Fast Anomaly Detection in SmartGrids via Sparse Approximation Theory

Marco Levorato[*†] and Urbashi Mitra[†]
∗ Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305 USA
† Dept. of Electrical Engineering, University of Southern California, Los Angeles, USA.
e-mail: levorato@stanford.edu, ubli@usc.edu.

***Abstract*— The SmartGrid (SG) is a complex system connecting physical components (*e.g.*, human, weather, power plants) and logical components (*e.g.*, control algorithms, communication infrastructure, protocols). The large number of components and the interactions between the individual components induce an extremely intricate behavior of the overall system. Detecting anomalies in the behavior of the system requires a large number of observations and is unpractical. A novel learning and estimation framework to analyze stochastic processes over graphs associated with SG systems is proposed. The critical observation behind the proposed framework in that these systems induce an underlying sparse structure which enables dimension reduction via compressed sensing-like schemes. Numerical results show that the *compression* approach proposed herein reduces by orders of magnitude the number of observations required to detect an anomalous behavior of the SG.**

I. Introduction

The need for a shift in the energy production and consumption model dictated by environmental and social challenges is generating an intense research effort to define the energy grid of the future, *i.e.*, the SmartGrid (SG) [1]. Intelligent buildings and load control [2], energy markets [2], and distributed and renewable-based energy production [3] make the behavior of the SmartGrid extremely dynamic and complex. It is envisioned that a large sensor network will be deployed to provide measurements and data at a fine time scale to monitoring and control stations distributed over the grid. Thus, the SG is a large system interconnecting millions of intelligent entities producing, selling, consuming energy, sensing the grid and actively making decisions. Ensuring robustness and reliability of this complex and heterogeneous system is a major technological challenge.

Prompt detection of *anomalies* is critical to enable effective countermeasures and avoid grid failures. Prior work on anomaly detection focused on detecting failures in the physical structure of the grid (*e.g.*, [4]) or on particular aspects of the overall SG system (*e.g.*, cyber attacks [5]). In this paper, we seek the identification of anomalies in the stochastic behavior of the overall SG system. We propose a broader definition of anomaly as malfunctions of physical entities of the grid (lines, production sites, *etc.*), but also as unexpected or unforeseen behavior of production and consumption potentially leading to failure. Traditional analysis and estimation algorithms cannot deal with the enormous complexity of the SG system, and would fail to provide timely identification especially given the complex behavior of the system [6].

We consider a local sector of the grid where energy production sites (traditional fossil fuel plants and renewable-energy farms) are interconnected to smart loads (residential and commercial buildings) by transmission and distribution lines. Fossil fuel power plants have a production profile that is planned a day ahead by the utility. The energy produced by renewable energy production sites (either owned by the utility or individual households) is a function of local weather factors, such as wind strength and solar illumination. Smart loads are intelligent buildings whose energy consumption is controlled by automated systems that determine the activation of appliances and electric devices in the building based on consumer choices and habits, weather conditions and energy prices. Smart meters installed in the buildings and sensors provide measurements and data to monitor and control stations. Data are used to build a stochastic model of the system and determine pricing and production strategies, as well as to track the current state of the grid. A cost function is defined to penalize potentially unstable and dangerous states of the grid. The stochastic model based on prediction is used to compute the discounted value function associated with the cost function [6]. The value function captures the long-term average cost, which is associated with the probability that the system will hit high cost states.

