0.5124	



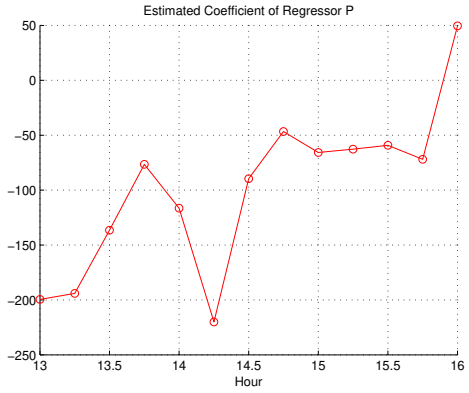
(a) The probability plot of ϵ for normal distribution (Kurtosis = 3.1809).

(b) The plot of $\hat{Q}$ over Q ($r_{\hat{Q}Q}$ = 0.7160).

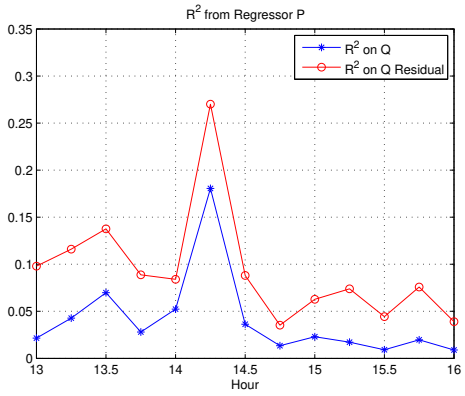
Figure 2.6: The plots of prediction error ϵ [1]. (© 2015 IEEE)



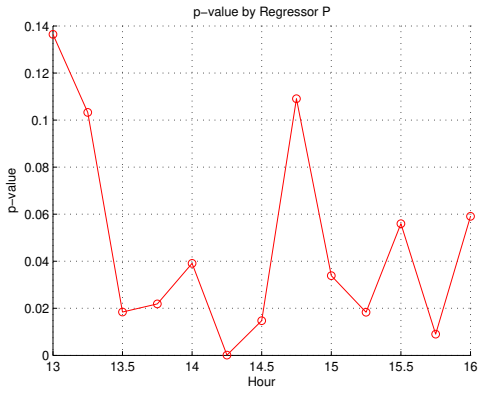
(a) Demand response delay.



(b) The estimated coefficient of regressor price spike.



(c) R^2.



(d) p-value.

Figure 2.7: The estimation results and the measurement of the goodness of fit (Fig. 2.7(c) Fig. 2.7(d)). Each point is measured over 30-minute intervals [1]. (© 2015 IEEE)

2.2.1.1.9 Summary of Prior Work In this section, I have introduced my prior work that poses the problem of modeling price responsive demand at wholesale level. Based on the empirical data acquired from ERCOT, I propose a dynamical transfer function approach to modeling such behavior. Empirical study suggests that (1) *the price responsiveness of demand may have qualitatively different behavior during "normal price" and "peak price" periods;* and (2) *there exists a demand response delay consequent on a high price surge.* The first finding can be reasonably interpreted as saying that there is little incentive for increasing power consumption at low price since we do not have efficient energy storage yet. On the other hand, this finding is in line with the observation in financial markets that a financial market tends to react more sensitively to bad news than good news [35]. The second finding shows that there exists a certain "inertia" in consumption so that it takes a certain time delay to reduce power consumption after a peak price observation.

Perhaps more important than the two specific findings is the potential value of the very approach of employing transfer functions for flexible demand modeling. This modeling approach offers many more degrees of freedom in characterizing the salient nature of power consumers as compared with classical econometric modeling of price elasticity.

2.2.1.2 A Study of Consumer Behavior in Response to Day Ahead Prices (DAP)

In Section 2.2.1.1, we have seen that there is a delayed demand response after a price spike, indicating that there is an inertial in demand. Accepting this fact and assuming that a customer is rational, the reasonable reaction of the customer is to respond and adjust its consumption based on price prediction. However, the verification of the hypothesis that demand is responsive to predicted prices requires the availability of such a price prediction.

Day Ahead Prices (DAP) can be a good source of such a price prediction available to most stake holders. While the current ERCOT electricity market is equipped with a day-ahead market for actual energy purchase, in 2008 there were day ahead markets only for ancillary services in ERCOT. Ancillary services stands for services to maintain system reliability upon the request of the ISO. For this reason, there was no actual power transacted in the day ahead market in 2008, so

that, strictly speaking, there was no economic reason for the customers to respond to DAP in 2008. Thus, it would be a valid evidence that customers are actually responsive to price anticipation if we can indeed verify the fact that customers are well responsive to DAP in this 2008 dataset.

Another point to note here is that the verification actually supports the hypothesis only if the DAP is somewhat different from RTP; otherwise it is not clear where the verified good responsiveness comes from, while it is meaningless if DAP is very different from RTP because this contradicts the assumption that DAP is a good predictor of RTP. Our data of the year 2008 from customers in Houston shows that the correlation between DAP and RTP is about 0.5 which indicates that it could be an appropriate source for the verification.

Since the verification is ongoing work, I conclude this section with Figure 2.8 which is a sample demonstrating how demand is responsive to DAP. What we can observe here is that demand is responsive to the overall shape of the price series especially in peak hours, but the responsiveness varies over time.

2.2.2 Preliminary Model on Consumer Behavior

In Section 2.2.1.1, we observe that (1) power demand responds to high price surges while it does not exhibit significant response to modest price changes; (2) there exists a demand response delay consequent on a high price surge. On the other hand, another important observation we have seen in both Sections 2.2.1.1 and 2.2.1.2 is that the responsiveness of power demand to both RTP and DAP varies over time. While such observations are explicable in a broad sense, it looks as if these conflict with the rational consumer assumption. One of the key assumptions underlying economics is that every consumer seeks the least available expense for a given amount of consumption. Hence, it is a crucial to examine whether the given power consumption data can be rationalized. Such a rationalization problem can be described as a process to demonstrate the existence of a coherent optimization problem of which the solution appears as the given result. Once such an optimization problem is identified, the identification will provide us the understanding of the mechanism of consumer behavior as a price responsive system. In this section, I propose a model attempting to explain the observed phenomena, which delineates the optimization problem

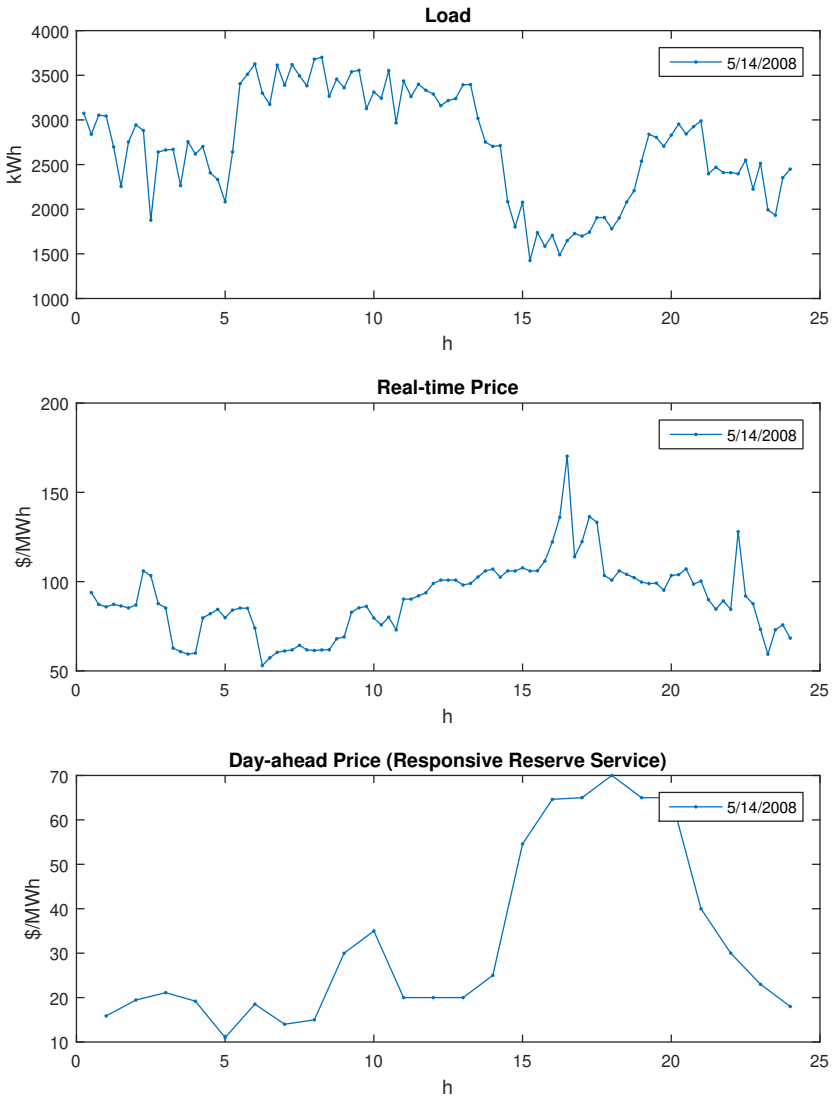


Figure 2.8: A sample of load, RTP, and DAP series on May 14, 2008. The load and RTP are from Houston, and the DAP is the price of the responsive reserve service.

for the system to be estimated.

2.2.2.1 An "Appliance" Usage Model

The identification of an optimization problem requires the identification of both an objective function and the set of feasible solutions. From the economic point of view, the objective function seems to be obvious, i.e., to minimize the total expenditure of the consumption. However, it is

not clear whether the consumption does actually minimize the total expenditure from the given data set. The correlation between the 10-month-series of price and consumption from 30 randomly picked consumers in Texas, who explicitly indicated themselves as price responsive customers, does not go below -0.1, while the average of the correlation between one-day-series of price and consumption does not go below -0.6 for both DAP and RTP. Figure 2.8 shows an example why such a phenomenon happens. While demand and price are typically negatively correlated when it is a peak hour period, in off-peak hours there is a near zero or even positive correlation on many days.

To rationalize the data, it is worth setting up a hypothesis, i.e., there exists some constraints which restrain the set of feasible solution, before we jump to the conclusion that consumers tend to be less rational in off-peak hours. This raises an interesting question: *Can we estimate the unknown constraint set of an optimization problem from the price as an independent variable which determines the objective function, with the demand as a dependent variable which is an optimal solution of the optimization problem?*

Below I formalize a model of consumption that attempts to answer the above question. For the estimation process, it is necessary to set up a model about the structure of the constraints as an initial step. Consider a consumer using an "appliance" such as a dishwasher. Accepting the idea that such constraints may be related to our natural "appliance" usage pattern, it becomes necessary to set up a model for appliance usage to identify the constraints. As a beginning, let us call the usage of an appliance for a *task* as an atomic load.

Definition 2.2.1 (Task)**.** *A task* T is a tuple (q, S, a) where,

- Load level $q \in \mathbb{R}^+$, this is the power drawn by the load while it is active;
- Session $S \equiv [t^s, t^d] \subset \mathbb{Z}$, a time interval which indicates the feasible time period over which the activation of T is feasible. Here, t^s is the *release time* of T and t^d is *deadline*. We assume the duration of $|S|$ does not exceed 24h, with $|S|$ denoting the number of (typically 15 min or 1 hour) time slots in S, i.e., $|S| = |t^d - t^s + 1|$;

- Activeness constraint $a \in \mathbb{Z}^+$, denoting the length of the total active time period of the task which needs to be mandatorily accomplished. Note that $a \leq |S|$. The total energy consumption of T is aq.

The main motivation of proposing the structure of a task as shown in Definition 2.2.1 is to set up the relationships between all the loads in each time slot. The real-time retail pricing advocates claim that the load on each time slot should be regarded as a distinct commodity because of the real-time characteristic of loads. However, the traditional demand-supply curve model does not capture how these loads are related to each other, because this model is specialized to depict the relationship between saturated demand, supply, and market clearing price, while the explanation of the dynamics of the demand or the relations with other goods is not the main issue the model deals with. The main purpose of introducing the task model is to manifest such a relationship between all the loads on each time slot. Once a task T is given, we can easily infer that the load on time slots not in S and the load on time slots within S have an independence relationship; any load change on one time slot does not affect that of the other one. However, between the two loads on the time slots within S, we can infer that they are substitutes of each other; if one time slot in the past is also active, it is less likely the another slot in the future is active, due to the activeness constraint a. On the other hand, if two time slots are both within S and adjacent, it is likely that the loads on them are in a complementary relationship if T has a delay, especially if both belong to one atomic job.

Example 1. I indicate below some examples of appliance usage.

- Running a dishwasher: $(q = 1.8kW, S = 1 \text{ day}, a = 2h)$
- Charging a mobile phone: $(q = 0.006kW, S = [\text{low battery alert, departure time}], a = 3h)$
- Running a simulation on a desktop computer: $(q = 1kW, S = [5\ PM, 8\ AM], a = 1h)$

2.2.2.2 Price Responsiveness of a Task to Real Time Prices (RTP)

In this section, I show how the proposed appliance usage model can potentially provide an understanding of dynamic demand response to RTP that we have observed in Section 2.2.1.1. I

presume that completion of a task T produces a certain level of utility $u(T)$, while the cost from energy consumption is aq. At first, it is natural to assume that a customer executes a task if and only if it is worth doing it. To address this, we also introduce a penalty for aborting a task, and denote it by $c \geq 0$. This could be a contract cancellation fee if T is a contractual task.

Assumption 2.2.1. A task T subsists if and only if the *expected net utility of its completion* is greater than the cost of its abortion.

While the introduction of Assumption 2.2.1 seems to be natural, it has a crucial implication: *the exact net utility cannot be realized until its completion*. This is because the available information to a customer is limited to real time electricity prices of past and present time slots, not those of the future, so that the one can only estimate the cost of the task, but not know it with certainty. Assumption 2.2.1 can be described in terms of the following two conditions, the initiation condition and the abortion condition of a task.

Proposition 2.2.2 (Initiation Condition)**.** A customer initiates a task if and only if there exists A such that the expected total cost is less than the utility of the task, i. e.,

$$\exists A \subset S \text{ s.t. } u(T) - a \cdot q \cdot E_A[P_t^{RTP}] > 0, \tag{2.5}$$

where $u : \{T_i : i \in \mathbb{Z}\} \to \mathbb{R}^+$ is the utility function, $A \subset S$ is a set of time slots of active power consumption by T so that $|A| = a$, and $E_A[P_t^{RTP}]$ denotes the expected power price during A.

Proposition 2.2.3 (Abortion Condition)**.** A customer aborts a task at time t_0 if and only if the utility of the task minus the minimum expected total cost of the remaining period of A is greater than the penalty for abortion,

$$u(T) - a^f \cdot q \cdot E_{A^f}[P_t^{RTP}] < -c, \quad \forall A^f, \tag{2.6}$$

where, $A^p = A \cap \{t : t < t_0\}$ is the set of past active time slots, $A^f \subset (S \cap \{t : t \geq t_0\})$ is the set of future active time slots so that $a = a^p + a^f$ where $a^p = |A^P|$ and $a^f = |A^f|$, and $c \geq 0$ is the penalty

for aborting T, e.g., there could be a contract cancellation fee if T is a contractual task.

The maximum net utility of the completion of the task T is the total utility of task T minus the minimum expected total cost, while the net utility from the abortion of the task T is zero utility minus the the total cost from the past active time slots as well as the additional penalty. Thus, a customer aborts the task T if the net utility of the completion of the task T (LHS of the inequality (2.7)) is less than the net utility from the abortion of the task T (RHS of the inequality (2.7)) for all possible $A \subset S$. i.e.,

$$\forall A \subset S \text{ such that } A^p \subset A, \;\; u(T) - aqE_A[P_t^{RTP}] < 0 - \sum_{t \in A^p} qP_t^{RTP} - c. \tag{2.7}$$

This can be restated as,

$$\begin{aligned} &\forall A^f \subset (S \cap \{t : t \geq t_0\}), \\ &u(T) - (\sum_{t \in A^p} qP_t^{RTP} + a^f qE_{A^f}[P_t^{RTP}]) < - \sum_{t \in A^p} qP_t^{RTP} - c \end{aligned}$$

so that

$$u(T) - a^f qE_{A^f}[P_t^{RTP}] < -c, \;\; \forall A^f \subset (S \cap \{t : t \geq t_0\}),$$

which is the inequality (2.6) in Proposition 2.2.3.

Here, the utility is an internal concept that cannot be measured. The initiation of tasks will appear in a stochastic manner depending on the prices in a real situation. However, the important implication of Propositions 2.2.2 and 2.2.3 is that, using both propositions, (1) We can restate the task abortion condition without involving an unmeasurable utility function, as we show below and, (2) We can gain understanding on how the observed delays in given data could be explained by shedding light on where the inertia of demand comes from.

Lemma 2.2.4. An initiated task T is aborted at t_0 if the following condition is met,

$$\forall A^f,\ \ a^f q E_{A^f}^{ex-post}[P_t^{RTP}] - aqE_A^{ex-ante}[P_t^{RTP}] > c \tag{2.8}$$

where $E_A^{ex-ante}[\cdot]$ is an expectation over A obtained before t^s and $E_{A^f}^{ex-post}[\cdot]$ is an expectation over A^f obtained at t_0.

Proof. An initiated task T should always meet the inequality (2.5) by Proposition 2.2.2,

$$u(T) - a \cdot q \cdot E_A^{ex-ante}[P_t^{RTP}] > 0. \tag{2.9}$$

On the other hand, multiplying -1 on both sides of the inequality (2.6) allows us to restate (2.6) as

$$\forall A^f, -u(T) + a^f q E_{A^f}^{ex-post}[P_t^{RTP}] > c. \tag{2.10}$$

Adding (2.10) to the inequality (2.9) reduces to (2.8). □

Lemma 2.2.4 states that a task is not aborted unless there is a significantly erroneous prediction at the time of the task initiation. This situation is not likely to happen when it is off-peak hours, because $a^f \leq a$ so that the condition (2.8) could be satisfied only when $E_{A^f}^{ex-post}[P_t^{RTP}]$ is significantly greater than $E_A^{ex-ante}[P_t^{RTP}]$.

However, our observations indicate that a price spike is likely to occur soon after a price spike occurrence, which may yield a substantial difference between $E_{A^f}^{ex-post}[P_t^{RTP}]$ and $E_A^{ex-ante}[P_t^{RTP}]$.

Observation 1. A price spike is likely to be occurred soon after a price spike occurrence if it is in peak hours.

Figure 2.9 shows a comparison of the estimated conditional probability of price spike in different situations, based on the obtained data from Houston. Each figure in Figure 2.9 shows a different time period. We can easily check from the Figure 2.9 that the conditional probability of a price spike after the occurrence of a price spike quickly reduces in off-peak hours, so that I infer that the price spike at off-peak hours is not likely to cause a task abortion. However, we also observe that

the conditional probability of a price spike after the occurrence of a price spike during peak-hours remains at a significantly higher level than the probability of price spike without any conditioning. From this, I surmise that the price spikes in peak-hours are highly likely to induce task abortion. Another notable point from Lemma 2.2.4 is that the relative position of the price spike occurrence within a session S is also a crucial factor which determines whether the task T is to be aborted. The abortion of a task T caused by a price spike is likelier if the price spike occurs near the beginning of the session S. Such a relationship is captured in the following Theorem 2.2.5.

Theorem 2.2.5. A task is aborted in response to a price spike at time t_0 if the following condition is satisfied,

$$\forall A^f, \quad \frac{a^f}{a} > \frac{E_A^{ex-ante}[P_t^{RTP}]}{E_{A^f}^{ex-post}[P_t^{RTP}]} + c^d \tag{2.11}$$

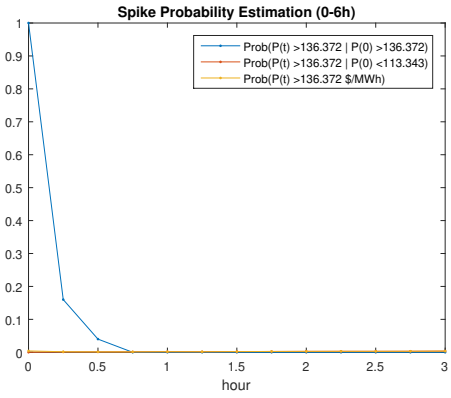
where $c^d = \dfrac{c}{aqE_{A^f}^{ex-post}[P_t^{RTP}]}$.

Proof. This is a simple rearrangement of Lemma 2.2.4. □

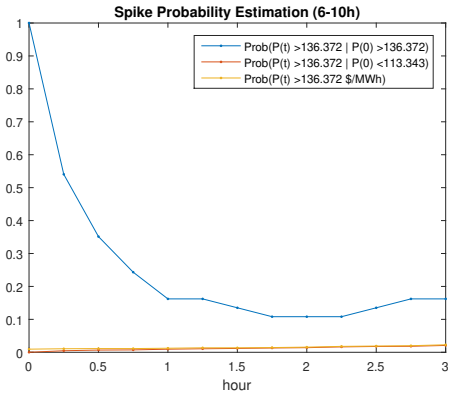
Theorem 2.2.5 shows that a greater a^f or $E_{A^f}^{ex-post}[P_t^{RTP}]$ induces a task abortion. This implies that if a price spike in peak hours, making a greater difference between $E_A^{ex-ante}[P_t^{RTP}]$ and $E_{A^f}^{ex-post}[P_t^{RTP}]$, occurs in the earlier part of S, then it is more likely to be aborted.

Besides the explanation of Theorem 2.2.5 showing why the task is only responsive to price spike, Theorem 2.2.5 also provides a candidate cause for the observed delay in response from the given data in Section 2.2.1.1. The following example illustrates how this induces a demand response delay.

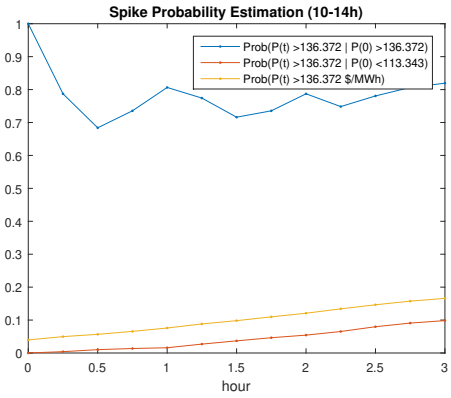
Example 2. Suppose there are two tasks T_1 and T_2, such that T_1 has a session starting at $t_s^{T_1}$, and T_2 has a session ends at $t_d^{T_1}$. Consider two scenarios as follows. In one scenario, there is a price spike at time slot t_0, which causes T_1 to be aborted, but not T_2. In the other, there is no price spike. Then, Figure 2.10 shows how the aggregated load $Q_t = Q_t^{T_1} + Q_t^{T_2}$ differs in the two scenarios. We can observe that Q_t in Figure 2.10(a) shows the pattern of a delayed response, while the session termination of S_1 is not revealed in Figure 2.10(b).



