Managing load deferability to provide power regulation

Federico Bliman, Andres Ferragut and Fernando Paganini

Universidad ORT Uruguay

Abstract—**Providing frequency regulation services in power networks has become an important part of network operation, traditionally carried out by fast responding generators. In this paper we consider regulation services from the demand side, through a new actor in the power market: a demand aggregator that manages a large number of consumer loads. The aggregator exploits the deferability of certain loads to control the consumption profile and thus reduce regulation needs, or even provide regulation services to others. We analyze this control through macroscopic ODE models inspired by queueing systems, where a fluid state represents load quantities. Two versions are considered: a one-state model that tracks the entire load population, and a two-state version that separately tracks quantities of currently deferrable and non-deferrable loads. The control input is the fraction of deferrable loads that are active, and is controlled using a combination of feedforward for tracking of a reference signal, and feedback to reduce the impact of random fluctuations. The performance of such controllers is evaluated by simulation using regulation signals from real networks.**

I. Introduction

Due to limitations in storage, demand and supply must always match in an electric power system. This task is carried out by system operators (SO) with the aid of markets involving different time scales, from long-term contracts to real-time. Market transactions involve energy and ancillary services (reserves and regulation, see [9]), the latter being summoned close to real-time. Of these services, frequency regulation occurs at the shorter time scale, its objective being to keep the frequency as close to nominal as possible (50Hz or 60Hz), as a consequence of matching demand and supply. Providers of this service must respond to a signal sent by the operator every few seconds and adapt their generation output to the reference. Certain fast responding generators like gas or hydro turbines which are already providing nominal power to the system usually perform frequency regulation.

Regulation needs, which traditionally came from fluctuations in demand, are now increased by the rise of renewable energy sources. Energy coming from wind or sun is non-dispatchable and the operators must cope with sudden changes in its energy output. This tendency is making regulation more expensive and it may even complicate operation in systems where wind energy moves above the 20% share [7]. An alternative to such supply-side regulation is to use the flexibility of the loads for this task, which are becoming controllable with the deployment of smart-grid technologies [14]. For this to become feasible, a large number of loads must be controlled to make a significant effect in the grid. Under this scenario a new actor comes into play, a load *aggregator* (e.g. [3], [4]). The function of the aggregator is to manage a cluster of loads from its clients and use them to provide services such as frequency regulation to the grid.

We briefly survey some related work in this area. The control of the ON-OFF cycle in Thermostatically Controlled Loads (TCLs) to provide ancillary services is investigated in [5], [6], [10]. In particular, [6] shows this approach can provide significant regulation capability in a practical situation with real data from California. More relevant to this paper is another line of work [11], [15] that exploits the time deferability of generic loads, characterized by arrival times, deadlines, and power and energy requirements. In particular [15] investigates different options for *scheduling* such deferrable loads, comparing classical approaches from processor scheduling (earliest deadline first, least laxity first [8]) with a model predictive control proposal tailored to the power setting.

In this paper we build upon our previous work [1], where we introduced a fluid model of a cluster of loads and studied its frequency regulation capabilities. This method gives excellent regulation tracking; now, as reviewed in Section II, to guarantee load deadlines it must be combined with a sophisticated scheduling algorithm. In Section III we introduce an alternative fluid model, which isolates the population of loads with expired *laxity*, guaranteeing them service and only deferring the remaining loads. Although this is found to slightly reduce our regulation impact, it provides a far simpler means to keep strong load deadlines, amenable to decentralized implementation. Incorporating a feedforward/feedback control in Section IV, the capability of the system to track real-world regulation signals is investigated by simulation. In Section V we analyze the performance of the system.

II. Fluid models of load aggregation

Consider a load aggregator that manages a large set of loads from its customers. Each load is characterized by a certain amount of required energy and a given power consumption. As discussed before, we assume that loads need not be serviced immediately, but instead they have a certain amount of *laxity* in time to complete the request. We now propose a model for this load aggregator and analyze how deferring service of requests may be used to reduce frequency regulation demand, or even become a provider of ancillary services.

