2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Controlling big, diverse, nonlinear load aggregations for grid services by adjusting device setpoints

Kevin J. Kircher, *Member, IEEE,* Yuan Cai, Leslie K. Norford, and Steven B. Leeb, *Fellow, IEEE*

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683195

Abstract— **Many electrical loads seek to maintain a measurement, such as a temperature, pressure, flow rate, fluid level or charge state, near a setpoint. In some cases, setpoints can be adjusted slightly without noticeably affecting quality of service. Small setpoint adjustments have an indirect effect on power use that, when aggregated over a large number of loads, can be significant. This paper develops a framework to provide services to the power grid by adjusting device setpoints. The framework has several practical advantages: it scales to very large load aggregations; accommodates a wide variety of loads, including those with nonlinear behavior; and requires little sensing or communication and no private information. The framework involves (1) learning a model to predict aggregate power under baseline operation, (2) exciting the system to identify a model relating setpoint perturbations to aggregate power perturbations, and (3) embedding baseline predictions and the perturbation model in load-shifting optimization. Simulations of a 50,000-load, 115-MW aggregation in the Texas storms of February, 2021, suggest that this framework can reduce peak demand, arbitrage dynamic energy prices or carbon intensities, and provide utility demand response or wholesale ancillary services.**

I. Background and Motivations

Between 2008 and 2018, global electricity production from the wind and sun grew by factors of 5.7 and 47.3, respectively [1]. As power systems integrate more of these variable, uncertain renewables, the need for grid-balancing services increases [2]. Generators have traditionally provided these services, but aggregations of controllable loads also can [3]. Recent research has demonstrated the potential of a wide variety of loads to provide a similarly wide variety of services. For a few examples, load aggregations can respond to dynamic energy prices [4], [5], provide operating reserve [6], [7], [8], limit peak demand [9], [10], [11], and curtail load for distribution-level demand response [9], [10], [12]. Commonly-studied loads include electric vehicles [4], [9]; air conditioners, heat pumps, refrigerators and water heaters in residential buildings [5], [6], [11]; heating, ventilation and air conditioning (HVAC) systems in commercial buildings [7], [10]; and water pumps [8], [12].

This paper develops a unifying control framework that incorporates all of the loads and services mentioned above. The scope includes any load that normally maintains a measurement near a setpoint, but can tolerate small changes to that setpoint. In this framework, illustrated in Fig. 1, a central controller sends a normalized setpoint perturbation signal to all controllable loads. Each load scales this signal according to its current flexibility, then adjusts its setpoint accordingly. This indirectly affects each load's power use, causing an aggregate response that can be shaped to provide grid services.

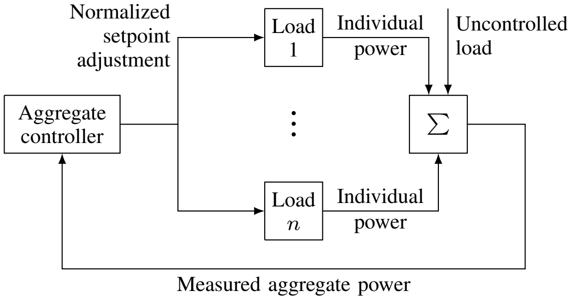

Fig. 1. One normalized setpoint adjustment is sent to all controlled loads. Only the aggregate power, which includes uncontrolled loads, is measured.

The idea of adjusting device setpoints to shape aggregate load is not new. For example, one research thread has demonstrated the potential to provide distribution-level demand response [13] and frequency regulation [14], [15] by adjusting room temperature setpoints in commercial buildings. That approach is generalized here to accommodate a wider variety of loads and services.

Other methods exist to provide grid services from big, diverse load aggregations. In [16], Kraning *et al.* used distributed optimization to coordinate thermal loads, appliances, and electric vehicles, as well as generators and energy storage, under network constraints. In [5], [17], virtual battery models were developed to control aggregations of air conditioners, heat pumps, refrigerators and water heaters for frequency regulation and price-based load shifting. Subsequent work extended this approach to commercial building HVAC systems [18], as well as appliances and electric vehicles [19].

Of the methods mentioned above, [16] and [19] accommodate particularly large and diverse load aggregations. However, these methods assume linear load models and exact knowledge of model parameters. A complementary body of work has used reinforcement learning to avoid these restrictive assumptions. Wang and Hong reviewed this literature in [20], including applications to HVAC systems, water heaters, refrigerators, windows, lights, distributed generators, and energy storage. While reinforcement learning can handle diverse and nonlinear loads, its scalability suffers from the curse of dimensionality: computational complexity grows exponentially with the dimensions of the state and action spaces. Tellingly, only 14% of the papers reviewed in [20]

K.J. Kircher (kircher@mit.edu) and S.B. Leeb are with the Department of Electrical Engineering and Computer Science and Y. Cai and L.K. Norford are with the Department of Architecture, all at the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

978-1-6654-3659-5/21/$31.00 ©2021 IEEE

considered more than ten control points.

The approach in this paper has several practical advantages over existing methods. It requires no device-level information or models, so it applies to a wide variety of devices, accommodates nonlinear loads, and protects user privacy. Its sensing, computation and communication requirements are independent of the number of loads, so it scales to very large aggregations and could be deployed at relatively low cost. The only required measurement is the aggregate power, which could include both controlled and uncontrolled loads. This measurement could be read from a single meter on a building, neighborhood, or entire distribution grid; or it could be the sum of many individual meters. At the device level, only a network connection and the ability to adjust a setpoint are needed. Communication requirements are low: network-wide, one number is broadcast per time step.

