Envelope Modeling for Renewable Energy Intermittency and Capacity

Xiaoyue Jiang, Geoffrey Parker, Ekundayo Shittu
Tulane Energy Institute
Tulane University
New Orleans, LA

Efforts to integrate renewable energy resources into power operations have faced many challenges, largely driven by variation in the output of the dominant technologies. A fundamental problem has been the lack of a consistent methodology to quantify the capacity contribution and system impact of intermittent generating resources such as wind and solar so as to enable the effective evaluation, management and compensation of these resources. The electricity system can be viewed as a special supply chain system with zero-tolerance of delay and rapidly growing but still very limited storage resources. To analyze the system, we adapt an envelope-based modeling method from telecommunications engineering that is inspired by Network Calculus (NetCal) for deterministic queuing systems. The variability of electricity supply and demand can be described by upper and lower-envelopes and their Legendre conjugates, which quantify not only the variability across different time scales, but also characterize the tradeoff between any assigned capacity value and its corresponding Quality-of-Service (QoS) measures of performance. In particular, the QoS performance leads to an intuitive interpretation through storage resources. The envelope modeling leads to the definition of two QoS-based capacity metrics: Guaranteed Capacity and the Best-Effort Capacity, whose conceptual and numerical properties are analyzed and compared against existing capacity metrics. For illustration, the proposed methods are applied to real data from the California Independent System Operator (CAISO). We explicitly quantify the capacity contribution (via the notion of the Best-Effort Capacity) of wind during peak hours and its negative system impact at night. The same envelope-characterization further shows the capacity value of storage resources.

Key words: Intermittency, Capacity, Smart-Grid, Envelope Modeling, Network Calculus (NetCal), Legendre Transform, Quality-of-Services (QoS) Guarantees, and Best-Effort.

1. Introduction

Renewable power resources offer the potential of a lower emissions alternative for electricity generation compared to conventional fossil fuel-based generators. In the last decade, the share of electricity produced from these resources has been steadily increasing across the globe, largely as as a result of government policies promoting their advancement. However, the increased penetration of renewable energy creates unprecedented challenges for system operators and utility companies, mainly due to the intermittent nature of the major renewable power resources such as wind and solar. In the absence of subsidies, many of the technologies would not be deployed due to cost and engineering concerns.

Unlike conventional sources of electricity, the supply from intermittent renewable resources is highly variable, (thus non-dependable), inflexible (thus non-dispatchable) and difficult to predict. These undesirable features bring engineering and economic challenges to existing systems. For instance, as the system faces increasing production variability, the system baseload drops and the average production cost increases. Further, due to subsidies that renewable suppliers receive, it can be economic to produce even when locational marginal prices are negative, indicating that additional power is undesirable at certain times.

To deal with these issues of integration, a variety of rules have been formulated across utility systems ranging from wind curtailment in favor of hydroelectric production to charging renewable power producers for extra operating reserves to be maintained on the utility system. However, these rules can appear to be arbitrary and potentially conflicting. The industry lacks a standardized method to assess and manage the impact of intermittent resources on the electric grid. At the core of this problem is the difficulty of properly evaluating the capacity contribution of an intermittent power resource and compensating it adequately and fairly. Intermittent renewables are well accepted as valuable **energy** resources, but their performance as a **capacity** source has not been convincing.

From an operations research perspective, the electricity system can be viewed as a special supply chain with nearly-zero tolerance of delay, negligible storage to buffer supply/demand mismatch, inelastic demand, and extremely high standard for reliability. These special characteristics explain the current emphasis on power (generation and consumption in megawatts, MW) instead of energy (cumulative flows in megawatt-hours, MWh) in the analysis on resource adequacy and system reliability. However, technological progress as well as progress made in market design and operation is making the emphasis on instantaneous production and consumption less critical. Indeed, technological improvement and growing investment in storage, demand response, and smart-grid technologies imply increasing storage capability, increasing elasticity in demand, richer and faster information collection and sharing, more effective control means, and more efficient market action/reaction. In such a context, a comprehensive characterization of electricity supply and demand is becoming more critical and relying on single-valued capacity metrics discards valuable information that can help market formation as well as system operation.

In this paper, we develop a new way to model capacity based on the concept of Quality of Service (QoS) performance guarantees. Essentially, our modeling methodology is to characterize each type of generating resources in terms of both **quantity** and **quality**. The quantity attribute is evaluated based on the amount of power the generating facility can produce. The quality attribute refers to a resource's quality of service (QoS). It describes the match/mismatch pattern between power demand and supply, which is profoundly affected by intermittency of a generating resource, as well as the availability and utilization of supplementary resources including storage and demand response. The notion of QoS reflects a view of capacity as a metric of service rather than that of equipment.

Our goal is to contribute to the literature on three fronts. First, we apply envelopes from the theory of network calculus (NetCal) to the management and operation of renewable energy. To our knowledge, this has not yet appeared in the literature. We adopt the transition from the time domain to the conjugate domain via the Legendre transformation, and we substitute application-specific concepts such as arrival and service curves to generic vocabularies such as upper and lower envelopes. Second, on the application front, the envelope method we develop can capture intermittency at different time scales. This is a significant deviation from existing capacity metrics and leads to a direct characterization that places a value (positive or negative) on the specific variability of a given intermittent resource. Third, we present a unifying model and method for generating resources that considers differences in dispatchability.

The rest of the paper is organized as follows. In Section 2, we review related literature. We follow in Section 3 with a discussion of existing capacity evaluation methods and metrics. In Section 4, we present the fundamentals of NetCal, and introduce the concepts of envelopes and their Legendre conjugates. We follow this conceptual introduction with an envelope-based capacity modeling framework in Section 5. We discuss some applications to real data, courtesy of the California Independent System Operator (CAISO). We quantitatively characterize the tradeoff between capacity and QoS performance of system load and generating sources through upper and lower envelopes. This characterization enables us to define the two types of capacity metrics for intermittent sources—guaranteed capacity and the best-effort capacity that we cover in subsection 6. The latter allows us to compare our metric with a widely used measure of capacity, Effective Load Carrying Capacity (ELCC), with the real data. Section 7 concludes.

2. Related Literature

The enactment of the Public Utilities Regulatory Policy Act (PURPA) in 1978 stimulated electricity generation from co-generation and renewable energy facilities (Joskow et al. 1989). Additional policies by the Federal Energy Regulatory Commission (FERC), such as the Energy Policy Act of 1992, expanded the wholesale transactions of renewable energy (Joskow 2006). However, the increase in renewable energy generation has created significant challenges including discontinuities in output, unstable system costs, and operational complexities. The operations research community has responded to some of these hurdles. In particular, risk management and the vertical integration of utilities has attracted attention (Aıd et al. 2011, Joskow 2005). However, few models address the problems of comprehensively valuing the capacity of intermittent resources.

Hobbs and Pang (2007) address the generator's profit maximization problem using a Cournot model of competition among electricity generators on a transmission network. Their model nicely addresses how price caps transform affine demand curves into piecewise linear functions in liberalized electricity markets, but less attention is paid to the influence of technological differences in supply and the capacity value of intermittent resources. Kamat and Oren (2002) study the pricing problem of efficiency-motivated instruments in the electricity industry and find that the volatility from price jump behavior affects forward and option prices. They show that including volatilities in supply capacities such as renewable sources will further aggravate the stochastic volatility in spot prices. In their analysis of long-run markets for electric power and emissions permits, Zhao et al. (2010) present a complementarity formulation on the efficiency of alternative systems for emissions allowances. We argue that their finding of less investment distortion under emissions allowance allocation by energy sales may be influenced when the capacity values of renewable resources are factored in. They offer support for this view by concluding that the absence or presence of capacity markets, amongst others, can influence outcomes.

Powell and Oren (1989) use a social planning model that determines the investment rates of nondepletable and depletable energy production to show that the price of energy exceeds the operating cost of the nondepletable energy sources, and capacity investments in nondepletable energy only yield outputs at the price-cost equilibrium. We posit that when the measures of QoS for the nondepletable sources, guaranteed and best-effort capacities, are included, these conditions at equilibrium may no longer hold.

Using QoS as a performance measure is well studied in multiple literatures. For example, in wireless transmission, Ata (2005) minimize the long-run average energy consumption subject to a QoS constraint—expressed as an upper bound on the packet drop rate. In queuing applications, Mandelbaum and Zeltyn (2009) study staffing in a system operating in a many-server configuration with general customer patience distributions (Bassamboo and Randhawa 2010) with the objective of satisfying a QoS constraint. While the domains may be different, our adoption of QoS as a measure of capacity supply performance is consistent with earlier applications. Maglaras and Zeevi (2005), in a model of service systems, considers our QoS measures—“guaranteed” processing rate and “best-effort” type service—as two nonsubstitutable services to a market of heterogenous users. With the service providers objective of maximizing revenues, they find that real-time congestion notification results in increased revenues. Our analysis draws on these concepts to characterize the performance measures of intermittent renewable capacities.

Our methodology draws from the theory of Network Calculus (Cruz 1991a,b) in telecommunication systems. Other applications of this theory can be found in queing systems (Chang 2000), the Internet (Le Boudec and Thiran 2001), manufacturing systems (Bose et al. 2006), and supply chain systems (Jiang and Parker 2012). In Section 4, we discuss the fundamentals of this theory. The novel application in this paper to valuing renewable capacity of intermitent sources of energy follows in Section 5.

3. Metrics

The most common electric power capacity metrics are installed capacity (ICAP), unforced capacity (UCAP) or capacity factor, effective load carrying capability (ELCC) or capacity credit, variants of time-period based methods (i.e., averaged capacity over specific time interval during the day), and exceedance method (e.g., median, $70\%, 95\%$ percentiles) (NERC 2011). There is also the concept of “guaranteed capacity,” a metric under which intermittent resources are given a zero capacity rating and then treated as negative loads that offer energy but not capacity. Roughly speaking, the ICAP, UCAP and guaranteed capacity metrics correspond, respectively, to the maximum, the mean, and the minimum of the power supply.

Whenever a value in the unit of MW is quoted, readers will naturally match one of these capacity metrics based on both context and their personal judgment, which can cause confusion and error in communication. For instance, when investment in renewables is under consideration, ICAP is the implied choice of metrics. When renewable and, in particular, wind penetration is defined as the proportion of a power source on a system, expressed as a percentage, both ICAP and UCAP (effectively, average of energy flow) have been used but they give different numbers. Various degrees of subjectivity are involved in the definition and choices of capacity metrics. For time-based metrics, the specification of peak period is subjective. The choice between UCAP or exceedance (i.e. percentile) methods and, for the latter, the specific choice of exceedance level, are subject to debate and/or compromise. In addition, there is a wide range of experience and backgrounds of professionals involved. It is difficult for some professionals to assign any positive capacity to wind given the fact that wind resources do not blow all the time. Others will assign a positive value to wind given all the energy contributions of wind to the grid. To add to the confusion, people note the existence of negative locational marginal price (LMP) and even negative capacity of wind as an evidence of intermittency, which, at a high-level view, shifts the debate from the focal point of summer afternoon peak hours to the the middle of the night when baseload suppression becomes a main grid challenge.