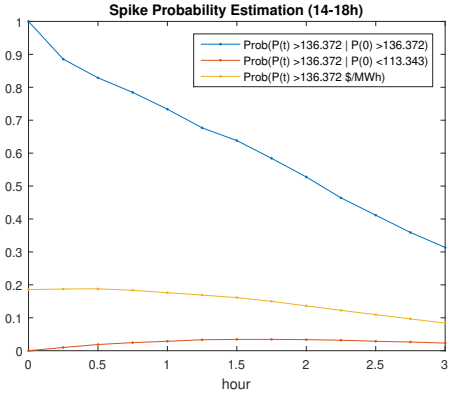
(a) 0:00 AM - 6:00 AM.



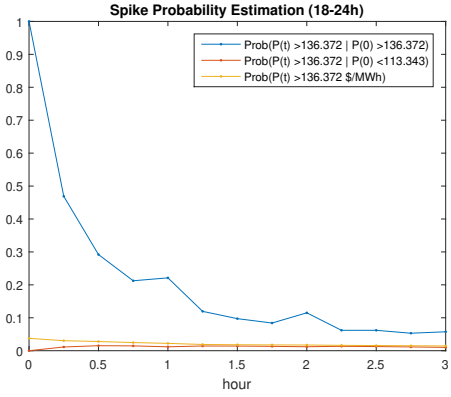
(b) 6:00 AM - 10:00 AM.



(c) 10:00 AM - 2:00 PM.

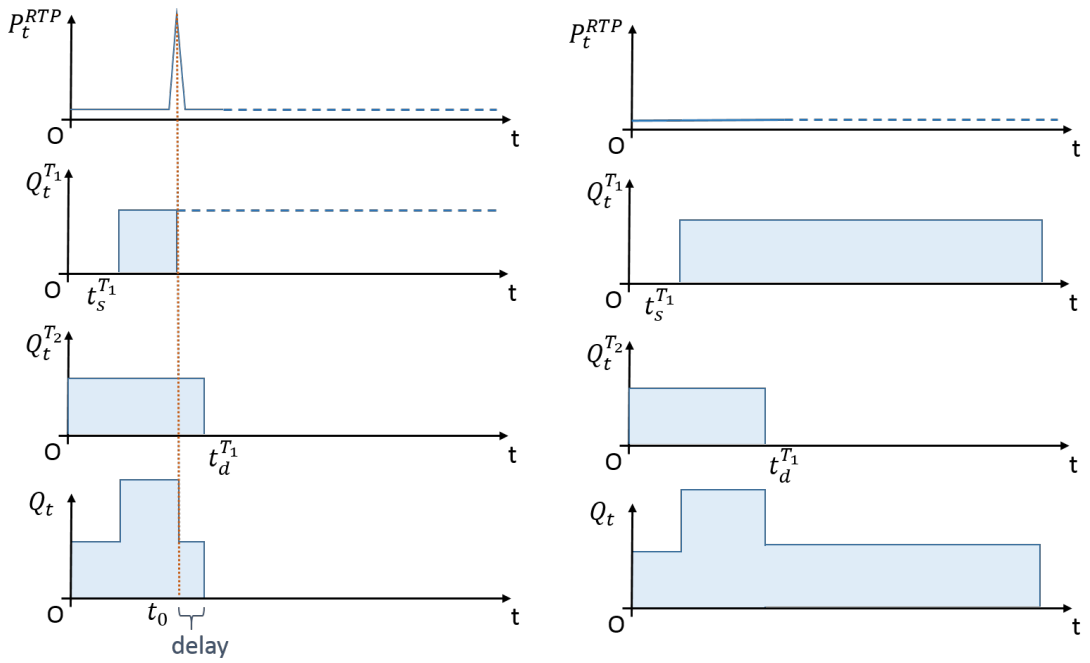


(d) 2:00 PM - 6:00 PM.



(e) 6:00 PM - 0:00 AM.

Figure 2.9: The conditional probability of price spike occurrence compared for different time periods



(a) The scenario with a price spike.

(b) The scenario without any price spike.

Figure 2.10: Load patterns with same tasks under two different price scenarios, one with a price spike, and another without a price spike.

As in Example 2, aggregated loads of multiple tasks, with different abortion chances determined by various sessions and activeness constraints, may aggregately exhibit a delayed response to price spike observed in Section 2.2.1.1.

Corollary 2.2.6. A load is responsive to P_t^{RTP} if the following conditions are met:

1. Every task has its session $S = [t^s, t^d]$ within a time slot so that $t^s = t^d$, and
2. The distributions of $u(T_i)$ and q_i are time invariant.

Proof. Since the minimal time length we can recognize is one time slot by the definition of a task (Definition 2.2.1), $a = a^f = 1$ and $A = A^f = \{t\}$. Thus, a task T_i is initiated and activated at time t only if $u(T_i)/q_i > P_t^{RTP}$ by Proposition 2.2.2.

Take two time slots t_1 and t_2 for which $P_{t_2}^{RTP} < P_{t_1}^{RTP}$. Let i be the index of the tasks in t_1 and j be the index of the tasks in t_2 such that $u(T_i) = u(T_j)$ and $q_i = q_j$ if $i = j$, indicating that $\{T_i\}$ and $\{T_j\}$ meet the condition (2) in Corollary 2.2.6.

Then, we can easily check that $u(T_i)/q_i > P_{t_2}^{RTP}$ if $u(T_i)/q_i > P_{t_1}^{RTP}$, whereas the reverse is not necessarily true when $P_{t_2}^{RTP} < u(T_i)/q_i < P_{t_1}^{RTP}$. This implies that $I \subset J$, from the fact that $u(T_i)/q_i = u(T_j)/q_j$ if $i = j$, where $I := \{i : \frac{u(T_i)}{q_i} > P_{t_1}^{RTP}\}$ and $J := \{j : \frac{u(T_j)}{q_j} > P_{t_2}^{RTP}\}$.

Therefore, $Q_{t_2} > Q_{t_1}$ where $Q_{t_1} = \sum_{i \in I} q_i$ and $Q_{t_2} = \sum_{j \in J} q_j$ are total loads for corresponding time slots, since $q_i = q_j$ if $i = j$. □

Typically, in an economic market it is assumed that every trade brings utility to both market participants, suppliers and consumers. For demand response, Corollary 2.2.6 suggests that RTRP may work well in a situation satisfying both the following conditions:

- The utility from consumption is realized at every market clearing point, and
- The demand characteristic for price is nearly time-invariant.

In contrast, demand may not be well responsive if utility necessarily involves multiple market clearing points. From this point of view, electricity has a peculiar feature in comparison to other commodities; consumer utility from electricity consumption comes from the completion of a task so that if the activeness constraint a is longer than the market clearing period, the traditional concept of price elasticity is not well defined.

2.2.2.3 Price Responsiveness of a Task to Day Ahead Prices

The appliance usage model in Section 2.2.2.1 along with Propositions 2.2.2 and 2.2.3 in Section 2.2.2.2 implies that there is a fundamental reason for a task to be responsive to price prediction rather than the actual RTP: a task is initiated or aborted based on the prediction of overall cost by the task, unless the actual price signal indicates that the prediction of price is likely to be significantly erroneous. From Theorem 2.2.5 in Section 2.2.2.2, I have shown that the responsiveness of demand to RTP, or demand inertia, critically depends on the predetermined active slots A in a task T. Although the determination of active slots A from S of a task T for the minimization of the expected expenditure crucially depends on the individual customer's price prediction, obtaining such information is generally not possible in practice, as previously mentioned in Section 2.2.1.2.

Hence, I assume that DAP represents the customer's price prediction in this section. However, unless the obtained DAP data is from an actual day ahead energy transaction market, DAP may significantly differ from RTP. While a day ahead ancillary service market may be a good prediction source for RTP, there is no reason that the ancillary service price should match up with RTP. Hence, to assume DAP as a predictor for RTP, I assume that a day series of RTPs preserves the overall *shape* of DAP for the day even though the actual price level may deviate.

Assumption 2.2.7. Consider a rearrangement of time slots so that resulting one-day series of $\{P_t^{DAP}\}$ are monotone increasing. We assume that the same rearrangement of time slots of one-day series of $\{P_t^{RTP}\}$ also results in monotone increasing.

Proposition 2.2.2 implies that a customer should find $\min_{A\subset S,\ |A|=a} a\cdot q\cdot E_A[P_t^{RTP}]$ of a task T before the initiation of T to compare this to $u(T)$. The optimal solution of the cost minimization problem for the task T is described in the following lemma.

Lemma 2.2.8 (Optimal load scheduling for a task T)**.** Let $A^* := \arg\min_{A\subset S,\ |A|=a} a\cdot q\cdot E_A[P_t^{RTP}]$ and $A^*_{DAP} := \{t : \text{time index of the } i\text{th smallest element of } \{P_t^{DAP} : t \in S\} \text{ where } i = 1,...,a\}$, both assumed to be unique minimizers. Then, $A^* = A^*_{DAP}$.

Proof. By Assumption 2.2.7, A^*_{DAP} and the set $A^*_{RTP} := \{t : \text{time index of the } i\text{th smallest element of } \{P_t^{RTP} : t \in S\} \text{ where } i = 1,...,a\}$ have exactly same elements so that $A^*_{DAP} = A^*_{RTP}$. To verify the proposition $A^* = A^*_{RTP}$, suppose there exists an A' such that $aqE_{A'}[P_t^{RTP}] < aqE_{A^*_{RTP}}[P_t^{RTP}]$ so that $\sum_{t\in A'} P_t^{RTP} < \sum_{t\in A^*} P_t^{RTP}$.

Since $|A^*_{RTP}| = |A'| = a$, there exists a time slot $t' \in S\setminus A^*_{RTP}$ such that $P_{t'}^{RTP} < P^{RTP}_{\sup A^*_{RTP}}$. This conflicts with the definition of A^*_{RTP}. Therefore, $A^* = A^*_{RTP} = A^*_{DAP}$. □

Once A^* is obtained, an active power consumption is *scheduled* at each time slot in A^* for the task T. The following theorem states the condition when the load scheduled for a task T is not responsive to DAP, to provide an idea on the price responsiveness of load to DAP.

Theorem 2.2.9. Consider two identical tasks T_1 and T_2 initiated on two different days d_1 and d_2 (i.e., there may exist a specific time t_0 such that $P^{DAP}_{t_0 \text{ in } d_1} \neq P^{DAP}_{t_0 \text{ in } d_2}$) respectively. Scheduled loads for tasks T_1 and T_2 are same if and only if $A^*_1 = A^*_2$.

Proof. Set $T_1 := (q_0, S_0, a_0)_{d_1}$ and $T_2 := (q_0, S_0, a_0)_{d_2}$. The scheduled load by T_1 is $Q^{T_1}_t = q_0 \cdot \mathbf{1}_{A^*_1}(t)$, and the scheduled load by T_2 is $Q^{T_2}_t = q_0 \cdot \mathbf{1}_{A^*_2}(t)$, where $\mathbf{1}_X(t)$ is an indicator function. Therefore, $Q^{T_1}_t = Q^{T_2}_t$ if and only if $A^*_1 = A^*_2$. □

Theorem 2.2.9 can be considered as the corresponding version of Theorem 2.2.5 for DAP. Since Theorem 2.2.9 asserts that $Q^{T_1}_t = Q^{T_2}_t$ regardless of their DAPs as far as equality of A^*_1 and A^*_2 is concerned, the load scheduled for a task is not responsive to any difference between $\{P^{DAP}_t\}_{d_1}$ and $\{P^{DAP}_t\}_{d_2}$ unless $A^*_1 \neq A^*_2$. Moreover, the load assigned to task T is not responsive to any price decrements within A^* nor increments within $S \setminus A^*$. Hence, it can be inferred from Theorem 2.2.9 that the relationship between DAPs and loads would exhibit nonlinearity. However, the identification of a task T would become clearer if T is routinely initiated at the same time everyday so that S is almost invariant, or at least stationary over days. This would help precise load prediction from DAP as well as task identification.

The difference between two sessions of multiple tasks initiated in a same day may appear as a time-varying price responsiveness of load. The following example illustrates how DAP delineate the load dynamics if there are multiple tasks initiated.

Example 3. Suppose a series of DAP $\{P^{DAP}_t, t \in \{1, 2, 3, 4, 5, 6, 7\}\}$ is announced as follows,

$$P^{DAP}_t = (0, 1, 8, 9, 8, 1, 0),$$

and a customer initiates its tasks as follows,

$$T_1 = (q_1 = 1, \;\; S_1 = [1, 5], \;\; a_1 = 4),$$
$$T_2 = (q_2 = 1, \;\; S_2 = [2, 6], \;\; a_2 = 4),$$
$$T_3 = (q_3 = 1, \;\; S_3 = [3, 7], \;\; a_3 = 4).$$

Then, optimal load scheduling for each task is,

$$Q_t^{T_1} = (1, 1, 1, 0, 1, 0, 0),$$
$$Q_t^{T_2} = (0, 1, 1, 0, 1, 1, 0),$$
$$Q_t^{T_3} = (0, 0, 1, 0, 1, 1, 1).$$

Hence, the total load from the aggregation of the initiated tasks is,

$$Q_t = \sum_i Q_t^{T_1} = (1, 2, 3, 0, 3, 2, 1).$$

The shape of P_t^{DAP} and load Q_t is provided in Figure 2.11.

Remark. Despite the simple problem setting in example 3, we can clearly see that the pattern of the load Q_t in Figure 2.11 has all the peculiar features present in the load data from a customer in Houston shown in Figure 2.8, such as time-varying price responsiveness, especially the positive correlation between price and load at off-peak hours as well as the sudden load drop at peak hours.

Another interesting point in Example 3 is that, the correlation between the load Q_t and P_t^{DAP} is 0.2, which is a positive value. The positive correlation between load and price throughout whole time period as well as the strong positive correlation between load and price in off-peak hours are both counterintuitive results which are not explicable by the traditional consumption model. Our proposed model, however, gives an idea how to rationalize the pattern we can frequently observe in real world load data. It supports the claim that such a behavioral pattern is an optimal choice

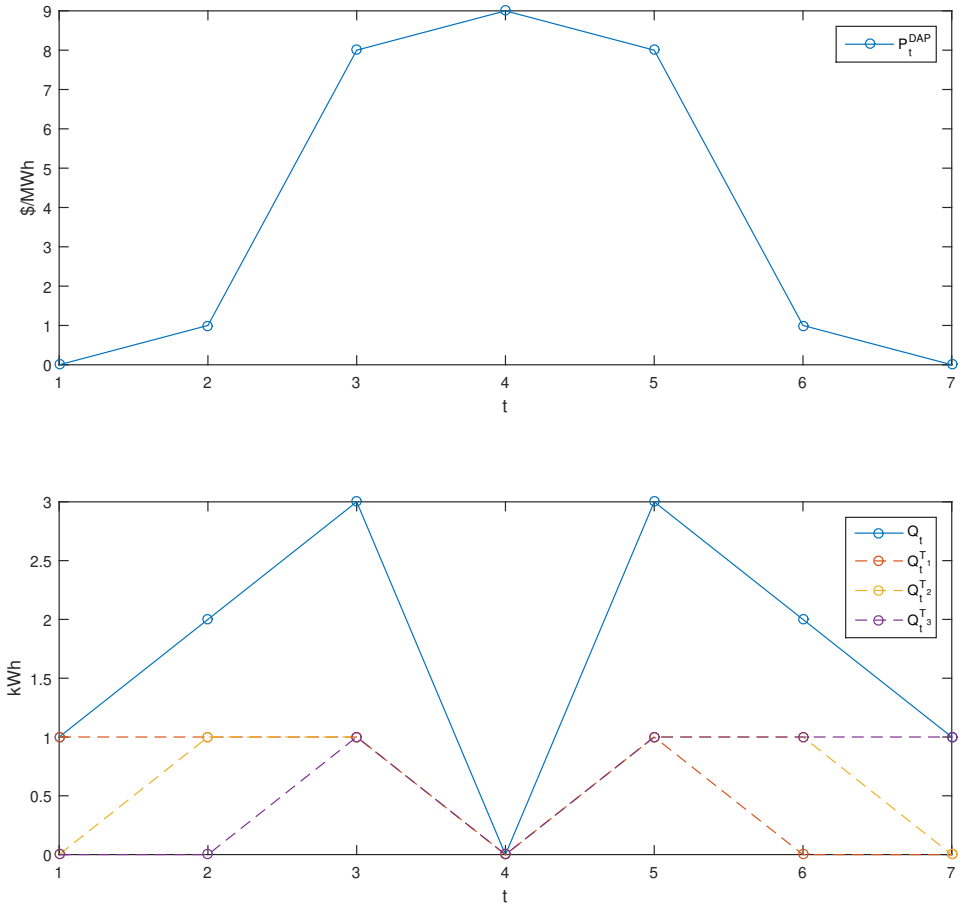


Figure 2.11: The DAP and optimal load scheduled for the given tasks T_1, T_2, and T_3.

selected by a rational customer, by providing us insight about the underlying structure in such a pattern.

Corollary 2.2.10. If $a = |S|$, any price change does not involve load change unless the task T is aborted.

Proof. If $a = |S|$, $A^* = S$ so that A^* is always same for any price change. □

There could be a critical task, e.g., the tasks of life supporting devices in hospitals, which would appeared as extremely price inelastic loads. Corollary 2.2.10 shows the ability of the proposed appliance usage model for describing such loads.

2.3 Potential Further Extension

As this work aims modeling and closing the loop around price responsive demand data, potential extension of his study may include the empirical quantitative modeling of price responsive consumers based on the proposed "appliance" usage model in Section 2.2.2.1, and the design of optimal pricing as well as market efficiency analysis based on the proposed consumer behavior model.

2.3.1 Construction of a Quantitative Price Responsive Consumer Behavior Model from Empirical Data

The potential next step of this study can be the development of a mechanism to systematically identify and construct a quantitative model of consumer behavior from past history. Unlike what has been believed so far, our study of consumers' behavior with respect to real time price changes based on empirical data on price-responsive loads in ERCOT area suggests the following:

- Load is a prescheduled quantity based on price prediction, rather than an instantaneously price-responsive quantity in real time.
- A prescheduled load is only affected by real time price shocks that severely deviate from the price prediction.

Therefore, it is crucial to develop a dynamic model between the load as a prescheduled quantity, and the DAP as an RTP predictor. To address the nonlinearity between consumption and DAP suggested from Theorem 2.2.9 in Section 2.2.2.3, a suitable choice one may consider is to introduce an artificial neural network. A possible multilayer perceptron structure can be deduced from the proposed appliance usage model as follows. By the appliance usage model, the load Q_t can be described as,

$$Q_t = \sum_i q_i \mathbf{1}_{A_i}(t),$$

where the T_i's are the task initiated at time t, with each $T_i := (q_i, S_i, a_i)$. Hence, $\hat{Q}_t$, the predicted

Q_t, can be achieved by taking the expectation of Q_t over the T_i's,

$$\hat{Q}_t = \sum_i q_i \Pr[t \in A_i].$$

Here, $\Pr[t \in A_i]$ may vary by S_i and a_i, as well as $\{P_\tau^{DAP} : \tau \in S_i\}$. Meanwhile, it is reasonable to assume that S_i is independent of DAP itself, due to the assumption in Section 2.2.2.3 that S is stationary over days. Here, I emphasize the following key assumption for the availability of effective off-line estimation: Most power consumption is from the tasks which are routinely initiated by necessity in daily life before the exact electricity price is known. Hence, $\hat{Q}_t$ can be rewritten as follows,

$$\hat{Q}_t = \sum_i q_i \Pr[t \in S_i] \cdot \Pr[t \in A_i | t \in S_i, P^{DAP}], \tag{2.12}$$

where P^{DAP} is a given one-day series of DAP. Assuming $\Pr[t \in S_i]$ is an estimable constant p_{it}, and assuming that $\Pr[t \in A_i | t \in S_i, P^{DAP}]$ takes the form of a logistic function of DAP, $\hat{Q}_t$ in (2.12) can be reduced to,

$$\hat{Q}_t = \sum_i q_i p_{it} \cdot \frac{1}{1 + e^{\mathbf{w_{it}} \cdot P^{DAP}}} = \sum_i v_{it} \cdot \frac{1}{1 + e^{\mathbf{w_{it}} \cdot P^{DAP}}}, \tag{2.13}$$

where $v_{it} := q_i p_{it}$ is a scalar to be estimated, and $\mathbf{w_{it}}$ is a weight vector corresponding to P^{DAP}.

In fact, as can be seen this results in a two-layer perceptron with one hidden layer z_{it}:

$$\hat{Q}_t = \sum_i v_{it} z_{it} \text{ where } z_{it} := \frac{1}{1 + e^{\mathbf{w_{it}} \cdot P^{DAP}}}. \tag{2.14}$$

Hence, the training process to obtain v_{it} and $\mathbf{w_{it}}$ can be performed by backpropagation, and optimal prices can be in a similar fashion through, for example, a gradient descent method, once training is done.

As mentioned in Section 2.2.2, and specifically in Section 2.2.2.1, the main purpose of proposing the appliance usage model is to delineate the structure of the constraints of an optimization problem that we ultimately aim to identify. Hence, a further step of my future work is to develop

the methodology for identifying an underlying optimization problem which will provide a unified view on consumer behavior in response to price. This would ultimately rationalize the given data by linking the observed demand response to RTP. The arguments made here is extended and reinterpreted in Chapter 3.

2.3.2 Market Efficiency Analysis and Optimal Pricing Design

Market is a dynamical system that is designed to proceed toward an optimal state as its equilibrium. However, such a process necessarily requires a certain amount of time to reach its equilibrium. While dynamic modeling and control on the generation side in power systems has been well understood, the understanding of dynamic behavior on the demand side in response to price has been unclear. Our study of consumers' behavior with respect to real time price changes based on empirical data from price-responsive loads suggests that a load is basically determined by price prediction, and only affected by real time price shocks with delay. Such behavior features imply that frequent price changes do not necessarily bring economic efficiency in the sense of social welfare maximization.

This idea provides important guidance in designing two fundamental factors in time-varying retail electricity prices – the frequency and timeliness, where (1) *Frequency of Price* is the frequency at which retail prices change, and (2) *Timeliness of Price* is the time lag between when a price is set and when it is effective [18]. It is generally assumed among economists that RTRP with high frequency and just-in-time timeliness would be ideal in terms of economic efficiency in the electricity market, as RTRP is an attempt to get more accurate signals closely reflecting the actual supply/demand status in the market. However, my inference based on my work is that both arguments are not necessarily right. The inherent delayed responsive nature of loads with high price volatility exacerbates the predictability of price, thereby making demand less responsive to RTRP, which worsens economic efficiency. Consumers which are more exposed to market volatility stiffen their demand to be more inelastic and tend to be more conservative due to the inertial nature of demand. This suggests that there exists a trade-off between controllability of demand and observability of markets, so that there may exist an optimal frequency and timeliness which is

not extreme for optimal pricing design. This also supports the importance of relatively long-term contract markets such as day-ahead electricity markets. Market efficiency should be reanalyzed taking into consideration the trade-off between the controllability of demand and the observability of the market.