This paper is organized as follows. §II describes the control framework at a high level. A variety of example loads are discussed in §III. §IV presents simulation results for an aggregation of these loads providing several grid services during the Texas blackouts of February, 2021. §V summarizes the paper and discusses possible extensions.

II. Control Framework

The control framework in this paper involves a discrete time span $\mathcal{K} = \{1, \dots, K\}$, time index $k \in \mathcal{K}$, aggregate power measurements $P(k)$ (kW), and normalized setpoint perturbations $u(k) \in [-1, 1]$. At each time k, an aggregate-level controller decides $u(k)$ and broadcasts it to all loads. Each load scales $u(k)$ in proportion to its current flexibility and adds it to its baseline setpoint. This scaling is done such that $u(k) = 0$ corresponds to baseline operation, $u(k) = \pm 1$ gives the highest or lowest acceptable setpoint perturbation, and all else being equal, each load uses more power when $u(k)$ increases.

For example, a heating load that tolerates deviations of magnitude $\varepsilon(k)$ (°C) from an ideal load temperature $T^\star(k)$ (°C) implements the perturbed setpoint $T^{\text{set}}(k) = T^\star(k) + \varepsilon(k)u(k)$. With this scaling, the setpoint $T^{\text{set}}(k)$ equals its baseline value $T^\star(k)$ when $u(k) = 0$ and reaches its highest or lowest acceptable value when $u(k) = \pm 1$. Power use increases with $u(k)$, as required. Similarly, a cooling load implements $T^{\text{set}}(k) = T^\star(k) - \varepsilon(k)u(k)$, so that the setpoint decreases (power increases) as $u(k)$ increases. The tolerance $\varepsilon(k)$ may vary over both loads and time. A water heater might tolerate ± 4 °C swings at any time, while a heat pump or air conditioner might tolerate ± 1 °C while its building is occupied and ± 2 °C while unoccupied. If a device is unavailable at time k, it simply sets $\varepsilon(k) = 0$.

Each device is assumed to be capable of determining a setpoint perturbation tolerance that respects its operational constraints, such as protecting equipment or maintaining quality of service. Device-level controllers should react stably to setpoint changes within this tolerance. Beyond this, no assumptions are made about loads. Dynamics could be nonlinear, stochastic, coupled across loads, high- or infinite-dimensional, or entirely unknown. Devices could be on/off, multi-stage or variable-speed.

A. Baseline Prediction

The proposed control framework has three phases. In the first phase, a model is learned to predict aggregate power under baseline operation. This is a time-series forecasting problem. The required data are historical aggregate power measurements and predictive features such as weather conditions or the hour, weekday or season. Model options include neural networks, linear time-series models, regression trees, and support vector machines. In [21], Yildiz *et al.* compared all of these methods and found that a feedforward neural network best predicted baseline power. That model structure is used here, with one hidden layer, ten neurons, and the outdoor temperature and hour of day as features.

B. Perturbation System Identification

In the second phase, the system is excited to learn how the aggregate power responds to setpoint perturbations. The system is excited by implementing a sequence of nonzero setpoint perturbations. This drives the measured powers $P(k)$ away from the baseline powers $\hat{P}(k)$ (kW). Data from the excitation phase are then used to train a model that predicts $P(k) - \hat{P}(k)$ based on current and past values of $u(k)$.

Unlike the baseline model, the perturbation model is restricted here to be linear in the setpoint perturbations:

$$P(k) - \hat{P}(k) = \sum_{i=1}^{m} a_i(k)u(k - m + i) + e(k). \qquad (1)$$

The linearity restriction is justified to an extent by Taylor's theorem, which suggests that output (power) perturbations should respond approximately linearly to small input (setpoint) perturbations. Nevertheless, (1) is only an approximation of the true input-output behavior, which is nonlinear in general. Linearity is imposed here for tractability of optimization in the third phase. Unmodeled nonlinearities influence the model errors $e(k)$; the optimization is designed to be robust to these errors.

The structure of (1) is a finite impulse response model. This structure is a standard instrument in system identification [22], although the time-varying coefficients $a_i(k)$ used here are non-standard. Time-varying coefficients are used here to accommodate two unique characteristics of electrical loads. First, load behavior often varies with weather conditions. Air conditioners, for example, are typically turned off in cool weather and therefore unresponsive to setpoint perturbations. Second, load behavior often varies with the time of day. A typical electric vehicle charger in a home, for example, will actively charge overnight but be disconnected during most days.

In the examples in this paper, the time-varying model coefficients $a_i(k)$ have the form

$$a_i(k) = \sum_{j=1}^{p} \beta_{ij} f_{ij}(k).$$

Here the β_{ij} are time-invariant model parameters that are fit to training data. The time-varying features $f_{ij}(k)$ are known functions of the weather conditions and hour of day. With this form of the $a_i(k)$, the sum in (1) becomes

$$\begin{aligned}&\sum_{i=1}^{m} u(k-m+i)\sum_{j=1}^{p}\beta_{ij}f_{ij}(k)\\ &= u(k-m+1)f_1(k)^\top\beta_1+\cdots+u(k)f_m(k)^\top\beta_m\\ &= \begin{bmatrix} x_1(k)^\top & \dots & x_m(k)^\top\end{bmatrix}\begin{bmatrix}\beta_1\\ \vdots\\ \beta_m\end{bmatrix},\end{aligned}$$

with the definitions

$$f_i(k)=\begin{bmatrix}f_{i1}(k)\\ \vdots\\ f_{ip}(k)\end{bmatrix},\ \beta_i=\begin{bmatrix}\beta_{i1}\\ \vdots\\ \beta_{ip}\end{bmatrix}$$
$$x_i(k)=u(k-m+i)f_i(k).$$