All of these current metrics have their shortcomings when it comes to measuring the capacity of intermittent resources. For example, UCAP has traditionally been adopted as the primary metric for evaluating generator capacity. For conventional resources, capacity factors are relatively high (around 80% or more) and the variability of real-time capacity is relatively small. Therefore, the mean performance as quoted by UCAP does not deviate too much from the instantaneous performance. However, such a metric does not work well for wind or solar as the high variability of these resources can neither be captured by the mean, nor can it be ignored.

As a result of problems with UCAP and ICAP, there has been a shift towards evaluating capacity contribution using ELCC. This metric essentially is a reliability-oriented estimate of generation capability. ELCC is obtained by replacing the generating source under consideration by an equivalent generator of constant capacity that maintains the same system reliability standard. The most commonly used reliability measure for this purpose is loss of load expectation for which the target value is typically chosen as 1 day in 10 years (Milligan and Porter 2008). ELCC is frequently used in practice to value wind capacity by various entities including ERCOT, MISO, PacifiCorp, Colorado PUC/Xcel Energy and Quebec Balancing Authority Area (Milligan and Porter 2006, 2008, Gross and Organization, Holttinen et al. 2007), and NERC (2011).

There are several issues with an ELCC-based evaluation: (1) Due to the system nature of the ELCC metric, renewable resources are commonly evaluated at the resource class level. For instance, ERCOT calculated wind ELCC as 8.7% in 2007 (Milligan and Porter 2008). And, this value is subject to change from year-to-year. This is an issue because a class-level evaluation is not sufficient for some individualized incentive mechanisms. (2) The value of ELCC is system-dependent and is neither transparent nor intrinsic to the resources themselves. In particular, the value decreases monotonically as the penetration rate of intermittent resources goes up. Such a well-known negative correlation of capacity value and penetration level is driven by the dependence of renewables on the rest of the system for backup. (3) ELCC measures how the generating resource performed *ex post*, but not what it can promise *ex ante*. In other words, it reflects the notion of “best effort” as opposed to that of performance (or equivalently, Quality of Service) guarantees. This is not a trivial distinction. In essence, it is the capability of guaranteeing a quality of capacity service level or lack of it that differentiates conventional resources from renewables. To be more concrete, what can be counted on for wind capacity in the day-ahead market is substantially different from what is going to be actually realized, given the relatively low precision of wind forecasting techniques. The forecast error is inherent to the intermittency nature of wind resource which is, however, not modeled in the ELCC model. Finally, as will be show more explicitly in Section 6, ELCC effectively treats power supply/demand of all time instances in isolation. This is inadequate for modeling the dynamics and, consequently, the capacity contribution of storage resources.

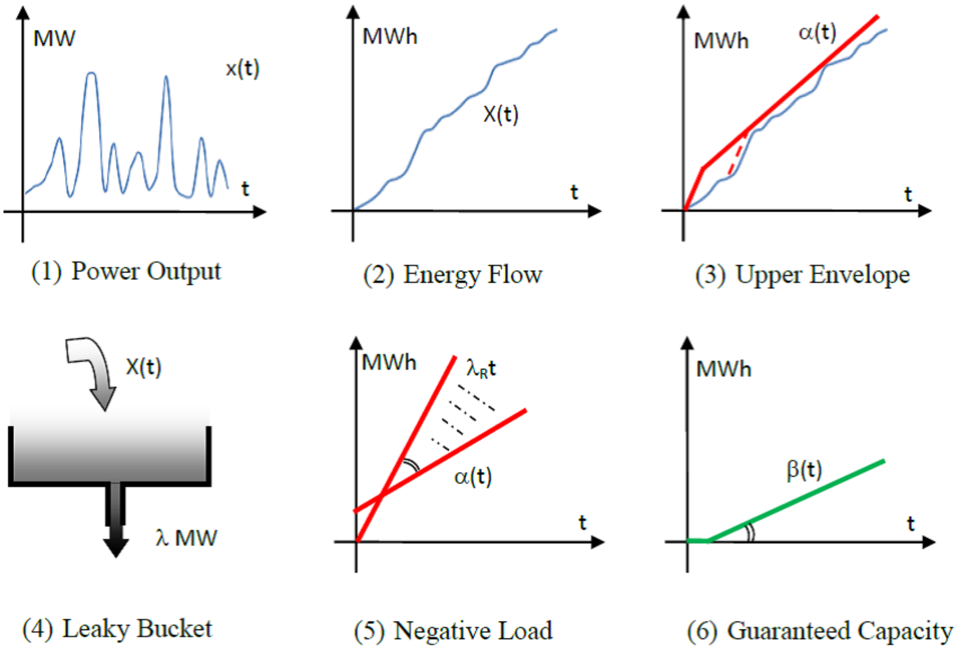
From the view of QoS performance guarantee, evaluating the capacity of intermittent resources at the class level, as in current practice, has multiple issues. Consider: if the performance of an individual generator cannot be differentiated from the rest of the generators in its class, little can be done to incentivize high performance of service and, similarly, to provide incentives to reduce the consumption of capacity during scarcity periods. The method we propose addresses this issue directly by evaluating the capacity contribution of individual generating resources (conventional and renewable) and rewarding their quality of service guarantees.

A fundamental deviation from existing approaches is that we model the cumulative power generation and consumption without loss of information. In this way, we capture not only the instantaneous power generation characteristics, but also the pattern of energy generation and consumption on different time scales. More importantly, the method applies not only to conventional and renewable generators, but also naturally incorporates those supplementary capacity resources including storage and demand response. Simply put, a generating resource matches supply and demand from the supply side; a demand resource does the same from the demand side; and a storage resource can do the same from both supply and demand sides at the expense of some energy conversion loss.

4. Methodology

The proposed methodology is inspired by the theory of Network Calculus (NetCal) that was developed for deterministic queuing systems and, in particular, for the Internet, see e.g. Chang (2000) and Le Boudec and Thiran (2001) for reference. The essence of NetCal is to model input and output flows by their respective upper- and lower-envelopes, through which to derive various QoS performance bounds. Below we will illustrate the basic ideas of the proposed capacity evaluation methodology with Figure 1. Details of formulation are presented in Subsections 4.1-2.

To describe the QoS characteristics of a given load, the first step is to convert the raw load data (Panel 1) from the power domain (MW against time) to the energy domain (MWh against time) (Panel 2). Then, an upper envelope is constructed to bound the energy flow from the above (Panel 3). The way to construct this upper envelope is effectively performed on the Legendre conjugate



(1) Power Output (2) Energy Flow (3) Upper Envelope

(4) Leaky Bucket (5) Negative Load (6) Guaranteed Capacity

Figure 1 Illustration of Capacity Modeling Procedure for an Intermittent Source

domain, where the net excessive electricity demand (in MWh) with respect to the constant demand flow at any given capacity level (in MW) is captured. This computational procedure corresponds to the leaky bucket mechanism (Panel 4) that can be implemented with an efficient recursive algorithm. In this way, any given demand flow is bounded from above by a convex upper envelope.

By further taking intermittent resources (for example, wind) as a negative load (Panel 5), lower envelopes (Panel 6) of wind power supply can also be obtained. Intuitively, the lower envelope of supply reveals the minimum output (in MWh) over any period of a given duration. In particular, when the wind resource claims a certain capacity credit, the Legendre conjugate of lower envelope gives the maximum net deficit in electricity production of this wind generator in comparison against a constant generator at the claimed capacity level. We view this maximum deficit as a quality indicator of the claimed capacity. In this way, the dependence of this wind generator on the system as backup can be explicitly quantified. More detailed discussion on the involved NetCal formulations is presented in Subsection 4.1. Applications are presented in Sections 5 and 6.

4.1. NetCal Fundamentals

Network Calculus (NetCal) originated in the field of telecommunication (Cruz, 1991 a,b) and has evolved as a convenient method for analyzing the QoS issues in deterministic queuing systems and the Internet (Chang 2000, Le Boudec and Thiran 2001). The basic theory of NetCal is founded on the concept of min-plus convolution operation, $\otimes$, and de-convolution, $\oslash$, both acting on the space of cumulative flows and their envelope functions: $F = \{f : f$ left-continuous, non-decreasing, and $f(0) = 0$, where $\otimes$ is defined by $f \otimes g(t) = \inf_{0\leq s\leq t}\{f(s) + g(t-s)\}$, and $\oslash$ by $f \oslash g(t) = \sup_{s\geq 0}\{f(t+s) - g(s)\}$. For a service node S with input, output flows $X, Y \in F$, NetCal models make three fundamental assumptions: (1) Causality property: $X \geq Y$; (2) Arrival curve property: $X \oslash X \leq \alpha$, or equivalently $X \leq X \otimes \alpha$; and (3) Service curve property: $Y \geq X \otimes \beta$. Functions α and $\beta \in F$ are called the arrival and service curves, which are, respectively, the upper- and lower-envelopes for the input and service flows. The intuition behind a service curve can be explained through a slightly stronger version of it, known as strict service curve: the minimum service of node S during any busy period of duration $t \geq 0$ is no less than $\beta(t)$.

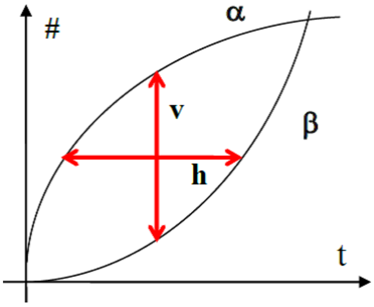


Figure 2 NetCal backlog and delay bounds

At the single-node level, the main concern is how to bound the three major QoS performance measures - backlogs: $Q(t) = X(t) - Y(t)$, (virtual) delay: $D(t) = \inf\{d : X \oslash Y(-d) \leq 0\}$, and the output bound: $Y \oslash Y$. For a system made out of multiple nodes, an immediate task is to determine the service curve of a system with two nodes in tandem. The NetCal theory offers elegant bound results on these interested performance measures: Backlog bound: $Q(t) \leq \alpha \oslash \beta(0) := v$; Delay bound: $D(t) \leq \inf\{d : \alpha \oslash \beta(-d) \leq 0\} := h$; Output bound: $Y \leq \alpha \oslash \beta$; Node concatenation: $Z \geq X \otimes (\beta_1 \otimes \beta_2)$, where X, Z are the input, output flows of a tandem system of two nodes whose service curves are β_1 and β_2, respectively. Figure 2 here shows how arrival and service curves can jointly determine the backlog and delay bounds as respectively the maximal vertical and horizontal gaps between the two.