Since we aim to analyze the macroscopic behavior of the system, we characterize loads by their average parameters: let Q_0 denote the average energy request and p_0 the requested power. Service requests arrive into the system at a given rate λ (in requests per time unit), which we assume constant for the period of interest: as regulation happens at a very fast time scale we can consider that the demand does not change in the time-window we study. The load aggregator may estimate the value of λ and Q_0 and therefore its average required power $p^* = \lambda Q_0$ and purchase it in the day-ahead market. The actual demand will differ from the purchased power, so regulation services must be procured. If the aggregator could control user demands to match the power profile purchased in advance, it could avoid paying for regulation services. It could even receive a payment for regulation services if loads could be controlled to follow a regulation signal sent from the SO.

Let $\tau := \frac{Q_0}{p_0}$ denote the average time required to service a request. To characterize the deferability of loads we introduce a time-window parameter h that represents the mean deadline for service. If $h \geq \tau$ it means that loads are deferrable. In such a situation, the load aggregator may choose to serve only a fraction $u \in [0, 1]$ (which we will call *service level*) of the loads at a given time, deferring the service of the remaining ones. Finally, let $n(t)$ denote the number of requests at disposal of the load aggregator at a given time. Its evolution can be traced by the following dynamic state space model from [1]:

$$\dot{n}(t) = \lambda - \frac{1}{\tau} n(t) u(t), \tag{1a}$$

$$p(t) = p_0 n(t) u(t). \tag{1b}$$

The above fluid model, although quite simple, captures the essential dynamics of a cluster of loads, suitable for predicting the average power the cluster of clients will consume. However, since regulation services rely of fine-grained matching of supply and demand, it is important to also characterize deviations from this average. Deviations arise because of the randomness in load arrivals and departures, and can be accommodated in the preceding model through a noise disturbance.

A detailed stochastic model of the load queue would involve a random process of arrivals, e.g. Poisson of intensity λ, and a model of randomness in load sizes, for instance an exponential distribution of mean Q_0. The latter would introduce random variations in departure times. Consistently with our fluid differential equation model we will represent both arrival and departure variations by a white noise signal v injected in (1a), with power spectral density $S_v(\omega) \equiv 2\lambda$ (see [1] for background on justifying this modeling step).

To analyze the effect of such noise on the power output, which determines the regulation requirements, we will linearize the dynamics around the operating point for $u(t) \equiv u^*$; the linear model with injected noise is:

$$\dot{\delta n} = -\frac{1}{\tau} u^* \delta n + v, \tag{2a}$$

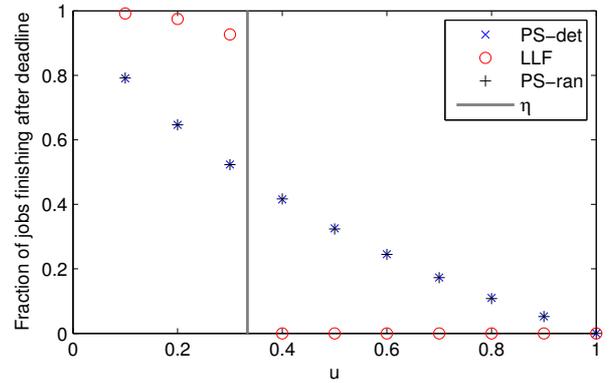$$\delta p = p_0 u^* \delta n; \tag{2b}$$

Fig. 1. Fraction of jobs completed after the deadline with varying u^*, and different scheduling policies.