Therefore, the full perturbation model (1) can be written in standard linear regression form, $y = X\beta + e$, where

$$y=\begin{bmatrix}P(1)-\hat{P}(1)\\ \vdots\\ P(K)-\hat{P}(K)\end{bmatrix},\ \beta=\begin{bmatrix}\beta_1\\ \vdots\\ \beta_m\end{bmatrix},\ e=\begin{bmatrix}e(1)\\ \vdots\\ e(K)\end{bmatrix}$$
$$X=\begin{bmatrix}x_1(1)^\top & \dots & x_m(1)^\top\\ \vdots & & \vdots\\ x_1(K)^\top & \dots & x_m(K)^\top\end{bmatrix}.$$

The full model (1) is trained offline in a batch fit by forming the data matrix X and target vector y, then computing the least-squares estimate $\hat{\beta} = (X^\top X)^{-1}X^\top y$ of the parameter vector β. The model could also be trained online using a linear Kalman filter, but this is not investigated here.

Predicting the power perturbation $P(k) - \hat{P}(k)$ at some future time k requires evaluating the trained model coefficients

$$\hat{a}_i(k)=\sum_{j=1}^{p}\hat{\beta}_{ij}f_{ij}(k)$$

at that future time. For this reason, the model should only use basic features $f_{ij}(k)$ for which accurate forecasts are available. The examples in this paper use features

$$\begin{aligned}&f_{i1}(k)=1,\ f_{i2}(k)=\max\{0,\theta_0-\theta(k)\},\ f_{i3}(k)=\hat{P}(k)\\ &f_{i4}(k)=\sin(\pi k\Delta t/12),\ f_{i5}(k)=\cos(\pi k\Delta t/12)\\ &f_{i6}(k)=\sin(\pi k\Delta t/6),\ f_{i7}(k)=\cos(\pi k\Delta t/6)\end{aligned}$$

for all i. Forecasts of the outdoor temperature θ (°C) are readily available, as is the output $\hat{P}(k)$ of the baseline prediction model from §II-A. The hyperparameter θ_0 (°C) is hand-tuned.

To train the model (1) efficiently, the excitation signal should maximize the information content of the aggregate power response. This paper uses random binary input sequences, where each $u(k)$ is drawn independently from a discrete uniform distribution on $\{-1, 1\}$. These sequences maximize information content for input-constrained systems [22]. The hyperparameter m in (1), which determines the model's memory, is hand-tuned.

C. Load-Shifting Optimization

In the third phase, the baseline predictions $\hat{P}(k)$ and the perturbation model (1) are embedded in load-shifting optimization. The general problem is to decide setpoint perturbations $u(1), \dots, u(K)$ and aggregate powers $P(1), \dots, P(K)$ to

$$\begin{array}{ll}\text{minimize} & \mathcal{R}(f(u(1),\dots,u(K),P(1),\dots,P(K)))\\ \text{subject to} & P(k)=\hat{P}(k)+\sum_{i=1}^{m}a_i(k)u(k-m+i)\\ & \quad +e(k),\ k\in\mathcal{K}\\ & g(u(1),\dots,u(K))\preceq 0.\end{array}\tag{2}$$

In (2), f is the objective function and $\mathcal{R}$ is a risk measure, such as the expected value or the worst-case value, taken over the joint distribution of all uncertain inputs. The symbol '$\preceq$' denotes component-wise inequality. The vector-valued function g encodes the constraints $|u(k)| \leq 1$ and possibly others. For example, $\sum_{k\in\mathcal{K}} u(k) = 0$ ensures that perturbations are zero-mean, and $\underline{\delta} \leq u(k)-u(k-1) \leq \overline{\delta}$ constrains setpoint ramp rates. The deterministic input data are the past perturbations $u(2 - m), \dots, u(0)$ and the aggregate power baseline $\hat{P}(1), \dots, \hat{P}(K)$. The coefficients $a_i(k)$ and disturbance $e(k)$ may be random, so (2) is a stochastic optimization problem in general.

If the objective and constraint functions f and g are convex and the risk measure $\mathcal{R}$ is convex and nondecreasing, then (2) is a convex optimization problem. If, additionally, the joint distribution of all uncertain inputs is known and $\mathcal{R}$ can be evaluated exactly, then (2) can be reduced to a deterministic convex problem and solved to global optimality in polynomial time using interior-point methods [23]. If some distributional information is unknown or $\mathcal{R}$ cannot be evaluated exactly, then sample-based methods such as sample-average approximation [24], [25] or scenario convex optimization [26] can be used. The optimization can also be implemented in a model predictive control (MPC) framework. In MPC, at each time step the uncertain inputs are predicted over a receding planning horizon, a version of (2) is solved, and only the first resulting setpoint perturbation is implemented. The system then evolves and the process repeats. By continually updating forecasts and state estimates, MPC can improve performance over static optimization.

Examples in §IV of this paper use scenario convex optimization. In this approach, S samples are generated independently from the joint distribution of all uncertain inputs. These samples are indexed by $s \in \mathcal{S} = \{1, \dots, S\}$. The risk measure $\mathcal{R}$, which in these examples is the worst-case value, is approximated by the sample-wise maximum. The resulting problem is to

$$\begin{array}{ll}\text{minimize} & \max_{s\in\mathcal{S}} f(u(1),\dots,u(K),P_s(1),\dots,P_s(K))\\ \text{subject to} & P_s(k)=\hat{P}(k)+\sum_{i=1}^{m}a_{i,s}(k)u(k-m+i)\\ & \quad +\ e_s(k),\ k\in\mathcal{K},\ s\in\mathcal{S}\\ & g(u(1),\dots,u(K))\preceq 0.\end{array}\tag{3}$$

Probabilistic optimality guarantees based on the sample size S can be obtained from [26].