4.2. Envelopes and their Legendre Conjugates

The essence of NetCal theory is the use of upper- and lower-envelopes to describe the involved flows and from which to derive the performance bounds that are of interests. The bound-based performance measure is tied to a view of worst-case analysis, very much compatible to high reliability standard in the field of electricity. However, the NetCal models together with their original interpretations do not directly apply to the power systems. First, a major QoS measure in queuing system, delay, is to a large extend irrelevant in power system as the demand has to be satisfied moment-to-moment or otherwise be dropped. Consequently, it is the notion of capacity in MW instead of MWh that becomes much more critically relevant here than to general queuing systems. Secondly, as for the other QoS measure, backlog, for the same reason of zero-tolerance of delay, it would be more naturally interpreted as mismatch between supply and demand and must ultimately be absorbed by parallel virtual or physical storage resources. Thus, it is the parallel rather than serial configuration, and the depletion rather than overflow of the storage that are more relevant. Third, the performance of a service node such as a wind turbine can often be directly observed and modeled based on its actual output. In other words, information is available for calculating the **strict** service curve. Recall that the original definition of service curve requires the coupling between input and output flows, the strictness property allows for simplified construction of the lower envelope.

We present below a modified NetCal analysis based on a construction first outlined in Jiang (2008) under the name of CT-NetCal. It turns out that this version of NetCal nicely adapts to the special needs arisen from the power system context. To avoid unnecessary confusion that may be caused by by using standard queuing terminologies such as **arrival** and **service curves**, and not to overload the existing electricity jargons including **demand** and **supply curves** that carry different meanings, we will refer to those NetCal curves as upper- and lower-envelopes. We will describe the proposed method as an envelope-modeling method rather than a network calculus method. As a notational convention, for flow X, we denote its upper- and lower-envelopes as α_X and β_X, respectively. With this non-application specific vocabulary, we can re-interpret Figure 2 as an envelope-based characterization of the (mis-) matching pattern of a supply-demand pair.

It is crucial to realize that additional analytic insights can be gained by applying Legendre transform to the envelopes. Figure 3 below shows how. In fact, by translating these enveloping curves into their corresponding Legendre conjugates, the two performance bounds (v and h) can be simply calculated as the sum of corresponding conjugates. Effectively, these supply and demand flows are separately "scanned" and evaluated against a family of constant flows parameterized by the reference capacity level λ. Deviations from the reference flows thus depicts the variation pattern of the flows. More fundamentally, we will show in this subsection that the Legendre transform can be applied directly to the original flows as a way to construct the envelopes, not just as a computational method to determine the gaps between given envelopes.

More specifically, let flows L, $G \in F$ be the demand and supply flows, respectively. Define

$$\begin{aligned} \overline{Q}^L_\lambda(t) &:= L \oslash C_\lambda(t,t) &&= \sup_{s\le t}\{L(s,t) - C_\lambda(t-s)\} \\ \underline{Q}^G_\mu(t) &:= C_\mu \oslash G(t,t) &&= \sup_{s\le t}\{C_\mu(t-s) - G(s,t)\}, \end{aligned} \tag{1}$$

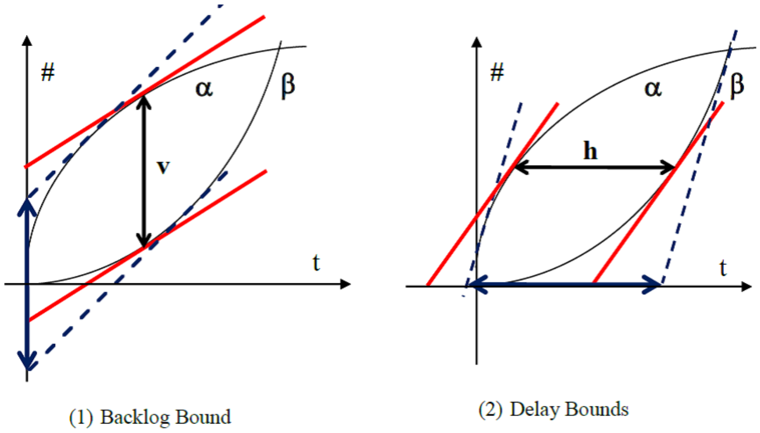


Figure 3 NetCal bounds as additive Legendre conjugates

where for arbitrary function f, $f(s,t) := f(t) - f(s)$, and C_μ is a constant flow such that $C_\mu(t) := \mu t$. We make the following two enveloping assumptions: for all $\lambda, t \geq 0$,

$$\begin{array}{ll} \text{Upper-Envelope Conjugate: } & \overline{Q}^L_\lambda(t) \leq \overline{\mathcal{L}}[\alpha_L](\lambda), \\ \text{Lower-Envelope Conjugate: } & \underline{Q}^G_\lambda(t) \leq \underline{\mathcal{L}}[\beta_G](\lambda). \end{array} \tag{2}$$

By definition, $L(s,t) \leq \lambda(t-s) + \overline{\mathcal{L}}[\alpha_L](\lambda)$ and $G(s,t) \geq \lambda(t-s) - \underline{\mathcal{L}}[\beta_G](\lambda)$ for all $\lambda \geq 0, t \geq s \geq 0$. Consequently, for all $t \geq s \geq 0$, we can define an upper envelope for flow L and a lower envelope for G, such that

$$\begin{array}{l} \alpha_L(t-s) := \inf_{\lambda \geq 0}\{\lambda(t-s) + \overline{\mathcal{L}}[\alpha_L](\lambda)\} \geq L(s,t), \text{ and} \\ \beta_G(t-s) := \sup_{\lambda \geq 0}\{\lambda(t-s) - \underline{\mathcal{L}}[\beta_G](\lambda)\}^+ \leq G(s,t). \end{array} \tag{3}$$

Proposition 1 **(Three Bounds in Conjugate Form)**. *For a queuing system with input, service, and output flows L, G and Y that are enveloped respectively by α_L and β_G as in Equations (3), performance bounds can be represented in terms of the Legendre conjugates as follows: For all $\lambda \geq 0$,*

$$\begin{array}{ll} \textit{Backlog Bound:} & Q(t) \leq \overline{\mathcal{L}}[\alpha_L](\lambda) + \underline{\mathcal{L}}[\beta_G](\lambda) \\ \textit{Delay Bound:} & D(t) \leq [\overline{\mathcal{L}}[\alpha_L](\lambda) + \underline{\mathcal{L}}[\beta_G](\lambda)]/\lambda \\ \textit{Conjugate Output Bound:} & \overline{Q}^Y_\lambda(t) \leq \overline{\mathcal{L}}[\alpha_L](\lambda) + \underline{\mathcal{L}}[\beta_G](\lambda) \end{array} \tag{4}$$

where Q, D are respectively the queue size, the (virtual) delay as defined earlier, and $\overline{Q}^Y_\lambda(t)$ is the conjugate output flow.

Proof. The proofs for all three bounds are similar. We will prove the output bound to illustrate the method. In fact,

$$\begin{array}{rl} \overline{Q}^Y_\lambda(t) & = \sup_{0 \leq s \leq t}\{Y(s,t) - \lambda(t-s)\} \\ & \leq \sup_{0 \leq s \leq t}\{L(t) - Y(s) - \lambda(t-s)\} \\ & \leq \sup_{0 \leq s \leq t}\{L(t) - L \otimes \beta_G(s) - \lambda(t-s)\} \\ & \leq \sup_{0 \leq u \leq s \leq t}\{L(t) - L(u) - \beta_G(s-u) - \lambda(t-s)\} \\ & = \sup_{0 \leq u \leq s \leq t}\{[L(u,t) - \lambda(t-u)] + [\lambda(s-u) - \beta_G(s-u)]\} \\ & \leq \sup_{0 \leq u \leq t}\{L(u,t) - \lambda(t-u)\} + \sup_{0 \leq u \leq s}\{\lambda(s-u) - \beta_G(s-u)\}\} \\ & \leq \overline{\mathcal{L}}[\alpha_L](\lambda) + \underline{\mathcal{L}}[\beta_G](\lambda). \quad \blacksquare \end{array} \tag{5}$$

To achieve the tightest backlog and delay bounds, optimal capacity levels λ^* can be identified by solving the corresponding convex optimization problems, for instance, solution to $\inf_{\lambda\geq 0}\{\overline{\mathcal{L}}[\alpha_X](\lambda)+\underline{\mathcal{L}}[\beta_G](\lambda)\}$ would yield the minimum backlog bound. Note that the optimal capacity levels that minimize the backlog and delay bounds in Equation (4) are, in general, different. For the detailed presentation of convexity involving Legendre transform, please see Rockafellar (1970).

This version of NetCal formulation brings several major advantages over the original NetCal formulations: First, the bounds and the envelopes are represented simply as the summations of Legendre conjugates, as opposed to the highly nonlinear operators $\otimes$ and $\oslash$ involved in NetCal. Such a property is especially important for extending the deterministic NetCal models to their stochastic counterparts. In this way, one can scan through each rate level λ and characterizes the performance bounds as summations of random variables, and there is no need to deal directly with stochastic processes, nor nonlinearity, which represents a significant reduction of complexity. Second, the strictness of service curve enables β_G to be constructed without convolving with X as needed in the general service curve definition. Moreover, a simple recursive leaky bucket implementation exists by using the notion of negative load (illustrated in Figure 1 and formulated in Section 5). As a result, the conjugates are convex in rate λ, we translate the derivation of performance bounds into convex optimization problems. Finally, the “scanning” rate (slope) and the corresponding Legendre conjugates can be naturally interpreted as capacity and storage size, respectively. Implication of this interpretation will be elaborated in details throughout the rest of the text.

5. Intermittency Modeling

As can be seen from Equation (3), the upper- and lower-envelopes quantifies the maximum (minimum) demand (supply) of electricity on any given time scale. Equivalently, their conjugates characterize the excess (shortage) of electricity when the corresponding demand (supply) is compared against a constant demand (supply). This valuable information enables us to derive a suite of important properties. In Subsection 5.1, we present some fundamental properties of the upper- and lower-envelopes. In Subsection 5.2 we will first present the modeling method as illustrated in Figure 1, based on which we will discuss the system impact of intermittent resources. Computational algorithms are developed and implemented to enable numerical studies in Subsection 5.2. The numerical examples are based on real data from CAISO including 10-minute system load and aggregated wind data in Year 2010, and 1-minute system load, production of four wind farms, one solar farm and one geothermal plant in Years 2007, 2008, and 2009 from 1st June to 30th September. Some basic statistics of these generating resources are summarized below.