Calculations carried out in [1] show that the variance of the consumed power is given by the expression

$$E\left[(\delta p)^2\right] = p^* p_0 u^*. \tag{3}$$

From (3) we see that we could potentially reduce the output variability as much as we want, eliminating the need for *acquiring* regulation to serve these random loads. This comes at the cost of further deferring loads ($u^* \to 0$), which may not meet their deadlines. This issue does not appear transparently in the previous model, which only specified the fraction of loads being served, but does not provide information on the specific scheduling algorithm used by the aggregator. To get more insight into this question, in [1] we studied by simulation the effect of three scheduling algorithms on keeping deadlines:

- *Equal sharing:* The load aggregator chooses to serve all present jobs with power $p_0 u^*$. While some loads may not allow this kind of service, it serves as a reference point for analysis. It corresponds to the Processor Sharing (PS) discipline of queueing theory.
- *Random:* The load aggregator chooses a fraction u^* of the available jobs at random. This policy is easy to implement in a decentralized environment, by distributing the value of u^* and having the loads choose whether to become active or not based on a local random variable.
- *Least-Laxity-First (LLF) [8]:* Here, the load aggregator chooses a fraction u^* of the loads ordered by decreasing *laxity*, i.e. the remaining amount of time before the job needs to become active in order to meet its deadline. This requires a centralized scheduling at the aggregator level.

In Fig. 1 we compare the fraction of loads that miss their deadlines as a function of the fixed service level u^*. As we can see, egalitarian policies show a smooth decrease of missed deadlines as service level increases, whereas laxity-based mechanisms show a sharp decrease in missed deadlines as soon as the service level satisfies $u^* \geq \frac{\tau}{h} =: \eta$.

Fig. 2 shows the power output variability for the three algorithms and different values of u. The main conclusion here is that power regulation requirements are agnostic to the individual scheduling, and fall well within the predictions
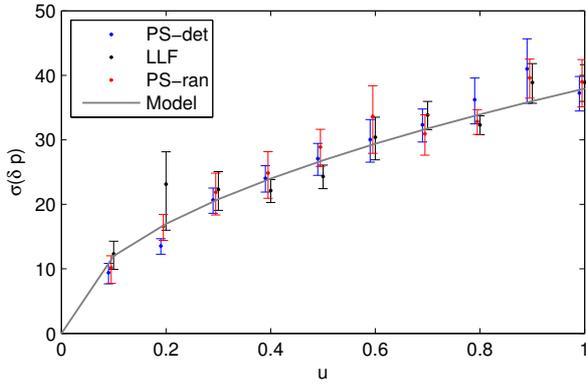
Fig. 2. Variability of power output as a function of u^*, for different scheduling policies, for model (1).

of the simple dynamical model (2). We could potentially reduce the consumed power variability down to $\sqrt{p^* p_0 \eta}$, if we are able to implement LLF which implies knowing detailed information and taking actions on individual loads.

III. SEPARATING THE CRITICAL POPULATION

Since scheduling loads in real time may be difficult to achieve in practice, we now propose a second approach to this problem, based on a more detailed classification of the set of loads. Let $n(t)$ denote now the number of loads that at time t may still be deferred, i.e. those that still have remaining *laxity* (spare-time). A fraction $u(t)$ of these loads are assumed active. Loads not being served consume their laxity, and may reach a point when meeting the deadline requires turning on the load immediately. We propose this course of action, which can be implemented in a decentralized fashion (e.g. a thermal load which decides to start consuming power since the temperature has become too low), and denote by $m(t)$ the number of loads active because of expired laxity. Setting $L = h - \tau$ the mean laxity, a dynamic model for this new system is:

$$\dot{n}(t) = \lambda - \frac{1}{\tau}n(t)u(t) - \frac{1}{L}n(t)(1 - u(t)), \quad (4a)$$

$$\dot{m}(t) = \frac{1}{L}n(t)(1 - u(t)) - \frac{1}{\tau}m(t), \quad (4b)$$

$$p(t) = p_0(n(t)u(t) + m(t)). \quad (4c)$$

Now $u(t)$ represents the fraction of loads with positive laxity that are being served. Loads can exit the first queue in two ways: a fraction of the loads, represented by the term $\frac{1}{\tau}n(t)u(t)$ get completely served before their deadline; the rest, represented by $\frac{1}{L}n(t)(1 - u(t))$, are automatically turned on when they run out of laxity and move from n to m. The departure rate from the second queue is represented by $\frac{1}{\tau}m(t)$.