III. Example Loads

This section models heat pumps, resistance space heaters, air conditioners, heat-pump water heaters, resistance water heaters, refrigerators, water pumps, and electric vehicles. The models and parameters in this section are not used in the control framework, but are presented here to demonstrate the framework's broad applicability. They are also used in simulations in §IV.

Each device has electric power constraints of the form

$$\underline{P}_\ell \leq P_\ell(k) \leq \overline{P}_\ell,$$

where ℓ indexes devices and $\underline{P}_\ell$ and $\overline{P}_\ell$ (kW) are the device's power limits. Each device seeks to track a user-specified setpoint that is time-varying in general. The power required to perfectly track this setpoint is denoted $P_\ell^\star(k)$. In the simulations in §IV, $P_\ell^\star(k)$ is calculated by discretizing the load's continuous-time dynamics (presented in the subsections below), then solving the discrete-time dynamics for the power required to drive the load from its current state to the desired setpoint at the next time step. During baseline operations, each device is assumed to either perfectly track its setpoint (using power $P_\ell^\star(k)$), or to saturate at a capacity limit:

$$P_\ell(k) = \max\left\{\underline{P}_\ell, \min\left\{\overline{P}_\ell, P_\ell^\star(k)\right\}\right\}.$$

These saturation nonlinearities make all of the load models in this section nonlinear. Several of the models have additional nonlinearities.

A. Heat Pumps and Air Conditioners

Conditioned spaces are modeled here as second-order thermal circuits with two states: the temperatures T and T_m (°C) of the indoor air and the building's lumped thermal mass. The air temperature dynamics are

$$C\dot{T}(t) = \frac{\theta(t) - T(t)}{R} + \frac{T_m(t) - T(t)}{R_m} + \dot{Q}_c(t) + \dot{Q}_e(t).$$

The input signals are the outdoor air temperature θ (°C), thermal power $\dot{Q}_c$ (kW) from controlled heating or cooling equipment, and exogenous thermal power $\dot{Q}_e$ (kW) from the sun, lights, plug loads and body heat. The parameters are the thermal capacitances C and C_m (kWh/°C) of the indoor air and thermal mass, the thermal resistance R (°C/kW) between the indoor and outdoor air, and the resistance R_m (°C/kW) between the indoor air and mass. The mass dynamics are

$$C_m\dot{T}_m(t) = \frac{T(t) - T_m(t)}{R_m}.$$

Each heat pump and air conditioner is modeled through its coefficient of performance (COP) η (-), which depends in general on indoor and outdoor temperatures and device loading. Heat pumps and air conditioners use power $P = \pm\dot{Q}_c/\eta$, with the plus sign for heating and minus for cooling. Resistance heaters are modeled as heat pumps with $\eta = 1$.

This paper simulates heat pumps with COPs of 2.5–3.5 serving single-family homes with 170–200 m^2 of floor area. The capacitance C is set by multiplying the floor area by a ceiling height of 2.9–3.1 m to calculate the indoor air volume. This volume is multiplied by the density and specific heat of air, then multiplied by 2–4 to account for other, tightly-coupled material. The resistance R is tuned to a 14–21 kW thermal load under steady design conditions of -10 °C outdoor, 19–25 °C indoor, and 4–7 W/m^2 from plug loads and body heat. The capacity $\overline{P}$ is oversized relative to design load by 50–75%, resulting in 4–8 kW. The ratios C_m/C and R_m/R are 10–12 and 5–7, based on the fits in [27].

At each time step, the solar thermal power is set by multiplying the global solar irradiance on a horizontal surface (in W/m^2) by the floor area, then rescaling to peak at 2–4 kW. Thermal powers from plug loads, lights and body heat at each time step are the products of floor area and the respective intensities of 0.5–2, 1.5–5 and 0.5–2 W/m^2. Half of the buildings have constant 19–25 °C air temperature setpoints. Their setpoint perturbation tolerances are ±1 °C at all hours. The other half of the buildings have setpoints in the same range during most hours, but 2–4 °C lower overnight. Their tolerances are ±1 °C normally and ±2 °C overnight.

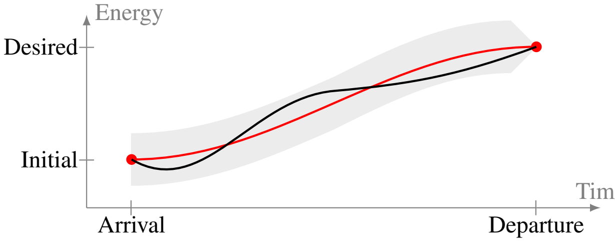

Fig. 2. Baseline (red) and perturbed (black) energy setpoint trajectories for a charging electric vehicle. The gray area contains all acceptable trajectories.