Table 1 Basic statistics (in mean MW) of system load and generating resources (Years 2007-2010)

	LOAD	Geothermal	Solar	Wind A	Wind B	Wind C	Wind D
2007	30750	1111	116	233	178	198	167
2008	30836	1068	122	219	128	154	153
2009	29824	1043	143	245	183	208	233
2010	26035	N/A	N/A	625			

5.1. Envelope Fundamentals

To keep the presentation aligned with the form of data we analyze, from now on we focus on flows that are defined on a discrete time horizon $\mathcal{Z}^+ = \{0, 1, 2, ...\}$, where the index is interpreted as equal-spaced time epochs. More concretely, a flow X is represented as $\{X_0 \equiv 0, X_1, X_2, ..., X_i, ...\}$. Equivalently, we can represent flow X by its difference flow $\mathcal{X} := \{x_1, x_2, ...\}$, where $x_i = X_i - X_{i-1}$ for all $i \geq 1$. In the context of electricity applications, $\mathcal{X}$ typically models a power flow, where x_i is the power supply (or demand) at time $t = i$, measured in MW; the corresponding flow X is the cumulation of power flow from time 0, which is nothing but the cumulated electricity from time $t = 0$ to i measured in MWh. Obviously, Flow X is monotonically increasing if and only if flow $\mathcal{X}$ is non-negative.

Also, to keep a deterministic spirit of the study, instead of involving probability structures along with heavy-weighted assumptions such as ergodicity, we assume the following **Cesàro summability** for all supply and demand flows. Basically, we only ask flows under consideration to have a mean value.

Assumption (Convergence of Mean Partial Sum) For any given difference flow $\mathcal{X} = \{x_i, i = 1, 2, ...\}$. Then there exists a constant $\mu_X < \infty$ such that

$$\lim_{t\to\infty} \Sigma_{i=1}^t x_i/t = \mu_X. \tag{6}$$

Proposition 2 below establishes a theoretical basis for constructing the upper- and lower envelopes and the associated Legendre conjugates. The proof is listed in the appendix.

Proposition 2 **(Cesàro Summability)**. *For any cesàro summable flow $\{x_i\}, i \geq 1$, define for all $\lambda > \mu_X$,*

$$\overline{\mathcal{L}}[\alpha_X](\lambda) := \sup_{0<j\leq i} \{(X_i - X_j) - \lambda(i - j)\}, \tag{7}$$

and for $t \geq 0$, denote inf *by* $\wedge$, sup *by* $\vee$, *and*

$$\alpha_X(t) := \bigwedge_{\lambda\geq 0} [\lambda t + \overline{\mathcal{L}}[\alpha_X](\lambda)]. \tag{8}$$

We claim

(i) For all $\lambda < \mu_X$*,* $\overline{\mathcal{L}}[\alpha_X](\lambda) = \infty$*; and for all* $\lambda > \mu$*,*

$$\overline{\mathcal{L}}[\alpha_X](\lambda) < \infty. \tag{9}$$

(ii) On (μ, ∞)*,* $\overline{\mathcal{L}}[\alpha_X](\lambda)$ *is convex and monotonically decreasing in* λ*.*

(iii) On $[0, \infty)$*,* $\alpha_X(t)$ *is concave and monotonically increasing in* t*,*

(iv) For any arbitrary semi-differentiable function f*, denote* f'_+ *and* f'_- *as its respective right and left derivatives, we have*

$$\begin{array}{rl} [\alpha_X]'_+(0) & = \sup_{t\geq 0} x(t) \\ \lim_{t\to\infty}[\alpha_X]'_+(t) & = \lim_{t\to\infty}[\alpha_X]'_-(t) = \mu_X. \end{array} \tag{10}$$

Intuitively, Proposition 2 states that as long as a demand flow has a well-defined mean value, then for any reference capacity that is larger than the mean, the net deficit is finite; Therefore, nontrivial upper-envelope can be constructed. In addition, both the envelope and its conjugate possess nice convexity property. In particular, the slope of the envelope at time 0 and ∞ is nothing but the maximum and mean rate of the flow. In this way, we can say the capacity and energy aspects of a demand flow can be explicitly represented by the two ends of the envelope. For simplicity, we will symbolically write $\alpha'_X(0) = \max_t x_t$, $\alpha'_X(\infty) = \mu_X$, and leave the rigorous interpretation to Proposition 2 (iv).

As a preview of some numerical results and findings, we note:

1. For each generating resources, we can construct the lower envelope to quantify the guaranteed supply at different time scales; the corresponding Legendre conjugates explicitly quantify the tradeoff between an assigned capacity value and its implied QoS performance. The same kind of tradeoff relation exists for the demand flow. As for numerical insights, the envelope characterization appears to be robust across multiple years and, as a confirmation of general understanding, the upper-envelope of the system load up to a 4-month scale is dominated by June-September, (the peak months).

2. We define and distinguish two types of capacity metrics, the guaranteed capacity and the best-effort capacity. The latter views the intermittent resources as a negative load to the system load. Instead of requiring moment-to-moment dominance, it captures the extent of envelope reduction of the system load. In return, it provides a more favorable view of intermittent sources than is seen when focusing only on their worst-case performance.

3. By identifying the conceptual linkage between conjugate curves and storage resources, careful design and operation of integrated renewable and storage systems can provide a higher level of guaranteed capacity. Moreover, demand response can be incorporated into the resource portfolio, whose specification and execution can be naturally coordinated with storage operations.

5.2. Capacity-Performance Tradeoff

The lower envelope of supply quantifies the dependence of intermittent resources on the rest of the system. As an objective representation, it captures the guaranteed performance at each and every time scale, thus is free from subjective emphasis on capacity (instantaneous supply) or on energy (long-term average), or any specific choice in between (as in time-period based methods). As an intrinsic characterization, it is independent of the system condition (load or other generating resources). We will present below the envelope characterization of demand and supply which quantifies the tradeoffs between the claimed capacity level (in MW) and the corresponding excess and deficit (in MWh).

5.2.1. Demand characterization We first develop a computational procedure to construct an upper-envelope α_X of any given load X. By definition, for any duration $L \geq 0: X(s, s+L) \leq \alpha_X(L)$. Clearly, α_X quantifies the maximal demand of electricity from a customer during any period of duration L.

The Legendre Transform based NetCal formulation presented in Subsection 4.2 in fact implies a simple queuing mechanism called Leaky Bucket. As depicted in Figure 4, the Leaky Bucket mechanism not only provides an intuitive interpretation to quantities of interest, it also induces a recursive algorithm for computing $\overline{\mathcal{L}}[\alpha_X](\lambda) := \sigma := \lambda\mathcal{T}$ that is visualized as the bucket size, and consequently for computing α_X. More specifically, for each $\lambda \geq 0$, a leaky bucket of leaking rate λ corresponds to a queuing system with a constant server C_λ. Denote the input flow as X. By Lindsey's equation, the output flow is given by $X \otimes C_\lambda$, thus the queue at any given time t can be computed recursively by

$$\begin{aligned} \overline{Q}_\lambda^X(t) &= X(t) - X \otimes C_\lambda(t) \\ &= \sup_{s \leq t}\{X(s,t) - C_\lambda(t-s)\} \\ &= [\overline{Q}_\lambda^X(t-1) + X(t-1,t) - \lambda]^+, \end{aligned} \tag{11}$$

By Equation (11), the maximum queue size can be obtained by

$$\overline{\mathcal{L}}[\alpha_X](\lambda) := \sup_{t \geq 0}\{\overline{Q}_\lambda^X(t)\}, \tag{12}$$

which is finite thus is well-defined for all $\lambda > \mu_X$ according to Proposition 2. In reference to Figure 4, it is clear that as long as the bucket size is no less than $\overline{\mathcal{L}}[\alpha_X](\lambda)$, the input flow X would not experience any blockage, i.e., the buffer behind the flow controller is empty all the time.

In the meantime, Equation (12) implies for all $s \leq t$ on $\mathcal{R}^+$, $\alpha_{X,\lambda}(t-s) := C_\lambda(t-s) + \overline{\mathcal{L}}[\alpha_X](\lambda)\} \geq X(s,t)$, thus qualifies $\alpha_{X,\lambda}$ as an upper-envelope of flow X. By further piecing together the whole family of envelopes parameterized by leaking rate λ, we construct

$$\begin{aligned} \alpha_X(t-s) &:= \wedge_{\lambda \geq 0}\alpha_{X,\lambda}(t-s) \\ &= \wedge_{\lambda \geq 0}\{C_\lambda(t-s) + \overline{\mathcal{L}}[\alpha_X](\lambda)\} \\ &\geq X(s,t). \end{aligned} \tag{13}$$

The inequality qualifies α_X as an upper-envelope for load X.

The above procedure corresponds to Panels (1)-(4) of Figure 1. The representation of bucket size $\overline{\mathcal{L}}[\alpha_X](\lambda) = \lambda\mathcal{T}$, or $\mathcal{T} := \overline{\mathcal{L}}[\alpha_X](\lambda)/\lambda$ in the unit of hour(s) gives a (capacity) scale-free QoS index. Intuitively, this index tells the following fact: while the constant generator C_λ **cannot** cover all the load instantaneously (as the peak load may be higher than λ, thus system or local backup is needed), generator C_λ **can** cover all the load within $\mathcal{T}$ hours lag time.

By Proposition 2, α_X and $\overline{\mathcal{L}}[\alpha_X]$ are concave and convex, respectively. Skipping technical details regarding differentiability, we note the slope of α_X at the origin, $\alpha'_X(0)$, corresponds to the peak power demand; the slope at the infinity, $\alpha'_X(\infty)$ represents the long-term average of power demand. For any value λ in between, i.e., $\alpha'_X(\infty) < \lambda_X \leq \alpha'_X(0)$, λ can be interpreted as the effective capacity of this load with QoS at level $\overline{\mathcal{L}}[\alpha_X](\lambda) < \infty$. This means it can be guaranteed that the maximal net electricity demand beyond a constant demand flow C_λ on any given time interval would never exceed $\overline{\mathcal{L}}[\alpha_X](\lambda)$, which is a finite value.

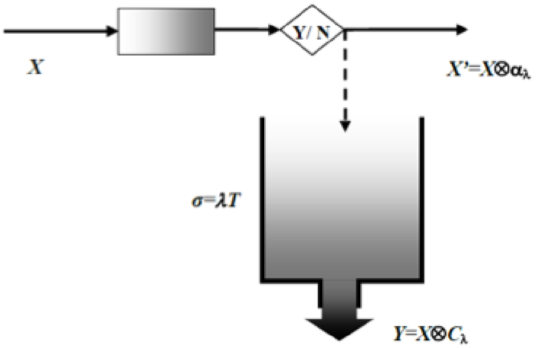


Figure 4 Leaky bucket with rate λ and bucket size $\sigma = \lambda\mathcal{T} = \overline{\mathcal{L}}[\alpha_X](\lambda)$.

Remark 1 (**Intuitions behind Legendre Conjugates**). As an analogy from a daily life context, the leaky bucket acts like a credit card. The expense flow on the card corresponds to the load. While the card holder may purchase goods without actually paying money on the spot as long as the net balance on card does not exceed the credit limit, he or she will have to pay all the expenses at a later point of time. The leaky bucket mechanism portraits a perfectly disciplined card holder who on every time period pays a constant amount, λ or the exact amount of the remaining balance whichever is smaller (no deposit in advance). For high enough λ, this card holder will always be able to repay every purchased item on the card within a finite period of time, $\mathcal{T}$.