Analyzing consumed power variability in this model is somewhat more complicated. We will again linearize the model for a fixed $u \equiv u^*$ and add the noise sources with the

same assumptions as in (2). The linear model is the following:

$$\dot{\delta n} = -\frac{1}{\tau}u^*\delta n - \frac{1}{L}(1 - u^*)\delta n + v_1 - v_2 - v_3, \quad (5a)$$

$$\dot{\delta m} = \frac{1}{L}(1 - u^*)\delta n - \frac{1}{\tau}\delta m + v_3 - v_4, \quad (5b)$$

$$\delta p = p_0(u^*\delta n + \delta m). \quad (5c)$$

Now v_1 stands for the arrival noise in n, v_2 and v_3 for the two departure noises. v_3 is also the arrival noise for m, and v_4 the departure noise for m.

To calculate the variability in the consumed power we will resort to the state space representation of the system:

$$\begin{bmatrix} \dot{\delta n} \\ \dot{\delta m} \end{bmatrix} = \overbrace{\begin{bmatrix} -\frac{u^*}{\tau} - \frac{1-u^*}{L} & 0 \\ \frac{1-u^*}{L} & -\frac{1}{\tau} \end{bmatrix}}^{A} \begin{bmatrix} \delta n \\ \delta m \end{bmatrix} + B \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}; \quad (6a)$$

$$\delta p = \underbrace{\begin{bmatrix} p_0 u^* & p_0 \end{bmatrix}}_{C} \begin{bmatrix} \delta n \\ \delta m \end{bmatrix}. \quad (6b)$$

Here $u(t) \equiv u^*$ is fixed again and the inputs (w_i) are i.i.d. white noises of *unit* power spectral density. To map these normalized processes to the corresponding v_i we invoke the variance growth of a Poisson process, which is equal to its intensity. For noises v_2, v_3 which represent departures from the n queue, the relative intensity is "thinned" by the probability of departing from the system (denoted by α), or respectively the probability $1 - \alpha$ of moving to the second queue. Departures of the latter queue also must occur at a rate $\lambda(1 - \alpha)$, fixing the scaling for v_4. The resulting B matrix is:

$$B = \begin{bmatrix} \sqrt{\lambda} & -\sqrt{\alpha\lambda} & -\sqrt{(1-\alpha)\lambda} & 0 \\ 0 & 0 & \sqrt{(1-\alpha)\lambda} & -\sqrt{(1-\alpha)\lambda} \end{bmatrix}.$$

To calculate α explicitly we compute the probability that the job is served (at level u^*) before the laxity expires:

$$\alpha = P\left[\frac{\tau_k}{u^*} \leq \frac{L_k}{1 - u^*}\right] = \frac{\frac{u^*}{\tau}}{\frac{u^*}{\tau} + \frac{1-u^*}{L}},$$

where we invoked the exponential distribution of τ_k, L_k with respective means τ, L.

We have now a stable state-space system driven by vector valued white noise. In steady-state, the covariance matrix Q of the state is (see e.g. [2]) the solution to the Lyapunov equation

$$AQ + QA^T + BB^T = 0,$$

and the resulting variance of the output p is $E\left[(\delta p)^2\right] = CQC^T$. Carrying out the calculations for the given matrices results in a final result for the variance of consumed power:

$$E\left[(\delta p)^2\right] = p^* p_0 \left[1 - \frac{1}{\frac{1}{1-u^*} + \frac{\tau}{Lu^*}}\right]. \quad (7)$$