We propose a novel anomaly detection technique for SmartGrids based on sparse approximation theory and wavelets tailored to our scenario. Our baseline observation is that the graphical structure of the stochastic process associated with the behavior of technology-driven and artificial systems is *compressible*, meaning that it admits a concise representation in a proper domain. A concise representation allows the construction of compressed sensing-like algorithms detecting the behavior of the SG and estimating the value function from a small number of observations. We underscore that prior work proposed the use of compressed sensing paradigms to retrieve the structure or properties of the *physical graph* of networks, that is, the graph composed by the terminals and the physical connections between them (*e.g.*, [7]). In contrast, our work centers on the analysis, and compression, of the *logical graph* of the SG, where vertices are logical state of the system and edges are state transitions. In the proposed framework, Sparse Group Least Angle Selection and Shrinkage Operator (LASSO) [8] is used to select the model for the behavior of the system based on the observation of a trajectory of the system. This first step enables the detection of anomalies

This work was supported in part by the National Science Foundation (NSF) under Grant CCF-0917343.

$$\Gamma(\mathbf{l}'|\mathbf{c},\mathbf{w}',p') = P(\mathbf{L}(t{+}1){=}\mathbf{l}'|\mathbf{C}(t){=}\mathbf{c},\mathbf{W}(t{+}1){=}\mathbf{w}',P(t{+}1)){=}p') \tag{1}$$

at the structural level of the stochastic process determining the trajectory of the system. Then, for the selected model, LASSO is employed to estimated the value function and detect anomalies in the expected long-term behavior of the system.

The rest of the paper is organized as follows. Section II describes the stochastic model and the estimation problem. The learning and detection algorithm is presented in Section III. Section IV presents numerical results.

II. Stochastic Model and Estimation Problem

We define a time indexing $t{=}1, 2, 3, \ldots$, and the time interval $[t\ell, (t+1)\ell]$ corresponds to "time t". The probability of an event is denoted by $P(\cdot)$. A collection of variables $\mathbf{W}(t){=}\{W_i(t)\}_{i=1,\ldots,N_w}{\in}\mathcal{W}$ describes the weather conditions at time t. The variable $W_i(t)$ can be associated, *e.g.*, with wind strength or solar illumination in a geographic area of the grid. We capture time variations in the behavior of the consumer by defining the variable $\mathbf{C}(t) = \{C_i(t)\}_{i=1,\ldots,N_c}{\in}\mathcal{C}$. $\mathbf{C}(t)$ is a collection of *logical states* associated with the individual consumer (or aggregates of consumers). Each logical state $C_i(t)$ expresses factors influencing the energy consumption of a building/load (*e.g.*, presence of the consumer at home, electric vehicle under charge, etc.). Analogously, we define the collection of variables $\mathbf{F}(t){=}\{F_i(t)\}_{i=1,\ldots,N_f}{\in}\mathcal{F}$ and $\mathbf{R}(t){=}\{R_i(t)\}_{i=1,\ldots,N_r}{\in}\mathcal{R}$ associated with the level of energy production of fossil fuel power plants and renewable-based farms. The variable $P(t){\in}\mathcal{P}$ defines the price of energy at time t.

In order to capture correlation and burstiness in the weather conditions and consumer behavior we model the sequences $[\mathbf{W}(t)]_{t=1,\ldots}$ and $[\mathbf{C}(t)]_{t=1,\ldots}$ as Markov chains. Note that variables tracking time and slowly varying states of metereological conditions or consumer behavior can be included in the model to improve its accuracy. The transition probabilities of the chain modeling the weather are denoted by $p_w(\mathbf{w}'|\mathbf{w}){=}P(\mathbf{W}(t{+}1){=}\mathbf{w}'|\mathbf{W}(t){=}\mathbf{w})$. Similarly, the logical state of each smart load evolves according to the transition probabilities $p_c(\mathbf{c}'|\mathbf{c}){=}P(\mathbf{C}(t{+}1){=}\mathbf{c}'|\mathbf{C}(t){=}\mathbf{c})$.