5.2.2. Supply characterization The task at hand is to construct the lower-envelope β_G of any given generating source G, i.e., for any $L \geq 0: G(s, s+L) \geq \beta_G(L)$, which quantifies the minimal supply of electricity from a generating source during any period of duration L.

We are proposing the following so-called "**Negative Load**" method, or"Neg-Load" in short, to construct the lower envelope.

By viewing the supply flow G as a negative load, the leaky bucket mechanism involved in determining the upper-envelope α_X is readily adopted to find the lower-envelope β_G. More specifically, let us choose a reference load C_R, where R can be an arbitrary rate no less than the maximum instantaneous power supply of the resource under consideration. In this way, one obtains a net load of $(C_R - G)$, whose upper- envelope can be determined by passing the net load flow through a leaky bucket. Denote the leaking rate as $(R-\lambda)$ and the corresponding maximum deficit as $\overline{\mathcal{L}}[\alpha_{C_R-G}](R-\lambda)$. Then, the resulting lower-envelope of the supply becomes $C_\lambda - \overline{\mathcal{L}}[\alpha_{C_R-G}](R-\lambda)$ (see Figure 1, Panels (5)-(6)).

Denote for all $\lambda \geq 0$,

$$\begin{aligned}\underline{\mathcal{L}}[\beta_G](\lambda) &:= \overline{\mathcal{L}}[\alpha_{C_R-G}](R-\lambda)\\ \beta_{G,\lambda} &:= C_\lambda - \underline{\mathcal{L}}[\beta_G](\lambda)\\ \beta_G &:= \textstyle\bigvee_{\lambda\geq 0}\beta_{G,\lambda}.\end{aligned} \tag{14}$$

We now show the above constructions, in particular, $\underline{\mathcal{L}}[\beta_G]$, are independent to the choice of the reference level R, thus the resulting lower-envelope β_G is well-defined.

Proposition 3 **(Lower-Envelope via Net-Load)**. *(i) For any supply flow G, β_G defined in Equation (14) gives an independent to the choice of the reference level for all $R \geq \sup_{t\geq 0} g_t$, where $g_t = G_t - G_{t-1}$.*

(ii) More explicitly, for all $\lambda < \mu_G$, $t \geq 0$,

$$\underline{\mathcal{L}}[\beta_G](\lambda) = \bigvee_{s\geq t}\{\lambda t - G(s,t)\} < \infty \tag{15}$$

$$\beta_G(t) = C_R(t) - \alpha_{C_R-G}(t). \tag{16}$$

In particular, $\beta_G(t-s) \leq G(s,t)$ for all $s \leq t$, which qualifies β_G as a lower envelope of flow G.

Proof. See Appendix.

Note that Equation (16) can be rewritten in a more symmetric manner

$$\beta_G + \alpha_{C_R-G} = C_R. \tag{17}$$

Thus, the earlier discussion on the linkage between capacity metrics with the lower envelope can be carried over for the upper envelope.

More specifically, β_G and $\underline{\mathcal{L}}[\beta_G]$ are both convex. Moreover, $\beta'_G(0)$, which can often be 0 for wind and solar, corresponds to the minimum of instantaneous power supply; $\beta'_G(\infty)$ gives the long-term average of power supply. For any value λ in between, i.e., $\beta'_G(0) \leq \lambda < \beta'_G(\infty)$, λ can be interpreted as the effective capacity of this source with QoS of $\underline{\mathcal{L}}[\beta_G](\lambda) < \infty$. This means, it can be guaranteed that the maximal net electricity supply below a constant flow C_λ would be no more than $\underline{\mathcal{L}}[\beta_G](\lambda)$, which is a finite value.

Figure 5 summarizes how conventional capacity metrics could be represented as tangents on the upper- and lower-envelope curves.

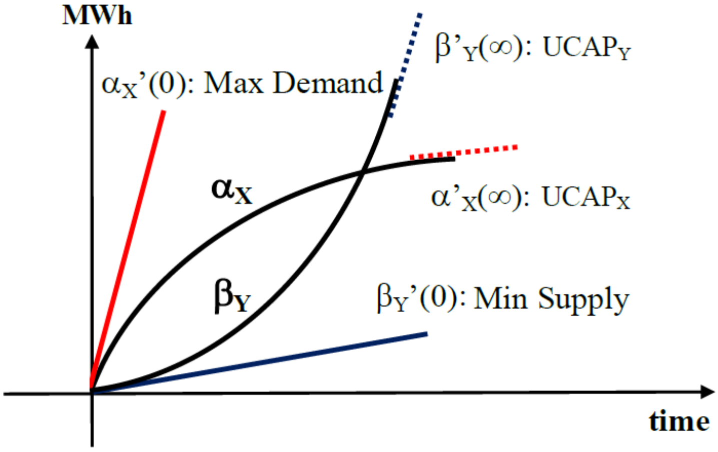


Figure 5 ICAP and UCAP as tangents on envelopes

Remark 2 (**Translation Property of Legendre Conjugates**). Proof of Proposition 3 (i) involves a simple but important property called "translation" property for Legendre conjugates:

$$\overline{\mathcal{L}}[\alpha_{X+C_\mu}](\lambda) = \overline{\mathcal{L}}[\alpha_X](\lambda + \mu). \tag{18}$$

As a matter of fact, it offers a natural way to extend the envelope modeling from monotone flows to all flows with bounded rate $\sup_t |x_t| < M$, which is a minimal restriction in practice.

5.3. Numerical Examples with CAISO Data

The above computational procedures are implemented and applied to a set of real data for four wind farms, one solar plant, and one geothermal plant within the California ISO region. Captured in Figure 6 are one-minute power output over an one-month period in the summer (July 2009). The curves in Panel 1 are the service curves of wind, solar, and geothermal resources. The corresponding conjugates are displayed in Panel 2, where the maximum net deficit of each resource as an increasing function of capacity is displayed. The service curves are all normalized according to the resources' capacity factor (i.e., the overall mean). In particular, the diagonal with a slope 100% corresponds to the service curve of a reference constant generator. In this way, the service curve on X-axis becomes dimensionless; whereas Y-axis has unit of hour interpreted as virtual delay. Therefore, the deviation of the service curves from the diagonal represents a scale-free QoS aspect of the resource, which consequently manifests the dependence of the intermittent resources on the system as backup.

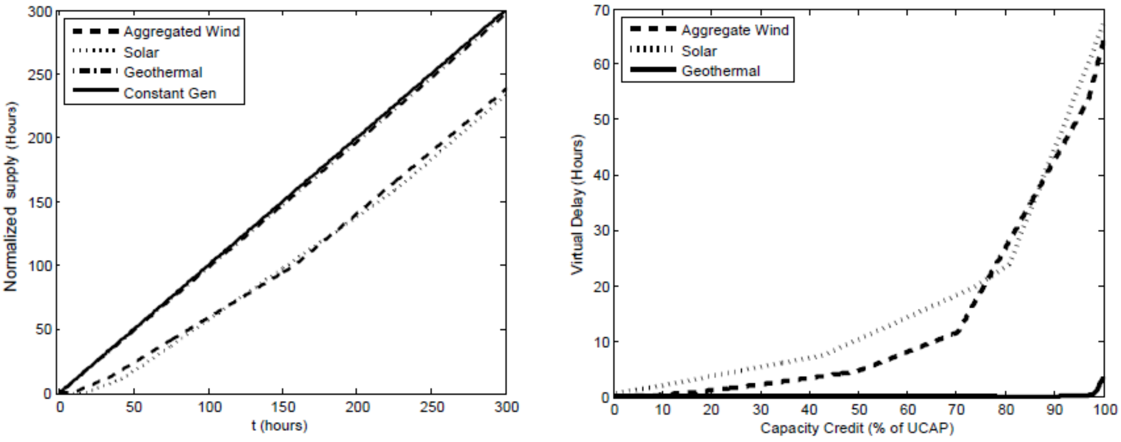


Figure 6 Lower Envelopes of Wind, Solar and Geothermal and their conjugates

In parallel, we can also construct the upper-envelope of the system load to study the demand pattern for capacity. Figure 7 below illustrates some of the numerical results. The solid line corresponding to the envelope based on the summer peak period from June-September, and the dotted curve uses the annual data. From the result, we observe that the summer curve overlaps with the annual curve. The curves demonstrates expected properties including concavity, and convergence of the slopes in the large time-scale regime.

By no means can upper-enveloping only be applied to load or lower-enveloping applied to supply. In fact, Figure 8 below compares the conjugate upper- and lower-envelopes of the system load and the net load due to wind power injection, which reveals valuable information on system impact of wind. In particular, the shift of right wing of the conjugate curve quantifies the reduction of the system peak load due to wind power, which manifests wind's capacity contribution to the system; in the meantime, the shift of left wing quantifies the reduction of the system base load, which unfortunately confirms the negative impact on the system.

What we did not anticipate is that the net load has a higher maximum deficit than the load; Moreover, the left- and right- wings of both loads are quite symmetrical, both around the mean and the wings.

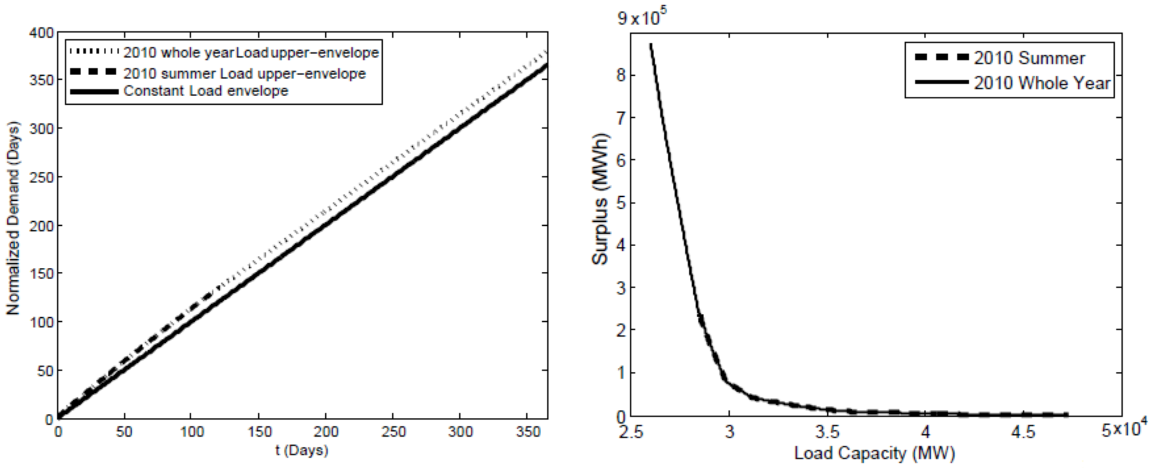


Figure 7 Upper envelope of the Load and its corresponding conjugate curvs

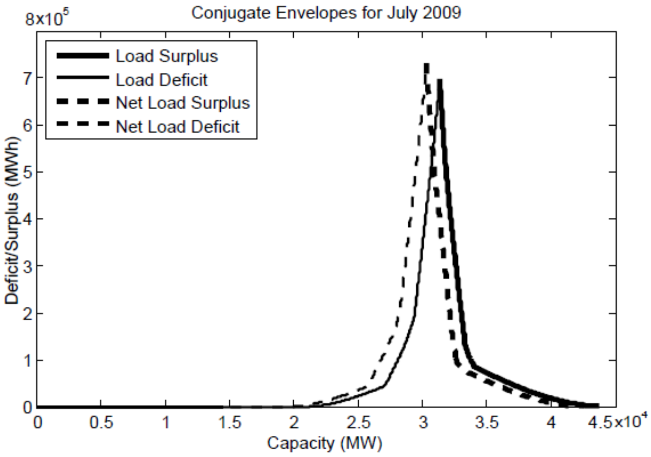


Figure 8 Conjugate envelopes of Load and Net Load for July 2009 (where the Net Load is Load Less the Aggregated Wind.)