This model behaves in a different way to the one in Section II: only the deferrable portion of the load population is within the scope of the service level u^*, hence this parameter can take
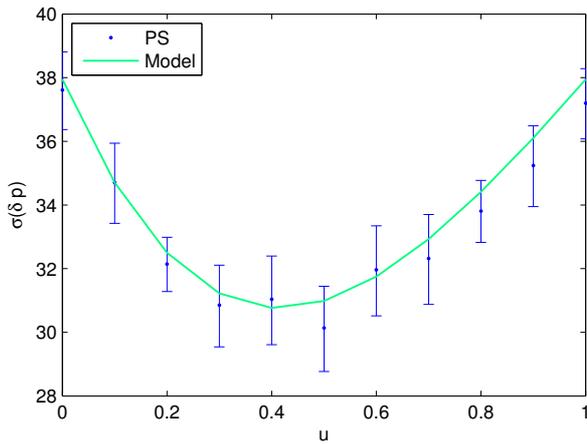
Fig. 3. Variability of power output as a function of u^* for model (4).

Fig. 4. Tracking a real life reference signal.

any value in $[0, 1]$, without violating job deadlines, In fact we find from (7) that both extremes 0 or 1 have the same effect on the output variance, which becomes $E\left[(\delta p)^2\right] = p^* p_0$. This makes sense because the only difference between both cases is that for $u^* = 1$ loads are turned on when they arrive, whereas for $u^* = 0$ they do after their laxity expires, but in any case the time in service is the same, hence the steady-state output variability is the same.

Choosing an intermediate value of u^* allows us to lower the variance but with a lower bound, which is the price to pay for keeping all deadlines. The optimal value of u^* that minimizes the variance can be calculated to be:

$$u^*_{opt} = \frac{\sqrt{\tau}}{\sqrt{L} + \sqrt{\tau}}. \tag{8}$$

We show in Fig. 3 the power variability of the loads predicted by the model for the case $\eta = 1/3$, and simulation results for the randomly varying system (PS). The main conclusion of this analysis is that, with a simpler mechanism that does not resort to scheduling, we can nevertheless reduce the regulation requirements. By comparing the results of Figs. 2 and 3, we can see that although we cannot reduce the power variability as much as in the one state model with $u^* = \eta$, by carefully choosing the service level u^* of the second system, we can still achieve a significant reduction in the power variability, but now with deadlines automatically attained.

IV. PROVIDING REGULATION BY ADAPTING DEFERABILITY

In this section we would like to go one step further and design a controller that will allow the aggregator not only to reduce its regulation requirements, but also to provide this service to the grid. This is done by actuating on the control variable $u(t)$, in response to a command signal $r(t)$ sent by the system operator. Such a control for model (1) was already designed in [1], we now present a similar controller for the two state model.
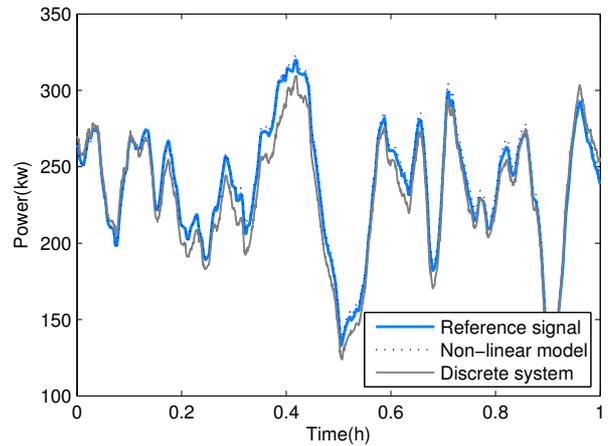
The transfer function of the linearized version of the plant (4) is:

$$G_{up}(s) := \frac{\hat{\delta p}}{\hat{\delta u}} = \frac{p_0 n^* s}{s + \frac{u^*}{\tau} + \frac{1-u^*}{L}}. \tag{9}$$

Note that despite this being a second order state-space system, the result has first order reflecting a non-minimal realization.

As a first solution to our tracking requirement we used the inverse of this plant as a feedforward controller. We tested the tracking capability of the loads with this controller simulating the response of the system to a real life regulation signal taken from the PJM Interconnection, a Regional Transmission Organization in the United States [12]. We compare the reference signal against the prediction of model (4) and a discrete system simulation which better represents the real system. In the latter loads arrive at random times and we schedule them in a random way according to the signal u, until they are served or they run out of laxity and are turned on automatically. The results of this simulation in Fig.4 show that the system is able to set the consumed power very close to the reference.