3. PRICE RESPONSIVE LOAD MODELING WITH CAUSAL ANALYSIS

In this chapter, we are concerned with generic data-centric modeling of a price responsive demand from given data, extending the arguments provided in Section 2.2. Throughout this chapter, the problem we are concerned with is specified, and the theoretical justification of our proposed problem-tackling methodology is expounded. In Section 3.1, we elucidate the fundamental problem of the modeling electricity demand – the impossibility of consistent modeling with non-experimental data alone regardless of its sample size. This necessitates the invocation of some untested assumptions prior to any observational studies. In the following subsections, we thoroughly examine the theoretical premises, and establish required parsimonious postulates for consistent modeling of demand response. In Section 3.2, we propose the abstract consumer behavior framework that rests on the equivalence between the consumer behavior portrayed in contemporary economic theories and Shannon's secrecy system framework with a minor modification. In Section 3.3, we propose a novel neural model representation, the Stochastic Neuron, for an effective instantiation of the above consumer behavior framework. This is followed by the proof that the proposed demand response model backed by the proposed consumer behavior framework enables the consistent modeling of price-responsive electricity demand. Subsequently in Section 3.4, we refine the proposed *most rational account principle* for effective and irreducible model representation by developing a measure of rationality.

3.1 The Fundamental Problem of Consumer Behavior Modeling in Electricity Consumption

In this work, we address the problem of modeling the effect of real-time price on individual electricity load given a data set $\mathcal{D} := (X, Y)$, where X and Y are the sets of observed price sequences and load sequences respectfully. Generally speaking, this task falls into an identification problem of "$y = g(x)$" given a full knowledge of $P(\mathbf{x}, \mathbf{y})$[1], the joint probability distribution of

[1]Full knowledge on $P(\mathbf{x}, \mathbf{y})$ is the ultimate knowledge one can achieve from data regardless of the data size.

price $\mathbf{x}$ and load $\mathbf{y}$. However, the underlying difficulty of the electricity consumption identification problem arises from the existence of *confounders*, which are the common factors that influence both the electricity price and the load behavior, such as weather condition, and the practical infeasibility of maintaining full information of potential confounders on each consumer individual. It is well known from the slogan "correlation does not imply causation" that such confounding bias precludes the disentanglement of causation and spurious association in purely passive observational studies [36]. Thus, in principle, any problem-tackling strategies solely relying on associational inference techniques could be erroneous in regard to the identification problems with confounders unless we have full information of them, as they may infuse ambiguity into the identification tasks.

For the purpose of articulation, it is convenient to illustrate the problem with the tool established by Pearl and his collaborators: structural causal modeling [37], which is a formal causal framework that unifies graphical causal analysis [38], counterfactual analysis [39], and structural equation modeling (SEM) [40]. The starting point of this framework is to model the system as a directed acyclic graph (DAG) called a *causal diagram* of a *Markovian model* or a *causal Bayesian network*. A Markovian model consists of a causal graph represented as a directed acyclic graph (DAG) $G = (V, E)$ over a vertex set V, which represents the set of random variables, and an edge set E of ordered vertex pairs, which indicates the causal influence between the pair of random variables. The term "Markovian" comes from the interpretation of the causal model graph G that each variable is independent of all its non-descendants conditioned on its parent variables in G. Such a structural property permits one to construct the joint distribution $P(V)$ of all variables $V := \{V^{(i)}\}_{i=1}^{n}$ via a modular configuration, i.e, it can be factorized as follows:

$$P(V) = \prod_{i=1}^{n} P(V^{(i)}|Pa(V^{(i)})), \tag{3.1}$$

where $Pa(V^{(i)})$ denotes the parent node set of $V^{(i)}$ in G. Hence, the full description of a Markovian model $\mathcal{M}$ is defined as the following tuple:

$$\mathcal{M} = \left\langle V,\ G_V,\ \{P(V^{(i)}|Pa(V^{(i)})) : V^{(i)} \in V\} \right\rangle, \tag{3.2}$$

where (i) V is a set of the variables in the system of interest, (ii) G_V is a DAG with the nodes corresponding to the elements of V, and (iii) $P(V^{(i)}|Pa(V^{(i)}))$ is the conditional probability of variable $V^{(i)}$ given $Pa(V^{(i)})$ in G_V. $P(V^{(i)}|Pa(V^{(i)})) = P(V^{(i)})$ if $Pa(V^{(i)}) = \phi$.

Traditionally, SEM has been the main workhorse for causal effect analysis in many scientific disciplines. SEM represents the causal influence as a set of functional equations of the form

$$v^{(i)} = g_i(Pa(v^{(i)}), \epsilon^{(i)}), \tag{3.3}$$

where $Pa(v^{(i)})$ stands for the set of variables which directly determine the value of $v^{(i)}$, and $\epsilon^{(i)}$ denotes unknown or unmeasured factors causing errors. The interpretation of $g_i(\cdot)$, which is called *law* in science, is standard in natural and social sciences; it is a description of the data generating process that assigns a value to $v^{(i)}$ by a mechanism, in response to the values $Pa(v^{(i)})$ and $\epsilon^{(i)}$ taken on by external intervention. The identification of $g_i(\cdot)$ has been of primary interest in most scientific studies since it is not meaningful to attempt to identify the causal effect of an arbitrary $v^{(j)}$ on $v^{(i)}$ without the specification of $g_i(\cdot)$. We use the notation $do(v^{(j)} = v_0)$ to denote the external intervention of assigning a value v_0 to $v^{(j)}$, which is named as *do-operator* by Pearl [38].

Converting SEM to the corresponding causal diagram G can be straightforwardly done by taking each $v^{(i)}$ as $V^{(i)}$ and placing an arrow toward $V^{(i)}$ from each member of $Pa(V^{(i)})$. As the mathematical operation $do(V^{(j)} = v_0)$ implies the simulation obtained by fixing $V^{(j)}$ to v_0, it can be described as the following two-step operation in SEM: (i) delete $g_j(\cdot)$ from $\{g_i : i = 1, ..., n\}$, and (ii) replace it with an equation $v^{(j)} = v_0$. Equivalently, it can be defined as the following sequential operation in a causal diagram creating the submodel $\mathcal{M}_{V^{(j)}=v_0}$: (i) the removal of all incoming edges to $V^{(j)}$ from $Pa(V^{(i)})$ keeping the rest of the model unchanged, and (ii) the assignment of $V^{(j)} = v_0$. Hence, the post-intervention distribution $P_{v^{(j)}}(V) := P(V|do(V^{(j)} = v^{(j)}))$ is given

by [36]:

$$P_{v^{(j)}}(V) = \begin{cases} \prod_{V^{(i)} \in V \setminus \{V^{(j)}\}} P(V^{(i)} | Pa(V^{(i)})) \\ \qquad\qquad V \text{ consistent with } V^{(j)} = v^{(j)}, \\ 0 \qquad\qquad \text{Otherwise.} \end{cases} \tag{3.4}$$

It is worth noting that $P_{v^{(j)}}(V) \neq P(V|v^{(j)})$ in general, providing the reaffirmation that correlation does not imply causation. Causal effects on an arbitrary variable set $Y \subset V$ can be obtained through appropriate marginalization of Eq. (3.4), i.e., $P_{v^{(j)}}(Y) = \sum_{(V \setminus \{V^{(j)}\}) \setminus Y} P_{v^{(j)}}(V)$.

The notable implication of structural causal modeling is that the full specification of G_V explicates the prerequisite tacit information for the consistent computation of the "causal query" $P_x(Y)$ for any disjoint $X, Y \subset V$ from the probabilistic premise $P(V)$ given G_V, if one has a full observation of the model variables. The gold standard for the estimation of causal effect among scientists has been the *randomized controlled experiment*: after the subjects to be manipulated are randomly chosen from the population of interest to neutralize the effect of confounding factors, the experimenter applies direct manipulation on the subjects. In a variety of studies, however, this methodology is infeasible due to the issues of practicality or ethics. Instead, one inevitably needs to rely on non-experimental observations to infer causal effects. Structural causal modeling provides a tool for hypothetical simulation of randomized controlled experiments to infer causal effects from statistical information obtained from intervention-free behavior.

In many cases, however, it is unrealistic to presume that one has full observation of V, as most real world systems in practice do not permit one to have full observation of all system variables. If there are unobserved confounders affecting two or more variables in V, it may prohibit the modular configuration (3.4) so that the inferability of a causal effect could be questionable. Thus, the determination of "identifiability" of a causal query $P_x(Y)$ of interest is the first step of most scientific studies when the observations are made under partial information on $\mathcal{M}$. The main problem dealt with in this work, the identification of causal effect of energy price on its consumption, also falls into this category in the sense that we do not have any information on the confounders other than

their presence. Taking into account the ignorance of some variables in V by taking $V = O \cup N$, where O and N are the sets of observable and unobservable variables respectively, the causal effect of $do(X = x)$ on Y is said to be *identifiable* from $P(O)$ in G_V if $P_x(Y)$ is *uniquely determined* from $P(O)$ in any causal model which induces G_V. The formal definition of causal identifiability is as follows [36]:

Definition 3.1.1 (Identifiability)**.** Let $X, Y \subset O$ be disjoint sets of observable variables. Given a causal diagram G_V, the causal query $P_x(Y)$ is *identifiable* if we have $P^{(1)}(O) = P^{(2)}(O) \Rightarrow P_x^{(1)}(Y) = P_x^{(2)}(Y)$ for any two models $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ sharing a common G_V, where $P^{(i)}(\cdot)$ is the associated probability distribution of $\mathcal{M}^{(i)}$ and $P_x^{(i)}(\cdot)$ is the post-interventional distribution of $\mathcal{M}_x^{(i)}$ by $do(X = x)$.

A line of literature by Tian and Pearl [41], Shpitser and Pearl [43], and Huang and Valtorta [44] provides a complete graphical characterization of models for the identifiability of causal queries in a *semi-Markovian model*, which is a Markovian model $\mathcal{M}$ with unobserved variables such that each unobserved variable in $\mathcal{M}$ is a root node with exactly two observed children. In fact, these works completely close the causal identifiability decision problem for the Markovian model with arbitrary unobserved variables, since Tian and Pearl present in their work [42] a conversion method for an arbitrary Markovian model to a semi-Markovian model, preserving its causal effect identifiability properties. As semi-Markovian models structurally assume an unobserved confounder has two observed children, it is convenient to introduce a bidirected edge to denote the unknown confounding effect. That is, if there is an unobserved confounder $N^{(k)}$ affecting both observables $O^{(i)}$ and $O^{(j)}$, i.e., $O^{(i)} \leftarrow N^{(k)} \rightarrow O^{(j)}$, then it is depicted as a dashed bidirectional edge $V^{(i)} \dashleftarrow\dashrightarrow V^{(j)}$ in the causal diagram of the semi-Markovian model. The key result from [41] used in this work is the following lemma:

Lemma 3.1.1 (Tian and Pearl [41])**.** $P_x(O)$ is identifiable if and only if there is no bidirectional path connecting X to any of its children in G_V.

Using the language of structural causal modeling, the modeling task of demand response in

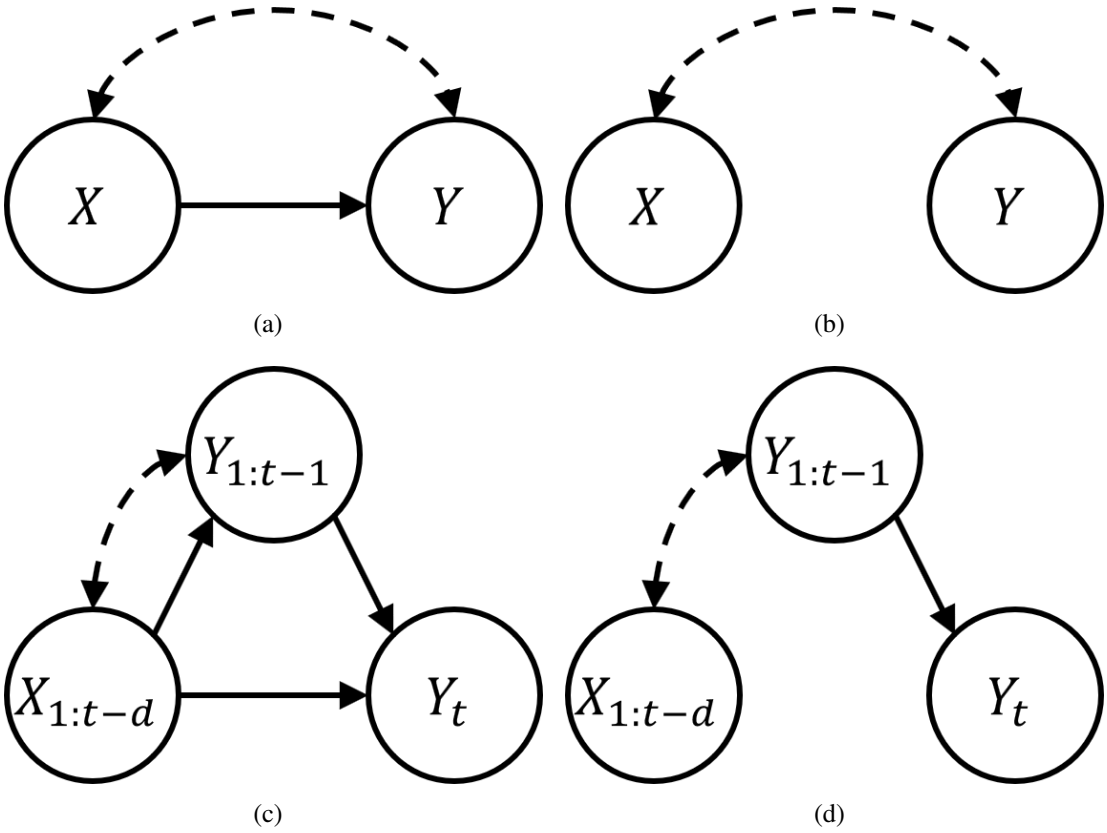


Figure 3.1: (a) The causal diagram of the simplest demand response model. (b) The causal diagram presuming that there is no causal effect of X on Y but spurious correlation in between those. (c) The expanded causal model of Fig. 3.1(a) for time series representation of demand response. (d) The expanded causal model of Fig. 3.1(b) for time series representation of the hypothesis that there is no causal effect of $X_{1:t}$ on $Y_{1:t}$.

electricity consumption can be articulated as the identification problem of $P_x(Y)$ in $V = O \cup N$ where $O = \{X, Y\}$ and N are the unknown factors, where X is a variable representing an electricity price, and Y denotes the electricity consumption. The simplest causal diagram of the demand response model we consider is depicted in Fig. 3.1(a). While the direction of the edge (X, Y) can be justified from an essential assumption in microeconomics that an individual consumer is a price-taker, one may cast a doubt on the existence of the edge (X, Y), since the causal diagram may be as depicted in Fig. 3.1(b).

The Chap. 2 and our previous work [1] statistically validates the dismissal of the model in Fig. 3.1(b), where X and Y are related only through a confounding factor, by showing the causal

effect of X on Y from the given data of a consumer in Texas. As the main goal of the paper [1] was to construct a dynamic model of demand response, it begins by temporally separating X and Y into time series representations $X_{1:t}$ and $Y_{1:t}$, and showing that the consumer exhibits $X_t \perp\!\!\!\perp Y_t$ while $Y_t \not\perp\!\!\!\perp Y_{1:t}$, where $\perp\!\!\!\perp$ denotes independence. With these independence/non-independence relationships in hand, it then proceeds by making some further postulates as follows: (i) $Y_{1:t-1}$ have a direct effect on Y_t. E.g., whether to turn on a dish washer with one hour cycle may directly affect the electricity usage 15 minutes later. (ii) $Y_{1:t-1}$ *blocks* the effect of the confounders which potentially affect both $X_{1:t-d}$ and Y_t, where $d > 1$. E.g., the past weather condition prior to the time $t - d$ may not be informative for the inference of Y_t if one knows the air conditioner's operational status at the time $t - 1$. These extra postulates, along with a *stationarity*[2] assumption permit one to expand the causal models in Fig. 3.1(a) and 3.1(b) to those in Fig. 3.1(c) and 3.1(d), respectively. By verifying $X_{1:t-d} \not\perp\!\!\!\perp Y_t | Y_{1:t-1}$ via statistical hypothesis testing, the existence of the edge $(X_{1:t-d}, Y_t)$ is substantiated, which implies that the causal relations in Fig. 3.1(c) as well as Fig. 3.1(a) are the correct ones.

While our previous work [1] shows that the simplest demand response causal model is a valid form, its structure highlights the fundamental *unidentifiability* problem of modeling electricity consumption.

Theorem 3.1.2 (The Fundamental Problem of Consumer Behavior Modeling in Electricity Consumption)**.** The simplest demand response model in Fig. 3.1(a) is unidentifiable.

Proof. This is deduced directly from Lemma 3.1.1 as the model in Fig. 3.1(a) is semi-Markovian. □

It is not a difficult task to construct a toy example illustrating the unidentifiability in Theorem 3.1.2. Consider two different models $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ with the same causal diagram depicted in Fig. 3.1(a). Regarding the data generation process from the unobservable variable N, suppose $X = N$ for both models but let $Y^{(1)} = X + 2N$ for $\mathcal{M}^{(1)}$, and $Y^{(2)} = 2X + N$ for $\mathcal{M}^{(2)}$. Then,

[2]This is a common assumption in time series analysis. It also means that the causation remains intact after time shift.

it can be easily checked that the joint distributions of the observable variables in both models are the same, i.e., $P^{(1)}(X,Y) = P^{(2)}(X,Y)$, but $P_x^{(1)}(Y) = P(x+2N)$ and $P_x^{(2)}(Y) = P(2x+N)$, so that $P_x^{(1)}(Y) \neq P_x^{(2)}(Y)$. Shpitser and Pearl also present another illustrative example in their work [43] to show the unidentifiability of Fig. 3.1(a) when all variables in the model are Boolean.

The provided example clearly illustrates why the determination of $P_x(Y)$, i.e., $y = g(x)$, is fundamentally impossible solely from the obtained $P(X,Y)$. Even the expansion of Fig. 3.1(c) to include a link (X,Y) does not help to resolve unidentifiability. For this reason, our previous work [1] takes a conservative approach to model selection; it presents the model with the least price causality in the sense that it chooses the demand response model with the least influence of price on load among all the possible models. The above fundamental unidentifiability implies that the task of modeling of price responsive consumer behavior inevitably necessitates some extra knowledge or untested assumption of the data-generating process, and the model cannot be determined from the data alone regardless of sample size, nor from probabilistic premise. In the next section, we first examine some preexisting studies in modern economics mainly focusing on consumer theory and behavioral economics to aid in making appropriate structural assumptions to construct an identifiable causal model. Subsequently, we propose an abstract framework for consumer behavior models which enlightens the key structural aspects of consumer behavior for identifiable causal model construction.

3.2 The General Framework for Consumer Behavior Models

Most theoretical demand response literature widely adopts the neoclassical viewpoint of the microeconomic consumer model, which is considered as a standard approach accepted in modern economic theory, taking a consumer as a selfish solver of the following optimal choice problem (OCP) [46]:

$$u^* := \max_{\vec{y} \in B(\vec{x},b)} u(\vec{y}) \tag{3.5}$$

where $u(\cdot) : \mathbb{R}^k_+ \to \mathbb{R}$ is a utility function, $(\vec{x}, \vec{y}) \in (\mathbb{R}^k_{++}, \mathbb{R}^k_+)$ are the sequences of prices and loads respectively for a short term period which may be influenced by each other[3], and $b \in \mathbb{R}_{++}$ is a budget constraint. Both $\vec{x}$ and b determine the set of affordable load profile choices $B(\vec{x}, b) := \{\vec{y} \in \mathbb{R}^k_+ : \vec{x} \cdot \vec{y} \leq b\}$. The optimal solution for the problem (3.5) is called *Walrasian demand correspondence* defined to be $y^*(\vec{x}, b) := \operatorname{argmax}_{\vec{y} \in B(\vec{x}, b)} u(\vec{y})$ because $y^*(\vec{x}, b)$ is not necessarily unique in general.

The essential foundation of consumer theory, which eventually leads Walrasian demand correspondence into a *demand function* via guaranteeing a unique optimal load profile at any price given budget, is (i) the consistency, named *consumer rationality* by economists, which indicates the transitiveness and completeness of the preference; (ii) the *local nonsatiation*; and (iii) the *strict convexity* of her preference [46]. The presupposition of rational consumers implies that a consumer's choices are the manifestation of her *rational* preference so that the observed consumer choices are the series of her *revealed preference*. As a result of efforts to *rationalize* the observed behavior, the traditional axiomatic approach on the basis of the principle of revealed preference specifies a family of *rationalizable* utility functions as those which rationalize an observed consumer behavior, e.g., the set of *continuous*, *strongly monotone*, and *strictly quasi-concave* functions, as is stated in Afriat's theorem [46].

In addition to the above properties for rationalizable utility functions, it is often convenient to posit the *differentiability* of utility function enabling us to analyze consumer behavior via standard tools in calculus. A well known example is Roy's identity which formulates the demand function directly from the derivatives of *indirect utility function* $v(\vec{x}, b) = u^*$, i.e., $y^* = -\nabla_{\vec{x}} v / \frac{\partial v}{\partial b}$. More importantly, it is known that *twice differentiability* of the utility function has a crucial implication beyond its analytic convenience when we take a whole exchange economy into consideration beyond the mere consumer side, in the sense that the uniqueness and stability of equilibria of an economy is premised upon the smoothness of demand function as well as the corresponding utility function [48]. While a utility function is in fact an unobservable entity, another notable result on differentiable rationalizable utility functions is that they always generate unique homeomorphic

[3] $\mathbb{R}^k_+$ denotes the non-negative orthant, and $\mathbb{R}^k_{++}$ denotes the strictly positive orthant.

demand functions; specifically, the differentiability of a rational utility function implies the existence of an open $X \subseteq \mathbb{R}^k_{++}$ and a *unique bijective demand function* $y^\star(\cdot, b) : X \to \text{int}(B(\vec{x}, b))$, where $\text{int}(\cdot)$ denotes the interior of a topological set [49]. Hence, the invertibility of demand from its homeomorphy is one of the core interests of economists and market operators in (i) empirical identification and estimation of demand; and (ii) theoretical analysis of the stability and uniqueness of Walrasian equilibrium prices [50].

A key limitation of the neoclassical axiomatic approach to modeling consumer behaviors is that the assumption of *perfect information* is at the heart of the optimal choice problem formulation [47]. Although the standard neoclassical viewpoint suggests an alternative formulation of (3.5), viz., the *expected utility theory* constructed by von Neumann and Morgenstern [51] for the analysis of decision making under risk which supposes a consumer choice is a maximizer of expected utility, it still presumes that a consumer has a perfect knowledge of her preference over the choice space, which is endogenous and consistent, since this is the crucial premise of consumer rationality. Another important limitation of the neoclassical approach is that it was never intended to be a realistic model of human cognition; it originally emerged as a normative theory rather than a descriptive one [52]. The prediction solely based on an optimal choice problem formulation would be hardly accurate so that it may require a deliberate manual process of filtering, adjustment and calibration, or even extra assumptions.