It is natural to question at this point if the envelope-based characterization is robust across multi-year periods. With the apparent annual cycle in mind, we examined wind data of Julies from 2007 to 2010, and the results are plotted in Figure 9.

We make several additional comments on the numerical results exhibited in Figures 6-9.

1. The upper envelopes and their respective conjugates are concavely increasing and convexly decreasing as expected. It is also verified that the slope at the original corresponds to the peak (i.e., the max) demand, whereas the slopes at the large scale approaches the capacity factor (i.e., the mean), a fact becomes most apparent from the 300-hour time scale onwards.

2. Unsurprisingly, geothermal out is nearly constant. This is explicitly demonstrated by the small deviation of its lower-envelope from the diagonal.

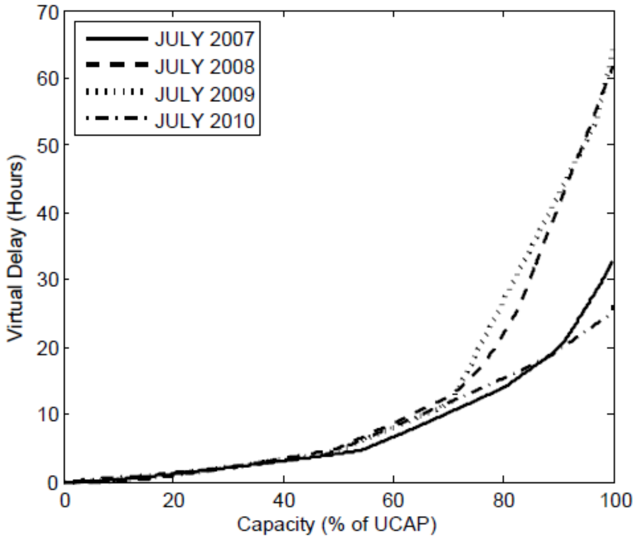


Figure 9 August Wind capacity pattern

3. For wind power generation, the big gap between its lower-envelope and the diagonal indicates the capacity factor is a significant overestimate for intermittent resources from a QoS point of view.

4. The solar envelope starts with an absolute zero-capacity period of approximately 17 hours, attributable mostly to the diurnal effect. The same diurnal effect induces a strong correlation between peak system load and solar output which supposedly warrants solar a favorite capacity rating. Putting these two seemingly contradictory thoughts side-by-side, we conclude that the definition of capacity should focus on the matching performance of supply and demand instead of just supply or demand in isolation. In other words, the goal is to focus on service rather than specific resources or equipment. This view point lays the conceptual foundation for developing quality-of-service based definition of capacity metrics to be presented in the next section.

5. We observe that the lower-envelopes (normalized by capacity factor) for the wind supply cross multiple years exhibits a consistent pattern. In fact, they are almost on top of each other up to 40% of the capacity factor, which corresponds to a time scale of 10 hours). Robustness is also observed for system load across multiple years.

In summary, numerical examples suggest that an envelope-based characterization provides a robust representation of QoS pattern of intermittent sources. Further analysis of these envelopes, more importantly, on the relationship between various envelopes, will extract critical information about intermittency and lead to ways to assess the contribution and impact of intermittent generating resources to the system.

6. Capacity Modeling

While the tradeoff between capacity and QoS performance can be quantified through the envelopes or equivalently, the corresponding conjugates, a question remains: what exact capacity value be assigned to a generating resource? We respond to this question in Subsection 6.1. In Subsection 6.2, we establish a direct connection between QoS performance and storage which enables us to confirm the significance of storage as a capacity resource and quantify its impact.

6.1. Guaranteed vs. Best-Effort Capacity

We respond to the question of capacity valuation with explicit definitions of two QoS-based capacity metrics: the guaranteed capacity, and a weak sense: the best-effort capacity.

Definition 1 (**QoS-Based Capacity Metrics**). Assume the supply of a resource is Y with a lower-envelope β_Y, and the system load on the same period X has an upper-envelope α_X. Without loss of generality, let $X \geq Y$. We call $(X - Y)$ the net load and as a convention, denote its upper-envelope as α_{X-Y}. Recall $C_\mu(t) = \mu t$ is a linear function with slope μ.

(i) (Guaranteed Capacity) Define $C^G := \max\{\mu | X - Y \leq X - \mu\}$ as the guaranteed capacity of resource Y.

(ii) (Best-Effort Capacity): Define $C^{BE} := \max\{\mu | \alpha_{X-Y} \leq \alpha_{X-C_\mu}\}$ as the best-effort capacity of resource Y given the load and the net load have upper-envelopes α_X and α_{X-Y}, respectively.

Remark 3 (**Meanings of QoS-Based Capacity Metrics**). We call C^G the guaranteed capacity due to the fact that the dominance of Y over C_μ is moment-to-moment: at any epoch, the resource generates no less electricity than a constant generator at level C^G. It is easy to see that $C^G := \max\{\mu | X - Y \leq X - \mu\} = \max\{\mu | \mu \leq Y\} = \min_t Y(t)$. In other words, the guaranteed capacity defined as such is nothing but the minimum value of the supply flow Y. Alternatively, we also have $C^G = \sup\{\lambda | \underline{\mathcal{L}}[\beta_Y](\lambda) = 0\} = \beta_Y'(0)$.

Differing from the best-effort capacity, the guaranteed capacity is an intrinsic property, i.e., it is independent to the load or other resources. It is easy to see that the best-effort capacity is no less than the guaranteed capacity, i.e., $C^{BE} \geq C^G$. In particular, if the load is a constant, $X = C_\mu$, then $C^{BE} = C^G = \min Y$.

The essence of the best-effort capacity is to evaluate the capacity contribution as reduction of the system load. In particular, we have

$$\begin{aligned} C^{BE} &:= \max_\mu\{\mu | \alpha_{X-Y} \leq \alpha_{X-C_\mu}\} \\ &= \max_\mu\{\mu | \overline{\mathcal{L}}[\alpha_{X-Y}] \leq \overline{\mathcal{L}}[\alpha_{X-C_\mu}]\} \\ &= \max_\mu\{\mu | \overline{\mathcal{L}}[\alpha_{X-Y}](\lambda) \leq \overline{\mathcal{L}}[\alpha_X](\lambda - \mu) \text{ for all } \lambda \geq 0]\}. \end{aligned} \tag{19}$$

The last equality shows, seeing from the conjugate domain, reduction of the load by a constant generator corresponds a horizontal translation of the conjugate curve towards the left. By definition, the best-effort capacity corresponds to the maximal horizontal translation until the shifted load curve touches the curve of the net load $(X - Y)$. Figure 10 illustrates the intuition behind the best-effort capacity using the aggregated wind data for a typical mid-week day on July 1st of 2009. It is worth noting that the best-effort capacity is achieved when the loads are evaluated on a positive time scale, which shows the information captured by envelopes at time scales away from both zero and infinity can be and indeed are relevant to capacity evaluation.

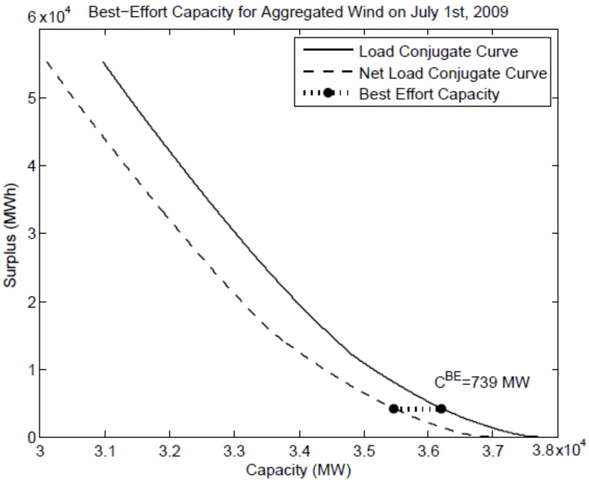


Figure 10 The Best-Effort capacity and the associated Load and Net-Load Conjugates for July 1st, 2009

In comparison with guaranteed capacity, best-effort capacity does not guarantee that the intermittent resource outperforms a constant generator at any specific moment. Instead, it requires dominance between their respective envelopes. This is a weaker requirement that implies a potentially more generous assignment of capacity to intermittent resources.

Formal distinction of these two capacity metrics provides a meaningful and quantitative explanation for why wind reasonably can and should be assigned with a higher capacity value than what it can guarantee in advance. In particular, wind can have a positive capacity value even if its minimum power output is zero. Intuitively, this situation occurs when the minimum wind power supply does not overlap with the time when system peak load occurs. As a consequence, the unavailability of wind power during non-critical periods does not necessarily cause problem to system resource adequacy. Thus, in this circumstance, wind is not penalized with a zero capacity rating under the best-effort capacity definition. It is clear at this point the best-effect capacity represent an *ex post* perspective that is shared by all existing metrics such as ELCC and time-based

UCAP variants. On the contrary, guaranteed capacity represents an *ex ante* perspective, which reflects the insurance/guarantee nature of capacity.

The availability of CAISO data enables us to determine the best-effort capacity of individual or aggregated resources based on the gap between the load and net load envelopes. As an illustration, we compute the daily best-effort capacity of wind and solar. For daily capacity, we use all of the one-minute data in a day to construct the load and net load curves, from which to compute their respective (best-effort) capacity values for that day.

Practical insights can be gained by cross-examining multiple intermittent sources with different metrics. We pick a typical day in the summer, here July 1, 2009, and then calculate the best-effort capacity values for the aggregated wind and for solar as 739.36 and 251.42 MW, respectively. The 24-hour mean (capacity factor) is 1056.23 MW and 147.87 MW. OVer the peak period from 2pm to 6 pm, the max, min, mean and 70% exceedance are respectively 908.20, 552.39, 722.99, and 660.33 MW for wind, and 338.18, 271.28, 311.74, and 305.14 MW for solar.