Still, there are some differences between the reference and the output, which may not be tolerable if we want to be regulation providers. To improve tracking we consider adding feedback to the design, to compensate for the deviations introduced by randomness in arrivals and departures. In Fig. 5 we depict such a controller, using state feedback of the variables δn and δm.
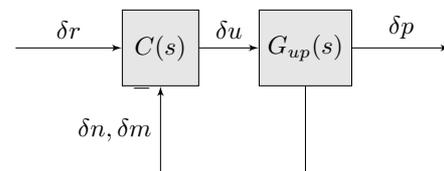

Fig. 5. Controller design for tracking the regulation signal

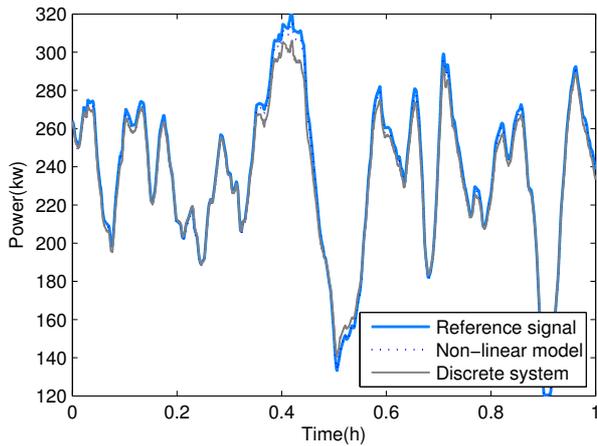Now the controller is the sum of two terms, one that tracks

Fig. 6. Tracking a reference signal for the system with noise rejection.

Fig. 7. Performance for different values of feedback.

$r(t)$, plus a noise reducing term with inputs $n(t)$ and $m(t)$. The final form of the controller is, in Laplace transform notation:

$$\delta u = \frac{s + \left(\frac{u^*}{\tau} + \frac{1-u^*}{L} \right)(1-a)}{p_0 n^* s} \delta r - \frac{u^*}{n^*} a \left(\delta n + \frac{\delta m}{u^*} \right). \tag{10}$$

The parameter a fixes the feedback term for noise reduction, being 0 for no feedback. Setting $a = 1$ would make the system internally unstable, so we choose a strictly less than 1. In our simulation experiments we used $a = 0.7$.

The last simulation we present, shown in Fig. 6, illustrates how the system is capable of tracking the same signal we used before. We see there is a notorious improvement in tracking after we add the noise reducing feedback.

V. Performance Analysis

As mentioned in the beginning of this paper, the SO must ensure enough regulation resources at every time. In order for a generator, an aggregator, or other actor in the system to provide this service they must demonstrate their capability to follow a signal, which usually consists of a test that measures their accuracy and delay when performing this task. This qualification test sets a minimum threshold for being able to provide frequency regulation, but, from the operator point of view, the higher the score the better.

It is of our interest to analyze how our system performs in different conditions and compared to other regulating resources. For this purpose we will use the performance score used by PJM [13], which depends on the delay, the correlation and the precision of the response compared to the reference signal.

$$\frac{\text{Performance}}{\text{Score}} = \frac{1}{3} \left(\frac{\text{Delay}}{\text{Score}} + \frac{\text{Correlation}}{\text{Score}} + \frac{\text{Precision}}{\text{Score}} \right).$$