Triggered by the seminal works of Kahneman and Tversky [53, 54], which offer a descriptive alternative to expected utility theory based on a psychological background for decision making under uncertainty that is labeled as *prospect theory*, behavioral economists provide a variety of descriptive heuristic alternatives of the orthodox neoclassical approach leaning on laboratory experiments and empirical evidence. While the number of cognitive and behavioral errors identified and described by behavioral economists is large and constantly growing [55], their main viewpoints may be divided into two groups. The first group of critiques attacks the *rationality* of a consumer, initiated by the works of Kahneman, Tversky, and Thaler [53, 54]. A large body of literature in behavioral economics has shown that consumers perceive utility as being in flux which

leads to one's behavior appearing as *irrational*, exhibiting various distortion patterns. This point of view argues that a consumer is prone to have numerous types of "error" that its manifestation is always inconsistent and context-dependent due to (i) social and cultural factors, as well as (ii) her selection of sources of information, and (iii) her unawareness of the way it is processed [56]. An example is the work of Thaler [57], who first applied a behavioral economics approach to describe consumer behavior, and showed an inertial effect in consumer choices through various experiments, and concluded that a consumer is uncertain about her utility and she tends to cling to choices she previously made to avoid potential regret. Although one may consider a more sophisticated and complex utility structure with multiple latent states representing exogenous changes, e.g., by taking a consumer as an optimal multi-armed bandit problem solver to choose 'exploration' or 'exploitation' at each time step, which is an active area of study in reinforcement learning these days, Sunstein and Thaler [58] point out that the term 'preference' may be inappropriate under strong context-dependency [55], stating that: "If the arrangement of alternatives has a significant effect on the selections customers make, then their true preferences do not formally exist".

On the other hand, the second point of view attempts to reconcile empirical observations with neoclassical theory by presupposing consumers' *willingness* to be rational. Rather than abandoning the concept of rationality, the studies consider the problem of implementing it in practice. Simon proposes the concept of *bounded rationality* in his work [59] to denote the entire range of restrictions that crop up on both the informational and computational sides that prevent people from behaving as in the normative ideal in neoclassical theory. As Simon indicates "most people have reasons for what they do" [59], he argues that individuals choose from among the best options available to them, given a certain encoded utility function and a set of constraints, which may be either physical and informational. Another body of literature including Stigler's work [60] introduces the concept of *information cost* so that it is rational to stop processing or collect further information if the cost of information processing outweighs the expected gain by it. This viewpoint also suggests that rationality may be deemed not as a dichotomous but as a continuous variable so that one can refer to degrees of it [56]. In fact, recent studies suggest that the first viewpoint at-

tacking the concept of rationality can be explained by the interpretation of the bounded rationality as a consumer possibly having a bounded insight of her own utility, for biological reasons such as the information processing capability limitation of a brain. As an example, an empirical study in neuroscience demonstrates that the human brain encodes the subjective utility (value) of an observed item in a compressed form, where the precision with which neural representations encode is proportional to the frequency with which the item is actually encountered [61].

The idea of potential extra constraints on the consumer choice space in addition to those defined by price and budget is also applicable to models of electricity consumption. Besides the potential uncertainty or fluctuation in utility function from various exogenous factors such as weather, *ex ante* demand response to uncertainties in future prices, a peculiar non quasi-concave utility structure resulting from the nature of an appliance usage, and limited load controllability no better than simple on-off switches, could all derail the rational load behavior ideal of the standard model (3.5) equipped with the rational preference.

Although empirical evidence extensively studied by behaviorists well demonstrates the limitation of neoclassical standard model, developing an alternative paradigm of consumer choice encompassing various empirical observations from diverse domains is a challenging task. This is because, behavioral economics lacks a general conceptual foundation that coherently synthesizes a variety of observations [55]. Unlike economics, which has been dominated by a single paradigm for a long time, psychology has multiple paradigms without any one school prevailing as the mainstream. Moreover, no systematic attempts have been made within behavioral economics to assess the frequency of various cognitive and behavioral errors. Affected by this cultural background, behavioral economics does not suggest any general theory of decision making over cognitive distortions, and appears to be just a catalog of psychological phenomena of observations made on an occasion-by-occasion basis [55]. The existence of a multitude of opportunistic heuristics from behavioral economics leads to the following critical questions for which it is difficult to get clear answers when we attempt to translate the results from behavioral economics into a particular field such as the study of demand response in electricity markets: “Which methods and concepts to be

imported from behavioral economics are universally applicable?" and "How to decide whether an observation from lab experiments in a specific field is transferable to another field?" No shared understanding has been evolved in behavioral economics for answering these questions [56]. As a consequence of this difficulty, a policy recommendation based on behavioral economics is neither conclusive nor solid without a thorough empirical validation of a candidate set of heuristics for a specific field. Such validation also could be a highly demanding task, as the list of identified heuristics is too vast and it spans a broad range of different conditions. According to one of the lists which is far from complete, the number of the heuristics was already almost 50 in 2009 [62].

Although the traditional neoclassical viewpoint presents an elegant normative explanation on how demand should behave stemming from the postulate of rational behavior, a large body of literature in behavioral economics criticizes that it lacks descriptive power based on empirical evidences. The core ideas permeating the behaviorists' works can be summarized as follows: (i) a consumer generally faces arbitrary physical or institutional restrictions in her choices in addition to those prescribed by the prices and her budget; (ii) a consumer also faces informational restrictions due to her limited capability to observe and process information about her utility, available choices, and even a full bundle of prices, all of which brings about distortion during her information processing; and (iii) the consumer's choice is context dependent and inconsistent; so that a consumer hardly or never makes an optimal choice anticipated from a normative consumer. On the other hand, the lessons from a variety of heuristics developed for different domains regarding bounded rationality suggest that to pick and impose a specific presumptive hypothetical heuristic belief may not appropriate for a domain such as electricity consumer behavior. It may instead be desirable to drop or relax normative or heuristic beliefs that may be unnecessary or outside of our interests, and instead attempt to posit a flexible prior on a clean slate.

Adopting the viewpoint from [61] that a consumer or a decision maker is an information processor of limited capability, we can generalize a consumer's decision making process as a sequential pair of encoding and decoding process, i.e., $\vec{y} = g_d(g_e(\vec{x}))$, where $g_e : \mathbb{R}^k_{++} \to X'$ is an encoding operation, $g_d : X' \to \mathbb{R}^k_+$ is an decoding operation, and X' is an arbitrary vector space which rep-

resents the space of encoded prices, which reflects the whole body of encoded information for her decision making, including her subjective utility and the observable prices. We may interpret such a decomposition as follows: (i) a consumer first takes the price sequence as an input and encodes it into her internal representation, namely an *encoded price*, then (ii) she makes her choice and transforms the encoded price into a realization such as a consumption bundle or a load process as an output. If we let the framework be totally free without any specific prior restriction on the space of the pair of encoder and decoder, it may possess superfluous flexibility in the sense that there could be infinitely many encoder-decoder pairs (g_e, g_d) describing a common consumption behavior. However, the assumption of a normative consumer, i.e., a rational consumer with a rationalizable and differentiable utility, imposes certain restrictions on the framework such that (i) an encoder $g_e(\cdot)$ and the correspondent decoder $g_d(\cdot)$ are both invertible (homeomorphic) functions, and (ii) the decoder is uniquely determined by a given encoder.

This observation implies that the proposed model structure resembles Shannon's secrecy system [63] depicted in Fig. 3.2. Shannon's secrecy system is composed of three components, an encoder, a decoder, and a key source. If a system receives a message M from an information source, the encoder performs a functional operation $M' = g(M; K)$ to generate an encoded message or a cryptogram M', where K is a key randomly generated by the key source, and $g(\cdot, K)$ is an invertible function. Then, the decoder operation $M'' = g^{-1}(M'; K)$ which transforms M' space to M'' space, recovers the original message M. The similarity between Shannon's secrecy system and the proposed framework for a normative consumer behavior can be readily checked as follows: (i) both $g_e = g(\cdot, K)$ and $g_d = g^{-1}(\cdot, K)$ are invertible for all K; (ii) if an encoder $g(\cdot, K)$ is fixed by a given key K, then the decoder $g^{-1}(\cdot, K)$ is uniquely determined; and (iii) the key K is latent in the sense that it may be present but is supposed to be invisible to others, resembling the subjective utility function of a consumer. Thus, Shannon's secrecy system with minor modifications, which has a pair of homeomorphic encoder and decoder, allowing the decoder a general $g_d(\cdot, K)$ instead of $g^{-1}(\cdot, K)$, gives an equivalent model of a normative consumer.

One may see that the only stochastic part in Shannon's secrecy system, the key source com-

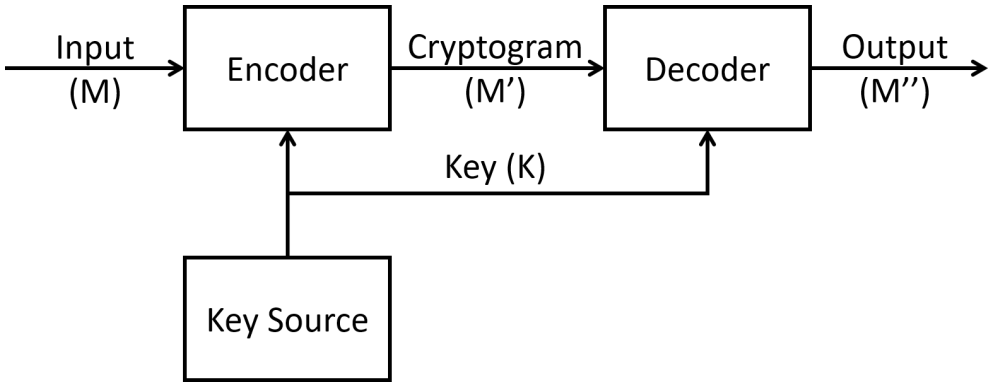


Figure 3.2: A schematic diagram of Shannon's general secrecy system [63].

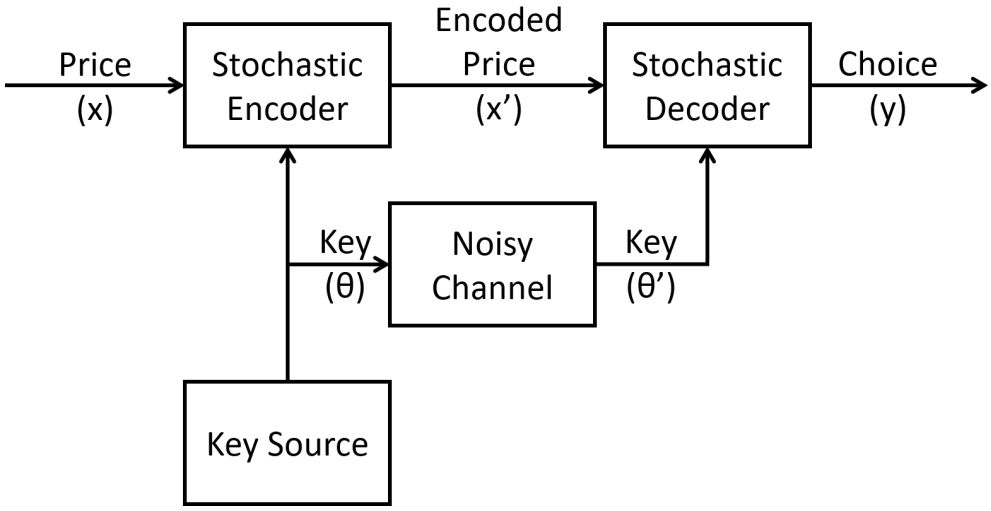


Figure 3.3: A schematic diagram of the proposed consumer behavior framework.

ponent that determines an active encoder-decoder pair, has a similar role to the utility function in consumer theory in the sense that it materially determines a demand function in neoclassical consumer theory. In the language of consumer theory, such stochasticity can be translated as follows: (i) the natural fluctuation in a consumer utility excited by exogenous environmental factors, and (ii) an *epistemic* uncertainty from its latent nature, affected by common environmental factors, drive price change such as weather condition in electricity consumption. We emphasize that the structural presumption that K blocks the influence of the confounder on the output load plays a crucial role for the identifiability of consumer behavior modeling. We address this in Theorem 3.3.9 in detail.

In addition to the epistemic uncertainty dwelling in the key source, relaxing the rationalizable and differentiable utility condition taking account of bounded rationality could be implemented by presuming extra uncertainties on each part of the modified secrecy model as follows, which is depicted in Fig. 3.3:

1. *The physical and institutional constraints* may arbitrarily limit the available choices or the controllability of the load. Thus, it is reasonable to posit a stochastic decoder [64] instead of a deterministic one to reflect such uncertain constraints.

2. Like physical/institutional constraints, *the informational constraints* can be implemented by the replacement of the deterministic encoder by a stochastic encoder [64], and the possibility of information distortion suggests that the encoder may not necessarily be a bijection.

3. *The inconsistency of preference* implies that the key θ which determines a stochastic encoder may not uniquely determine the corresponding stochastic decoder. We may consider implementing this implication by the insertion of a noisy channel between the key source and the decoder, which reflects a random distortion of θ to θ'.

However, presuming such multiple sources of uncertainty infuses ambiguity into the model framework: how to break down the observed noise into three distinct uncertainty sources in practice? We first note that it is reasonable to believe that the stochastic encoder and decoder are conditionally independent given a key θ, i.e., $g_e \perp\!\!\!\perp g_d \mid \theta$ or equivalently, $P(g_e, g_d|\theta) = P(g_e|\theta)P(g_d|\theta)$, unless there is a strong reason to posit that there is a certain interrelationship between informational constraints and physical constraints. On the other hand, we admit that a consumer has willpower to be rational so that it is reasonable to choose the most rational explanation if there are multiple possible accounts. This belief is in line with the idea that it is most probable to be a right representation if we find a representation of uncertainty decomposition having the least noisy channel in Fig. 3.3 among numerous ways of uncertainty decomposition keeping conditional independence between the stochastic encoder and decoder. Hence, another core problem of interest throughout this work, besides the model identifiability problem, is to find the model representation that provides a *most*

rational account from empirical data, which is articulated in the following sections. Using the language of machine learning, this goal could be paraphrased as the extraction of stochastic features that well factorize the demand response system.

The following examples from different types of electricity load usage for a consumer who is facing real-time electricity price may clarify *the principle of the most rational account*:

1. *The scenario for interruptible and deferrable "dishwahser" type loads*: One enjoys the benefit from a dish washer only after the completion of the running cycle. This inevitably involves a price forecast for the remaining running cycle period to decide whether to turn a dishwasher on or off. Among two options the consumer has, it is rational to keep it turned on only if the risk of total energy cost for the remaining running cycle exceeding the expected utility from the cleaned dishes is sufficiently low. For this type of load, the actual fluctuating real-time price may not be very informative to capture the load behavior, but the most rational account principle suggests that there exists a representation of her latent price forecast that almost determines the load behavior. This could be restated as the statistical parameter θ of a stochastic encoder, indicating a denoiser with epistemic uncertainty on the ways in which the actual prices are smoothed, *almost determines* a unique bijection that transforms the representation of smoothed prices to the observable load behavior. The measured load of a smart meter installed in a typical home could be a mixture of these types of loads with various load profiles e.g., a dish washer, a laundry machine, a coffee maker machine, and an electric oven, etc. Such a scenario can be captured by rephrasing the previous statement for the single dish washer case in a loosened way that a stochastic encoder almost determines a stochastic decoder, which in turn maps an encoded price to a load behavior taking into account the uncertainty on which appliances are on demand.

2. *The scenario for duty cycle controlled "air conditioner" type loads*: Most modern thermostats apply a hysteresis control (also known as a bang-bang control), so that a consumer is not capable of precisely controlling the load itself, but can only manipulate the reference room temperature. This indicates that actual prices may be useful for estimating the aver-

age proportion of time the air conditioner is in operation, but they may not be informative enough to precisely predict the current level of an air conditioner load. However, the most rational account principle implies that it is reasonable to believe there is a homeomorphic map from electricity price to her current reference room temperature setting, and such map almost determines the stochastic behavior of the air conditioner based upon reference temperature. This could be paraphrased by saying that a stochastic decoder, taking the reference temperatures to the actual load with epistemic uncertainty introduced by the hysteresis control, is determined by a bijection that transforms the price into the reference temperature. In the case of a consumer with various thermostats for multiple rooms, boilers, refrigerators etc., the deterministic encoder in the single air conditioner scenario could be extended to a stochastic encoder for the mapping from the price to the mixture of reference temperature settings for the thermostats for various locations and purposes.

On the other hand, we emphasize that it may not be common to have a perfect determinism between the stochastic encoder and decoder, since there may exist an *aleatory* uncertainty that cannot be reduced further, which could be interpreted as either irrationality, or the consumer's imperfect knowledge of her utility. Hence, how bijective the map between the stochastic encoder and the stochastic decoder is, could be used as a measure of rationality.

In summary, inspired by our observation that the normative neoclassical consumer model can be boiled down to a modified Shannon's secrecy system model, we propose an abstract framework for consumer behavior on the basis of the modified Shannon's secrecy system with uncertainty components, giving consideration to behaviorists' critiques. As the ambiguity on the representation of uncertainty remains, we attempt to break such ambiguity by the separation of the potential epistemic and aleatory uncertainties, and posit a minimum aleatory uncertainty assumption, namely the proposed principle of "most rational account".

A detailed description of the implementation will be provided in the following two sections. For the implementation of the pair of encoder and decoder, we will consider the application of a neural representation to take advantage of the full expressivity of a universal approximator to drop

all unnecessary assumptions for bottom-up model construction. To describe the innate stochasticity of the pair of encoder and decoder with neural representation, we propose a novel neural representation in Section 3.3, *a stochastic neuron*, which has a more concise form with an equivalent expressivity to a Bayesian Neural Networks with symmetric priors. Thereafter, in Section 3.4, we discuss how to measure rationality to eventually find the "most" rational account from the observed data. As we interpret the noisiness of the channel in Fig. 3.3 as an indicator of irrationality, we postulate that the measure of rationality is an innate channel property uniquely determined by the given channel itself. Based on the intuition that the demand response of a rational consumer is a *causal effect* driven by the price dynamics, we consider the use of causality measure as a measure of rationality in Section 3.4.

3.3 The Stochastic Artificial Neuron

In this section, we discuss how to represent the pair of stochastic encoder and decoder in detail. We consider a neural representation due to its ability to approximate arbitrary functions. In the past, a collection of results [65] [66] has shown that the conventional standard *feedforward network* (FFN) with one hidden layer is a universal approximator in the sense that it can approximate any continuous function of real variables arbitrarily well. Based on their flexibility and expressivity, neural networks have dramatically gained a great deal of popularity and are being successfully applied across a remarkably diverse range of problem domains such as medicine, finance, engineering, and energy. However, it is well known that the standard FFNs suffer from several drawbacks, and extensive efforts have been made to overcome them. First, from a Bayesian perspective, the training of a standard FFN is equivalent to a *maximum likelihood estimate* (MLE) for its weights, which is susceptible to overfitting. The most popular known remedy for this problem are the standard regularization techniques such as weight decay [67], which are essentially equivalent to the *maximum a posteriori* (MAP) estimate for weights inducing various specific priors on the weights. Aside from the traditional explicit regularization approaches, there are widely adopted implicit variants of these, such as the dropout technique [68] which prevents overfitting by injecting a noise during the training; and batch normalization [72] which is an operator that

normalizes the layer inputs within each mini-batch for the purpose of stabilizing the input distribution during training. Besides those, traditional methods to avoid overfitting such as bootstrap aggregating (bagging) [69, 70], and early stopping in the iterative training process [71], are also widely used.

Second, the determination of the proper neural architecture is a challenging task since it involves an expensive global search process. As a core topic in the study of *automated machine learning* (AutoML), this task addresses the problems of finding the optimal numbers of layers and neurons in a neural network, and the types of activation functions. While neural network architectures were typically designed by experts in a painstaking and ad hoc fashion in the past, there has been a recent surge of interest in *network architecture search* (NAS), and many black-box methodological strategies for NAS have been proposed, e.g., based on evolutionary algorithms, random search, Bayesian optimization, evolutionary methods, reinforcement learning, and gradient-based methods [73, 75]. Regarding the size of the network, Neal's work [81] breaks the common beliefs in the machine learning community by introducing Bayesian learning on an FFN with a hidden layer of infinite number of hidden units, and showing the following findings [74]: (i) neural networks with a very large number of hidden units can avoid overfitting via Bayesian learning, and (ii) such huge networks are still computationally feasible to be numerically optimized. Extending Neal's idea, Le Roux and Bengio introduce continuous neural networks [74] by replacing the sum over neurons by an integral over neurons with different continuous weight functions assigned to a neuron. Philipp and Carbonell introduce nonparametric neural networks [75] as an alternative to black-box based NAS for deep networks, which is a non-probabilistic framework to reduce the heavy computation required for typical black-box NAS approaches.

Last but not least, conventional standard FFNs are not suitable for modeling the uncertainty associated with their model parameters and the predictions they make. For that, the most common way is to consider a full Bayesian treatment. Given a dataset $\mathcal{D} \coloneqq \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{n}$, where $\mathbf{X} \in \mathbb{R}^{n \times d_x}, \mathbf{Y} \in \mathbb{R}^{n \times d_y}$, let $(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ with $\phi(\cdot)$ is a nonlinear activation. A *Bayesian FFN*

(BFN) with single hidden layer [76] is defined as

$$BFN(\mathbf{x}) \coloneqq \phi(\mathbf{W}^{(x)}\mathbf{x})^{\top}\mathbf{W}^{(y)} \tag{3.6}$$

where $(\mathbf{W}^{(x)}, \mathbf{W}^{(y)})$ are random weight matrices on $(\mathbb{R}^{m\times d_x}, \mathbb{R}^{m\times d_y})$.[4] There are well known notable advantages of such treatment [81]. In addition to its representation capability with respect to the model uncertainty, a BFN provides a unified view of various techniques devised for overfitting avoidance in standard FFN training, as the training of a standard FFN is deemed to be just a point estimate of a BFN obtained by various estimation techniques such as MLE or MAP. Another important main, and commonly believed, benefit of Bayesian modeling is that it is immune to overfitting. This implies that controlling model complexity based on the size of the collected data is in fact a theoretically irrelevant idea, and that practical matters including the computational expense should be the only reasons for limiting the network size [81]. However, in terms of practicality, such BFNs have several shortcomings, as follows: (i) the Bayesian setting of the state of the art of a complex deep neural net with over millions of parameters involves complicated high dimensional integrals which render its estimation and prediction cost prohibitively expensive, forcing us to resort to poor approximations; and (ii) Such models of highly complex representation are still incomprehensible, like other complex neural networks.