A summary comparison is provided in Table 2.

Table 2 Mean Daily Wind and Solar Capacity (in MW) under Multiple Metrics (Years 2007-2010)

W, S	**BE**	**UCAP**	**PK UCAP**	**PK 70% Exceedance**	**PK Max**	**PK Min**
2007	629, 176	779, 116	688, 278	449, 270	1581, 346	29, 0
2008	507, 189	656, 123	546, 289	308, 279	1376, 346	42, 0
2009	694, 216	872, 143	749, 287	441, 283	1735, 347	13, 0
2010	803	961	933	707	1909	8

Note that for Wind the mean daily best-effort capacity is comparable with the daily UCAP over peak hours. The histogram of the daily wind capacity during Year 2010 depicted in Figure 11 reveals further details of the strong similarity between Wind's best-effort and UCAP capacity metrics.

In contrast to Wind, Solar's best-effort capacity is closer to its 24-hour mean than to the on-peak mean, which is indeed not an accident. In fact, the computational procedure also shows that the critical time scale to determine Wind capacity is usually rather short, which indicates that the intermittency pattern of wind is most constraining at small time scales (minute-hour). On the other hand, the intermittency of solar is most-critical at larger scales (hour-day), which coincides with industry experience and commonsense. Such a difference makes Solar and Wind good candidates for resource bundling, meaning they can complement one another over different time scales. As an energy resource, Wind does well over large time scales. With strong correlation between solar resources and the system load, Solar acts as a good capacity resource over small time scales which are most likely to occur during peak hours on sunny days.

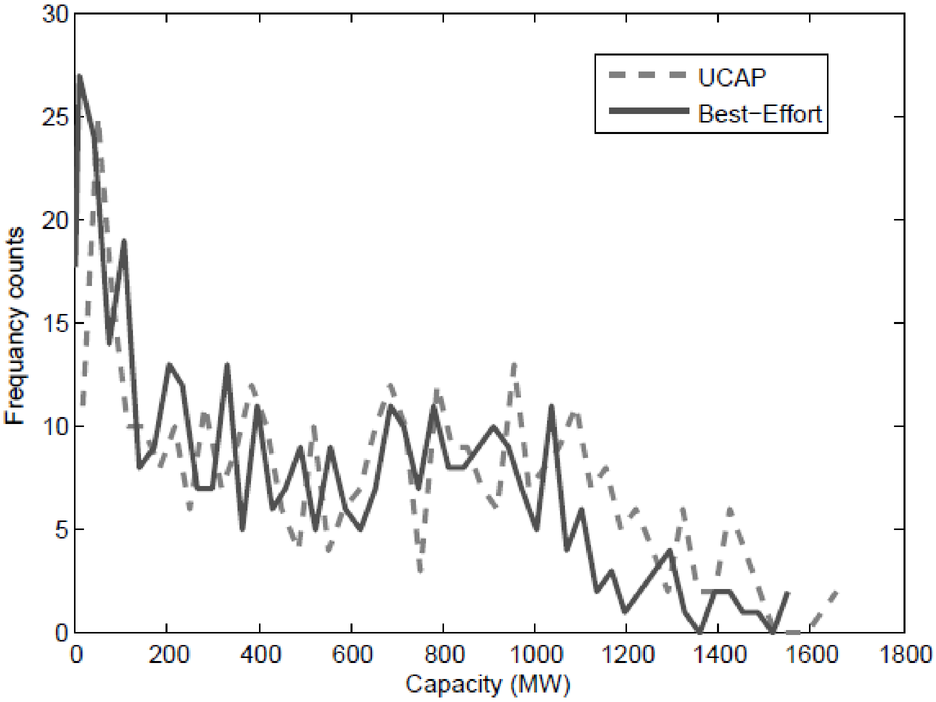


Figure 11 Histograms of the On-Peak UCAP and the Best-Effort capacity in Year 2010

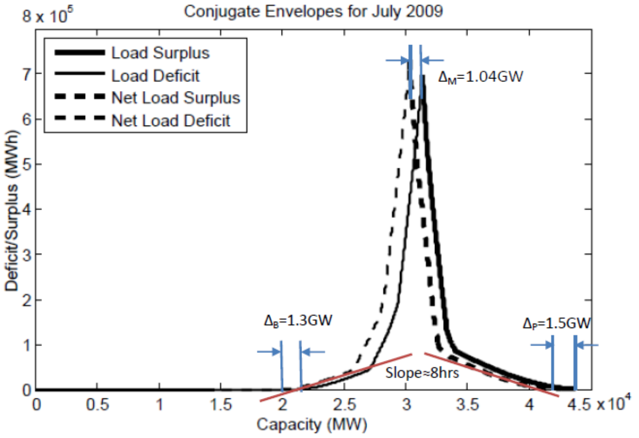


Figure 12 Impact of Wind as presented by comparison between Load and Net Load conjugates

6.2. Storage as A Capacity Resource

At this point, it should be clear that conjugate envelopes can be understood from a new perspective – storage. We now re-examine Figure 8 from a storage perspective. To set the stage, the left-shift of the peaks in the plot shows that the integration of wind generation displaced fossil-based generation by $(UCAP_L - UCAP_{NL}) * 24h/d * 31d = 776$ GWh during a one-month period (in July

2009). In addition, by looking at the the two-tails, we find that the minimum size of ideal storage to maintain the exact minimum baseload of the load is about 3.8GWh, which can be utilized during peak hours to reduce the peak load capacity to 40.9 GW from 42.5 GW (for net load) or 43.8 GW (for the load). Given that a typical utility-scale pumped storage unit, for example HELM in California, has both pumping and generating capacity of about 1GW, the above baseload and peakload guarantees require less than two pumped storage unit of the capacity of HELM (max(42.5-40.9,21.4-20.1)=1.6GW), each of which needs to be charged (or discharge) over a period lasting no more than 2 hours. This time scale that fits well within daily storage operating cycles. Figure 12 illustrates the discussion.

Also, the slope at the two tails correspond to approximately 8 hours, which corresponds to intra-day storage operation. Effectively, this shows that by performing intra-day load shift operations, adequate storage can effectively squeeze the net load down to a range of 25-33 GW as compared to the original 20-44 GW range. In summary, the above numerical example clearly demonstrates the benefit of storage resources for capacity management, and the great potential for further storage deployment.

7. Concluding Remarks

In this paper, we present an envelope-based modeling methodology to analyze the QoS and capacity value of intermittent renewable energy resources. The envelope method takes off from a worst-case analysis stand point and sets the target on QoS performance guarantees. This approach fits the extremely high reliability standards in the power industry.

The proposed envelope method aims at unifying the capacity modeling methods for generation and storage resources on the supply side and the load from the demand side based on a different perspective. Rather than focusing on the resource itself, we suggest that it is the service that a resource provides or the load consumed that is more important. This perspective provides a solid foundation to the notion of quality of service, which quantifies variability of supply and demand on multiple time scales and characterizes how well supply and demand are (mis-)matched. Objective and fair evaluation of capacity value can thus be derived from the QoS implications.

On the methodological front, the envelope method grows out of the theory of network calculus (NetCal). The sole application of NetCal to date has been in the telecommunications field. By switching the NetCal models from the time domain to the conjugate domain via the Legendre transform, and by replacing application-specific terminologies (most representatively, arrival and service curves) to generic vocabularies such as upper- and lower-envelopes, the Legendre-transform based envelope modeling method enables us to switch back and forth between domains in order to find the easiest way to discover and demonstrate desired results.

Different capacity metrics have been used to characterize specific aspects of the power system. In particular, ICAP is used for capacity planning/fixed cost evaluation, UCAP for evaluating the average production capability, and ELCC for reliability assessment. For traditional generating resources, e.g., a fossil fuel based generator, its ICAP, UCAP and ELCC enjoy relatively simple quantitative relations. For instance, UCAP can be understood as ICAP subject to a discounting factor, which is basically a reliability measure, and it typically ranges between 80-90%. This factor can be further used to derive the ELCC at the aggregated level using analytical or simulation methods.

However, the close quantitative connection between these metrics no longer holds due to major technological advances that have taken place in recent years. For example, the growth in wind and solar resources introduces high variability in power supply and complicates the relationship between ICAP, UCAP and ELCC. There is also a large expected growth of storage resources, both at the household level due to the maturity of electric vehicle technology, and at the grid level because of new technologies such as flow batteries as well as existing pumped storage technology which can now be built underground, dramatically increasing the number of places that can deployed the technology. Capacity models must therefore include storage resources, which is beyond the scope of the current metrics.

IT-based technologies, discussed under the broad title of “smart-grid” will enable more sophisticated communication and control between the supply and demand, and thus an increased elasticity from the demand side that will impose additional challenges to capacity modeling on the demand side. The envelope method establishes a direct linkage between capacity modeling and optimal planning and operation of various resources. This linkage brings together evaluation, operation/management and compensation of resources in a consistent manner.

Appendix

Proof of Proposition 2.

(i) By Cauchy test of convergence, we have $\forall \epsilon > 0,\ \exists N(\epsilon) > 0$, such that $\forall i > j \geq N(\epsilon)$

$$-\epsilon/2 < \frac{X_i}{i} - \frac{X_j}{j} < \epsilon/2;\ \frac{X_i}{i} < \mu + \epsilon/2. \tag{20}$$

Therefore,

$$\begin{aligned} & \tfrac{X_i}{i} - \tfrac{X_j}{j} < \epsilon/2 \\ \Longleftrightarrow\ & \tfrac{\Sigma_{k=j+1}^{i} x_k + X_j}{i} < \epsilon/2 + \tfrac{X_j}{j} \\ \Longleftrightarrow\ & \Sigma_{k=j+1}^{i} x_k/i < \epsilon/2 + X_j[1/j - 1/i] \\ \Longleftrightarrow\ & \Sigma_{k=j+1}^{i} x_k < i\epsilon/2 + (i-j)X_j/j. \end{aligned}$$

We now estimate $\Sigma_{k=j+1}^{i} x_k - (\mu+\delta)(i-j)$, where

$$\begin{aligned} \Sigma_{k=j+1}^{i} x_k - (\mu+\delta)(i-j) &< i\epsilon/2 + (i-j)[X_j/j - (\mu+\delta)] \\ (\Sigma_{k=1}^{i} - \Sigma_{k=1}^{j})x_k - (\mu+\delta)i + (\mu+\delta)j &< i\epsilon/2 + (i-j)[X_j/j - (\mu+\delta)] \\ \Sigma_{k=1}^{i} x_k - i(\mu+\delta) &< i\epsilon/2 + (i-j)[X_j/j - (\mu+\delta)] \\ &\quad + (\Sigma_{k=1}^{j} x_k - j(\mu+\delta)) \\ &< i\epsilon/2 + (i-j)[(\mu+\epsilon/2) - (\mu+\delta)] \\ &\quad + j((\mu+\epsilon/2) - (\mu+\delta)) \\ &< i[\epsilon - \delta] < 0 \text{ for all } \epsilon < \delta. \end{aligned}$$

Therefore, we have

$$\begin{aligned} & \sup_{N(\epsilon)\leq j\leq i}(\Sigma_{k=j+1}^{i} x_k - (\mu+\delta)(i-j)) < 0 \\ \Longrightarrow\ & \sup_{1\leq j\leq i}(\Sigma_{k=j+1}^{i} x_k - (\mu+\delta)(i-j)) = \overline{L}[\alpha_X](\mu+\delta) < \infty. \end{aligned}$$

(ii) The concavity of the upper envelope function α_X is rooted from the concavity of affine functions $\lambda t + \overline{L}[\alpha_X](\lambda)$ that generates α_X, and the fact the minimum of concave functions is still concave.