B. Electric Vehicles

Electric vehicle battery dynamics are modeled here as

$$\dot{E}(t) = -rE(t) + P_c(t).$$

The state E (kWh) is the energy stored in the battery, r (1/h) is the dissipation rate, and P_c (kW) is the chemical charging power. The electric power exchanged with the grid is

$$P(t) = \begin{cases} P_c(t)/\eta_+ & \text{if } P_c(t) \geq 0 \text{ (charging)} \\ \eta_- P_c(t) & \text{if } P_c(t) < 0 \text{ (discharging)}, \end{cases}$$

where η_+ and η_- (-) are the charging and discharging efficiencies. The piecewise electric power curve makes this model nonlinear. The battery has energy capacity $\overline{E}$ (kWh) and charging and discharging power limits $\underline{P}$ and $\overline{P}$, with $\underline{P}$ negative if vehicle-to-grid discharging is allowed. Each day, the vehicle arrives at time t_a with energy E_a and desires energy E_d by its departure time t_d. Between arrival and departure, the battery seeks to track an energy setpoint trajectory from E_a to E_d, as illustrated in Fig. 2. This trajectory could be decided by the vehicle's on-board software to optimize charging efficiency, battery health or other criteria. The tolerance for perturbations about the setpoint trajectory may vary with time, and shrinks to zero at t_d to ensure that the vehicle departs with its desired charge.

The batteries simulated here dissipate 1–3% of charge in 24 hours. Energy capacities are 60–80 kWh. Charging and discharging capacities are 7–11.5 and 2.5–3.5 kW. Efficiencies are 85–95%, arrival times are 4–8 PM, and departures are 6–9 AM. Vehicles arrive with batteries 20–40% full and seek to depart 70–90% full. Energy setpoint trajectories are linear, with setpoint perturbation tolerance

$$\varepsilon(t) = \begin{cases} \lambda \min\left\{E_a, \overline{E} - E_a\right\} & \text{if } t \in [t_a, t_d) \\ 0 & \text{otherwise.} \end{cases}$$

The tunable parameter $\lambda \in [0, 1]$ is set here to 0.25.

C. Water Heaters and Refrigerators

Water heaters and refrigerators are modeled here as first-order thermal circuits with dynamics

$$C\dot{T}(t) = \frac{\theta(t) - T(t)}{R} + \dot{Q}_c(t) + \dot{Q}_e(t).$$

For both devices, θ is the temperature of the surrounding air and $\dot{Q}_c$ is the controlled thermal power. For water heaters, T is the water temperature, C is the water capacitance, R is the resistance between the water and surrounding air, and $\dot{Q}_e$ is the thermal power from water withdrawals. For refrigerators, T is the inside temperature, C is the inside capacitance, R is the resistance between the inside and surrounding air, and $\dot{Q}_e$ is the thermal power from door-openings, food additions, *etc*. Water heaters and refrigerators use electric power $P = \pm\dot{Q}_c/\eta \in [0, \overline{P}]$ (kW), with the plus sign for water heaters and minus for refrigerators. Resistance water heaters have $\eta = 1$; for heat-pump water heaters, η is significantly higher.

The water heaters simulated here have cylindrical tanks with 0.19–0.38 m^3 volumes and 0.3 m radii. The capacitance is the product of the density, specific heat and volume of water. The resistance is set by dividing an R-value of 6–8 °F·ft^2/BTU/h (1060–1410 °C·m^2/kW) by the tank's vertical surface area. Heat-pump water heaters, 25% of the units, have COPs of 2–2.5. The other 75% are resistance units with $\eta = 1$. Thermal capacities are 4–5 kW. Water temperature setpoints are 43–54 °C and perturbation tolerances are 4 °C. Thermal powers of 0.25–1.3 kW from water use occur during morning and evening busy periods, which are randomized over units. The mean withdrawal is 5.4 kWh per day.

The refrigerators simulated here have $C = 0.4$–0.8 kWh/°C and $R = 80$–100 °C/kW. Electrical capacities are 0.7–0.85 kW and COPs are 1.5–2.5. Door-openings and food additions during busy morning and evening periods, which are randomized over units, cause thermal power injections of 25–35% of thermal capacity. Temperature setpoints are 2–3.5 °C and perturbation tolerances are 1 °C.

D. Water Tanks and Pumps

Water storage tank dynamics are modeled here as

$$\dot{h}(t) = \frac{q_c(t) - q_e(t)}{A}.$$

The state $h \in [0, \overline{h}]$ (m) is the water level in the tank, A (m^2) is the tank's cross-sectional area, q_c (m^3/s) is the inflow from a controlled pump, and q_e (m^3/s) is the outflow from uncontrolled water withdrawals. The pump uses power

$$P(t) = \frac{\rho g h(t) q_c(t)}{\eta} \in [0, \overline{P}],$$

where $\rho = 997$ kg/m^3 is the density of water, $g = 9.8 \times 10^{-3}$ km/s^2 is the acceleration of gravity and η (-) is the pump efficiency. The $h(t)q_c(t)$ product makes this model nonlinear.

The water tanks simulated here are cylinders with 5-6 m radii and 2.2–2.6 m heights. At each time step, the outflow is 0–100% of the worst-case outflow, which absent pumping would drain the tank in 4.5–5.5 hours. Design inflows completely refill tanks in four hours under the worst-case outflow. Pumps are 20–40% oversized relative to design inflow. Pump efficiencies are 70–90%. Each pump seeks to maintain the water level at a setpoint of 95% of the tank height and tolerates ±5% perturbations about this level.

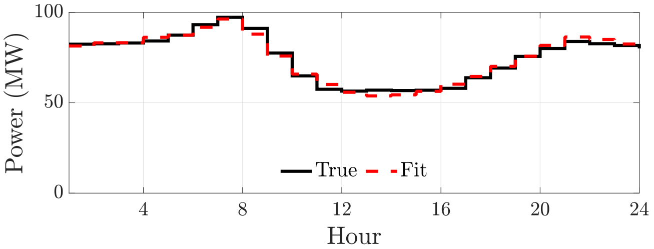

Fig. 3. True (black) and predicted (red) aggregate power under baseline operation on a typical validation day.

