Deep Learning for Energy Markets

Michael Polson∗ and Vadim Sokolov†

First Draft: April 2018
This Draft: August 2018

Abstract

Deep Learning (DL) provides a methodology to predict extreme loads observed in energy grids. Forecasting energy loads and prices is challenging due to sharp peaks and troughs that arise from intraday system constraints due to supply and demand fluctuations. We propose deep spatio-temporal models and extreme value theory (DL-EVT) to capture the tail behavior of load spikes. Deep architectures, such as ReLU and LSTM can model generation trends and temporal dependencies while EVT captures highly volatile load spikes. To illustrate our methodology, we use hourly price and demand data from the PJM interconnection for 4719 nodes and we develop a deep predictor. DL-EVT outperforms traditional Fourier and time series methods, both in- and out-of-sample, by capturing the nonlinearities in prices. Finally, we conclude with directions for future research.

Keywords: Deep Learning, PJM Interconnection, EVT, Machine Learning, Locational Marginal Price (LMP), Peak prediction, Energy Pricing, Smart Grid, LSTM, ReLU.

∗Michael Polson is a researcher at Bates White. email:michael.alan.polson@gmail.com
†Vadim Sokolov is Assistant Professor in Operations Research at George Mason University. email:vsokolov@gmu.edu
1 Introduction

Predicting load and wholesale energy prices to on the energy grid is essential for the economic operation of grid resources. Electricity grids operate without large amounts of storage, so the generation of energy (supply) within the system must always match the demand of energy (load). As supply must constantly adapt to meet changes in load, accurate predictions are essential to making informed short and long-term generation decisions. Accurate anticipation of fluctuations in load, especially sharp fluctuations, would remove certain flexibility constraints, allowing for efficient deployment of generation and grid resources.

Electricity price prediction is challenging due to a number of complex factors that impact the intraday grid conditions, which create highly volatile price spikes. We develop Deep Learning multi-layer networks to capture nonlinearities and the spatio-temporal patterns in energy prices and demand.

Energy price prediction has been traditionally achieved economic models based on firm behavior. Data-driven analytics, on the other hand, uses large data sets of prices and machine learning techniques to uncover price patterns. Until now data-driven models were not flexible enough to capture the extreme nonlinearities in the price the dynamics. Recently, deep learners (DL) have shown empirical success in large datasets forecasting problems with high dimensional nonlinearities. Long-short-term memory (LSTM) provides a framework for building spatio-temporal model [Dixon et al., 2017, Polson and Sokolov, 2017b].

The key to efficient electric grid management is to understand the peak loads. At the day-to-day level, over or underestimating the peak load can be costly to ratepayers. Overestimating the peak will cause the system too much generation in reserve. Underestimating the peak will cause the system to call upon costly, but flexible, sources of energy to quickly meet the demand. Day-to-day prediction is further complicated by the increase in renewable energy, whose pattern of generation does not always match the system’s pattern of demand. This imbalance in supply and demand patterns adds to the volatility of the system’s energy prices, further complicating prediction [Varaiya et al., 2011, Hogg and Klugman, 1983].

Forecasting supply and demand with a standard deep learning model fails to address the importance of peak prediction. Simple deep learning models predict the mean level of the dependent variable, in this case demand or energy price. Furthermore, squared loss is used to fit the model, which implicitly assumes the normal distribution of the errors. Therefore, a simple model would be well suited for predicting the system’s average demand or energy price, not the peaks and a simple model would fail to capture the true, fat-tailed distribution of the dependent variables.

Modeling extremes has a long history in financial risk management [Poon et al., 2003]. Incorporating extreme value theory (EVT) into deep learning allows us to capture the tail behavior of the price and load distribution. In the context of energy markets this is a crucial as peaks are the central component of interest in the market. We demonstrate how deep learning EVT approaches can be applied to model electricity prices, and we provide an improvement over traditional deep learning approaches, which focus on capturing the mean. Our work builds on that of Sigauke et al. [2013] who develop probabilistic EVT
model and [Shenoy and Gorinevsky, 2014] who use generalized linear model, with EVT errors to model electricity demand.

Davison et al. [2012] develop a spatial statistical model for the extremes of a natural process. Peaks are modelled as an exceedance of a certain threshold. EVT provides the framework for the prediction of these exceedances and for markets, it predicts the frequency of energy price exceeding a certain threshold [Davison and Smith, 1990]. The exceedance over a threshold allows to measure risk associated with high prices [Smith, 2002]. On the theoretical side we show how to incorporate deep learning into EVT.