The energy consumption is influenced by weather conditions (*e.g.*, heating and cooling), logical state of the user and energy price. To capture this dependence, we define the statistics of the sequence $[\mathbf{L}(t)]_{t=1,\ldots}$ as in Eq. (1). The production of renewable energy is function of the weather conditions. In particular, we define $\Phi(\mathbf{r}|\mathbf{w}){=}P(\mathbf{R}(t){=}\mathbf{r}|\mathbf{W}(t) = \mathbf{w})$. The fossil fuel-based energy production and energy price are functions of the state of the system. We define $\Psi(\mathbf{f}|\mathbf{s}){=}P(\mathbf{F}(t)){=}\mathbf{f}|\mathbf{S}(t)){=}\mathbf{s})$ and $\Omega(p|\mathbf{s}){=}P(P(t){=}p|\mathbf{S}(t){=}\mathbf{s})$. The overall state of the system is $\mathbf{S}(t){=}(\mathbf{C}(t), \mathbf{R}(t), \mathbf{F}(t), \mathbf{W}(t), \mathbf{L}(t), P(t)){\in}\mathcal{S}$. The transition probabilities $p(\mathbf{s}', \mathbf{s}){=}P(\mathbf{S}(t{+}1){=}s', \mathbf{S}(t){=}s)$ are estimated based on weather forecast, technical specifics and past data sent by the smart meters and automated management systems.

The cost function $\rho(\mathbf{s})$, $\forall\mathbf{s}{\in}\mathcal{S}$, expresses the *healthiness* of the system in state $\mathbf{s}$. For instance, imbalanced production and load is undesirable and is assigned a high cost. The expected discounted long-term cost (value function) [6] from state $\mathbf{S}(t)$ is given by

$$V(\mathbf{S}(t)) = E\big[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}\rho(\mathbf{S}(\tau))\big],\ \forall\mathbf{S}(t){\in}\mathcal{S}, \tag{2}$$

where $\gamma{\in}(0, 1)$ is the discount factor and $E[\cdot]$ denotes expectation]. V is independent of t, and we can fix $t{=}0$. The value function is intimately connected with the transition probability matrix of the Markov chain. We have

$$V(\mathbf{s}) = \rho(\mathbf{s}){+}\sum_{\tau=1}^{\infty}\sum_{\mathbf{s}'\in\mathcal{S}}\gamma^{\tau}p(\mathbf{s},\mathbf{s}')^{\tau}\rho(\mathbf{S}(\tau)). \tag{3}$$

The function V is the fixed point solution of the system of equations

$$V(\mathbf{s}) = \rho(\mathbf{s}) + \sum_{\mathbf{s}'\in\mathcal{S}}\gamma\ p(\mathbf{s},\mathbf{s}')V(\mathbf{s}'),\ \forall\mathbf{s}{\in}\mathcal{S}. \tag{4}$$

The functions Γ, Ψ and Ω influence the transition probability of the system, and thus, the value function. These functions are designed to ensure that the probability that states associated with the normal functioning of the grid high cost states are reached is small (*i.e.*, small values of the value function in the former class of states).

The value function V is computed based on the estimated statistics and functions defining the behavior of the system and is periodically recomputed. In order to detect anomalies in the operations of the SmartGrid, the control stations observe the state of the grid and *estimate the value function $\tilde{V}$ associated with the current operations of the grid*. Significant differences between the predicted and the actual value functions correspond to differences between the predicted and the real functioning of the grid, which may result in instability and failure. In particular, a larger value of the value function associated with states in which the grid normally operates means that the stochastic process will hit the future high cost states with a larger probability than that predicted. However, due to the large size of the state space of the system, an accurate estimation of the value function requires an extremely large number of observations. More importantly, traditional estimation techniques (*e.g.*, Reinforcement Learning, *etc.*) requires the process to hit all the states multiple times, including those potentially leading to unstable behavior.