IV. Simulations

The simulations in this section are set in winter in Texas. They include 10,000 heat pumps, water heaters, refrigerators, water pumps and electric vehicles, for 50,000 loads total. The time step Δt is one hour. Within each load class, model parameters vary from load to load to emulate diversity in device ages, sizes, efficiencies, usage patterns, *etc.* Parameters are drawn independently and uniformly from the ranges in §III. Simulations use Austin weather data for 2012 and 2015 (the historical years available at [28]) and the extreme storm from February 12–20, 2021, that caused widespread power outages. The 2012 and 2015 data are used for training and validation, respectively. Results are presented for the 2021 storm. The monetary values in these results are highly atypical, as energy and ancillary service prices spiked during the storm, but they illustrate the value that a load aggregation could provide during emergencies.

Under baseline operation, the aggregate load averages 77 MW and peaks at 115 MW. Fig. 3 shows the true and predicted baseline power on a typical validation day. The baseline model's training and validation R^2 values are 0.98 and 0.97. To fit the perturbation model, the system is excited for 30 days using random binary setpoint perturbations. Fig. 4 shows the true and predicted power perturbations on a typical validation day. The perturbation model's training and validation R^2 values are 0.95 and 0.92. The perturbation model is significantly less accurate than the baseline model, but still proves useful for optimization.

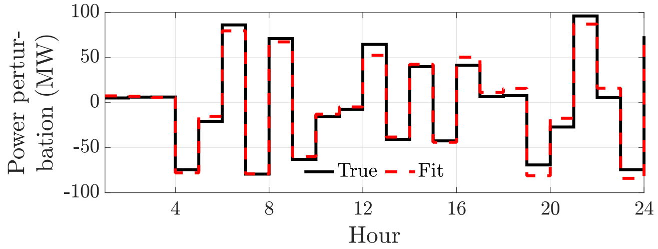

Fig. 4. True (black) and predicted (red) aggregate power perturbation under excited operation on a typical validation day.

A. Minimizing Peak Demand

In this example, the goal is to minimize the peak aggregate power used in a day. An accurate weather forecast is assumed to be available, so all uncertainty lies in the perturbation model errors $e(k)$. These errors are treated as independent zero-mean Gaussian random variables with the validation error standard deviation of 5.9 MW. The optimization uses a worst-case risk measure over $S = 100$ samples:

$$\begin{aligned} \text{minimize}\quad & \max_{s\in\mathcal{S}}\max_{k\in\mathcal{K}} P_s(k) \\ \text{subject to}\quad & P_s(k)=\hat{P}(k)+\textstyle\sum_{i=1}^{m} a_i(k)u(k-m+i) \\ & \quad + b(k)+e_s(k),\ \ k\in\mathcal{K},\ \ s\in\mathcal{S} \\ & |u(k)|\le 1,\ \ k\in\mathcal{K} \\ & \underline{\delta}\le u(k)-u(k-1)\le\overline{\delta},\ \ k\in\mathcal{K} \\ & \textstyle\sum_{k\in\mathcal{K}} u(k)=0. \end{aligned} \tag{4}$$

Fig. 5 shows optimization results for February 20, 2021. In the top plot, the red curve is the baseline aggregate power prediction. The black curve is the actual aggregate power, calculated by simulating the true system dynamics under the optimal setpoint perturbations from (4). Each gray curve is a scenario of what the aggregate power might have been under the optimal setpoint perturbations and a different sample from the model error distribution. The true power (black curve) stays within the gray 'cloud' of power scenarios, suggesting that the optimization is robust to model errors. The bottom plot shows the optimal setpoint perturbations. Peak demand is reduced by 'discharging' ($u < 0$) the aggregation during the morning peak and 'charging' ($u > 0$) after. In this simulation, peak demand is reduced by 16%, from 115 to 97 MW.

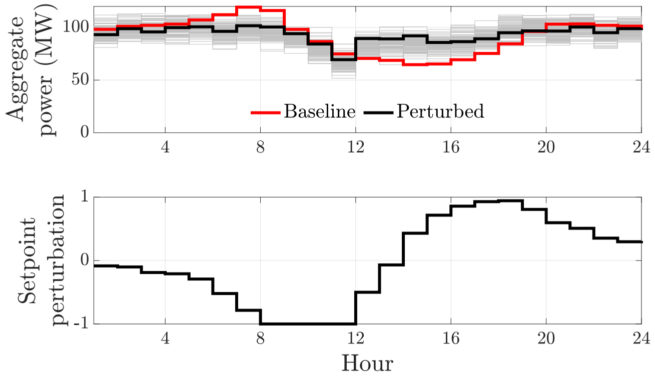

Fig. 5. Aggregate power (top) and setpoint perturbations (bottom). Peak demand on this day is reduced by 18 MW (16%).

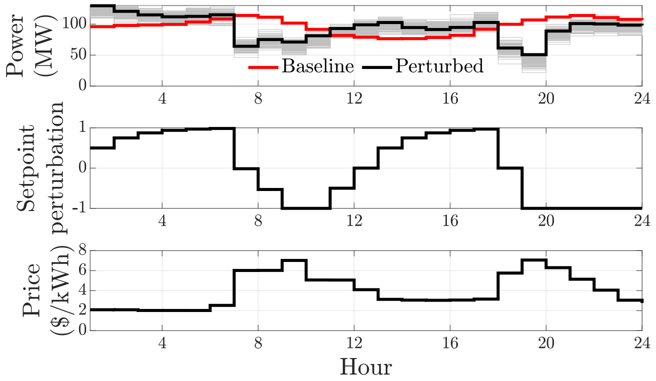

Fig. 6. Aggregate power (top), setpoint perturbations (middle) and energy prices (bottom). Energy costs on this day are reduced by $1.1 million (12%).