The rest of our paper is organized as follows. Section 1.1 provides connections to previous work. Section 2 describes the energy market for electricity and the PJM interconnection. Section 3 discusses traditional deep learning models. Section 4 extends DL models using extreme value theory (EVT). Section 5 provides algorithms for load and price prediction for PJM. Finally, Section 6 concludes with directions for future research.

1.1 Connection to Previous Work

Data-driven energy pricing models to forecast hourly locational marginal price (LMP) has been previously studied [Catalão et al., 2007, Hong and Hsiao, 2002, 2001, Kim, 2015]. Hong and Hsiao [2001] proposed neural networks to predict LMPs in the PJM interconnection. Mandal et al. [2007] use neural networks to improve the performance, and Catalão et al. [2007], Kim [2015] predict LMPs in Nord Pool, an energy spot market located across countries in Northern Europe. Wang et al. [2017] predict prices at various hubs in the American Midwest with a stacked denoising auto-encoder exploiting local information to improve the predictive performance. Modeling of wind generated electricity was considered by Hering and Genton [2010]. Our analysis extends the functional data analysis approach for electricity pricing developed by Liebl [2013].

In another line of research, Cottet and Smith [2003] and Wilson et al. [2018] develop a random effects Bayesian framework to quantifying uncertainty in wholesale electricity price projections. Jónsson et al. [2013] forecast electricity prices while accounting for wind power prediction. Christensen et al. [2012] forecasts spikes in electricity prices. Heavy tails in electricity prices are modeled in Cottet and Smith [2003] with multivariate skew t-distributions. Benth and Schmeck [2014] address the non-Gaussian nature of the price data using Lévy process. Dupuis [2017] develop detrended correlation approach to capture the price dynamics within the New York section of the grid. Garcia et al. [2005] explains time-varying volatility in prices, using GARCH effects for one-day price forecasting. Li et al. [2007] developed fuzzy inference system to forecast prices on LMP spot markets. Subbayya et al. [2013] address the problem of model selection.

Our approach uses extreme value theory (EVT) in combination with deep learning. To our knowledge, this is the first time EVT-based deep learning model is developed and applied.
2 Energy Prices in PJM Interconnection

The PJM Interconnection is a regional transmission organization (RTO), which exists to create a competitive wholesale electricity market, coordinating numerous wholesale electricity producers and consumers in all or parts of 13 states located in America’s Mid-Atlantic and Great Lakes Regions as well as the District of Columbia.

PJM is broken into 20 transmission zones. Each zone is owned and operated by separate transmission owners who are responsible for designing and maintaining their portion of the system. Figure 1 shows the locations of the load nodes and zone boundaries of the PJM Interconnection. Individual utilities within PJM plan their use of resources around peak loads. Predicting the strength of these peaks, and when they occur, is integral to improving both short-term and long-term decision-making. Current methods for short-term prediction focus on neural networks (weather channel, PJM). note: discuss distribution of errors

PJM acts as a guarantor of system reliability and is responsible for preventing outages within the system. PJM operates the system at a cost-efficient level by coordinating generating plants operations, which are owned by numerous entities, to match the system’s demand. Operating the system includes ensuring real-time demand is met, maintaining a reserve capacity of generation, and, monitoring transmission lines, to prevent overloaded lines which could cause system failure [Cain et al., 2007].

![Figure 1: PJM Zone boundaries and node locations](image)

The PJM interconnection contains over 11,000 nodes for which an hourly day-ahead or real-time prices are reported. These nodes are specific generation or load locations,
aggregates of various locations, regions, or points of interconnection with areas outside of PJM. We examine price at the 4,700 load nodes across the system.

Within the PJM Interconnection nearly all wholesale electricity is bought or sold through bilateral contracts. The remainder is bought or sold on the two bid-based electricity markets PJM operates: the day-ahead (DA) and real-time (RT) markets. In the day-ahead market, market participants submit bids or offers to buy or sell energy to the scheduling operator (PJM). The operator uses the bid and offers to determine the day-ahead LMP, which reflects the expected cost of energy, congestion and transmission loss needed to provide electricity at a location given expected system constraints.

The real-time market operates in a similar way, but reflects the actual cost of providing electricity at a location given actual system constraints. Despite the comparatively smaller volume, the real-time market plays a central role in determining the price of all futures contracts as the futures contracts’ price depends on the expectation of the real-time market prices. The day-ahead market is a futures market that allows generators to enter agreements to provide electricity for the coming day.