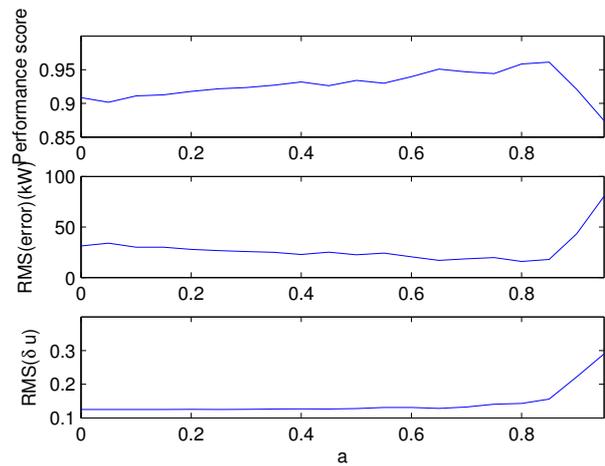Delay and correlation scores are calculated together at the point where the response has maximum correlation with the time-shifted reference. The correlation value is the corresponding score and the delay score is a linear function of the time shift, scoring 1 if the time shift is 0 and scoring 0 when the time shift is 5 minutes. Precision score is calculated based on the relative error in the signal 1-norm:

$$1 - \frac{\|error\|_1}{\|reference\|_1}.$$

The minimum score for qualifying to participate in the regulation market is 0.75. As a reference we have that the average score for steam generators is slightly above 0.75, hydro generators score slightly higher around 0.8, batteries are one of the best resources scoring over 0.9, whereas other demand response resources score over all the range from 0.7 upwards.

A. Results

In Fig.7 we evaluate the performance and control effort of our system for different values of the feedback gain a. We plot the PJM performance score, along with the RMS value of the error (reference-output), a simpler measure of performance; and also the RMS value of δu, a measure of the control effort. We see that the performance increases and the error decreases as we increase the amount of feedback, as we would expect. This has a limit, around $a = 0.85$, given mainly because of the saturation of u. This saturation effect means the response cannot follow the reference signal and hence the error increases.

Another parameter that may affect the performance score is the amount of regulation provided. By this we mean the maximum absolute value the regulation signal can take, and we express it in terms of the nominal power. On one hand we would expect the performance to worsen as we provide more regulation because we move away farther from equilibrium, so the non-linearities get larger. On the other hand we have the intrinsic noise at the output that we study in section II. This noise is present independently of the amount of regulation offered, so in relative terms it gets smaller as we offer more
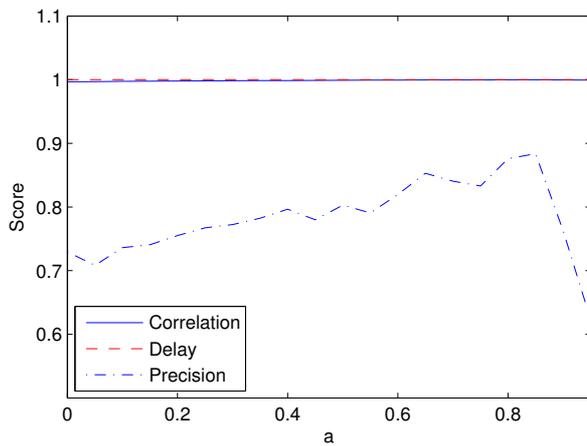
Fig. 8. Broken down performance score for different values of feedback.

regulation. Since the precision term in the performance score is inversely proportional to the amount of regulation provided, the effect of this noise in the score diminishes as we increase the regulation offered.

In Table I we present the performance score of the system (for $a = 0.7$) along with the RMS value of the error, and the relative error, for different values of offered regulation. For most of the range, as the amount of offered regulation increases we see that the RMS value of the total error increases, but in relative terms it becomes smaller; consistently, the performance score increases with the offered regulation, up to a certain point.

When the amount of offered regulation is equal to the nominal power of the system, the performance score begins to degrade, and so does the relative error. This can be explained by noting that the control variable is reaching its saturation limits. Consider for example the case when the regulation signal takes the extreme value $\delta r(t_1) = -p^*$ for a certain time t_1, so the power output should be 0 at that moment. Our system cannot track that reference, because even setting $u(t_1) = 0$, this does not control the loads with expired laxity. So if $m(t_1) > 0$ at that time, the power output will be positive.