B. Minimizing Energy Costs or Carbon Emissions

In this example, the goal is to minimize the worst-case cumulative energy cost over $S = 100$ scenarios,

$$\max_{s\in\mathcal{S}} \Delta t \sum_{k\in\mathcal{K}} \pi(k)P_s(k), \tag{5}$$

subject to the constraints in (4). The energy prices $\pi(k)$ ($/kWh) are assumed to be known in advance.

Fig. 6 shows optimization results for February 14, 2021. Prices are from the Electric Reliability Council of Texas (ERCOT) day-ahead energy market. The aggregation is 'charged' during the relatively low-price periods and 'discharged' during the morning and evening price spikes, as can be seen from the bottom two plots. This perturbs aggregate power (top plot, black curve) below the baseline (red curve) during price spikes. In this simulation, the energy cost is reduced by 12%, from $9.2 to 8.1 million.

The method in this section can also minimize carbon emissions by replacing $\pi(k)$ in (5) with the carbon intensity of electricity, $\mu(k)$ (kg/kWh). Alternatively, costs and emissions can be jointly optimized by specifying a carbon price π_c ($/kg) and replacing $\pi(k)$ with $\pi(k) + \pi_c\mu(k)$.

C. Maximizing Flexibility Revenue

Distribution utilities and transmission system operators value the ability to rapidly decrease load by a pre-determined amount in response to a dispatch call. This service is typically called emergency demand response at the distribution level and spinning, synchronized or replacement reserve at the transmission level. To provide these services, a load aggregator must determine in advance the flexibility that it can offer. If accepted, these offers become binding commitments. This gives rise to the problem of maximizing capacity revenue subject to the constraint that offers can be reliably delivered, even if they are all accepted and dispatched.

One approach to this problem decides (downward) flexibility offers $\underline{f}(k) \geq 0$ (kW) and setpoint perturbation scenarios $u_s(k) \in [-1, 1]$ corresponding to model error scenarios

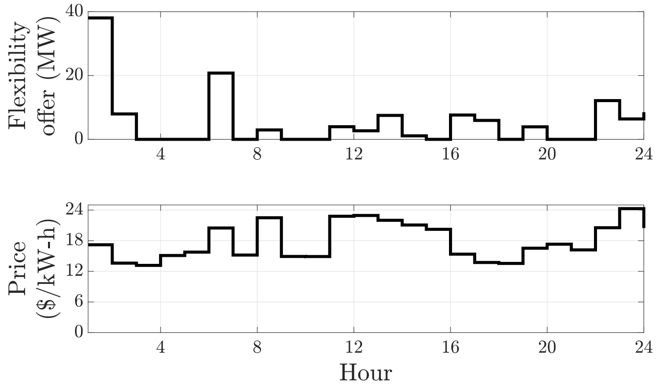

Fig. 7. Flexibility offers (top) and prices (bottom) in ERCOT's replacement reserve market. Revenue on this day is $2.4 million.

$e_s(k)$. The objective is to maximize flexibility revenue,

$$\Delta t \sum_{k \in \mathcal{K}} \underline{\pi}(k) \underline{f}(k), \tag{6}$$

where $\underline{\pi}(k)$ ($/kW-h) is a flexibility price, treated here as deterministic. In each scenario s, the power $P_s(k)$ is defined by the perturbation model (1) with u_s and e_s replacing u and e. To ensure that offers can be delivered despite model errors, curtailed load in each scenario should meet or exceed the flexibility offer:

$$\hat{P}(k) - P_s(k) \geq \underline{f}(k). \tag{7}$$

Fig. 7 shows the results of maximizing the flexibility revenue (6) subject to (1), (7) and $|u_s(k)| \leq 1$ with $S = 100$ scenarios. Prices are from the ERCOT replacement reserve market on February 17, 2021. Offers (top plot) are high during some but not all price spikes. This is because flexibilities are coupled over time. For example, a possible curtailment in the morning reduces the capacity to curtail in the afternoon. Due to very high prices, the aggregation earns $2.4 million on this simulated day.

The method in this section can optimize flexibility offers for other grid services. For example, some grid operators value the capacity to increase load. This can be handled by changing (7) to

$$P_s(k) - \hat{P}(k) \geq \overline{f}(k), \tag{8}$$

and maximizing $\Delta t \sum_{k \in \mathcal{K}} \overline{\pi}(k) \overline{f}(k)$, where $\overline{\pi}(k)$ ($/kW-h) and $\overline{f}(k)$ (kW) are the upward flexibility price and offer. Other grid operators value the capacity to symmetrically increase or decrease load. This can be handled by imposing both (7) and (8) and maximizing

$$\Delta t \sum_{k \in \mathcal{K}} \overline{\underline{\pi}}(k) \min \left\{ \overline{f}(k), \underline{f}(k) \right\},$$

where $\overline{\underline{\pi}}(k)$ ($/kW-h) is the symmetric flexibility price.

V. Summary and Extensions

This paper has proposed a general framework for providing a variety of grid services by adjusting device setpoints in load aggregations. The framework involves baseline power prediction, perturbation system identification, and load-shifting optimization. In all three phases, problem dimensions are independent of the number of loads, so this framework can handle gigawatt-scale aggregations. This paper has discussed a variety of candidate loads, including some with nonlinear behavior. Simulations of these loads during Texas' extreme storm of February, 2021, have shown the value that the framework could provide to the grid.