Generators can fulfill obligations to provide energy through physically producing electricity, or purchasing electricity on the real-time market. Multiple factors, such as unexpected maintenance, may cause a generator to fulfill their obligation through purchases on the real-time market, rather than generation. These factors, or risks, in part cause significant volatility in real-time markets compared to the day-ahead market [FERC, 2014].

Prices in the real-time market are a function of the cost to produce energy and system constrains, such as congestion in transmission lines. When these constraints are binding, prices differ across locations in the PJM Interconnection to reflect the relative ease of delivering energy to a non-congested location and the relative difficulty of delivering energy to a congested location. Therefore, each node (or location) has an associated locational marginal price (LMP), which reflects the price of the marginal unit of electricity delivered to that given location. LMPs are important price signals in the day-ahead and real-time market, which inform short-term decisions, as well as long-term investments and bilateral agreements [Cain et al., 2007].

2.1 Local Marginal Price Data (LMP)

Locational Marginal Pricing is used to price energy on the PJM market in response to changes in supply and demand and physical constraints of the hardware. LMP accounts for the cost to produce the energy, the cost to transmit the energy within PJM RTO, and the cost of energy lost due to resistance as the energy is transported across the system. LMP data is available at www.pjm.com [PJM]. Our study used price data, which includes real-time and day-ahead hourly prices from 1/1/2017 through 12/31/2017. Load prices represent the cost of providing electricity at a given location. The price reflects the system’s load (demand), generation and limits of the transmission system. The system’s constraints can affect locations asymmetrically, causing variations in price across different location. Hub prices are a collection of these locational prices intended to reflect the uncongested price of electricity.

LMPs have three components: energy, congestion and marginal loss. The energy component reflects the price of electricity, called system marginal price (SMP). SMP is calcu-
lated bases on the current dispatch (supply) and the load (demand). SMP is calculated both in day ahead and real time. The congestion component is greater then zero when one of the constraints at a given node are violated. Constraints occur when delivery limitations prevent use of least-cost generator, for example a higher cost generator closer to load must be used to meet the demand if transmission constraints are present. The congestion price is calculated using the shadow price, which is the value of the dual variable (price of violating a binding constrain) in the optimization problem that governs the grid. When none of the constraints are active, all of the congestion prices are zero.

The marginal loss component reflects the cost of transmission and other losses at a given location. Losses are priced according to marginal loss factors which are calculated at a bus and represent the percentage increase in system losses caused by a small increase in power injection or withdrawal.

| Variable name        | Description                                                                 |
|----------------------|-----------------------------------------------------------------------------|
| TotalLMP             | total cost, reflects Energy + congestion + marginal loss                      |
| CongestionLMP        | congestion component of the LMP, can be $+\text{ve}$ or $-\text{ve}$       |
|                      | Value is relative to the energy component                                   |
| MarginalLMP          | marginal loss component of the LMP                                           |

Table 1: Description of LMP variables.

2.2 Nonlinearities in Prices and Demand

The dynamics of the energy prices are nonlinear due to congestion component in the Locational Marginal Price. The congestion price represents the cost of violating a binding constraint in the linear program that models optimal generation strategy. The congestion price is payed by the load (consumer) to the generator (the producer). The congestion prices are calculated in real time and a day ahead. The constraints are the results of several physical limits of the electric grid and include thermal limits due to thermal capability of power system equipment, voltage limits and stability limits. Figure 2 shows the temporal patterns in the load data and relations between price and load variables.
We see that relations are non-linear.

3 Deep Learning

Deep learning model can efficiently approximate high-dimensional functions $Y = f(X)$. A deep network, denoted by $\hat{y}(x)$, is defined by hierarchical layers

$$
\begin{align*}
x_0 &= x, \quad x_l = \sigma(W_l x_{l-1} + b_l) \\
Y &= \sigma(x_L) \\
\sigma(x) &= \text{ReLU}(x) := \max(x, 0),
\end{align*}
$$

where $W_l \in \mathbb{R}^{n_l-1}$, $b_l \in \mathbb{R}$ and $n_l$ is the number of neuron in layer $l$.