Taking notice of the fundamental symmetry inherent in neural networks arising from its additive structure, we consider the reduction of a whole layer of a BFN to a more succinct form, a "random neuron", with an asymptotically zero sacrifice of its expressivity. This task is inspired from de Finetti's view that the consistency of an inference solely stems from the symmetry of observations, when we aim to encode information from a sequence of observed quantities of a random phenomenon. Our suggestion comes by tweaking the original idea of de Finnetti in the sense that symmetry in a BFN allows us to devise a consistent method for its compression, which leads to the idea of a stochastic neuron that we propose. Specifically, such symmetry of observations

[4]This definition is a simplified form of a BFN for less verbosity. However, all of the results presented in this section can be easily extended to the original definition of $BFN(\mathbf{x}) \coloneqq \mathbf{W}^{(y)\top}\phi(\tilde{\mathbf{W}}^{(x)}\tilde{\mathbf{x}}) + \vec{y}_0$, where $\tilde{\mathbf{x}} \coloneqq [\mathbf{x}^{\top}, 1]^{\top}$, $\tilde{\mathbf{W}}^{(x)} \coloneqq [\mathbf{W}^{(x)}, \mathbf{x}_0]$, $\mathbf{x}_0 \in \mathbb{R}^{m\times 1}$, and $\vec{y}_0 \in \mathbb{R}^{d_y\times 1}$.

implies *order irrelevance*, which can be articulated into *exchangeability* as follows:

Definition 3.3.1 (Exchangeable Random Variables)**.** A sequence of random variables $\{Z_i\} \coloneqq \{Z_1, ..., Z_m\}$ is said to be *exchangeable* if $P(\{Z_i\}) = P(\{Z_{\pi(i)}\})$ for every permutation π of $\{1, ..., m\}$.

The structural symmetry we make use of in this work originates from the commutative property of summation. Since $S_n = \sum_{i=1}^{m} Y_i$ is invariant to the exchange of Y_i, so is $P(\sum_i Y_i \leq s)$; i.e., $P(\sum_i Y_i \leq s) = \int_{\sum_i y_i \leq s} dF_{\{Y_i\}}(\{y_i\}) = \int_{\sum_i y_{\pi(i)} \leq s} dF_{\{Y_{\pi(i)}\}}(\{y_{\pi(i)}\})$ for all permutations π of $\{1, ..., m\}$, where $F(\cdot)$ is a cumulative distribution function (CDF).

When it comes to a BFN, it is fair to posit that all neurons are assigned equal priors if we have no reason to believe that the neurons are a priori different. We say that a BFN has a *neurosymmetric prior* if the priors of the row $\mathbf{w}_i$'s in the weight matrix $\mathbf{W} \coloneqq [\mathbf{W}^{(x)}, \mathbf{W}^{(y)}]$, as well as the interdependencies between them are identical with each other, i.e., $\mathbf{W}$ is *row-exchangeable*. Note that most well recognized regularization schemes used in a large body of literature share an underlying assumption of i.i.d. priors when we interpret them from the Bayesian point of view, which leads to the notion of neurosymmetric priors.

Definition 3.3.2 (Neurosymmetric Prior of a BFN)**.** A $BFN(\mathbf{x}) = \phi(\mathbf{W}^{(x)}\mathbf{x})^\top \mathbf{W}^{(y)}$ has a *neurosymmetric prior* on $\mathbf{W}$ if $P(\mathbf{W}) = P(\Pi\mathbf{W})$ for every $m \times m$ permutation matrix Π.

An interesting feature of the neurosymmetry of a BFN is that it is closed under belief updates, i.e., the neurosymmetry holds throughout the belief evolution on the weight space of the BFN driven by the learning process over incoming datasets.

Lemma 3.3.1. If a $BFN(\mathbf{x}) = \phi(\mathbf{W}^{(x)}\mathbf{x})^\top \mathbf{W}^{(y)}$ provided by a dataset $\mathcal{D}$ has neurosymmetric prior on $\mathbf{W}$, the posterior $P(\mathbf{W}|\mathcal{D})$ is also *neurosymmetric*.

Proof. Let $\mathbf{w}_i^{(x)}$ and $\mathbf{w}_i^{(y)}$ denote the ith row of $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(y)}$ respectively. Then, $BFN(\mathbf{x})$ is

invariant to the row exchange of $\mathbf{W}$, i.e.,

$$\begin{aligned}\phi(\mathbf{W}^{(x)}\mathbf{x})^{\top}\mathbf{W}^{(y)} &= \sum_{i=1}^{m}\phi(\mathbf{w}_i^{(x)}\cdot\mathbf{x})\mathbf{w}_i^{(y)} = \sum_{i}^{m} Y_i \\ &= \sum_{i=\pi(1)}^{\pi(m)} Y_i\end{aligned} \tag{3.7}$$

for all permutations π of $\{1,...,m\}$, where $Y_i := \phi(\mathbf{w}_i^{(x)}\cdot\mathbf{x})\mathbf{w}_i^{(y)}$. Thus, the likelihood of $BFN(\mathbf{x})$ is permutation invariant to its parameters, i.e.,

$$P(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}) = P(\mathbf{Y} \mid \Pi\mathbf{W}, \mathbf{X}) \tag{3.8}$$

for all permutation matrices Π. On the other hand, the LHS and RHS of equation (3.8) can be rewritten as follows by Bayes' rule:

$$\begin{aligned}P(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}) &= \frac{P(\mathbf{W} \mid \mathbf{X}, \mathbf{Y})P(\mathbf{Y}|\mathbf{X})}{P(\mathbf{W}|\mathbf{X})}, \text{ and} \\ P(\mathbf{Y} \mid \Pi\mathbf{W}, \mathbf{X}) &= \frac{P(\Pi\mathbf{W} \mid \mathbf{X}, \mathbf{Y})P(\mathbf{Y}|\mathbf{X})}{P(\Pi\mathbf{W}|\mathbf{X})},\end{aligned} \tag{3.9}$$

so that we obtain

$$P(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}) = P(\Pi\mathbf{W} \mid \mathbf{X}, \mathbf{Y}) \quad \forall\Pi, \tag{3.10}$$

since $\mathbf{W}$ and $\mathbf{X}$ are independent, and the $BFN(\mathbf{x})$ has neurosymmetric prior.

Hence, $P(\mathbf{W} \mid \mathcal{D})$ is invariant to the row exchange of $\mathbf{W}$. Here, note that the row-exchangeability of $\mathbf{W}|\mathcal{D}$ implies that $\{Y_i|\mathcal{D}\}$ is also exchangeable if $\phi(\cdot)$ is Borel measurable by Lemma 1 in [77]. □

We first consider the nonparametric version of BFN to show useful limiting properties.

Definition 3.3.3 (Infinitely Exchangeable Random Variables)**.** A stochastic process $\{Z_i\}_{i=1}^{\infty}$ is an *infinitely exchangeable random sequence* if the joint probability $P(\{Z_1,,...,Z_m\})$ is invariant to permuatation of the indices for any m.

Lemma 3.3.2. Consider a nonparametric BFN, $BFN^{\infty}(\mathbf{x}) \coloneqq \phi(\mathbf{W}^{(x)}\mathbf{x})^{\top}\mathbf{W}^{(y)}$, with neurosymmetric prior on $\mathbf{W} \in \mathbb{R}^{m\times(d^x+d^y)}$ with $m \to \infty$. If $BFN^{\infty}(\mathbf{x})$ is well defined, i.e., $E[|BFN^{\infty}(\mathbf{x})|] < \infty$, for all $\mathbf{x}$, $\mathbf{W}|\mathcal{D}$ is *infinitely row-exchangeable* for a given dataset $\mathcal{D}$.

Proof. Take $BFN^{\infty}(\mathbf{x}) = \sum_{i=1}^{\infty} Y_i \coloneqq \sum_{i=1}^{\infty} \phi(\mathbf{w}_i^{(x)} \cdot \mathbf{x})\mathbf{w}_i^{(y)}$. For an arbitraty $m_0 < \infty$, $\sum_{i=m_0+1}^{\infty} Y_i$ is also well defined if $BFN^{\infty}(\mathbf{x})$ is well defined. Hence,

$$\begin{aligned} BFN^{\infty}(\mathbf{x}) &= \sum_{i=1}^{\infty} Y_i = \sum_{i=1}^{m_0} Y_i + \sum_{i=m_0+1}^{\infty} Y_i \\ &= \sum_{i=\pi(1)}^{\pi(m_0)} Y_i + \sum_{i=m_0+1}^{\infty} Y_i, \end{aligned} \tag{3.11}$$

for all $m_0 \in \mathbb{N}$ since $\{Y_i\}_{i=1}^{m_0}$ is exchangeable. □

As the neurosymmetric belief of a BFN is invariant to belief updates, it is pointless to specify the chronological information of neurosymmetry explicitly.

Definition 3.3.4 (Neurosymmetric Bayesian Feedforward Neural Network)**.** The Neurosymmetric Bayesian Feedforward Neural Network (NBFN) is the BFN with neurosymmetric belief.

It is well known that infinitely exchangeable random variables are conditionally i.i.d. given their empirical distribution, by the de Finetti-Hewitt-Savage Representation Theorem. Such a conditionally i.i.d. property allow us to exploit some desirable features originally produced from i.i.d. random variables.

Lemma 3.3.3 (de Finetti-Hewitt-Savage for NBFNs)**.** For any nonparametric NBFN, the beliefs on rows $\{\mathbf{w}_i\}$ of $\mathbf{W}$ are *conditionally i.i.d.* on a random probability distribution F on $\mathbf{W}$ such that $F_m \xrightarrow{d} F$, where F_m is the empirical distribution of $\{\mathbf{w}_i\}_{i=1}^{m}$.

Proof. The nonparametric NBFN implies that $\{\mathbf{w}_i\}_{i=1}^{\infty}$ is infinitely exchangeable. The rest of the proof directly follows from the de Finetti-Hewitt-Savage representation theorem [78] by taking $Z_i = \mathbf{w}_i$, stating that a sequence $\{Z_i\}_{i=1}^{\infty}$ with its members in a real vector space (or more generally a Polish, or complete separable metric, space) is infinitely exchangeable if and only if each Z_i is

conditionally i.i.d. over F and $P(Z_i \leq z|F) = F(z)$, where $F = \lim_{m\to\infty} F_m$ with the random measure $F_m := m^{-1}\sum_{i=1}^{m} I_{Z_i\leq z}$, i.e. the empirical distribution of $\{Z_i\}_{i=1}^{m}$, and $I_{\{.\}}$ is the indicator function. □

Kolmogorov's Strong Law of Large Numbers (SLLN) states that the empirical mean of an i.i.d. sequence is an unbiased estimator for the expectation of an element of the sequence. The fact that SLLN holds for exchangeable sequences suggests that an NBFN is in fact an unbiased estimator of $E[Y_i]$ after rescaling depending on its size. This implies that the information on the belief of a neuron is sufficient to simulate a nonparametric NBFN, which may dramatically reduce the dimension of the parameter space we should work on.

Definition 3.3.5 (Stochastic Neuron)**.** Let $\mathbf{w} := (\mathbf{w}^{(x)}, \mathbf{w}^{(y)})$ be a random vector defined on $(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$. The *stochastic neuron* $SN(\mathbf{x})$ is defined as follows:

$$SN(\mathbf{x}) = \phi(\mathbf{w}^{(x)} \cdot \mathbf{x})\mathbf{w}^{(y)}$$

The following theorem shows that a nonparametric NBFN is an unbiased estimator of the expectation of an SN.

Theorem 3.3.4 (Strong Law of Large Numbers for NBFNs)**.** A nonparametric NBFN almost surely converges to a value $\hat{\mathbf{y}} \in \mathbb{R}^{d_y}$ as $m \to \infty$, if and only if an SN equals $\hat{\mathbf{y}}$ almost surely, i.e.,

$$\begin{aligned} m^{-1}\phi(\mathbf{W}^{(x)}\mathbf{x})^{\intercal}\mathbf{W}^{(y)} \to \hat{\mathbf{y}}\ a.s. \iff \\ E[\phi(\mathbf{w}^{(x)} \cdot \mathbf{x})\mathbf{w}^{(y)}] = \hat{\mathbf{y}}\ a.s. \end{aligned} \tag{3.12}$$

Proof. The conditional version of Kolmogorov's SLLN [79] is as follows: let $(\Omega, \mathscr{A}, P)$ be a probability space, and $\mathscr{F}$ a sub-σ-algebra of $\mathscr{A}$. If $\{Z_i\}_{i\geq 1}$ is $\mathscr{F}$-i.i.d., then

$$m^{-1}\sum_{i=1}^{m} Z_i \to Z^*\ a.s. \iff E[Z_1 \mid \mathscr{F}] = Z^*\ a.s. \tag{3.13}$$

The theorem still holds when we take $\mathscr{F}$ to be the σ-algebra in the probability space on which the

random probability distribution F in Lemma 3.3.3 is defined, and $Z_i = Y_i$, where $Y_i := \phi(\mathbf{w}_i^{(x)} \cdot \mathbf{x})\mathbf{w}_i^{(y)}$, so that $\sum_{i=1}^{m} Y_i = NBFN$, and $E[Y_1] = E[SN]$. □

Another elegant approach for the proof of the SLLN for exchangeable sequences is to use the backwards martingale convergence theorem, making use of the fact that $m^{-1}\sum_{i=1}^{m} Z_i$ is a backward martingale if $\{Z_i\}_{i\geq 1}$ is infinitely exchangeable; as shown in Examples 12.13 and 12.15 in [80]. Taking into account that an FFN is in fact a point estimate form of a BFN, Theorem 3.3.4 suggests that the $\hat{\mathbf{y}}$ predicted via a standard FFN trained to be a point estimate of an NBFN, and the prediction made by an FFN whose parameters are sampled from $P(\mathbf{w}|\mathcal{D})$ by an SN, are asymptotically the same.

It is notable that there is no particular restriction on the distribution of $\mathbf{y}|\mathbf{x}$ specified by an SN, while a nonparametric NBFN is in fact a Gaussian Process. It is well known that a nonparametric BFN with i.i.d. Gaussian priors is in fact a Gaussian process [81], which holds for deep BFNs with layers more than one [82] as well as those with an arbitrary i.i.d priors [83]. We now show that a nonparametric NBFN is also a Gaussian process,

Lemma 3.3.5 (The Conditional Multivariate Central Limit Theorem)**.** Let a sequence of random vectors $\{Z_i\}_{i\geq 1}$ be $\mathscr{F}$-i.i.d with $\Sigma_{\mathscr{F}} := E[(Z_i - E[Z_i|\mathscr{F}])(Z_{i'} - E[Z_{i'}|\mathscr{F}])^{\intercal} \mid \mathscr{F}]$ such that $\|\Sigma_{\mathscr{F}}\| < \infty\ a.s.$ Then,

$$(m\Sigma_{\mathscr{F}})^{-\frac{1}{2}}\big(S_m - E[S_m \mid \mathscr{F}]\big) \xrightarrow{d} \mathcal{N}(\vec{0}, I), \tag{3.14}$$

where $S_m = \sum_{i=1}^{m} Z_i$, and $\mathcal{N}(\vec{\mu}, \Sigma)$ is a multivariate normal distribution, with I the identity matrix.

Before proceeding to the proof, we note that the conditional version of the classical central limit theorem [84, 85] can be stated as follows: Let $\{Z_i\}_{i\geq 1}$ be $\mathscr{F}$-i.i.d with $\sigma^2_{\mathscr{F}} := E[(Z_1 - E[Z_1|\mathscr{F}])^2 \mid \mathscr{F}] < \infty\ a.s.$, and $S_m := \sum_{i=1}^{m} Z_i$. Then,

$$\frac{m^{-1}S_m - E[m^{-1}S_m \mid \mathscr{F}]}{m^{-\frac{1}{2}}\sigma_{\mathscr{F}}} \xrightarrow{d} N(0,1). \tag{3.15}$$

Here, $E[m^{-1}S_m|\mathscr{F}] = m^{-1}\sum_{i=1}^{m} E[Z_i|\mathscr{F}] = E[Z_1|\mathscr{F}]$.

The proof of the relation (3.15) follows from the original form of the conditional CLT:

$$E\Big[\exp\Big(\iota t\frac{m^{-1}S_m - E[m^{-1}S_m \mid \mathscr{F}]}{m^{-\frac{1}{2}}\sigma_{\mathscr{F}}}|\mathscr{F}\Big)\Big] \to e^{-\frac{t^2}{2}}\ a.s. \tag{3.16}$$

where $\iota^2 = -1$, using the dominated convergence theorem and Lévy's continuity theorem for characteristic functions [84, 85]. We can readily extend the relation (3.16) to the multivariate version by following standard arguments.

Proof. Let Z_i be an n dimensional random vector. Then

$$\begin{aligned}
&E[\exp\big(\iota\vec{t}^{\,\top}(m\Sigma_{\mathscr{F}})^{-\frac{1}{2}}\big(S_m - E[S_m \mid \mathscr{F}]\big)\big) \mid \mathscr{F}] \\
=&E\Big[\prod_{k=1}^{m}\exp\big(\iota\vec{t}^{\,\top}(m\Sigma_{\mathscr{F}})^{-\frac{1}{2}}\big(Z_k - E[Z_k \mid \mathscr{F}]\big)\big) \mid \mathscr{F}\Big] \\
=&\prod_{k=1}^{m}E\Big[\exp\big(\iota\vec{t}^{\,\top}(m\Sigma_{\mathscr{F}})^{-\frac{1}{2}}\big(Z_k - E[Z_k \mid \mathscr{F}]\big)\big) \mid \mathscr{F}\Big] \\
=&\Big(E\Big[\exp\big(\iota\vec{t}^{\,\top}(m\Sigma_{\mathscr{F}})^{-\frac{1}{2}}\big(Z_k - E[Z_k \mid \mathscr{F}]\big) \mid \mathscr{F}\Big]\Big)^m \\
=&\Big(1 - \frac{\vec{t}^{\,2}}{2m} + E\Big[o\big(\frac{\vec{t}^{\,\top}\Sigma_{\mathscr{F}}^{-1}\vec{t}}{m}\big)|\mathscr{F}\Big]\Big)^m \to e^{-\vec{t}^{\,2}/2}\ a.s.
\end{aligned} \tag{3.17}$$

as $E\big[o\big(\frac{\vec{t}^{\,\top}\Sigma_{\mathscr{F}}^{-1}\vec{t}}{m}\big)|\mathscr{F}\big] \to 0\ a.s.$ for a fixed $\vec{t} \in \mathbb{R}^n$. □

Theorem 3.3.6 (Central Limit Theorem for NBFN)**.** A nonparametric NBFN converges to a Gaussian Process.

Proof. If we take $\mathscr{F}$ to be the σ-algebra in the probability space on which the random probability distribution F and $Z_i = [Y_{ik}(\mathbf{x}_j)]_{j=1}^n$, then $P\big([y_{jk}(\mathbf{x}_j)]_{j=1}^n\big) = P\big([\sum_{i=1}^m Y_{ik}(\mathbf{x}_j)]_{j=1}^n\big)$ converges to a joint multivariate Gaussian for finite n arbitrary $\mathbf{x}_j$'s for any $k = 1, ..., d_y$ by Lemma 3.3.5, where $(\mathbf{x}_j, \mathbf{y}_j) := ([x_{jk}]_{k=1}^{d_x}, [y_{jk}]_{k=1}^{d_y})$ and $\mathcal{D} = \{(\mathbf{x}_j, \mathbf{y}_j) : j = 1, ..., n\}$. □

Aside from the expressivity in prediction, another matter of interest to us is the feature expressivity of SN concerning how well an SN is available to capture the features extracted by an NBFN. The classical Glivenko-Cantelli theorem states that the empirical distribution of an i.i.d. sequence converges to a common cumulative distribution function. We make use of the result given by Berti

and Rigo [86] proving that the Glivenko-Cantelli theorem holds for infinitely exchangeable sequences. We show that the empirical distribution of $\mathbf{w}_i$ provided by $\{\mathbf{w}_i\}_{i=1}^m$ in NBFN converges to a common probability distribution, namely, that of $\mathbf{w}$ in the SN.

Theorem 3.3.7 (Glivenko-Cantelli Theorem for NBFN)**.** For an NBFN, there exists a random probability measure $\tilde{P}$ on $\mathbf{w}$ such that $\|P(\mathbf{w}_{m+1}|\{\mathbf{w}_i\}_{i=1}^m) - \tilde{P}(\mathbf{w}|F_m)\|_{TV} \to 0\ a.s.$, where $\|P - \tilde{P}\|_{TV} \coloneqq \sup_A |P(A) - \tilde{P}(A)|$ is the total variation distance, and F_m is the empirical distribution from $\{\mathbf{w}_i\}_{i=1}^m$ so that $\tilde{P}(\mathbf{w} \leqq \vec{w}|F_m) = F_m(\vec{w})$.

Note that Theorem 3.3.7 suggests that F_m on $\mathbf{w}$ is sufficient to describe $\mathbf{w}_i$.

Proof. Berti and Rigo [86] extend the classical Glivenko-Cantelli theorem for exchangeable random sequences, showing that the difference between the predictive and empirical distributions converges to zero almost surely uniformly, i.e.,

$$\|P\big(Z_{m+1} \leqq z \mid \{Z_i\}_{i=1}^m\big) - F_m(z)\|_{TV} \to 0\ a.s. \tag{3.18}$$

for an exchangeable sequence $\{Z_i\}_{i\geq 1}$ with its members in a real vector space (Lusin space). The proof of the theorem 3.3.7 directly follows by taking $Z_i = \mathbf{w}_i$. □

Coming back to finite width NBFNs, finite exchangeability does not allow us to enjoy desirable results provided by infinite exchangeability in general [87]. However, we present a corollary without proof which shows that the joint distribution of exchangeable sequences may be approximated as the product of empirical distribution of the sequence, based on Theorem 13 by Diaconis and Freedman [88], which establishes the sharp bounds for the approximations of sampling $\mathbf{w}_i$, providing the bounds for the probabilistic difference of sampling with and without replacement.

Corollary 3.3.8 (Finite Version of de Finetti-Hewitt-Savage)**.** Consider an NBFN with M neurons. There exists a random probability measure $\tilde{P}$ on $\mathbf{w}$ such that $\|P(\{\mathbf{w}_i\}_{i=1}^m) - \prod_{i=1}^m \tilde{P}(\mathbf{w}|F_M)\|_{TV} \leq 1 - \frac{M^{-m}M!}{(M-m)!}$ for $m \leq M$, where F_M is the empirical distribution from $\{\mathbf{w}_i\}_{i=1}^M$ so that $\tilde{P}(\mathbf{w} \leqq \vec{w}|F_M) = F_M(\vec{w})$.

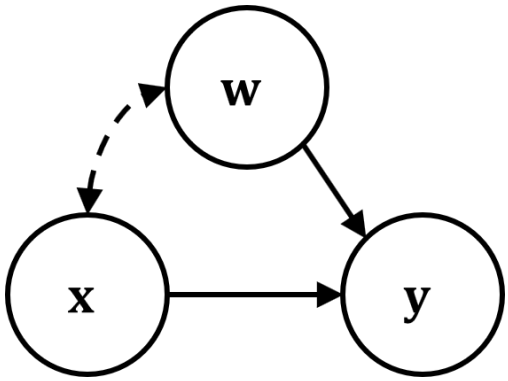


Figure 3.4: The causal diagram of the proposed consumer behavior model.

The Corollary 3.3.8 provides an upper bound on the potential discrepancy between the joint probability distribution on $\{\mathbf{w}_i\}$ in NBFN and the product of the empirical distribution of $\mathbf{w}$ in SN, which uniformly approaches to 0 as $M \to \infty$. Note that the Corollary 3.3.8 can be interpreted as the cost of the mean field approximation.