There are several opportunities to extend this work. First, the amount of data required to identify a useful perturbation model could be investigated and methods could be refined to make the best use of limited data. Second, this paper proposed a linear perturbation model to ensure that load-shifting problems could be solved to global optimality. A nonlinear perturbation model could be more accurate, but would require settling for locally-optimal load-shifting solutions. Whether this trade-off is worthwhile is an interesting question. Third, reinforcement learning could be investigated to balance trade-offs between exploring system behavior to refine the perturbation model, and exploiting the current model to maximize near-term rewards. Fourth, the control framework could be considered for grid services at time scales of minutes or seconds, where the closed-loop dynamics of device-level controllers become important. Finally, the control framework could be evaluated in hardware.

Acknowledgments

The authors gratefully acknowledge support from Exelon Corporation and the insights of Jake Jurewicz, Sunny Elebua, and the Corporate Strategy team at Exelon. The authors thank Adedayo Aderibole, Jasmina Burek, Subhro Das, Jeremy Gregory, Joe Higgins and Francis Selvaggio for thought-provoking discussions.

References

[1] International Energy Agency, "Renewables." [Online]. Available: https://www.iea.org/fuels-and-technologies/renewables

[2] B. Mohandes, M. El Moursi, N. Hatziargyriou, and S. El Khatib, "A review of power system flexibility with high penetration of renewables," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3140–3155, 2019.

[3] D. Callaway and I. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184–199, 2011.

[4] O. Sundstrom and C. Binding, "Flexible charging optimization for electric vehicles considering distribution grid constraints," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 26–37, 2011.

[5] J. Mathieu, M. Kamgarpour, J. Lygeros, G. Andersson, and D. Callaway, "Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 763–772, 2015.

[6] P. Douglass, R. Garcia-Valle, P. Nyeng, J. Ostergaard, and M. Togeby, "Smart demand for frequency regulation: Experimental results," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1713–1720, 2013.

[7] S. Palacio, K. Kircher, and K. Zhang, "On the feasibility of providing power system spinning reserves from thermal storage," *Energy and Buildings*, vol. 104, pp. 131–138, 2015.

[8] R. Menke, E. Abraham, P. Parpas, and I. Stoianov, "Demonstrating demand response from water distribution system through pump scheduling," *Applied Energy*, no. 170, pp. 377–387, 2016.

[9] S. Shao, M. Pipattanasomporn, and S. Rahman, "Grid integration of electric vehicles and demand response with customer choice," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 543–550, 2012.

[10] K. J. Kircher and K. M. Zhang, "Model predictive control of thermal storage for demand response," in *American Control Conference (ACC)*, 2015, pp. 956 – 961.

[11] K. Kircher, A. Aderibole, L. Norford, and S. Leeb, “Distributed peak shaving for small aggregations of cyclic loads,” *IEEE Transactions on Power Delivery*, 2021.

[12] K. Oikonomou, M. Parvania, and R. Khatami, “Optimal demand response scheduling for water distribution systems,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 5112–5122, 2018.

[13] N. Motegi, M. Piette, D. Watson, S. Kiliccote, and P. Xu, “Introduction to commercial building control strategies and techniques for demand response,” Lawrence Berkeley National Laboratory, Tech. Rep. 59975, 2007.

[14] G. Goddard, J. Klose, and S. Backhaus, “Model development and identification for fast demand response in commercial HVAC systems,” *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 2084–2092, 2014.

[15] I. Beil, I. Hiskens, and S. Backhaus, “Frequency regulation from commercial building HVAC demand response,” *Proceedings of the IEEE*, vol. 104, no. 4, pp. 745–757, 2016.

[16] M. Kraning, E. Chu, J. Lavaei, and S. Boyd, *Dynamic network energy management via proximal message passing*. Now Publishers, 2014.

[17] H. Hao, B. Sanandaji, K. Poolla, and T. Vincent, “Aggregate flexibility of thermostatically controlled loads,” *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 189–198, 2015.

[18] J. Hughes, A. Domínguez-García, and K. Poolla, “Identification of virtual battery models for flexible loads,” *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4660–4669, 2016.

[19] S. Barot and J. Taylor, “A concise, approximate representation of a collection of loads described by polytopes,” *International Journal of Electrical Power and Energy Systems*, vol. 84, pp. 55–63, 2017.

[20] Z. Wang and T. Hong, “Reinforcement learning for building controls: The opportunities and challenges,” *Applied Energy*, vol. 269, p. 115036, 2020.

[21] B. Yildiz, J. Bilbao, and A. Sproul, “A review and analysis of regression and machine learning models on commercial building electricity load forecasting,” *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.

[22] L. Ljung, “System identification,” *Wiley encyclopedia of electrical and electronics engineering*, pp. 1–19, 1999.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[24] S. Kim, R. Pasupathy, and S. G. Henderson, “A guide to sample average approximation,” in *Handbook of Simulation Optimization*. Springer New York, 2015, pp. 207–243.

[25] K. J. Kircher and K. M. Zhang, “Sample-average model predictive control of uncertain linear systems,” in *Conference on Decision and Control*, 2016, pp. 6234–6239.

[26] G. Calafiore, “Random convex programs,” *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3427–3464, 2010.

[27] J. M. Penman, “Second order system identification in the thermal response of a working school,” *Building and Environment*, vol. 25, no. 2, pp. 105–110, 1990.

[28] N. Merket, “Weather data for buildings energy simulations,” https://data.nrel.gov/submissions/128, 2020.