To add a time series component, we use Recurrent Neural Networks (RNNs) which are able to capture time series properties of electricity prices. Recurrent layers capture long term dependencies without much increase in the number of parameters. They learn temporal dynamics by mapping an input sequence to a hidden state sequence and outputs via a recurrent relations. Let $y_t$ denote the observed data and $h_t$ a hidden state, then

$$
\begin{align*}
y_t &= f(w^T_{h_z} h_t + b_z) \\
h_t &= f(w^T_h [x_t, h_{t-1}] + b_h).
\end{align*}
$$

The main difference between RNNs and feed-forward deep learning is the use of a hidden layer with an auto-regressive component, here $w_{h_t} h_{t-1}$. It leads to a network topology in which each layer represents a time step, and we index by $t$ in order to highlight the temporal nature.
A particular type of RNN, called LSTM (Long short-term memory) was proposed to address the issue of vanishing or exploding gradients in plain RNNs. A memory unit used in LSTM networks allows the network to learn what previous states can be forgotten \cite{Hochreiter1997, Schmidhuber1997}.

The hidden state will be generated via another hidden cell state $c_t$ that allows for long term dependencies to be “remembered”. Then we generate

$$\begin{align*}
\text{Output: } h_t &= o_t \ast \tanh(c_t) \\
 k_t &= \tanh(w_c^T [h_{t-1}, x_t] + b_c) \\
 c_t &= f_t \ast c_{t-1} + i_t \ast k_t \\
\text{State equation: } \begin{pmatrix} f_t \\ i_t \\ o_t \end{pmatrix} &= \sigma(w^T [h_{t-1}, x_t] + b).
\end{align*}$$

Where $\ast$ denotes point-wise multiplication. Then, $f_t \ast c_{t-1}$ introduces the long-range dependence. The states $(i_t, f_t, o_t)$ are input, forget and output states. Figure 4 shows the network architecture.

The key addition, versus RNN is the cell state $c_t$, the information is added or removed from the memory state via gates defined via the activation function $\sigma(x)$ and point-wise multiplication $\ast$. The first gate $f_t \ast c_{t-1}$, called the forget gate, allows to throw away some data from the previous cell state. The next gate $i_t \ast \bar{c}_t$, called the input gate, decides which values will be updated. Then, the new cell state is a sum of the previous cell state, passed through the forget gate selected components of the $[h_{t-1}, x_t]$ vector. Thus, the
vector $c_t$ provides a mechanism for dropping irrelevant information from the past, and adding relevant information from the current time step. The output is the result of the output gate $h_t = o_t \cdot \tanh(c_t)$, which returns $\tanh$ applied to the cell state, with some entries removed. The forget gate resolves the problem of vanishing gradient, which is the case when values of the gradient vector are close to zero. SGD optimization algorithm is straightforward to implement. See Section 3.1 for discussion.

Deep rectified linear units (ReLU) with long short term memory (LSTM) cells have become popular architectures as they can capture long-range dependencies and nonlinearities. They can efficiently approximate highly multivariate functions with small number of neurons at each layer [Bach, 2017, Schmidt-Hieber, 2017, Yarotsky, 2017].

3.1 Stochastic Gradient Descent for Deep Learning (SGD)

Once the activation functions, depth $L$ and size $n_1, \ldots, n_L$ of the learner have been chosen, the parameters, $\hat{W}$ and $\hat{b}$ are found by solving the following optimization problem

$$\minimize_{W, b} \frac{1}{N} \sum_{i=1}^{T} \mathcal{L}(Y_i, \hat{Y}_{W, b}(x_i)) + \phi(f_{DL}),$$

(1)

Which is a penalized loss function, where $(Y_i, x_i)^N_{i=1}$ is training data of input-output pairs, and $\phi(f_{DL})$ is a regularization penalty on the network parameters (weights and offsets). For example if, $\mathcal{L}(Y_i, \hat{Y}(x_i))$ is an $L_2$-norm we have a traditional least squares problem [Janocha and Czarnecki, 2017].

Most architectures employ regularization techniques to prevent the model from over-fitting training set data [Hinton and Salakhutdinov, 2006]. This improves the model’s predictive performance on data outside of the training set. Normally, a regularization penalty $\lambda(W, b)$ is added to the loss function to avoid over-fitting. Dropout, the technique of removing input dimensions in $x$ randomly with probability $p$, can also be used to further reduce the change of over-fitting during the training process [Srivastava et al., 2014].

The common numerical approach to find the solution of equation (1) is a form of stochastic gradient descent (SGD), generally referred to as back-propagation in machine learning. One caveat of back-propagation in this context is the complexity of the system to be solved, resulting in slow convergence properties. As a result, deep learning methods rely heavily on large computational power [Abadi et al., 2016, Cardoso].

We now show how to incorporate deep learning into extreme value theory (EVT) as described by [Coles et al., 2001].

