Curtailing Intermittent Generation in Electrical Systems

Owen Q. Wu
Stephen M. Ross School of Business
University of Michigan

Roman Kapuscinski
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1170
October 2011

Curtailing Intermittent Generation in Electrical Systems

Owen Q. Wu Roman Kapuscinski

Stephen M. Ross School of Business, University of Michigan, Ann Arbor, Michigan

October 14, 2011

Intermittent energy generation from renewable sources introduces additional variability into electrical systems, resulting in a higher cost of balancing against the increased variabilities. Ways to balance electrical systems include the use of flexible generation resources, storage operations, and a much less-used option: curtailing intermittent generation. This paper focuses on the impact of curtailing intermittent generation on the system's cost. We construct a model that captures the most critical components of the balancing cost in electrical systems. We find that curtailing intermittent generation may result in unexpected economic behaviors.

1. Introduction

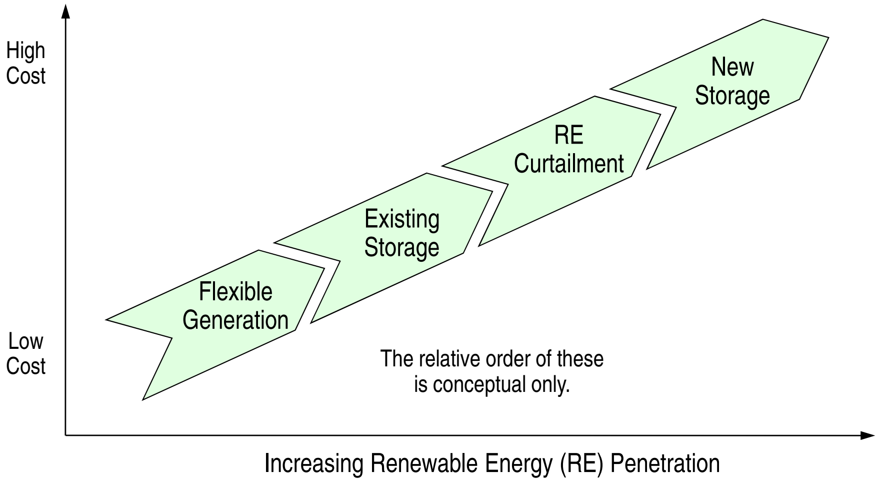
To reduce the environmental impact of electrical power generation, renewable energy sources increasingly have been integrated into the electrical systems. In the U.S., from 2002 to 2009, renewable energy generation capacity increased by 31 gigawatts and 96% of this growth was contributed by wind power (Energy Information Administration 2010). Wind power and other renewable sources such as solar and tidal power are intermittent, however, and bring additional variability into the electrical systems. For example, in regions operated by Midwest Independent System Operator (MISO), the actual wind power averages 3.1 gigawatts from January to April 2011 with a standard deviation of 1.7 gigawatts (based on the hourly wind data from MISO, see details in §6).

The intermittency of wind power and other renewable sources brings significant challenges to balance the electrical systems. Unlike the production-inventory systems in other manufacturing environments, demand and supply in the electrical systems must be constantly balanced. Figure 1 conceptually shows the available options for balancing the systems. Two commonly-used options are flexible generation and energy storage.

- Flexible generation resources are capable of adjusting their output to counteract the variabilities in demand and intermittent sources. Flexibility, however, is costly: More flexible generation units (e.g., gas-fired units) typically have higher marginal costs. The increasing penetration of intermittent generation requires the use of more flexible generation units, thereby increasing the

Figure 1: Operational options for balancing against intermittency

Adapted from National Renewable Energy Laboratory (2008)


High Cost
Low Cost
New Storage
RE Curtailment
Existing Storage
Flexible Generation
The relative order of these is conceptual only.
Increasing Renewable Energy (RE) Penetration

total operating costs of the flexible resources, known as the system's balancing cost.

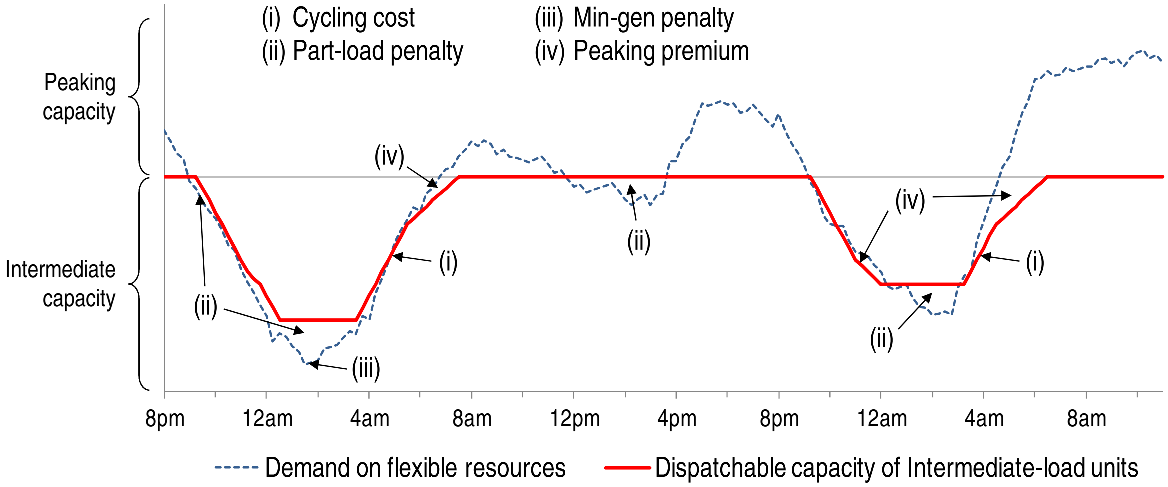
- Energy storage can buffer against variabilities and reduce the system's balancing cost, but to a limited extent at the current scale and technologies. In contrast to physical goods, electricity must be stored in other forms of energy and conversions often incur significant energy losses, e.g., the efficiency of hydroelectric pumped storage is 70-80%. Building new storage facilities is generally very expensive under the current technology.