III. Estimation and Detection Algorithm

Our critical observation is that *sparsity may also characterize the temporal-evolution of technology-driven systems, such as the SG* and functions of the statistics of the stochastic process modeling the behavior of the system find a concise representation if projected onto a proper basis. The Markov chain modeling the evolution of the logical state of the SG is represented as a graph, where vertices are the states in $\mathcal{S}$ and edges are state transitions with non-zero probability. The graph has a *regular* structure, which can be exploited to *compress* functions defined on the graph. By regular, we mean that the connectivity structure of a vertex is likely be similar to that of many others. As a consequence, the connectivity structure of the overall graph, and the distribution of the future state from each state, can be represented using a number of functions much smaller than the size of the state space. For instance, an increase of the solar irradiation will result in an increase of the energy generated by photovoltaic panels from all the states. Moreover, it is reasonable to assume that there exists classes of consumers with a similar behavior where the number of classes is much smaller than the number of total consumers. Analogously, renewable-based production sites may have a similar reaction to weather changes. In [9], we proved that compressed sensing-like algorithm can be effectively applied for the estimation of value functions in wireless network. We now extend this framework to model and anomaly detection in SG systems.

We use a linear approximation $\mathbf{V}{=}\mathbf{D}\mathbf{x}$ of the value function,[1] where $\mathbf{V}$ is a vector collecting the values $V(\mathbf{s})$, $\mathbf{x}$ is a vector of coefficients and $\mathbf{D}$ is the projection basis. The optimal vector $\mathbf{x}^*$ can be computed as the solution of

$$\mathbf{x}^* = \min_{\mathbf{x}} \|\boldsymbol{\rho}{-}(\mathbf{I}{-}\gamma\mathbf{P})\mathbf{D}\mathbf{x}\|_2^2, \tag{5}$$

where $\boldsymbol{\rho}$, $\mathbf{P}$ and $\mathbf{I}$ are the matrices collecting the values $\rho(s)$, the transition matrix and the identity matrix, respectively. We use a Diffusion Wavelets-based (DW) [11] set of basis functions as projection matrix $\mathbf{D}$. DW captures similarities of the operator $\mathbf{P}$ at different time scales (number of hops in the logical graph) and different locations (vertices of the graph). Due to the intimate connection between the value function and powers of the transition matrix, and the considerations on the connectivity of the logical graph mentioned before, DWs appear to be an excellent candidate for a *sparsifying* basis for the value function $\mathbf{V}$. The procedure to derive $\mathbf{D}$ from $\mathbf{P}$ is described in [11].

The projection onto the diffusion wavelet domain can accelerate the learning of the value function $\tilde{\mathbf{V}}$ from a sample path of state observations. In particular, the observed sample path $[(S(t)]_{t=0,\dots,T}$ is used to estimate $\tilde{\mathbf{P}}$.[2] The computation of the basis set $\mathbf{W}$ is based on the predicted behavior of the SG. However, anomalous behavior may result into different transition probability matrices, and thus, different matrices $\mathbf{D}$. Thus, as a first step, the algorithm detects the representation basis most suited to the trajectory in a set $\{\mathbf{D}_i\}_{i=1,\dots,I}$ by using sparse group Least Angle Selection and Shrinkage Operator (LASSO) [8]. Sparse group LASSO solves the optimization problem

$$\mathbf{y}^*(t) = \arg\min_{\mathbf{y}} \|\boldsymbol{\rho} - (\mathbf{I} - \gamma\tilde{\mathbf{P}}(t))\mathbf{D}\mathbf{y}\|_2^2 + \lambda_1 \sum_{i=1}^{I} \|\mathbf{y}_i\|_2 + \lambda_2\|\mathbf{y}\|_1, \tag{6}$$

where $\mathbf{D}{=}[\mathbf{D}_1,\dots,\mathbf{D}_I]$ and $\mathbf{y} = [\mathbf{y}_1,\dots,\mathbf{y}_I]$. Sparse Group LASSO promotes the concentration of non-zero coefficients in a small sub-set of the groups $\{y_i\}_{i=1,\dots,I}$ (group sparsity) via the regularization term $\sum_{i=1}^{I}\|\mathbf{y}_i\|_2$. Moreover, sparsity of the solution within the individual group is promoted by the regularization term $\|\mathbf{y}\|_1$. We select the basis set (model) $\mathbf{D}_{\mathbf{i}^*}$ associated with the $\mathbf{y}_i^*(T)$ with the highest energy $i^* = \arg\max_i \|\mathbf{y}_i^*\|_2^2$, where T is a pre-defined training time.