4 Deep Learning Extreme Value Theory (DL-EVT)

Suppose, that we have data denoted by $Y(s_i, t_j)$ at spatial locations $s_i, 1 \leq i \leq n$ and time $t_j, 1 \leq j \leq T$. We are interested in the exceedances of prices at a specific location $s$, $Y_t = \{Y(s, t)\}_{i=1}^{n}$ given that $Y_t \geq u$, where $u$ is a given threshold.
Following limiting results from EVT, the tail distribution of $Y_t$ follows a generalized Pareto (GP) distribution.

$$H(y \mid \sigma, \xi) = 1 - \left( 1 + \xi \frac{y - u}{\sigma} \right)^{-1/\xi}.$$  

The corresponding density function is given by

$$h(y \mid \sigma, \xi) = \frac{1}{\sigma} \left( 1 + \xi \frac{y - u}{\sigma} \right)^{-1/\xi - 1}, \quad \left( 1 + \xi \frac{y - u}{\sigma} \right) > 0, \quad \xi \neq 0$$

Here $(u, \sigma, \xi)$ are the location, scale and shape parameters, $\sigma > 0$ and $z_+ = \max(z, 0)$. $H(a)$ is called Generalized Pareto (GP) distribution. The Exponential distribution is obtained by continuity as $\xi \to 0$, and we have

$$\lim_{\xi \to 0} h(y \mid \sigma, \xi) = \sigma \exp(-\sigma(y - u))$$

Under this distribution, the mean value of the $y$ is $\sigma + u$.

Under the assumption of GP distribution for our dependent variable, the log-likelihood for a single observation $y$ is

$$\log \sigma(t) - \left( \frac{1}{\xi} + 1 \right) \log \left( 1 + \sigma(t) \xi (y - u) \right).$$

The loss function for our deep learning model model becomes

$$L(W) = -\frac{1}{n} \sum_{i=1}^{n} l(y_i \mid x_i, \xi, u, W, b).$$

The weights $W$ and offsets $b$ are learned using stochastic gradient descent (SGD) applied to the corresponding loss function.

DL-EVT framework models the size of the exceedance as a function of time and $\sigma(t)$, which depends on the recently observed values of demand (a local trend model). Linear regression GP were developed in [Davison and Smith, 1990, Beirlant et al., 2006]. To complete our specification for exceedance sizes we assume a functional form for $\sigma(t)$ that is a deep neural network. We introduce

$$\sigma(t) = F \left( \sum_{i=t-h}^{t} \sum_{j=1}^{p} W_j Y(s_j, i) \right), \text{ where } F = f_1 \circ \ldots \circ f_L$$

Here $F$ is a deep learner constructed via superposition of semi-affine univariate functions, see Polson and Sokolov [2017a, 2015, Dixon et al., 2017] for further discussion. In our setting, $Y = \{Y(s_j, i), i = 1, \ldots, p, j = t-h, \ldots, t\}$ is generally high dimensional set of
predictor variables. For example, a 2-layer network model

\[ \sigma(t) = f_1(W_1 Z(t) + b_1) \]
\[ Z(t) = f_2(W_2[Z_1(t-1), x(t)] + b_2), \]

where \( W_2 = [W_2^Z, W_2^X] \). The functions \( f_1, f_2 \) are univariate nonlinear activation functions, such as ReLU or tanh.

In general, a deep learning model recover the multivariate function (map) between a (high dimensional) set of inputs \( X = (x_1, \ldots, x_p) \in \mathbb{R}^p \) and a (low dimensional) output \( Y \). The map is denoted by \( Y = f(X) \) and constructed using training data of input-output pairs \( (X_i, f(X_i))_{i=1}^N \), that generalizes well for out-of-sample data. Deep learning uses compositions of functions, rather than traditional additive ones. By composing \( L \) layers, a deep learning predictor becomes

\[ \hat{Y} = F_{w,b}(x) = (f_{w_0,b_0} \circ \ldots \circ f_{w_L,b_L})(X) \]
\[ f_{w_l,b_l} = f_l(w_l x_l + b_l). \]

Here \( X = (x_1, \ldots, x_p) \). The set \( (W, b) = (W_1, \ldots, W_L, b_1, \ldots, b_L) \) is the set of weights and offsets which are learned from training data, and \( w_l \in \mathbb{R}^{p_l \times p_l} \) and dimensionality \( p_l \) is of the architecture specifications.

## 5 PJM Price and Demand Forecasting

We predict the price at each load point using the historical prices, demand and weather observations. There are a total of 4719 generating nodes in the system. Plot 5 shows that there are strong spatial correlation among prices at different zones. Zone aggregates several nodes. Thus, prices at nodes will be correlated as well. We use this spatial pattern to build our model.