In summary, we consider a neural representation for the pair of stochastic encoder and decoder in the proposed consumer behavior framework through this subsection. As previously developed various neural representations have a limited fit for our purpose, we propose a novel neural representation for that, namely a *stochastic neuron*, and demonstrate that it has good expressivity as powerful as a BFN. Specifically, we show the correspondence between SN and NBFN by limiting its width to infinity. The SN and NBFN framework can be readily extended to deep NBFNs on the basis of the results on the partial exchangeability but we do not treat it in this work as it is beyond our scope.

The proposed consumer behavior model, an incarnation of the consumer behavior framework via the representation of stochastic artificial neuron, renders the simplest elaboration of the unidentifiable causal model in Fig. 3.1(a) into the identifiable form shown in the following Theorem 3.3.9. This Theorem closes the fundamental problem of consumer behavior modeling introduced in Section 3.1 as follows:

Theorem 3.3.9 (The Identifiability of the Consumer Behavior Model)**.** The proposed model $\mathcal{M} = \left\langle V = O \cup N,\ G_V,\ \mathbf{y} = \phi(\mathbf{w}^{(x)} \cdot \mathbf{x})\mathbf{w}^{(y)} \right\rangle$ is identifiable, where $\mathbf{w} = [\mathbf{w}^{(x)}\mathbf{w}^{(y)}]$, $O = \{\mathbf{w}, \mathbf{x}, \mathbf{y}\}$,

$N = \{\boldsymbol{\nu}\}$ where $\boldsymbol{\nu}$ is the vector representation of all the unknown confounders affecting both $\mathbf{x}$ and $\mathbf{w}$, with G_V the causal diagram shown in Fig. 3.4.

Proof. Since the model $\mathcal{M}$ is semi-Markovian, $\mathcal{M}$ is identifiable by Lemma 3.1.1. Such identifiability can be readily proven by the direct computation of $P_{\vec{x}}(\mathbf{y})$. First. we can write

$$P_{\vec{x}}(V) = P(\boldsymbol{\nu})P(\mathbf{w}|\boldsymbol{\nu})P(\mathbf{y}|\mathbf{w},\vec{x}).$$

Here, even though we presume that we do not have any information on $P(\boldsymbol{\nu})$ and $P(\mathbf{w}|\boldsymbol{\nu})$, these terms effectively vanish due to the following marginalization:

$$\begin{aligned} P_{\vec{x}}(\mathbf{y}) &= \int_{\mathbf{w},\boldsymbol{\nu}} P(\mathbf{y}|\mathbf{w},\vec{x})P(\mathbf{w}|\boldsymbol{\nu})P(\boldsymbol{\nu})d\mathbf{w}d\boldsymbol{\nu} \\ &= \int_{\mathbf{w}} P(\mathbf{y}|\mathbf{w},\vec{x}) \int_{\boldsymbol{\nu}} P(\mathbf{w}|\boldsymbol{\nu})P(\boldsymbol{\nu})d\mathbf{w}d\boldsymbol{\nu} \\ &= \int_{\mathbf{w}} P(\mathbf{y}|\mathbf{w},\vec{x})P(\mathbf{w})d\mathbf{w} \\ &= E_{\mathbf{w}}[P(\mathbf{y}|\mathbf{w},\vec{x})]. \end{aligned}$$

Hence, we obtain the unique $P_{\vec{x}}(\mathbf{y})$ from $P(O)$. □

However, an important caveat to the utilization of the exchangeability is that it does not provide any guidance about setting priors. The key message from the de Finetti theorem is that there exists some prior which well represents the data, but the choice of priors remains as a subjective matter that requires further modeling assumptions to obtain nontrivial statements. In fact, setting up a prior is rather domain-specific in the sense that it has a strong linkage to the task of imposing a good structure *a priori* inherent in the problem of our interest, which is expected to play a central role as a right scaffolding for an effective analysis. We emphasize that this is where the proposed consumer behavior framework depicted in Fig. 3.3 and *the principle of the most rational account* come in. Regarding the choice of prior, we posit the most rational account principle as a minimal key assumption for an innate prior for the effort to keep parsimony in the plurality of assumptions, i.e., consumers are price-takers who tend to be rational so that a stochastic decoder is almost

determined by the stochastic encoder.

If we take an SN to be the pair of stochastic encoder and decoder, we can naturally interpret $\mathbf{w}^{(x)}$ be the stochastic component of the encoder, and $\mathbf{w}^{(y)}$ be that of the decoder. Taking the probability distributions of $\mathbf{w}^{(x)}$ and $\mathbf{w}^{(y)}$ to be $P(\mathbf{w}^{(x)};\vec{\theta}^{(x)})$ and $P(\mathbf{w}^{(y)};\vec{\theta}^{(y)})$ respectfully, where $\vec{\theta}^{(\cdot)}$ is the set of parameters for the corresponding probability distributions, then we may consider a hierarchical structure of random variables to separate the stochasticity of the encoder and the chance of choosing a stochastic encoder. The simplest form having such structure is the mixture distribution, i.e,

$$\begin{aligned} P(\mathbf{w}^{(\cdot)};\vec{\theta}^{(\cdot)}) = \sum_{k=1}^{l} \alpha_k^{(\cdot)} P_k(\mathbf{w}^{(\cdot)};\vec{\theta}_k^{(\cdot)}) \text{ such that} \\ \sum_{k=1}^{l} \alpha_k^{(\cdot)} = 1 \text{ and } \alpha_k^{(\cdot)} \geq 0,\ \forall k = \{1,...,l\}, \end{aligned} \tag{3.19}$$

where $\alpha_k^{(\cdot)}$ and $P_k(\cdot)$ are the weight and the probability distribution of a mixture component k, and l is the number of mixture components. In other words, $\alpha_k^{(\cdot)}$ represents the probability of choosing the stochastic encoder/decoder k, and $P_k(\mathbf{w}^{(\cdot)};\vec{\theta}_k^{(\cdot)})$ represents the internal stochasticity of the chosen encoder/decoder k. For Example, if we take a Gaussian mixture model for $\mathbf{w}^{(\cdot)}$, then $\mathbf{w}^{(\cdot)}|k \sim \mathcal{N}(\vec{\mu}_k^{(\cdot)}, \Sigma_k^{(\cdot)})$ so that $\vec{\theta}_k^{(\cdot)} = \{\vec{\mu}_k^{(\cdot)}, \vec{\Sigma}_k^{(\cdot)}\}$. Hence, the consumer behavior model with SN can be depicted as Fig. 3.5, where $K^{(\cdot)} \sim \text{Cat}(\vec{\alpha}^{(\cdot)})$ is the categorical random variable indicating the encoder/decoder k was chosen with parameter $\vec{\alpha}^{(\cdot)} = [\alpha_k^{(\cdot)}]_{k=1}^{l}$, and $A(K^{(x)}, K^{(y)})$ is a Markov kernel.

From de Finetti's and Bayesian perspective, the number of mixture components is the part of prior information that should be set subjectively. Here, the most rational account principle provides an essential guideline to deal with this subjectivity for consumer behavior modeling. Having an SN as a basic building block of the proposed consumer behavior model, the most rational account principle can be stated as follows: (i) The choice of decoder j and the choice of encoder i are "entangled" each other, i.e., the relation between $\{\vec{\theta}_i^{(x)} : i = 1,...,l\}$ and $\{\vec{\theta}_j^{(y)} : j = 1,...,l\}$ is bijection; and (ii) The stochasticities within an encoder and a decoder are mutually independent, i.e., $P(\mathbf{w}^{(x)}, \mathbf{w}^{(y)}; \boldsymbol{\theta}^{(x)}, \boldsymbol{\theta}^{(y)}) = P(\mathbf{w}^{(x)}; \boldsymbol{\theta}^{(x)}) P(\mathbf{w}^{(y)}; \boldsymbol{\theta}^{(y)})$ for all i and $j = 1,...,l$. We can take

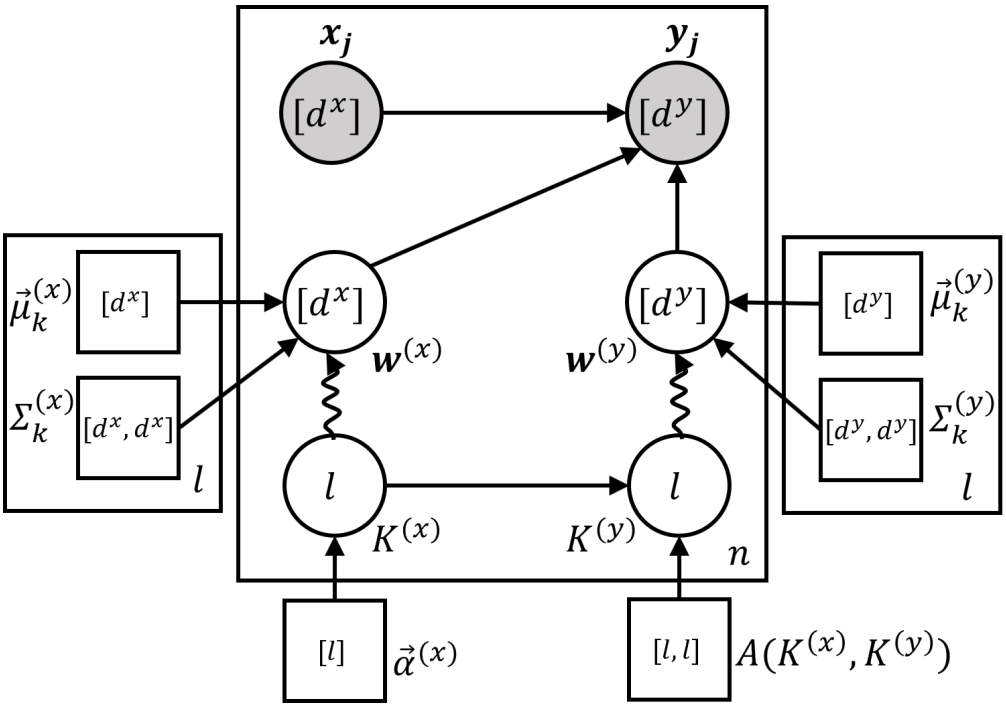


Figure 3.5: The plate diagram for the consumer behavior model with SN.

the number of mixture components as the number which produces a trained model satisfying both statements most. However, to quantifying how much a model fit to the most rational account principle is yet obscure, and should be necessarily formalized.

As we interpret the noisiness of the channel in the proposed consumer behavior framework, i.e., the Markov kernel $A(K^{(x)}, K^{(y)})$ in Fig. 3.5, as the indication of irrationality, the measure for the fitness of a model to to the most rational account principle could be used for the measure of rationality. The detailed explanation of how the concept of causality is employed in the principle of the most rational account is provided in Section 3.4.

In fact, the SN is mere a BFN with one neuron so that any training algorithm proposed for BFN in previous literature is applicable for training an SN. E.g., [89] proposes training algorithm to learn variational posterior for a Bayesian neural network with Gaussian mixture density. Instead of suggesting a new Bayesian style training algorithm, we focus on how a pretrained standard FFN can be successfully transformed to SN and vice versa in this work. We first obtain pretrained standard FFN via traditional methods. The arguments made in this section allow us to assume

that the weights of a neuron in a pretrained FFN is a normalized sample of an SN. Thus, we can directly estimate the weight uncertainty from the pretrained standard FFN. Prediction of an SN can be made in the inverse manner. We first obtain multiple output samples of $g(\mathbf{x})$ from the sampled weights of the SN from the estimated weight uncertainty. Then, taking average of the sampled output provides an unbiased prediction of SN, and the variance of the sampled output provides the output uncertainty.

3.4 The Rationality Measure and the Most Rational Account Principle

In this section, we address the problem of effective representation of the model uncertainty for the manifestation of the structural attribute which is brought by the proposed consumer behavior framework and elaborated to the most rational account principle (MRAP). This problem can be boiled down to the determination of the minimal sized latent key source in Fig. 3.3, which implies the least number of entangled pairs of stochastic encoder and decoder that well explains the variation of the demand behavior. As we provided in Sec. 3.3 that how an SN with mixture distribution defined on its parameter space well represents the proposed consumer behavior framework, we consider Gaussian mixture model (GMM) as the base distribution family that an SN as a demand function lies on.

GMM is a distribution representation mainly used for the following main purposes: (i) to provide a semiparametric density estimation to model unknown distributional shapes [91]; (ii) to provide a probabilistic clustering of the data that a component in the mixture model corresponds to a cluster [91]; (iii) to provide a probabilistic representation of piece-wise linear relationship such as trajectory or motion segmentation in robotics research [92]. Taking $\mathbf{w}^{(x)}$ and $\mathbf{w}^{(y)}$ be Gaussian mixtures, identifying the key source size in the proposed consumer behavior framework can be rephrased to the problem of finding the Gaussian mixtures of $\mathbf{w}^{(x)}$ and $\mathbf{w}^{(y)}$ with minimal components that meets the MRAP most.

From unsupervised learning perspective, the question of how many components to include in a GMM or a clustering task in general has been addressed by a large body of previous literature in machine learning and statistics communities [90, 91]. The approaches for the component number

determination of a GMM may vary by the context of the GMM usage, which may be differentiated by their particular interpretation of "Occam's razor" [93] for their specific application purpose. the mostly used approach in practice has been to use heuristic penalized log likelihood maximization criteria such as Akaike's Information Criterion (AIC) [94] and Bayesian Information Criterion (BIC) [95] to assess the order of a mixture model. It has been also common to use encoding length based criteria such as Minimum Description Length (MDL) principle [96] or Minimum Message Length (MML) principle [97], which are based on the idea that building the shortest code for a given dataset implies to build the best data generation model differing how to address the minimum encoding length principle [90]. It is noticeable that MDL mathematically coincides with BIC in spite of their clear conceptual difference [90]. Although numerous methods were proposed for components number selection in GMM in the past literature, none of them take account of MRAP as their main design objective is to estimate a generative model in the various flavours of unsupervised context regardless of its internal entanglement structure between its sub-vector[5].

To identify best GMM representation of an SN for demand response modeling in terms of MRAP, it is essential to define how to represent entanglement between stochastic encoders and decoders as the first step. Perhaps, a general communication system scheme introduced by Shannon [99] may be an illustrative causal model[6] offering a decent account of such entanglement, as the Shannon's communication system scheme is devised to formally describe the system that an observed outcome of which is a consequence of a transmitted symbol from the source of the system. The schematic similarity between the entanglement between stochastic encoders and decoders in the proposed consumer behavior framework and the communication channel model is depicted in Fig. 3.6 and Fig.3.7. We can check that a Markov kernel, i.e., 2×2 real matrix with

[5]We call a random vector $\mathbf{z}_{\text{sub}}$ a sub-vector of the random vector $\mathbf{z}_{\text{sup}} = [Z_1, ..., Z_n]$ if $\mathbf{z}_{\text{sub}} = [Z_{J_1}, ..., Z_{J_k}]$, where $J = \{J_1, ..., J_k\} \subseteq \{1, ..., n\}$, and $J_i < J_{i'}$ if $i < i'$.

[6]We clarify that the terminology "causality" used here has a slight conceptual difference with the term "causality" in Sec. 3.1 coined by Pearl. The "causality" in Sec. 3.1 is the causation which is represented by a Bayesian network putting causal direction in its center. The causality used in this section refers to the associational strength between two variables to express the entanglement between a stochastic encoder and decoder. This conceptually resembles the Granger causality [98] in the sense that the causality in this section refers to how an observation on one variable is informative to infer another unobserved variable, while this does not need to consider the physical time order of the events of each variable precisely as Granger causality does. In this section, we will interchangeably use "entangled" and "causal".

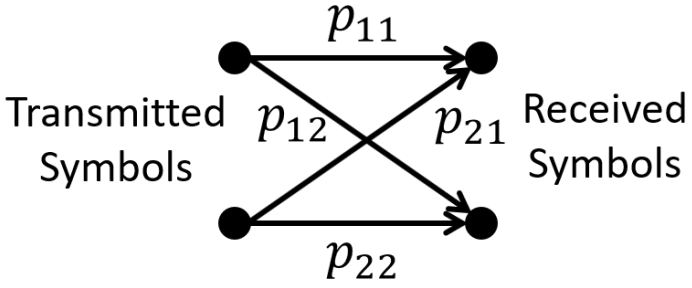


Figure 3.6: A schematic diagram of Shnnon's binary communication channel [99].

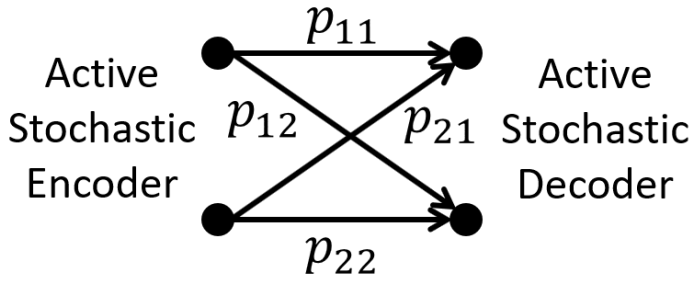


Figure 3.7: A schematic diagram of the entanglement between stochastic encoders and decoders proposed consumer behavior framework.

$\sum_j p_{ij} = 1$ in these cases, well represents both situations. We may deem that a system is perfectly causal if a noiseless channel is given, while an extent of channel noise may keep the system away from the maximal causality. This observation gives us an intuition that a Markov kernel, as $A(K^{(x)}, K^{(y)})$ in Fig. 3.5, is a decent choice to represent the entanglement between all possible stochastic encoder/decoder pairs, and we may set up a similar "entanglement measure" to the measure of the noiselessness for a communication channel, or other causality measures proposed in previous literature beyond information theoretic background.

The task of quantifying causality depends on the idea of how to measure the reduction of uncertainty in a target variable after observing another one [103]. In a large body of previous literature, numerous attempts to quantify causality have been made from both non-information theoretic background and information theoretic background. Perhaps, the most renowned and widely used non-information theoretic causality measure may be the p-*value* of the F-test statistic in the Analysis of Variance (ANOVA) technique developed by Fisher [34], which is a widely

used technique in a variety of hypothesis testing experiments suitable for verifying linear influences. Another renowned non-information theoretic causality measure is *Granger causality* [98], which is a hypothesis test to quantify variance reduction between time series. As the measure of the strength of an "arrow" in a Bayesian network, the *average causal effect* (ACE) proposed by Pearl [37] to quantify causal strength between binary variables. On the other hand, various causality measures are proposed from the informational theoretic viewpoint. The most basic information theoretic measure of the association between two random variables X and Y is the *mutual information* $I(X,Y) \coloneqq \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$. Numerous causality measures have been proposed by extending mutual information to time series analysis in various contexts. For example, *directed information* [100] and *transfer entropy* [101] applies mutual information to time series analysis by extending the argument of Granger causality. Ay and Polani [102] propose a causal dependence measure named *"information flow"* extending the spirit of ACE by Pearl from information theoretical viewpoint. Janzing et al. [103] establish the formal postulates on the desired properties that a proper causality measure should have, and propose a causality measure that well satisfies all of the established postulates, given by relative entropy distance (Kullback-Leibler divergence) between the distribution of target variables with and without intervention.

However, the mutual information based causality measure has an evident limitation as the measure of entanglement since an entanglement is solely determined by the Markov kernel $A(K^{(x)}, K^{(y)})$ or a communication channel, while mutual information is the quantity that not only determined by the communication channel itself but also the probability distribution of the source $P(X)$ as follows:

$$\begin{aligned} I(X,Y) &= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{P(y)} \\ &= \sum_x P(x) D_{KL}(P(y|x)||P(y)), \end{aligned} \tag{3.20}$$

where $D_{KL}(P(Z)||P'(Z)) \coloneqq \sum_z P(Z) \log \frac{P(Z)}{P'(Z)}$ is the Kullback-Leibler divergence. Shannon did not explicitly mentioned the concept of causality in his work [99] but he attempts to characterize a

channel's indigenous property through the concept of *channel capacity* by quantifying the strength of the strongest dependencies between the source and the outcome that the channel may demonstrate. In the sense of causality, the channel capacity can be deemed as a measure of *potential* influence, while mutual information is that of the *factual* causal influence [103].

Alongside the effort to understand and measure "consciousness" in neuroscience community, integrated information theory argues that a system's consciousness is determined by its causal properties of any physical system [104]. Hoel et al. [105, 106] attempt to extend the arguments to a general system analysis framework. As a part of their study, they propose a "channel-indigenous" measure of causality in Markov chain. By taking a Markov kernel as a state-to-state transition probability for all possible states in a complex system, they define a quantity named *effective information* defined as $EI(S) = \frac{1}{n} \sum_{s_C \in S_C} D_{KL}(P(S_E|s_C)||P^{(\max)}(S_E))$ to measure causality, where S_C and S_E are the sets of all possible states that could be cause and effect respectively, $P^{(\max)}(S_E)$ is the distribution on S_E when $P(S_C)$ is given as a uniform distribution over S_C. As the authors use the notation S_C and S_E to denote causal time difference only, $S = S_C = S_E$ for any given Markov chain, where S is the set of all states. n is the number of states in S, i.e., $n := |S_C| = |S|$. Note that effective information can be rephrased as $EI(S) = I(S_C, S_E)$ when S_C is uniformly distributed, while Shannon's channel capacity is defined as $\max_{P(X)} I(X, Y)$. In [105, 106], the authors argue that there is an optimal "resolution" to describe a system in the sense that a macroscale description sometimes can be more informative than a microscale description via assessing effective information of a system of different system description resolution. Specifically, they adjust the size of Markov kernel of a given system by controlling the number of states via grouping system components appropriately; reducing the number of states can be accomplished by merging multiple outputs from different components in a group into one value through reasonable ways, e.g., averaging, and take the value as an output of the group, i.e., a component of the system in the low-resolution system description. They attempt to find the optimal size of Markov kernel and optimal component grouping to maximize the effective information of a system description. However, these works do not show an efficient algorithm for optimal system component grouping than a brute-force search,

which limits the scalability of its application in practical use.

Perhaps, the simplest candidate measure of entanglement may be the absolute value on the matrix determinant operation. This is based on the fact that $|\det A| \leq 1$ for any Markov kernel A, with equality if and only if A is a permutation matrix [107], as a permutation matrix is the matrix representation of a perfectly noiseless channel. However, the absolute value of matrix determinant has some drawbacks to be a good measure of entanglement. (i) This may be too strict to be a proper measure for entanglement in the following sense; $|\det A| = 0$ does not imply that the set of possible stochastic encoders and decoders have zero entanglement. For example, if a Markov kernel A is a block diagonal formed matrix such that a block of A has zero determinant, then $|\det A| = 0$, though there are still some degrees of entanglement between stochastic encoders and decoders. That is, $|\det A| = 0$ does not imply zero-entanglement. (ii) Although we find a GMM representation of SN that its Markov kernel A is a permutation matrix, which may imply a perfect entanglement between stochastic encoders and decoders, we cannot conclude that the GMM representation of SN well satisfies MRAP. This is because, it does not imply the paired encoder and decoder are mutually independent.