TABLE I
PERFORMANCE FOR DIFFERENT VALUES OF OFFERED REGULATION.

Regulation offered (p.u)	Performance Score	RMS(error) kW	RMS(rel. err.)
0.2	0.875	16.3	0.339
0.4	0.937	18.1	0.188
0.6	0.957	18.9	0.131
0.8	0.964	23.7	0.123
1	0.961	37.4	0.155

To finish the performance analysis we show in Fig. 8 a breakdown of the performance score. We can see that the system has an excellent delay and correlation scores but is not

so good at precision. This can be explained by the fact that loads respond almost immediately as they have no inertia, but the uncertainty in the number of loads makes it hard to exactly follow the reference.

VI. CONCLUSIONS

In this paper we analyzed a new model for a load demand aggregator operating in the regulation market. We built over a previous model [1], seeking to improve its behavior in regard to load deadlines. We proposed a modification that takes into account this issue without resorting to complex scheduling algorithms. We showed that with this approach it is possible to reduce the amount of regulation required by the network, and further to become a regulation provider by adapting the service level of loads in real time.

Several lines of future work remain open: a more thorough controller design to take into account the constraints in the input signal, as well as to cope with the nonlinearities in the system can be performed. From the queueing perspective, it would be interesting to analyze more precise fluid models for the different scheduling mechanisms involved.

REFERENCES

[1] F. Bliman, A. Ferragut, and F. Paganini, "Controlling aggregates of deferrable loads for power system regulation," in *Proc. of the American Control Conference (ACC)*, 2015.

[2] J. Burl, *Linear Optimal Control: $\mathcal{H}_2$ and $\mathcal{H}_\infty$ methods*. Menlo Park, CA: Addison Wesley, 1999.

[3] EnergyPool, http://www.energy-pool.eu/en/.

[4] GoodEnergy, http://www.goodenergy.com.

[5] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "A Generalized Battery Model of a Collection of Thermostatically Controlled Loads for Providing Ancillary Service," in *Proc. of the 51st Allerton conference on Communication, Control and Computing*, 2013.

[6] ——, "Frequency Regulation from Flexible Loads: Potential, Economics, and Implementation," in *Proc. of the American Control Conference (ACC)*, 2014.

[7] H. Holttinen and et al., "Design and operation of power systems with large amounts of wind power," IEA wind, Tech. Rep., 2009.

[8] J. Hong, X. Tan, and D. Towsley, "A performance analysis of minimum laxity and earliest deadline scheduling in a real-time system," *IEEE Transactions on Computers*, vol. 38, pp. 1736–1744, 1989.

[9] B. Kirby, "Ancillary Services: Technical and Commercial Insights," Oak Ridge National Laboratory, Tech. Rep., 2007.

[10] S. Koch, J. Mathieu, and D. Callaway, "Modeling and Control of Aggregated Heterogeneous Thermostatically Controlled Loads for Ancillary Services," in *Proc. of the 17th Power Systems Computation Conference*, 2011.

[11] A. Nayyar, J. Taylor, A. Subramanian, K. Poolla, and P. Varaiya, "Aggregate Flexibility of a Collection of Loads," in *Proc. of the 52nd IEEE Conference on Decision and Control*, 2013.

[12] PJM, "Fast Response Regulation Signal," http://www.pjm.com/ markets-and-operations/ ancillary-services /mkt-based-regulation/ fast-response-regulation-signal.aspx.

[13] ——, "PJM Manual 12: Balancing Operations," http://www.pjm.com/ /media/documents/manuals/m12.ashx.

[14] F. Rahimi and A. Ipakchi, "Demand Response as a Market Resource Under the Smart Grid Paradigm," *IEEE Transactions on Smart Grid*, vol. 1, pp. 82–88, 2010.

[15] A. Subramanian, M. Garcia, D. Callaway, K. Poolla, and P. Varaiya, "Real-Time Scheduling of Distributed Resources," *IEEE Transactions on Smart Grid*, vol. 4, pp. 2122–2130, 2013.