![Color Key and Histogram](image)

**Figure 5: Correlation in Marginal Prices Among Zones**

Further, Figure 6 shows the relations between price and weather as well as price and
the demand. We see that demand is not always an accurate predictor for the price. This is due to the nonlinearities present in the system. High demand does not necessarily lead to high prices. We only observe the demand variable for the overall system and not on individual nodes. Thus, we cannot use spatial patterns in the demand variable to predict the prices.

![Figure 6: (a) Price vs Temperature (C). (b) Demand vs Price (C)](image)

We demonstrate our forecasting approach for a node named “KULLERRD138 KV T-2”, with id 48667, which is located in Clifton, NJ.

First, we try a traditional model for electricity prices, which uses Fourier series to describe the seasonal patterns and short-term time series dynamics modeled as an ARIMA terms. Here $y_t$ is decomposed as a sum, a deterministic Fourier term $f(t)$, and a stochastic component, $N_t$, leading to

$$y_t = a + f(t) + N_t,$$

where $f(t) = \sum_{k=1}^{K} \left[ \alpha_k \sin\left(\frac{2\pi kt}{m}\right) + \beta_k \cos\left(\frac{2\pi kt}{m}\right) \right],$

where $N_t$ is an ARIMA process. The number of terms of $K$ can be chosen by minimizing cross-validation. This allows: (i) any length seasonality, (ii) several seasonality periods. Smoothness of the seasonal term is governed by value $K$. The short-term dynamics is handled with an ARMA error. The only real disadvantage (compared to a seasonal ARIMA model) is that the seasonality is assumed to be fixed, the pattern is not allowed to change over time. In practice, seasonality is usually remarkably constant, so assumption generally holds except in applications with very long time series.

The in-sample fit of the Fourier model is shown on Figure [11]
This model captures the cyclical patterns in the prices but does not accurately capture the levels of the peak prices.

Figure 8 shows the out-of-sample prediction for the 32 hours of price observations for Fourier model with weather and demand predictors. Inclusion of predictors does not change the quality of forecasting peak prices. As we noted in our exploratory plots, demand is not a good predictor of a peak price.

Figure 8: Out-of-sample Prediction from linear model with ARIMA\(_{(2,0,0)}\) errors and Fourier predictors. Yellow is data and blue is the forecast with confidence intervals.
### Table 2: Out-of-sample performance of DL and Fourier models

| Method   | mse | mre  | mae | mape |
|----------|-----|------|-----|------|
| Fourier  | 26.6| 5.1  | 4   | 0.19 |
| LSTM     | 16.8| 4.1  | 2.4 | 0.09 |

Table 2: Out-of-sample performance of DL and Fourier models

![Figure 9: Comparison of Fourier and DL models](image)

**5.1 Demand Forecasting**

Electricity load forecasting is essential for designing operational strategies for electric grids. In presence of renewable energy sources short-term forecasts are becoming increasingly important. Many decisions, such as dispatch scheduling and demand management strategies are based on load forecasts [Taylor et al. 2007]. One hour-ahead forecasts are a key input for transmission companies on a self-dispatching markets [Garcia and Kirschen 2006]. Hourly behavior of electricity load is known to be non-stationary [Almeshaiei and Soltan 2011]. Since there is not much of a change in meteorological variables, it is typical to use univariate time series data for short-term load forecasting [Bunn 1982].

In this section, we analyze an hourly electricity load observations on the PJM interconnection. The data is available at [http://pjm.com/pub/account/loadhryr/2018.txt](http://pjm.com/pub/account/loadhryr/2018.txt). We use data for January 2016 - May 2018 period year of observations and use last 10 days of observations for testing. We use a local trend model that takes previous 24 observations (one day) to predict load in five hours.
Figure 10: Hourly electricity load on PJM interconnect in MW.

Figure 10(b) shows the hourly PJM interconnect load series from January 2016 to April 17, 2018. This data was used to train our DL-EVT model. The graph shows seasonal cycles. Figure 10(a) shows the shorter period (Jan-Feb 2016) of the same data and shows daily and weekly cycles. We can see that weekends have lower load levels compared to work days.

The following architecture is used to model the relations between previous load observations \(X\) and the scale parameter of the Generalize Pareto distribution \(\sigma\).

\[
X \xrightarrow{\tanh(W^{(1)}X + b^{(1)})} Z^{(1)} \xrightarrow{\exp(\tanh(Z^{(1)})}) \sigma(X)
\]

Figure 11: Our DL-EVT Architecture

where \(W^{(1)} \in R^{p \times 3}\) matrix. \(x \in R^p\) is the vector of recent observations of electricity demand, we used \(p = 24\) (1 day). To train the EVT model we only used the observations \(y_i > u\) with \(u = 31000\). We used the mean \(\sigma + u\) of the GP distribution as the point estimate for plotting Figure 12(b).