However, if we find a block diagonalized form of Markov kernel A, it may be informative to find an optimal component number of GMM of an SN as a DR model. For example, consider the Markov kernel A of an SN is given as follows when we take the component number $l = 4$:

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Then, we can easily conclude that the optimal component number $l = 2$ for following reasons:

1. The entanglement between the 1st and 2nd stochastic encoder/decoder pair, as well as the 3rd and 4th stochastic encoder/decoder pair are near zero. That is, e.g., having a clear knowledge of which one is activated among 3rd and 4th stochastic encoder is not informative

to determine which stochastic decoder will be activated among the 3rd one and the 4th one. In other words, the selection of activating 3rd and 4th components of stochastic encoder is independent with the selection of which of 3rd and 4th stochastic decoder to be activated.

2. The newly generated Markov kernel A' equals to the identity matrix I, if we group the 1st and 2nd stochastic encoder/decoder into a stochastic encoder/decoder, and so as 3rd and 4th stochastic encoder/decoder. A' exhibits perfect entanglement while the paired encoder and decoder remain almost mutually independent.

This example illustrate the idea that which mixture components are "lumpable" for better representation in terms of MRAP, when we have a "too high-resolution" on the distribution of the model. The key observation from this example is that if activating two different stochastic encoders results in the activation of the similar mixture of stochastic decoders, these encoders and decoders are lumpable. We extend this observation to develop a heuristic procedure to find the optimal number l of components of the mixture distribution for an SN as a demand function. Let our objective is to find an optimal l in terms of MRAP for a demand function $g(\mathbf{x}) = \phi(\mathbf{w}^{(x)} \cdot \mathbf{x})\mathbf{w}^{(y)}$ with mixture distribution $P(\mathbf{w}^{(\cdot)}) = \sum_{k=1}^{l} \alpha_k^{(\cdot)} P_k(\mathbf{w}^{(\cdot)})$, and a Markov kernel $A \in \mathbb{R}^{l \times l}$ such that $\vec{\alpha}^{(y)} = A\vec{\alpha}^{(x)}$, where $\vec{\alpha}^{(\cdot)} = [\alpha_k^{(\cdot)}]_{k=1}^{l}$. We assume that $(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^d, \mathbb{R}^d)$.

- Input parameters: $\epsilon > 0$, and the training dataset $\mathcal{D}$.
- Outputs: $P^{(\cdot)\prime} = \sum_{k=1}^{l^*} \alpha'_k P_k^{(\cdot)\prime}$ for x and y: Alternative mixture representation of $P(\mathbf{w}^{(x)})$ and $P(\mathbf{w}^{(y)})$.

1. Train a standard FFN $g_0(\mathbf{x}) = \phi(\vec{W}^{(x)} \cdot \mathbf{x})\vec{W}^{(y)}$ from the given training dataset $\mathcal{D}$, where $\vec{W}^{(\cdot)} \in \mathbb{R}^{m \times d}$.

2. Take an arbitrary sufficiently large $L \gg 3$, and estimate the mixture distributions $P(\mathbf{w}^{(x)})$ and $P(\mathbf{w}^{(y)})$ with the $l = L, L-1$ from $\vec{W}^{(x)}$ and $\vec{W}^{(y)}$ respectively. Obtain the Markov kernel A_L and A_{L-1}. Check if $|\det(A_L)| < \min(\epsilon, |\det(A_{L-1})|)$, and go to the next step if the condition is satified. Otherwise, take a larger L and repeat this step.

3. Obtain $\vec{\alpha}^{(y|k_x)} = A\vec{e}_{k_x}$ for $k_x = 1, ..., l$, where $\vec{e}_{k_x}$ is the unit vector that its k_xth component is 1 and all other components are 0. Calculate $P(\mathbf{w}^{(y)}|k_x) := \sum_{k=1}^{l} \alpha_k^{(y|k_x)} P_k(\mathbf{w}^{(y)})$ for all k_x.

4. Run k-mean clustering to on $\{P(\mathbf{w}^{(y)}|k_x) : k_x = 1, ...l\}$ with the distance metric $d(P', P'') := D_{KL}(P'||P'')+D_{KL}(P'||P'')$ where P' and P'' are probability distributions. Repeat this step with various k's.

5. Choose the optimal cluster number k^* as follows:

$$\begin{aligned} k^* = \arg\max & E_{\text{inter}}[d(P(\mathbf{w}^{(y)}|k_x = i), P(\mathbf{w}^{(y)}|k_x = j))] \\ & - E_{\text{intra}}[d(P(\mathbf{w}^{(y)}|k_x = i), P(\mathbf{w}^{(y)}|k_x = j))] \\ & \text{for all } i \neq j \text{ such that } i, j \leq l, \end{aligned} \tag{3.21}$$

   where $E_{\text{inter}}[d(P', P'')]$ is the average distance between the distributions P' and P'' in the same cluster and $E_{\text{inter}}[d(P', P'')]$ is the average distance between the distributions P' and P'' in different clusters.

6. From the step 5, we have k^* clusters on $\{P(\mathbf{w}^{(y)}|k_x) : k_x = 1, ...l\}$ in our hand, whose element can be identified by k_x. Take $l^* = k^*$. Approximate the mixture of the distributions in each cluster C_k to a Gaussian $P_k^{(x)\prime} = \arg\min D(\frac{1}{\alpha'_k} \sum_{k_x \in C_k} \alpha_{k_x}^{(x)} P_{k_x}^{(x)} || P_k^{(x)\prime})$, where $\alpha'_k = \sum_{k_x \in C_k} \alpha_{k_x}^{(x)}$, and a Gaussian $P_k^{(y)\prime} = \arg\min D(\frac{1}{|C_k|} \sum_{k_x \in C_k} P(\mathbf{w}^{(y)}|k_x) || P_k^{(y)\prime})$.

In the input parameters, ϵ is typically set as a small value (< 0.01) to be used as a threshold to find a sufficiently large initial l in step 2. In the output, $\alpha'_k = \alpha_k^{(x)\prime} = \alpha_k^{(y)\prime}$, since A^* is the identity matrix. The step 3 states the method to calculate the distribution of $P(\mathbf{w}^{(y)})$ if a stochastic encoder, i.e., a Gaussian component of GMM of $P(\mathbf{w}^{(x)})$, is activated (i.e., k_x). The step 4 is an implementation of the idea that some components from $\{P_{k_x}(\mathbf{w}^{(x)})\}$ are allowed to be lumped if the corresponding $P(\mathbf{w}^{(y)}|k_x)$ is similar. To define the similarity between two distribution, we define the distance metric $d(P', P'')$ of two probability distributions P' and P'' in step 4, as the Kullback-Leibler divergence D_{KL} is not symmetric. We attempt to find the components k_x's with

similar $P(\mathbf{w}^{(y)}|k_x)$ via k-mean algorithm. In the step 5, we show the proposed measure to quantify how a model representation meets the MRAP. The larger k^* indicates that the model meets the MRAP better.

Meanwhile, it is known that there is no closed form of D_{KL} for mixture distributions [108]. For this reason, we derive an approximated formula to calculate D_{KL} for mixture distribution for the proposed l^* selection procedure to be applicable. First, we derive an upper bound of the entropy of a mixture distribution. Here, we denote a mixture distribution $P = \sum_k \alpha_k P_k$. Then,

$$\begin{aligned} H(X) &= -\int (\sum_k \alpha_k P_k) \log \sum_{k'} \alpha_{k'} P_{k'} dx \\ &= -\sum_k \alpha_k \int P_k \log \sum_{k'} \alpha_{k'} P_{k'} dx \\ &= \sum_k \alpha_k \left(H_k + D_{KL}(P_k \| \sum_{k'} \alpha_{k'} P_{k'}) \right) \\ &= \sum_k \alpha_k \left(H_k + D_{KL}(P_k \| P) \right) \\ &\leq \sum_k \alpha_k \left(H_k + \sum_{k'} \alpha_{k'} D_{KL}(P_k \| P_{k'}) \right) \end{aligned} \tag{3.22}$$

by Jensen's inequality and the convexity of D_{KL}. Similarly, we can derive the approximated form of D_{KL}, where the approximation error is bounded in either sides, i.e., upper and lower bounded. Take another mixture distribution $Q = \sum_k \beta_k Q_k$. Then,

$$\begin{aligned} D_{KL}(P\|Q) &= \sum_k \int \alpha_k P_k \log \frac{\sum_{k'} \alpha_{k'} P_{k'}}{\sum_{k'} \beta_{k'} Q_{k'}} dx \\ &= \sum_k \alpha_k \left(-H_k - D(P_k\|P) + H_k + D(P_k\|Q) \right) \\ &= \sum_k \alpha_k \left(D(P_k\|Q) - D(P_k\|P) \right) \\ &\approx \sum_k \alpha_k \left(\sum_{k'} \beta_{k'} D(P_k\|Q_{k'}) - \sum_{k'} \alpha_{k'} D(P_k\|P_{k'}) \right). \end{aligned} \tag{3.23}$$

If we take P and Q as Gaussian mixtures, each mixture component P_k and Q_k are Gaussian distributions. The Kullback-Leibler divergence D_{KL} of two Gaussian distributions given as $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

and $\mathcal{N}(\boldsymbol{\mu}', \Sigma')$ are well known as follows:

$$D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma)||\mathcal{N}(\boldsymbol{\mu}', \Sigma')) = \frac{1}{2}\log\frac{\det\Sigma}{\det\Sigma'} + \frac{1}{2}\mathrm{Tr}(\Sigma'^{-1}\Sigma) + \frac{1}{2}(\boldsymbol{\mu}' - \boldsymbol{\mu})^{\top}\Sigma'^{-1}(\boldsymbol{\mu}' - \boldsymbol{\mu}) - \frac{d}{2} \quad (3.24)$$

where $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathbb{R}^d$.

We also note that this procedure can be applied to the state identification problem in imitation learning problems in Partially Observable Markov Decision Process (POMDP) setting. Specifically, consider the situation that a machine apprentice attempts to learn an optimal policy for an unknown environment which may be modeled in Markov Decision Process (MDP) through observing an expert's demonstration. However, suppose that their is a discrepancy of observability of state between the expert and the apprentice, i.e., apprentice has a limited observability on states, while an expert has full observability. The proposed procedure can help apprentice to identify minimal number of states in the MDP, which allow the apprentice to learn minimal policy which well explains the expert's demonstration. We leave the detailed implementation of this idea into future work.

3.5 Further Extensions

In the previous sections in this chapter, we have discussed on the identifiability of demand response as a fundamental problem of consumer behavior modeling. The other fundamental problems in price responsive electricity consumption modeling we pose are as follows:

1. *Data sparsity*: We expect the large-scale deployment advanced metering infrastructure (AMI) would avail real-time frequent load sampling, which is clearly better informative for system operators to detect mismatches between power supply and load. However, the proposed "Appliance" usage model in Sec. 2.2.2.1 suggests that the advance of AMI will not necessarily avail better learnability of consumer behaviors. This is because, increasing sampling frequency may only increase the dimensions of input and output of a consumer behavior model, while the total number of samples for a given period will remain fixed as we have defined the input and output as the sequences of prices and loads for a short term period which

may be influenced by each other in Sec. 3.2, e.g., the sample size would be always 365 if we take the period as one day, regardless of sampling frequency. Such increment of data dimension exacerbate data sparsity, as referred to the curse of dimensionality, exacerbates the learnability of model as it weakens statistical significance of a dataset. This eventually results the trained model far from robustness.

2. *Absence of dynamic account*: While the fundamental unidentifiability of demand response shown in Theorem 3.1.2 entails the establishment of a demand response model from pre-existing economic theories, the models premised on the static settings in economic theories lack dynamic systems account. Especially in real-time pricing setting, this may be unrealistic because the price and corresponding load is given in a sequential order. This necessitates to develop a dynamic model from a given demand response model represented as a stochastic neuron.

In this section, we introduce the methods to solve these problems.

3.5.1 Variational Meta-Learning for Multiple Sparse Datasets

Even though a dataset obtained from an individual may be sparse, we note that this problem may be mitigated if the dataset in hand possesses a large number of customers. To deal with the data sparsity problem, we consider meta-learning approach, because meta-learning, or learning-to-learn, has been known as a successful strategy in attacking problems that involve small amounts of data [109]. The key motivation of meta-learning is the observation that a human baby has an outstanding ability to learn and generalize new various concepts from few observed samples. This observation raises a highly noticed problem in artificial intelligence communities so called "few(one)-shot learning problem" [110], addressing how to utilize and transfer the knowledge from past learning tasks to learn a new task [111].

Inspired by the *principal component analysis* (PCA) method, we consider the construction of a code for the weight of a stochastic neuron $\mathbf{w}$, which aim to capture the model variance caused by different customer. Adopting a variational approach developed for training a variational au-

toencoder [112], we derive the evidence lower bound (ELBO) as follows by taking a dataset from a customer i as $\mathcal{D}_i$. Here, $\boldsymbol{\lambda}^{(l)} \in \mathbb{R}^{d_l}$ denotes the code sensitive to the customer change, while $\boldsymbol{\lambda}^{(g)} \in \mathbb{R}^{d_g}$ denotes the code insensitive to customer change, i.e., $\boldsymbol{\lambda}^{(g)} \perp\!\!\!\perp \mathcal{D}_i$, where $d_l + d_g = d$, the dimension of $\mathbf{x}$ and $\mathbf{y}$.

$$
\begin{aligned}
&\log \prod_i P(\mathbf{w}|\mathcal{D}_i) = \log \int \prod_i P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \mathcal{D}_i) P(\boldsymbol{\lambda}^{(g)}) d\boldsymbol{\lambda}^{(g)} \\
&= \log \int \prod_i P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \mathcal{D}_i) P(\boldsymbol{\lambda}^{(g)}) \frac{\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w})}{\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w})} d\boldsymbol{\lambda}^{(g)} \\
&= \log E_{\tilde{P}(g)} \left[\frac{\prod_i P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \mathcal{D}_i) P(\boldsymbol{\lambda}^{(g)})}{\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w})} \right] \\
&\geq E_{\tilde{P}(g)} \left[\log \frac{\prod_i P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \mathcal{D}_i) P(\boldsymbol{\lambda}^{(g)})}{\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w})} \right] \\
&\qquad (\because \text{Jensen's inequality }) \\
&= E_{\tilde{P}(g)} \left[\log \prod_i P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \mathcal{D}_i) \right] - D(\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w}) || P(\boldsymbol{\lambda}^{(g)})) \\
&=: \text{ELBO}^{(g)}.
\end{aligned}
\tag{3.25}
$$

We can further hierarchically derive $\text{ELBO}^{(l,g)}$ from $\text{ELBO}^{(g)}$ as follows:

$$
\begin{aligned}
&\text{ELBO}^{(g)} = \\
&= E_{\tilde{P}(g)} \left[\log \prod_i \int P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \boldsymbol{\lambda}^{(l)}, \mathcal{D}_i) P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_i) d\boldsymbol{\lambda}^{(l)} \right] \\
&\qquad - D(\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w}) || P(\boldsymbol{\lambda}^{(g)})) \\
&\qquad (\because \text{we posit } \boldsymbol{\lambda}^{(g)} \perp\!\!\!\perp \boldsymbol{\lambda}^{(l)}) \\
&\geq E_{\tilde{P}(g)} \Bigg[\sum_i E_{\tilde{P}(l)} \left[\log P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \boldsymbol{\lambda}^{(l)}, \mathcal{D}_i) \right] \\
&\qquad - D(\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathbf{w}) || P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_i)) \Bigg] \\
&\qquad\quad - D(\tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w}) || P(\boldsymbol{\lambda}^{(g)})) \\
&=: \text{ELBO}^{(l,g)},
\end{aligned}
\tag{3.26}
$$

where $\tilde{P}^{(g)} = \tilde{P}(\boldsymbol{\lambda}^{(g)}|\mathbf{w})$ and $\tilde{P}^{(l)} = \tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathbf{w})$. Obtaining $(\tilde{P}^{(l)}, \tilde{P}^{(g)}, P(\mathbf{w}|\boldsymbol{\lambda}^{(g)}, \boldsymbol{\lambda}^{(l)}, \mathcal{D}_i)) =$

$\arg\max \text{ELBO}^{(l,g)}$ allows us to train only $P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})$ defined on $\mathbb{R}^{d_l}$ instead of $P(\mathbf{w}|\mathcal{D}_{\text{new}})$ when we encounter a dataset $\mathcal{D}_{\text{new}}$ of a new customer, where we can set arbitrary $0 < d_l < d$.

$P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})$ can be obtained by solving

$$
\begin{aligned}
&\max_{\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})} E_{P(\boldsymbol{\lambda}^{(g)})}\Big[E_{P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})}\left[\log P(\mathcal{D}_{\text{new}}|\boldsymbol{\lambda}^{(g)},\boldsymbol{\lambda}^{(l)})\right] \\
&\qquad - D(\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})||P(\boldsymbol{\lambda}^{(l)}))\Big] \\
&= \max_{\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})} E_{P(\boldsymbol{\lambda}^{(g)})}\Big[\\
&\qquad E_{P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})}\Big[\log P(\mathcal{D}_{\text{new}}|\mathbf{w},\boldsymbol{\lambda}^{(g)},\boldsymbol{\lambda}^{(l)})P(\mathbf{w}|\boldsymbol{\lambda}^{(g)},\boldsymbol{\lambda}^{(l)})\Big] \\
&\qquad - D(\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})||P(\boldsymbol{\lambda}^{(l)}))\Big] \\
&= \max_{\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})} E_{P(\boldsymbol{\lambda}^{(g)})}\Big[\\
&\qquad\qquad E_{P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})}\left[\log P(\mathcal{D}_{\text{new}}|\mathbf{w})P(\mathbf{w}|\boldsymbol{\lambda}^{(g)},\boldsymbol{\lambda}^{(l)})\right] \\
&\qquad\qquad - D(\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})||P(\boldsymbol{\lambda}^{(l)}))\Big] \\
&\qquad (\because \mathcal{D}_{\text{new}}|\mathbf{w} \perp\!\!\!\perp (\boldsymbol{\lambda}^{(l)},\boldsymbol{\lambda}^{(g)})) \\
&= \max_{\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})} E_{P(\boldsymbol{\lambda}^{(g)})}\Big[\\
&\qquad E_{P(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})}\left[\log P(\mathcal{D}_{\text{new}}|\mathbf{w}) + \log P(\mathbf{w}|\boldsymbol{\lambda}^{(g)},\boldsymbol{\lambda}^{(l)})\right] \\
&\qquad\qquad - D(\tilde{P}(\boldsymbol{\lambda}^{(l)}|\mathcal{D}_{\text{new}})||P(\boldsymbol{\lambda}^{(l)}))\Big].
\end{aligned}
\tag{3.27}
$$

3.5.2 Conversion to a Dynamic Model from a Demand Response Model of Stochastic Neuron Representation

Suppose that we have a trained DR model represented via a stochastic neuron $\mathbf{y} = \mathbf{w}^{(y)}\phi(\mathbf{w}^{(x)} \cdot \mathbf{x})$. Our goal is to convert this model to a discrete time dynamic system with a finite horizon $T = d_x = d_y$. Given a series of the current and past prices $x_{1:t}$ and past loads $y_{1:t-1}$, our key interest is to predict the current load y_t, i.e., to compute $P(y_t|y_{1:t-1},\hat{\mathbf{x}}_t)$, where $\hat{\mathbf{x}}_t$ is the expected full

price vector $\mathbf{x}_t$ estimated by a customer at time t. The procedure of computation $P(y_t|y_{1:t-1}, \hat{\mathbf{x}}_t)$ is shown as follows:

1. Estimate the full price vector $\mathbf{x}$ a consumer may forecast at every time step t. The full price prediction $\hat{\mathbf{x}}_t = (x_{1:t}, x^*_{t+1:T})$ at time t can be obtained by the price forecast $x^*_{t+1:T} = \arg\max P(x_{t+1:T}|x_{1:t})$ given a prior distribution $P(\mathbf{x})$.

2. Update α_k by $\alpha_k = \frac{P_k(y_{1:t-1}|\hat{\mathbf{x}}_{t-1})}{\sum_k P_k(y_{1:t-1}|\hat{\mathbf{x}}_{t-1})}$, where

$$\begin{aligned} & P_k(y_{1:t-1}|\hat{\mathbf{x}}_{t-1}) \\ &= \int P_k(w^{(y)}_{1:t-1}\phi(\mathbf{w}^{(x)} \cdot \hat{\mathbf{x}}_{t-1})|\mathbf{w}^{(x)}, \hat{\mathbf{x}}_{t-1}) \\ &\quad P_k(\mathbf{w}^{(x)}|\hat{\mathbf{x}}_{t-1})d\mathbf{w}^{(x)} \\ &= \int P_k(w^{(y)}_{1:t-1}\phi(\mathbf{w}^{(x)} \cdot \hat{\mathbf{x}}_{t-1})|\mathbf{w}^{(x)}, \hat{\mathbf{x}}_{t-1}) \\ &\quad P_k(\mathbf{w}^{(x)})d\mathbf{w}^{(x)} \end{aligned} \tag{3.28}$$

since $\mathbf{w}^{(x)} \perp\!\!\!\perp \mathbf{x}_{t-1}$. Here, taking $z_{t-1} = \mathbf{w}^{(x)} \cdot \hat{\mathbf{x}}_{t-1} \in \mathbb{R}$ allows us to rewrite Eq. (3.28) as follows:

$$\begin{aligned} & P_k(y_{1:t-1}|\hat{\mathbf{x}}_{t-1}) \\ &= \int P_k(w^{(y)}_{1:t-1} = \frac{y_{1:t-1}}{\phi(\mathbf{w}^{(x)} \cdot \hat{\mathbf{x}}_{t-1})}|\mathbf{w}^{(x)}, \hat{\mathbf{x}}_{t-1}) \\ &\quad P_k(\mathbf{w}^{(x)})d\mathbf{w}^{(x)} \\ &= \int P_k(w^{(y)}_{1:t-1} = \frac{y_{1:t-1}}{\phi(z_{t-1})}|z_{t-1})P_k(z_{t-1})dz_{t-1} \end{aligned} \tag{3.29}$$

where

$$\begin{aligned} z_{t-1} &\sim \mathcal{N}(\vec{\mu}^{(x)}_k \cdot \hat{\mathbf{x}}_{t-1},\ \hat{\mathbf{x}}^\top_{t-1}\Sigma^{(x)}_k\hat{\mathbf{x}}_{t-1}), \text{ and} \\ w^{(y)}_{1:t-1}|z_{t-1} &\sim \mathcal{N}(\vec{\mu}^{(y)}_{k,1:t-1},\ \Sigma^{(y)}_{k,1:t-1}) \end{aligned} \tag{3.30}$$

since $w^{(y)}_{1:t-1} \perp\!\!\!\perp z_{t-1}$.