For the selected $\mathbf{D}_i$, the LASSO algorithm [12] can be then used to estimate the value function as

$$\mathbf{x}^*(t) = \arg\min_{\mathbf{x}} \|\boldsymbol{\rho} - \mathbf{A}(t)\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{7}$$

where $\mathbf{A}(t) = (\mathbf{I} - \gamma\tilde{\mathbf{P}}(t))\mathbf{D}_i$ and $t{>}T$. In the compressed sensing literature, the matrices $\mathbf{B}(t){=}(\mathbf{I}{-}\gamma\tilde{\mathbf{P}}(t))$ and $\mathbf{D}$ are generally referred to as the *sensing matrix* and the *representation matrix*, respectively.

Different from traditional learning approaches, at every new observation, the whole value function is updated. Intuitively, due to the projection of the DW subspace, an updated transition probability from one state to another updates the recovered structured in every "similar" transition, and at all the time scales involved. Note that the sensing matrix $\tilde{\mathbf{B}}(t){=}(\mathbf{I}{-}\gamma\tilde{\tilde{P}}(t))$ is a function of time and affected by estimation noise. In order to stabilize the solution $\tilde{\mathbf{x}}^*(t)$ over time, we adopt the LS CS algorithm proposed in [13]. At time t, LS CS makes use of the previous solution $\tilde{\mathbf{x}}^*(t{-}1)$ and the current estimated transition matrix $\tilde{\mathbf{P}}(t)$ to compute the new vector $\tilde{\mathbf{x}}^*(t)$. For a detailed description of LS CS interested readers are referred to [13]. The estimated vector $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{V}}{=}\mathbf{D}\tilde{\mathbf{x}}$ can be used to detect anomalies. Differences between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{V}}{=}\mathbf{D}\tilde{\mathbf{x}}$ and $\mathbf{x}$ and $\mathbf{V}$ indicate an anomalous functioning of the grid. An anomaly is detected if $\|\mathbf{V}{-}\tilde{\mathbf{V}}\|_2{>}\tau$, where τ is a predefined threshold.

IV. Numerical Results

We applied the proposed algorithm to a toy scenario with 360 states. The overall chain is composed by 5 sub-chains modeling weather conditions, energy price, consumer state, energy consumption and fossil fuel production.[3] The temporal evolution of the weather and consumer behavior are modeled as random walks. Referring to the indexing $\mathcal{W}{=}\{1,2,\dots,N_w\}$, we define the probabilities $p_w^+{=}p_w(j{+}1|j)$, $p_w^-{=}p_w(j{-}1|j)$ and $p_w^={=}p_w(j|j)$, with $p_w^+{+}p_w^-{+}p_w^={=}1$. Analogously, referring to the indexing $\mathcal{C}{=}\{1,2,\dots,N_c\}$, we define the probabilities $p_c^+{=}p_c(j{+}1|j)$, $p_c^-{=}p_c(j{-}1|j)$ and $p_c^={=}p_c(j|j)$, with $p_c^+{+}p_c^-{+}p_c^={=}1$. Small and large values of $W(t)$ are associated with low and high renewable energy production. The probabilistic maps Γ, Φ and Ω are not reported here due to space constraints. The consumption function Γ is defined such that high energy price drives consumption down. The pricing function Φ sets a high price if consumption is larger than production, whereas a small price is set to encourage consumption in the specular case. The fossil production function Φ is designed to slowly compensate imbalanced energy consumption and production. We use a simple cost function $\rho(S(t)){=}|F(t){+}R(t){-}L(t)|$ that penalizes imbalanced energy production-load. We remark that the model can be easily extended to more complex cases and behaviors.