Our DL-EVT model is compared with a vanilla deep learning model with standard mean squared error loss function. Figure 12 shows the resulting out-of-sample forecasts.
Figure 12: One our electricity load and its forecast for the period from Friday, April 27, 2018 to Monday, May 7, 2018

We can see that while a standard DL model captures both ups and downs in the load levels, the DL-EVT model does capture the location and level of the peak loads more accurately compared to the standard DL model.

6 Discussion

Deep learning combined with extreme value theory can predict peaks in electricity prices and demand. With availability of real-time data, computational power and machine learning pattern recognition tools such as deep learning, we now have the ability to better predict and manage energy generation and distribution. One of the goals of our is to demonstrate that an EVT extension of the standard DL framework is a viable option and is applicable to electricity data. DL-EVT performed well on in and out-of-sample forecasting of electricity prices and load.

Forecasting electricity prices is challenging as energy prices can spike due to supply-demand imbalances whilst simultaneously having long-range dependence. Deep ReLU LSTM models capture spikes with non-linear activation functions, scalable and can efficiently fit using SGD. For a grid of 4786 electricity load nodes we show how such models can fit in-sample with better accuracy then traditional time series models. There are a number for directions for future research. For extensions to multivariate time series data with spatio-temporal dynamics, see Dixon et al. [2017].

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. Tensorflow: A system for large-scale machine learning. In OSDI, volume 16, pages 265–283, 2016.
Eisa Almeshaiei and Hassan Soltan. A methodology for electric power load forecasting. *Alexandria Engineering Journal*, 50(2):137–144, 2011.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

Fred Espen Benth and Maren Diane Schmeck. Pricing futures and options in electricity markets. In *The Interrelationship Between Financial and Energy Markets*, pages 233–260. Springer, 2014.

Derek W Bunn. Short-term forecasting: a review of procedures in the electricity supply industry. *Journal of the Operational Research Society*, 33(6):533–545, 1982.

Collin Cain, Jonathan Lesser, and Bates White. A common sense guide to wholesale electric markets. Technical report, Bates White, Tech. Rep., Apr. 2007.[Online]. Available: http://www.bateswhite.com/media/publication/55_media.741.pdf, 2007.

Pedro Emanuel Almeida Cardoso. Deep learning applied to pmu data in power systems.

João Paulo da Silva Catalão, Sílvio José Pinto Simões Mariano, VMF Mendes, and LAFM Ferreira. Short-term electricity prices forecasting in a competitive market: A neural network approach. *Electric Power Systems Research*, 77(10):1297–1304, 2007.

Tim M Christensen, A Stan Hurn, and Kenneth A Lindsay. Forecasting spikes in electricity prices. *International Journal of Forecasting*, 28(2):400–411, 2012.

Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

Remy Cottet and Michael Smith. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849, 2003.

Anthony C Davison and Richard L Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 393–442, 1990.

Anthony C Davison, Simone A Padoan, and Mathieu Ribatet. Statistical modeling of spatial extremes. *Statistical science*, 27(2):161–186, 2012.

Matthew F Dixon, Nicholas G Polson, and Vadim O Sokolov. Deep learning for spatiotemporal modeling: Dynamic traffic flows and high frequency trading. *arXiv preprint arXiv:1705.09851*, 2017.

Debbie J Dupuis. Electricity price dependence in new york state zones: A robust detrended correlation approach. *The Annals of Applied Statistics*, 11(1):248–273, 2017.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pages 907–940, 2016.
Federal Energy Regulatory Commission FERC. Operator-initiated commitments in rto and iso markets. *FERC Docket No*, pages D14–14, 2014.

Maria P Garcia and Daniel S Kirschen. Forecasting system imbalance volumes in competitive electricity markets. *IEEE Transactions on Power Systems*, 21(1):240–248, 2006.

Reinaldo C Garcia, Javier Contreras, Marco Van Akkeren, and João Batista C Garcia. A garch forecasting model to predict day-ahead electricity prices. *IEEE transactions on power systems*, 20(2):867–874, 2005.

Amanda S Hering and Marc G Genton. Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105(489):92–104, 2010.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997.