As for the convexity of the conjugate envelope $\overline{L}[\alpha_X](\lambda)$ in λ, it can be easily shown by the inexchangeability between operators "sup" and "+": $\sup_{n\geq m\geq 1}\{(\lambda_1+\lambda_2)/2 - X(m,n)\} \leq \sup_{n\geq m\geq 1}\{\lambda_1 - X(m,n)\}/2 + \sup_{n\geq m\geq 1}\{\lambda_2 - X(m,n)\}/2$. As for the monotonicity of $\overline{L}[\alpha_X](\lambda)$ in λ, the proof is trivial. According to Eq. (8), it leads to desired concavity and mononotinicity of α_X as claimed in (iii).

(iv) By Property (i), for an arbitrary small $\delta > 0$, $\overline{L}[\alpha_X](\mu+\delta) < \infty$. By the definition of α_X, $\mu t \leq \alpha_X(t) \leq (\mu+\delta)t + \overline{L}[\alpha_X](\mu+\delta)$. Therefore, it is easy to see that there exists a finite $t_1 > 0$ such that for all $u > v > t_1$, $\mu < \alpha_X(u)/u, \alpha_X(u)/u < (\mu + 2\delta)$.

By the concavity of α_X shown in Property (ii), $\alpha_X(u)/u \leq (\alpha_X(u) - \alpha_X(v))/(u-v) \leq \alpha_X(v)/v$. It is thus clear that

$$\mu \leq \frac{\alpha_X(u)}{u} \leq \frac{\alpha_X(u) - \alpha_X(v)}{u-v} \leq \frac{\alpha_X(v)}{v} \leq \mu + 2\delta \tag{21}$$

Consequently, we have

$$\begin{aligned}\mu &\leq \alpha_X(u)/u \\ &\leq \underline{\lim}_{u\to v} \tfrac{\alpha_X(u)-\alpha_X(v)}{u-v} = \lim_{u\to v+} \tfrac{\alpha_X(u)-\alpha_X(v)}{u-v} \\ &\leq \lim_{v\to u-} \tfrac{\alpha_X(u)-\alpha_X(v)}{u-v} = \overline{\lim}_{v\to u} \tfrac{\alpha_X(u)-\alpha_X(v)}{u-v} \\ &\leq \alpha_X(v)/v \leq \mu + 2\delta.\end{aligned} \tag{22}$$

Therefore, we have

$$\underline{\frac{\mathrm{d}\alpha_X}{\mathrm{d}t}}(t) = \frac{\mathrm{d}_+\alpha_X}{\mathrm{d}t}(t) = \frac{\mathrm{d}_-\alpha_X}{\mathrm{d}t}(t) = \overline{\frac{\mathrm{d}\alpha_X}{\mathrm{d}t}}(t) = \underline{\lim}_{t\to\infty} \alpha_X(t)/t = \mu, \tag{23}$$

where, $\underline{\frac{\mathrm{d}\alpha_X}{\mathrm{d}t}}(\overline{\frac{\mathrm{d}\alpha_X}{\mathrm{d}t}})$ and $\frac{\mathrm{d}_+\alpha_X}{\mathrm{d}t}(\frac{\mathrm{d}_-\alpha_X}{\mathrm{d}t})$ are respectively the lower- (upper-) and right-(left-) derivatives of function α_X.

Finally, from the definition of α_X, at any given scale t, $\alpha_X(t)$ is either solely determined by one affine envelope (say, $\alpha_X(t) = \lambda t + b$), thus has a well-defined derivative $\frac{\mathrm{d}\alpha_X}{\mathrm{d}t}(t) = \lambda$, the leaking rate, or is on the intersection of two affine envelopes, whose slopes gives the right- and left- derivatives at point t. Therefore, independent to differentiability, the left- and right- sided derivatives of all points converge to μ when it approaches infinity. As a side trace, where will be maximally countable many non-differentiable points for α_X. ■

Proof of Proposition 3.

Claim (i) is a direct application of the "translation" property of Legendre Transform:

$$\overline{\mathcal{L}}[\alpha_{X+C_\mu}](\lambda) = \overline{\mathcal{L}}[\alpha_X](\lambda+\mu), \tag{24}$$

whose proof is trivial. More specifically, for any $\tilde{R} = R + \delta > R \geq \sup_{t\geq 0} g_t$,

$$\begin{aligned}&\overline{\mathcal{L}}[\alpha_{C_{\tilde{R}}-G}](\tilde{R}-\lambda) \\ &= \overline{\mathcal{L}}[\alpha_{C_{R+\delta}-G}](R+\delta-\lambda) \\ &= \overline{\mathcal{L}}[\alpha_{C_R-G}](R-\lambda),\end{aligned}$$

which confirms the desired reference independence.

Claim (ii) can be obtained by explicitly expand $\underline{\mathcal{L}}[\beta_G](\lambda)$ as follows

$$\begin{aligned}&\underline{\mathcal{L}}[\beta_G](\lambda) = \overline{\mathcal{L}}[\alpha_{C_R-G}](R-\lambda) \\ &= \sup_{t\geq 0}\{\overline{Q}^{C_R-G}_{R-\lambda}(t)\} \\ &= \sup_{s\leq t}\{[C_R - G](s,t) - (R-\lambda)(t-s)\} \\ &= \sup_{s\leq t}\{\lambda(t-s) - G(s,t)\}.\end{aligned}$$

As a consequence, for all $s \leq t$, $G(s,t) \geq \lambda(t-s) - \underline{\mathcal{L}}[\beta_G](\lambda)$. Nonnegativity of g_t ensures $G(s,t) \geq [\lambda(t-s) - \underline{\mathcal{L}}[\beta_G](\lambda)]^+ =: \beta_{G,\lambda}$, which consequently qualifies $\beta_{G,\lambda}$ as a lower envelope. In this way, we have constructed a family of lower envelopes for flow G, parameterized in rate $\lambda \geq 0$. Finally, by piecing them all together, we obtained a well-defined lower envelope β_G as follows:

$$\begin{aligned}\beta_G(t) &= \sup_{0\leq\lambda} \beta_{G,\lambda}(t) \\ &= \sup_{0\leq\lambda\leq R}[C_\lambda(t) - \underline{\mathcal{L}}[\beta_G](\lambda)] \\ &= C_R(t) - \inf_{0\leq\lambda\leq R}[C_R(t) - C_\lambda(t) + \underline{\mathcal{L}}[\beta_G](\lambda)] \\ &= [C_R - \alpha_{C_R-G}](t). \quad ■\end{aligned} \tag{25}$$

References

Aıd, R., G. Chemla, A. Porchet, N. Touzi. 2011. Hedging and vertical integration in electricity markets. *Management Science* **57**(8) 1438–1452.

Ata, B. 2005. Dynamic power control in a wireless static channel subject to a quality-of-service constraint. *Operations Research* 842–851.

Bassamboo, A., R.S. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research* **58**(5) 1398–1413.

Bose, A., X. Jiang, B. Liu, G. Li. 2006. Analysis of manufacturing blocking systems with network calculus. *Performance Evaluation* **63**(12) 1216–1234.

Chang, C.S. 2000. *Performance guarantees in communication networks*. Springer-Verlag New York Inc.

Cruz, R.L. 1991a. A calculus for network delay. I. network elements in isolation. *Information Theory, IEEE Transactions on* **37**(1) 114–131.

Cruz, R.L. 1991b. A calculus for network delay. II. network elements in isolation. *Information Theory, IEEE Transactions on* **37**(1) 114–131.

Gross, R., UKERC (Organization). 2006. *The Costs and Impacts of Intermittency: An assessment of the evidence on the costs and impacts of intermittent generation on the British electricity network*. UK Energy Research Centre.

Hobbs, B.F., JS Pang. 2007. Nash-cournot equilibria in electric power markets with piecewise linear demand functions and joint constraints. *Operations Research* **55**(1) 113–127.

Holttinen, H., VTT Bettina Lemström, F.P. Meibom, H. Bindner, et al. 2007. Design and operation of power systems with large amounts of wind power. *State-of-the-art report 2007* .

Jiang, X. 2008. New perspectives on network calculus. *ACM SIGMETRICS Performance Evaluation Review* **36**(2) 95–97.

Jiang, X, G. Parker. 2012. Modeling the bullwhip effects with network calculus. Working paper.

Joskow, P. 2005. Vertical integration. *Handbook of New Institutional Economics* 319–348.

Joskow, P. 2006. Markets for power in the united states: An interim assessment. *Energy Journal* **27** 1–36.

Joskow, P.L., D.R. Bohi, F.M. Gollop. 1989. Regulatory failure, regulatory reform, and structural change in the electrical power industry. *Brookings papers on economic activity. Microeconomics* **1989** 125–208.

Kamat, R., S.S. Oren. 2002. Exotic options for interruptible electricity supply contracts. *Operations Research* 835–850.

Le Boudec, J.Y., P. Thiran. 2001. *Network calculus: a theory of deterministic queuing systems for the internet*. springer-Verlag.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* 242–262.

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations research* **57**(5) 1189.

Milligan, M., K. Porter. 2006. The capacity value of wind in the united states: Methods and implementation. *The Electricity Journal* **19**(2) 91–99.

Milligan, M., K. Porter. 2008. Determining the capacity value of wind: An updated survey of methods and implementation. *Wind Power*.

NERC. 2011. Methods to model and calculate capacity contributions of variable generation for resource adequacy planning. *North American Electric Reliability Corporation*. Available at http://www.nerc.com/files/IVGTF1-2.pdf.

Powell, S.G., S.S. Oren. 1989. The transition to nondepletable energy: social planning and market models of capacity expansion. *Operations research* 373–383.

Rockafellar, R.T. 1970. *Convex Analysis, volume 28 of Princeton Mathematics Series*. Princeton University Press.

Zhao, J., B.F. Hobbs, J.S. Pang. 2010. Long-run equilibrium modeling of emissions allowance allocation systems in electric power markets. *Operations research* **58**(3) 529–548.