[1] Linear approximations have been widely used in the literature, see [10].

[2] The elements of $\tilde{\mathbf{P}}$ are obtained by counting the occurrences of a transition from $\mathbf{s}$ to $\mathbf{s}'$ and dividing by the occurrences of $\mathbf{s}$.

[3] Renewable energy production is assumed to be a deterministic function of the weather and is not included in the state space.

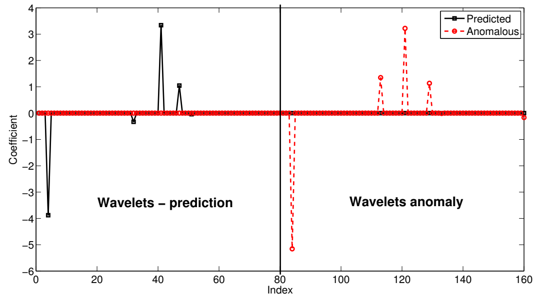

Fig. 1. $\mathbf{y}^*$ for the predicted and anomalous system. The non-zero coefficients concentrate in $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively.

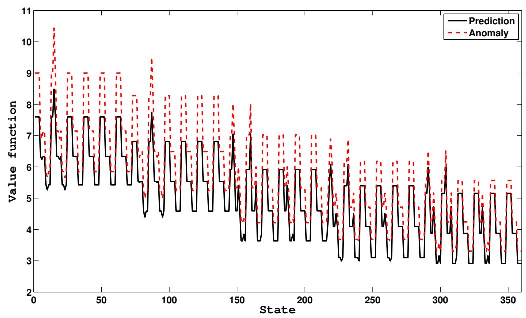

Fig. 2. Value function for the predicted and anomalous system.

We first show an example of model detection via Sparse Group LASSO. Fig. 1 shows the estimated $\mathbf{y}^*$ where $\mathbf{D}$=$[D_1, D_2]$ and $[D_1$ and D_2 correspond to DW computed for the predicted system and a system where a defective pricing and production strategy is leading to instability. It can be observed how the coefficients concentrate in the different portions of $\mathbf{y}^*$ in the normal and anomalous operating regime. Thus, Sparse Group LASSO enables the detection of the correct DW basis set for the estimation of the value function, as well as the detection of anomalies at the structure level of the transition matrix.

We now focus on the online learning of the value function and on the detection of anomalies in the value of the transition probabilities once the model has been detected. We define $\mathbf{V}_{an}$ and $\tilde{\mathbf{V}}_{an}$ as the real and estimated value function. If the system is running in an anomalous fashion, the transition probability matrix and/or the pricing and consumption function are different than those generating $\mathbf{V}$. We define Δ_{an}=$\|\mathbf{V}-\mathbf{V}_{an}\|_2$, $\tilde{\Delta}$=$\|\mathbf{V}-\tilde{\mathbf{V}}\|_2$ and $\tilde{\Delta}_{an}$=$\|\mathbf{V}-\tilde{\mathbf{V}}_{an}\|_2$. Thus, we expect $\tilde{\Delta}$ and $\tilde{\Delta}_{an}$ to converge to 0 and Δ_{an} as the estimate of the transition matrix, respectively.

Fig. 2 shows the value function associated with the predicted and anomalous system. The performance of the proposed algorithm is compared with that achieved if Q-learning is used to estimate the actual value function of the system [6]. In Fig. 3 shows Δ,$\tilde{\Delta}$ and $\tilde{\Delta}_{an}$ as a function of time for the proposed algorithm and traditional Q-learning. *The proposed algorithm shows an impressive convergence rate of the estimate of the value function*. In fact, based on a number of observations comparable to the number of states, $\tilde{\Delta}$ and $\tilde{\Delta}_{an}$ converge to the wanted values. Q-learning requires a large number of observations to accurately estimate the actual value function.