Robert V Hogg and Stuart A Klugman. On the estimation of long tailed skewed distributions with actuarial applications. *Journal of Econometrics*, 23(1):91–102, 1983.

Y-Y Hong and C-Y Hsiao. Locational marginal price forecasting in deregulated electricity markets using artificial intelligence. *IEE Proceedings-Generation, Transmission and Distribution*, 149(5):621–626, 2002.

Ying Yi Hong and Chuan-Yo Hsiao. Locational marginal price forecasting in deregulated electric markets using a recurrent neural network. In *Power Engineering Society Winter Meeting, 2001. IEEE*, volume 2, pages 539–544. IEEE, 2001.

Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv:1702.05659*, 2017.

Tryggvi Jónsson, Pierre Pinson, Henrik Aalborg Nielsen, Henrik Madsen, and Torben Skov Nielsen. Forecasting electricity spot prices accounting for wind power predictions. *IEEE Transactions on Sustainable Energy*, 4(1):210–218, 2013.

Mun-Kyeom Kim. A new approach to short-term price forecasting strategy with an artificial neural network approach: Application to the nord pool. *Journal of Electric Engineering and Technology*, 10:709–718, 2015.

Guang Li, Chen-Ching Liu, Chris Mattson, and Jacques Lawarrée. Day-ahead electricity price forecasting in a grid environment. *IEEE Transactions on Power Systems*, 22(1):266–274, 2007.

Dominik Liebl. Modeling and forecasting electricity spot prices: A functional data perspective. *The Annals of Applied Statistics*, 7(3):1562–1592, 2013.
P. Mandal, T. Senjyu, N. Urasaki, T. Funabashi, and A. K. Srivastava. A novel approach to forecast electricity price for PJM using neural network and similar days method. *IEEE Transactions on Power Systems*, 22:4, 2007.

PJM. Daily real-time LMP: August 1, 2017 through September 7, 2017. URL [http://www.pjm.com/markets-and-operations/energy/real-time/lmp.aspx](http://www.pjm.com/markets-and-operations/energy/real-time/lmp.aspx).

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but-not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Nicholas Polson and Vadim Sokolov. Bayesian analysis of traffic flow on interstate I-55: The LWR model. *The Annals of Applied Statistics*, 9(4):1864–1888, 2015.

Nicholas G. Polson and Vadim Sokolov. Deep Learning: A Bayesian Perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017a.

Nicholas G. Polson and Vadim O. Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17, June 2017b. ISSN 0968-090X. doi: 10.1016/j.trc.2017.02.024.

Ser-Huang Poon, Michael Rockinger, and Jonathan Tawn. Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *The Review of Financial Studies*, 17(2):581–610, 2003.

Jürgen Schmidhuber and Sepp Hochreiter. Long Short-Term Memory. *Neural Comput*, 9(8):1735–1780, 1997.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv:1708.06633*, 2017.

Saahil Shenoy and Dimitry Gorinevsky. Risk adjusted forecasting of electric power load. In *American Control Conference (ACC)*, 2014, pages 914–919. IEEE, 2014.

Caston Sigauke, Andréhette Verster, and Delson Chikobvu. Extreme daily increases in peak electricity demand: Tail-quantile estimation. *Energy Policy*, 53:90–96, 2013.

Richard L. Smith. Measuring risk with extreme value theory. *Risk management: value at risk and beyond*, 224, 2002.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Sankaranarayanan Subbayya, Jorjeta G Jetcheva, and Wei-Peng Chen. Model selection criteria for short-term microgrid-scale electricity load forecasts. In *Innovative Smart Grid Technologies (ISGT), 2013 IEEE PES*, pages 1–6. IEEE, 2013.
James W Taylor, Patrick E McSharry, et al. Short-term load forecasting methods: An evaluation based on european data. *IEEE Transactions on Power Systems*, 22(4):2213–2219, 2007.

Matus Telgarsky. Neural networks and rational functions. *arXiv:1706.03301*, 2017.

Pravin P Varaiya, Felix F Wu, and Janusz W Bialek. Smart operation of smart grid: Risk-limiting dispatch. *Proceedings of the IEEE*, 99(1):40–57, 2011.

Long Wang, Zijun Zhang, and Jieqiu Chen. Short-term electricity price forecasting with stacked denoising autoencoders. *IEEE Transactions on Power Systems*, 32(4):2673–2681, 2017.

AL Wilson, CJ Dent, and M Goldstein. Quantifying uncertainty in wholesale electricity price projections using bayesian emulation of a generation investment model. *Sustainable Energy, Grids and Networks*, 13:42–55, 2018.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.