3. Obtain the $P_k(y_t|y_{1:t-1}, \hat{\mathbf{x}}_t)$ as follows:

$$\begin{aligned} & P(y_t|y_{1:t-1}, \hat{\mathbf{x}}_t) \\ &= \int p(y_t|y_{1:t-1}, \mathbf{w}^{(x)}, \hat{\mathbf{x}}_t) p(\mathbf{w}^{(x)}|y_{1:t-1}, \hat{\mathbf{x}}_t) d\mathbf{w}^{(x)} \\ &= \int p(y_t|y_{1:t-1}, \mathbf{w}^{(x)}, \hat{\mathbf{x}}_t) p(\mathbf{w}^{(x)}) d\mathbf{w}^{(x)}, \end{aligned} \tag{3.31}$$

since $\mathbf{w}^{(x)}$ is independent with both $y_{1:t-1}$ and $\mathbf{x}_t$. Similar to the previous step, we can rewrite Eq. (3.31) as follows by taking $z_t = \mathbf{w}^{(x)} \cdot \hat{\mathbf{x}}_t \in \mathbb{R}$:

$$\begin{aligned} & P(y_t|y_{1:t-1}, \hat{\mathbf{x}}_t) \\ &= \int p(y_t|y_{1:t-1}, z) p(z_t) dz_t \\ &= \int p(w_t^{(y)} \phi(z_t)|w_{1:t-1}^{(y)} = \frac{y_{1:t-1}}{\phi(z_t)}) p(z_t) dz_t, \end{aligned} \tag{3.32}$$

where

$$z_t \sim \mathcal{N}(\vec{\mu}_k^{(x)} \cdot \hat{\mathbf{x}}_t,\ \hat{\mathbf{x}}_t^\top \Sigma_k^{(x)} \hat{\mathbf{x}}_t), \text{ and} \tag{3.33}$$

$$\begin{aligned} w_t^{(y)}|w_{1:t-1}^{(y)} \sim \mathcal{N}\Big(& \mu_{k,t}^{(y)} + \Sigma_{k,t}^{(y)\top} \Sigma_{k,1:t-1}^{-1(y)} (w_{1:t-1}^{(y)} - \mu_{k,1:t-1}^{(y)}), \\ & \sigma_{k,t}^{2(y)} - \Sigma_{k,t}^{(y)\top} \Sigma_{k,1:t-1}^{-1(y)} \Sigma_{k,t}^{(y)}\Big). \end{aligned} \tag{3.34}$$

Eq. (3.34) is derived from $P_k(\mathbf{w}^{(y)}) = \mathcal{N}(\vec{\mu}_k^{(y)}, \Sigma_k^{(y)})$ as we decompose it as follows:

$$\begin{bmatrix} w_{1:t-1}^{(y)} \\ w_t^{(y)} \end{bmatrix} \sim \mathcal{N}\Big(\mu_{k,1:t}^{(y)}, \begin{bmatrix} \Sigma_{k,1:t-1}^{(y)} & \Sigma_{k,t}^{(y)} \\ \Sigma_{k,t}^{(y)\top} & \sigma_{k,t}^{2(y)} \end{bmatrix}\Big), \tag{3.35}$$

where $\Sigma_{k,1:t-1}^{(y)} \in \mathbb{R}^{(t-1)\times(t-1)}$, $\Sigma_{k,t}^{(y)} \in \mathbb{R}^{(t-1)\times 1}$, and $\sigma_{k,t}^{2(y)} \in \mathbb{R}$ are the subblocks of $\Sigma_k^{(y)}$.

4. Obtain $P(y_t|y_{1:t-1}, \hat{\mathbf{x}}) = \sum_k \alpha_k P_k(y_t|y_{1:t-1}, \hat{\mathbf{x}}_t)$.

Such framework and the proposed dynamic load prediction procedure can be readily extended to continuous time dynamic system model via Gaussian Process $\mathcal{GP}(\cdot)$. If take $y(t) = \sum_k \alpha_k y_k(t)$ such that $y_k(t) = \phi(\int_{t=0}^{T} w_k^{(x)}(t)\hat{x}(t)dt) w_k^{(y)}(t)$ where $w_k^{(\cdot)}(t) \sim \mathcal{GP}(\mu_k^{(\cdot)}(t), \mathbf{K}_k^{(\cdot)})$, the proposed

procedure can be directly applied for the calculation of $P(y(t) \mid \{y(\tau) : \tau \text{ is a finite subset of } [0,t)\}, \hat{x}(t))$.

4. SUMMARY AND CONCLUDING REMARKS

Today's electricity power system and market is already a tremendously complex system. Including economic and technical characteristics of power systems as well as regulatory and political issues, there arise a number of issues to be considered for a successful market and smart grid design. Such complexity makes it extremely difficult to determine a priori whether a design will be a successful. Although there are many flourishing examples of power pools and exchanges in the energy market today, the inherent complexity of power systems and markets makes it extremely hard to adapt to novel technologies or innovative concepts. Demand Response is no exception. Though Demand Response is expected to be a key mechanism in the smart grid, it is unclear how to apply in the current energy market without a thorough understanding of demand sensitivity to price and sufficient understanding of its complexity. This study is intended to be an initial guide toward better understanding of markets and smart grid design.

REFERENCES

[1] J. An, P. R. Kumar, and L. Xie, “On transfer function modeling of price responsive demand: an empirical study," Proc. of *IEEE Power & Energy Society General Meeting (PESGM 2015)*, pp. 1-5, Denver, CO, Jul. 2015.

[2] M. O. Jackson, “The role of theory in an age of design and big data,” *The Future of Economic Design, edited by Jean-Franois Laslier, Her’ve Moulin, Remzi Sanver, and William S. Zwicker*, Forthcoming, Apr. 2018. SSRN: `https://ssrn.com/abstract=3031294`

[3] L. Xie, S. Puller, M. Ilic, and S. Oren, “Quantifying benefits of demand response and look-ahead dispatch in systems with variable resources," *PSERC Final Report M26*, Aug. 2013.

[4] H. Daneshi, and A. Srivastava, “ERCOT electricity market: transition from zonal to nodal market operation,” Proc. of *IEEE Power & Energy Society General Meeting (PESGM 2011)*, pp. 1-7, 2011.

[5] *Electricity Reliability Council of Texas (ERCOT)*, url: `http://www.ercot.com`.

[6] *ERCOT Board Members Overview and Orientation,* Electric Reliability Council of Texas (ERCOT), Mar. 2010. url: `http://www.ercot.com/content/news/presentations/2010/ERCOT_Board_Orientation,_2010.pdf`

[7] ERCOT, "The market guide: an introductory guide to how the electric reliability council of Texas (ERCOT) facilitates the competitive power market," *Electricity Reliability Council of Texas (ERCOT)*, Jan. 2005.

[8] S. Zhou, T. Grasso, and G. Niu, “Comparison of market designs," *Market Oversight Division Report from Project 26376, Rulemaking Proceeding on Wholesale Market Design Issues in the Electric Reliability Council of Texas,* Market Oversight Division, Public Utility Commission of Texas, Jan. 2003.

[9] C. Aubin, D. Fougere, E. Husson, and M. Ivaldi, “Real-time pricing of electricity for residential customers: econometric analysis of an experiment,” *J. Appl. Econometrics*, vol. 10, pp. 171-191. Dec. 1995.

[10] P. Joskow, and J. Tirole, “Reliability and competitive electricity markets,” *RAND J. Econ.*, vol. 38, pp. 60-84. 2007.

[11] Frank A. Wolak, “Managing demand-side economic and political constraints on electricity industry restructuring processes,” *Economic Reform in India,* Cambridge University Press, Cambridge, UK, ch. 13, pp. 455-498, 2013.

[12] S. Holland and E. Mansur, “The short-run effects of time-varying prices in competitive electricity markets,” *Energy J.*, vol. 27, pp. 127-155, Oct. 2006.

[13] S. Borenstein and J. Bushnell, “An empirical analysis of the potential for market power in California’s electricity market,” *J. Ind. Econ.*, vol. 47, pp. 285-323, Sep. 1999.

[14] S. Borenstein and J. Bushnell, “Electricity restructuring: deregulation or reregulation?," *Regulation*, vol. 23, no. 2, pp. 46-52, 2000.

[15] S. Borenstein, “The trouble with electricity markets: understanding Californias restructuring disaster,” *J. Econ. Perspect.*, vol. 16, pp. 191-211, 2002.

[16] S. Borenstein, and S. Holland, “On the efficiency of competitive electricity markets With time-Invariant retail prices,” *NBER Working Paper*, no. 9922, Aug. 2003.

[17] S. Borenstein, “The long-run efficiency of real-time electricity pricing," *Energy J.*, vol. 26, pp. 93-116, Apr. 2005.

[18] S. Borenstein, “Time-varying retail electricity prices: theory and practice," *Electricity Deregulation: Choices and Challenges,* The University of Chicago Press, Chicago, IL, pp. 317-357, 2005.

[19] S. Borenstein, ”Wealth transfers among large customers from implementing real-time retail electricity pricing," *Energy J.*, vol. 28, no. 2, pp. 131-149, 2007.

[20] H. Allcott, “Rethinking real-time electricity pricing," *Resour. Energy Econ.*, vol. 33, pp. 820-842, 2011.

[21] H. Allcott, “The smart grid, imperfect competition, and entry in electricity supply,” *NBER Working Paper*, no. 18071, May 2012.

[22] H. Allcott, “Real-time pricing and electricity market design,” *Working Paper*, 2013.

[23] W. Hogan, “Time-of-use rates and real-time prices,” *Working Paper*, Aug. 2014.

[24] K. Spees and L. B. Lave, “Demand response and electricity market efficiency," *Electr. J.*, vol. 20, no. 3, pp. 69-85, 2007.

[25] L. Xie, S. Puller, M. Ilic, and S. Oren, “Quantifying benefits of demand response and look-ahead dispatch in systems with variable resources," *PSERC Final Report M26,* Aug. 2013.

[26] A. Nayyar, M. Negrete-Pincetic, K. Poolla, and P. Varaiya, “Duration-differentiated energy services with a continuum of loads,” Proc. of *IEEE Conference on Decision and Control (CDC 2014)*, Dec. 2014.

[27] G. Sharma, L. Xie, and P. R. Kumar, “Optimal demand response for thermal inertial loads employing stochastic renewables: a privacy respecting architecture and its continuum scaling limit," Proc. of *IEEE Conference on Decision and Control (CDC 2014)*, Dec. 2014.

[28] M. Ilic, and J. Zaborszky, *Dynamics and control of large electric power systems,* Wiley, NY, 2000.

[29] I. A. Hiskens, “Dynamics of type-3 wind turbine generator models," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 465-474, Feb. 2012.

[30] F. Ding and T. Chen, “Identification of Hammerstein nonlinear ARMAX systems," *Automatica,* vol. 41, no. 9, pp. 1479-1489, Sep. 2005.

[31] M.H. Albadi, E.F. El-Saadany, ”A summary of demand response in electricity markets," *Electric Power Syst. Res.*, vol. 78, no. 11, pp. 1989-1996, Nov. 2008.

[32] F. Hayashi, *Econometrics,* Princeton University Press, 41 William Street, Princeton, NJ, 2000.

[33] P. R. Kumar and P. Varaiya, *Stochastic systems: estimation, identification and control,* Prentice Hall, Englewood Cliffs, NJ, 1986.

[34] R. Fisher, “Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk," *J. Agric. Sci.*, no. 11, pp. 107-135, 1918.

[35] B. Corgnet, P. Kujal, and D. Porter, “Reaction to public information in markets: how much does ambiguity matter?," *Economic J.,* Royal Economic Society, vol. 123, no. 569, pp. 699-737, Jun. 2013.

[36] J. Pearl, “Causal inference in statistics: an overview,” *Stat. Surv.*, vol. 3, pp. 96-146, 2009.

[37] J. Pearl, *Causality: models, reasoning, and inference,* 2nd ed., Cambridge University Press, New York, 2009.

[38] J. Pearl, “Causal diagrams for empirical research,” *Biometrika*, vol. 82, no. 4, pp. 669-688, Dec. 1995.

[39] D. B. Rubin, “Estimating causal effects of treatments in randomized and non-randomized studies,” *J. Educational Psychol.*, vol. 66, no. 5, pp. 688-701, 1974.

[40] A. Goldberger, “Structural equation models in the social sciences,” *Econometrica*, vol. 40, no. 6, pp. 979-1001, Nov. 1972.

[41] J. Tian and J. Pearl, “A general identification condition for causal effects,” Proc. of *the 18th AAAI Conference on Artificial Intelligence,* pp. 567-573, AAAI Press, Aug. 2002.

[42] J. Tian and J. Pearl, “On the identification of causal effects,” *Tech. Rep. R-290-L*, Department of Computer Science, University of California, Los Angeles, CA, 2002.

[43] I. Shpitser and J. Pearl. “Identification of joint interventional distributions in recursive semi-Markovian causal models,” Proc. of *the 20th AAAI Conference on Artifical Intelligence,* pp. 1219-1226, AAAI Press, Jul. 2006.

[44] Y. Huang and M. Valtorta, "On the completeness of an identifiability algorithm for semi-Markovian models," *Ann. Math. Artif. Intell.*, vol. 54, no. 4, pp. 363-408, Dec. 2008.

[45] J. Pearl and T. S. Verma, "A theory of inferred causation," Proc. of *the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR '91)*, pp. 441-452, San Mateo, CA, Apr. 1991.

[46] H. R. Varian, *Microeconomic Analysis*, 3rd ed., W. W. Norton & Company Inc., Mar. 1992.

[47] E. R. Weintraub, "Neoclassical economics," *The Concise Encyclopedia of Economics*, 1st ed., 1993. url: `http://www.econlib.org/library/Enc1/NeoclassicalEconomics.html`

[48] W. S. Neilson, "Smooth indifference sets," *J. Math. Econ.*, vol. 20, pp. 181-197, 1991.

[49] T. M. Diasakos and G. Gerasimou, "Preference conditions for invertible demand functions," *Discussion Paper Series*, School of Economics and Finance, University of St. Andrews, no. 1708, revised May 2017.

[50] S. Berry, A. Gandhi, and P. Haile, "Connected substitutes and invertibility of demand," *Econometrica*, vol. 81, no. 5, pp. 2087-2111, Sep. 2013.

[51] J. von Neumann and O. Morgenstern, *Theory of games and economic behavior*, Princeton University Press, Princeton, NJ, 1944.

[52] C. F. Camerer, "Behavioral economics," *Curr. Biol.*, vol. 24, no. 18, pp. R867-R871, Sep. 2014.

[53] D. Kahneman and A. Tversky, "Judgment under uncertainty: heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124-1131, 1974.

[54] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263-291, 1979.

[55] R. Kapeliushnikov, "Behavioral economics and the 'new' paternalism," *Russ. J. Econ.*, vol. 1, no. 1, pp. 81-107, Mar. 2015.

[56] A. Etzioni, "Behavioral economics: a methodological note," *J. Econ. Psychol.*, vol. 31, pp. 51-54, 2010.

[57] R. Thaler, "Toward a positive theory of consumer choice," *J. Econ. Behav. Organ.*, vol. 1, pp. 39-60, 1980.

[58] C. Sunstein and R. Thaler, "Libertarian paternalism is not an oxymoron," *University of Chicago Law Review*, vol. 70, no. 4, pp. 1159-1202, 2003.

[59] H. A. Simon, "Bounded rationality," *J. Eatwell, M. Milgate, & P. Newman (Eds.)*, The new Palgrave. N. Y.: W.W. Norton, 1987.

[60] G. J. Stigler, "The economics of information," *J. Political Econ.*, vol. 69, pp. 213-225, 1961.

[61] R. Polanía, M. Woodford, and C. C. Ruff, "Efficient coding of subjective value," *Nat. Neurosci.*, vol. 22, pp. 134-142, 2019.

[62] M. J. Rizzo and D. G. Whitman, "The knowledge problem of the new paternalism," *Brigham Young University Law Review*, vol. 4, pp. 905-968, 2009.

[63] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, pp. 656-715, 1949.

[64] P. Cuff and C. Schieler, "Secure source coding," *Information Theoretic Security and Privacy of Information Systems*, chap. 3, Cambridge University Press, Cambridge, 2017.

[65] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signals Syst.*, vol. 2, no. 4, pp. 303-314. Dec. 1989.

[66] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward Networks are universal approximators," *Neural Netw.*, vol. 2, pp. 359-366, 1989.

[67] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," Proc. of *Advances in Neural Information Processing Systems 4 (NeurIPS 1991)*, pp. 950-957, 1992.

[68] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: representing model uncertainty in deep learning,” Proc. of *the 33rd International Conference on Machine Learning (ICML 2016)*, PMLR, vol. 48, pp. 1050-1059, New York, NY, Jun. 2016.

[69] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123-140, Aug. 1996.

[70] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” Proc. of *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 6402-6413, 2017.

[71] Y. Yao, L. Rosasco, and A. Caponnetto, “On early stopping in gradient descent learning,” *Constr. Approx.*, vol. 26 no. 2, pp. 289-315, 2007.

[72] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” Proc. of *the 32nd International Conference on Machine Learning (ICML 2015)*, PMLR, vol. 37, pp. 448-456, 2015.

[73] T. Elsken, J. H. Metzen, and F. Hutter, “Neural architecture search,” *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, ch. 3, pp. 69-86, Springer, 2018.

[74] N. Le Roux and Y. Bengio, “Continuous neural networks,” Proc. of *the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, PMLR, vol. 2, pp. 404-411, 2007.

[75] G. Philipp and J. G. Carbonell, “Nonparametric neural networks,” Proc. of *the 5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, 2017.

[76] C. Bishop, “Bayesian methods for neural networks," *Tech. Rep. NCRG/95/009*, Neural Computing Research Group, Aston University, Jan. 1995. url: `https://www.microsoft.com/en-us/research/publication/bayesian-methods-for-neural-networks/`

[77] H. Chernoff and H. Teicher, “A central limit theorem for sums of interchangeable random variables,” *Ann. Math. Statist.*, vol. 29, no. 1, pp. 118-130, 1958.

[78] E. Hewitt and L. Savage, "Symmetric measures on cartesian products," *Trans. Amer. Math. Soc.*, vol. 80, pp. 470-501, 1955.

[79] D. Majerek, W. Nowak, and W. Zieba, "Conditional strong law of large number," *Int. J. Pure Appl. Math.* vol. 20, no. 2, pp. 143-157, 2005.

[80] A. Klenke, *Probability theory: a comprehensive course*, Translated from the 2006 German original, Universitext, Springer-Verlag London Ltd., London, United Kingdom, 2008.

[81] R. Neal, "Priors for infinite networks," *Tech. Rep.*, no. crg-tr-94-1, University of Toronto, 1994.

[82] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, "Gaussian process behaviour in wide deep neural networks," Proc. of *the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada, 2018.

[83] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as Gaussian processes," Proc. of *the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada, 2018.

[84] W. Grzenda and W. Zieba, "Conditional central limit theorems," *Int. Math. Forum*, vol. 3, no. 29-32, pp. 1521-1528, 2008.

[85] D. M. Yuan, L. R. Wei, and L. Lei, "Conditional central limit theorem for a Sequence of conditional independent random variables," *J. Korean Math. Soc.*, vol. 51, no. 1, pp. 1-15, 2014.

[86] P. Berti and P. Rigo, "A Glivenko-Cantelli theorem for exchangeable random variables," *Stat. Probabil. Lett.*, vol. 32, no. 4, pp. 385-391, Apr. 1997.

[87] P. Diaconis, "Finite forms of de Finetti's theorem on exchangeability," *Synthese*, vol. 36, pp. 271-281, 1977.

[88] P. Diaconis, and D. Freedman, "Finite exchangeable sequences," *Ann. Probab.*, vol. 8, pp. 745-764, 1980.

[89] C. Blundell, J. Cornebise, K. Kavukcuogle, and D. Wierstra, “Weight uncertainty in neural networks,” Proc. of *the 32nd International Conference on Machine Learning (ICML’15)*, vol. 37, pp. 1613-1622, Lille, France, 2015.

[90] M. A. T. Figueiredo and A. K. Jain, “Unsupervised learning of finite mixture models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381-396, Mar. 2002.

[91] G. J. McLachlan and S. Rathnayake, “On the number of components in a Gaussian mixture model,” *WIREs Data Min. Knowl.*, vol. 4, pp. 341-355, Sep., Oct. 2014.

[92] S. Krishnan, A. Garg, R. Liaw, B. Thananjeyan, L. Miller, F. T. Pokorny, and K. Goldberg, “SWIRL: a sequential windowed inverse reinforcement learning algorithm for robot tasks with delayed rewards,” *Int. J. Robotics Res.*, vol. 38, no. 2-3, pp. 126-145, Mar. 2019.

[93] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, “Occam’s razor,” *Inf. Process. Lett.*, vol. 24, pp. 377-380, Apr. 1987.

[94] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. 19, pp. 716-723, 1974.

[95] Schwarz G. “Estimating the dimension of a model,” *Ann. Stat.*, vol. 6, pp. 461-464, 1978.

[96] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[97] J. Oliver, R. Baxter, and C. Wallace, “Unsuperivsed learning using MML,” Proc. of *the 13th International Conference on Machine Learning (ICML 1996)*, pp. 364-372, 1996.

[98] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, pp. 424-438, 1969.

[99] C. E. Shannon, “A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, Jul., Oct. 1948.

[100] J. Massey, “Causality, feedback and directed information,” Proc. of *the International Symposium on Information Theory and Its Applications (ISITA 1990)*, Waikiki, Hawaii, 1990.

[101] T. Schreiber, “Measuring information transfer,” *Phys. Rev. Lett.*, vol. 85, pp. 461-464, 2000.

[102] N. Ay and D. Polani, “Information flows in causal networks,” *Adv. Complex Syst.*, vol. 11, pp. 17-41, 2008.

[103] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, “Quantifying causal influences,” *Ann. Stat.*, vol. 41, no. 5, pp. 2324-2358, 2013.

[104] G. Toroni, M. Boly, M. Massimini, and C. Koch, “"Integrated information theory: from consciousness to its physical substrate,” *Nat. Rev. Neurosci.*, vol. 17, no. 7, pp. 450-461, 2016.

[105] E. P. Hoel, L. Albantakis, and G. Tononi, “Quantifying causal emergence shows that macro can beat micro,” *Proc. Natl. Acad. Sci. U.S.A. (PNAS)*, vol. 110, no. 49, pp. 19790-19795, Dec. 2013.

[106] E. P. Hoel, “When the map is better than the territory,” *Entropy*, vol. 19. no. 5; 188, Apr. 2017. url: https://doi.org/10.3390/e19050188

[107] K. Goldberg, “Upper bounds for the determinant of a row stochastic matrix,” *J. Res. Natl. Stand. Sec. B.*, vol. 70B, no. 2, Apr.-Jun. 1966.

[108] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, “On entropy approximation for Gaussian mixture random vectors,” Proc. of *International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2008)*, TA2-3, Seoul, Korea, Aug. 20-22, 2008.

[109] S. Ravi and A. Beatson, “Amortized Bayesian meta-learning,” Proc. of *the 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, May 2019.

[110] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594-611, 2006.

[111] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni, “Generalizing from a few examples: a survey on few-shot learning,” *arXiv:1904.05046*, May 2019. url: https://arxiv.org/abs/1904.05046

[112] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," Proc. of *Advances in Neural Information Processing Systems (NeurIPS 2015)*, pp. 2575-2583, 2015.