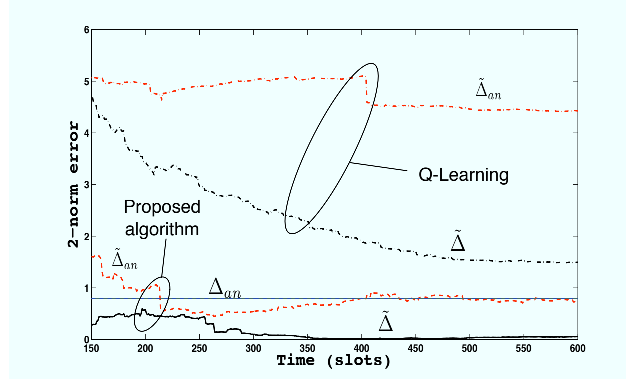

Fig. 3. Δ,$\tilde{\Delta}$ and $\tilde{\Delta}_{an}$ as a function of time. In the predicted model we have set p_w^-=0.2, p_c^-=0.4, p_w^-=0.5 and p_c^-=0.3. In the anomalous system, these probabilities are set to p_w^-=0.5, p_c^-=0.3, p_w^+=0.3, p_c^+=0.2. Thus, the anomalous system has a smaller production of renewable and a larger energy consumption on average.

Thus, after a short training period the difference between the predicted and estimated value function can be measured to detect anomalies in the behavior of the grid. The value of the threshold τ realizes a tradeoff between false alarms and correct detection rate.

V. Conclusions

A novel algorithm for the detection of anomalies in the behavior of SmartGrid systems was presented. The algorithm uses wavelet projection and a sparse approximation technique to estimate the value function of the actual system and compare it with the value function computed based on prediction. Numerical results show that the proposed algorithm detects the anomaly from a very small number of observations.

References

[1] C. Shuyong, S. Shufang, L. Lanxin, and S. Jie, "Survey on smart grid technology," *Power Sys. Tech.*, vol. 33, no. 8, pp. 1–7, 2009.

[2] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, "Residential Demand Response Using Reinforcement Learning," in *IEEE SmartGridComm*. IEEE, 2010, pp. 409–414.

[3] G. Chicco and P. Mancarella, "Distributed multi-generation: a comprehensive view," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 3, pp. 535–551, 2009.

[4] V. Calderaro, C. Hadjicostis, A. Piccolo, and P. Siano, "Failure identification in smart grids based on petri net modeling," *Industrial Electronics, IEEE Transactions on*, no. 99, pp. 1–1, 2011.

[5] R. Berthier, W. Sanders, and H. Khurana, "Intrusion detection for advanced metering infrastructures: Requirements and architectural directions," in *IEEE SmartGridComm*. IEEE, 2010, pp. 350–355.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Belmont, MA: Athena Scientific, 2001, vol. 2.

[7] M. Crovella and E. Kolaczyk, "Graph wavelets for spatial traffic analysis," in *INFOCOM 2003.*, vol. 3. IEEE, 2003, pp. 1848–1857.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *Arxiv preprint arXiv:1001.0736*, 2010.

[9] M. Levorato, U. Mitra, and A. Goldsmith, "Structure-based learning in wireless networks via sparse approximation," *submitted to the EURASIP Journal on Wireless Communications and Networking, available online: http://www-bcf.usc.edu/ levorato*, 2011.

[10] J. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. on Aut. Control*, vol. 42, no. 5, pp. 674–690, 1997.

[11] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and Comp. Harmonic Analysis*, vol. 21, no. 1, pp. 53 – 94, 2006.

[12] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," J. of the Royal Stat. Society., vol. 58, no. 1, pp. 267–288, 1996.

[13] N. Vaswani, "LS-CS-residual (LS-CS): Compressive Sensing on least Squares Residual," *IEEE Transactions on Signal Proc.*, vol. 58, no. 8, pp. 4108 – 4120, Aug. 2010.