A much less-used operational option for system balancing is curtailing intermittent generation, e.g., pitching the blades to reduce the wind turbine output when balancing against the excessive wind is too costly. In this paper we focus on understanding and quantifying the economic value of curtailment in electrical systems. In most of electricity systems, renewable energy is curtailed only in extreme situations, e.g., when excessive energy threatens system reliability (Rogers, Fink, and Porter 2010). Preliminary studies from National Renewable Energy Laboratory (Ela 2009) confirm the existence of the economic value of curtailing wind energy. Nevertheless, it seems uneconomical to refuse to take free wind energy yet produce the same amount using fossil fuels. Figure 1 suggests that renewable energy curtailment is a more expensive option than flexible generation and existing storage, because curtailment directly reduces the renewable energy penetration and increases the use of other resources. We develop a model that incorporates all three operational levers. We ask the following key research questions: When storage is absent, what drives the economic value of curtailment and is this value significant? When storage is present, would the storage operations significantly reduce the economic value of curtailment or even eliminate the need for curtailment? How much does the flexibility of the generation resources affect the value of curtailment?

To answer these questions, we need to understand the system's operating cost components that are crucial for curtailment decisions. Electricity demand is met by baseload units, intermediate-load units, and peaking units. Baseload units are inflexible and their output is assumed to be constant in the short run; intermediate-load units have some flexibility; peaking units are very flexible and expensive. Intermediate-load units and peaking units are jointly referred to as flexible resources and provide flexible generation (see Figure 1). Intermittent generation directly affects the operating cost of the flexible resources, i.e., the balancing cost.

The intermediate-load units have limited flexibility reflected by four balancing cost components: (i) *Cycling cost.* Starting up an intermediate-load unit requires extra fuel to bring the unit to operating condition and this process takes time. The solid curve in Figure 2 represents the intermediate capacity that is started and can be dispatched by the system operator to produce energy, i.e., the dispatchable capacity. (ii) *Part-load penalty.* Intermediate-load units are most efficient when producing at full capacity (i.e., all dispatchable capacity is dispatched). Operating at any lower load increases the average production cost; this extra cost is the part-load penalty. (iii) *Min-gen penalty.* In normal operating conditions, the part load should stay above a certain percentage of the dispatchable capacity, otherwise a minimum-generation (min-gen) penalty will be incurred. (iv) *Peaking premium.* The dispatchable capacity of the intermediate-load units cannot be adjusted instantaneously and, consequently, peaking units may be needed even if the demand on flexible resources is below the intermediate capacity – this occurs in the areas labeled as (iv) in Figure 2. Peaking units are much more costly to produce energy, although costs (i), (ii), and (iii) are negligible for them and they can adjust output very fast.

Figure 2: Balancing cost components of an electrical system (without storage)



The system operator balances the above four costs whether intermittent generation exists or not. For example, cycling cost can be reduced if fewer intermediate-load units cycle, but that increases either the peaking premium or the part-load and min-gen penalties. Intermittent generation resources introduce additional variability into the electrical systems. As a result, all of the four balancing cost components are likely to increase.

Curtailing intermittent generation may help mitigate the increase in the balancing cost. Curtailing intermittent generation encourages the use of intermediate-load units, reducing the amount of capacity cycling and need for peaking units. When intermediate-load units operate at very low load, curtailing intermittent generation reduces part-load and min-gen penalties. When renewable energy was initially introduced into most of countries and regions, however, the curtailment option was not considered. Because of its small initial scale and its benefit to the environment, renewable energy typically was accommodated generously by receiving priority in dispatch, as is still the case in many regions. The European Union Renewable Energy Directive (European Union 2009, Article 16) states, "Member States shall ensure that appropriate grid and market-related operational measures are taken in order to minimise the curtailment of electricity produced from renewable energy sources." An international survey of wind energy curtailment practices by Rogers et al. (2010) shows that although wind energy polices differ across countries and regions, one commonality is that in the absence of transmission constraints or grid reliability, whenever a wind generator is producing energy, the system will take the generation, effectively treating the use of wind power as a priority.

With the increasing penetration of renewable energy, studies show that intermittency causes extra system balancing cost (Swider and Weber 2007, Gross et al. 2006). Recently, Monsen and Norin (2011) surveyed wind policy changes passed or proposed in various regions. In two electricity markets in the U.S. (PJM and New York ISO), the new policy treats wind energy like other resources and requires these facilities to bid price curves indicating their willingness to be curtailed. In other areas, potential changes may offer economic incentives for wind and other intermittent sources to agree to the possibility of economic curtailment.

In this paper, we compare the economic curtailment policy with the priority dispatch policy from the perspective of the system's balancing cost. We refer to the cost reduction as the value of economic curtailment. The key findings of this paper are summarized below.

One might expect that economic curtailment helps reduce the extra balancing cost due to intermittent generation, but our results suggest a more significant impact: Economic curtailment may, in some situations, entirely offset the extra balancing cost due to intermittent generation, whether storage is present or absent. In other words, a system with intermittent generation and economic curtailment may incur a lower balancing cost than a system without any intermittent generation.

While storage operations reduce curtailment, we find it suboptimal to use the storage solely to reduce curtailment. When storage operations are jointly optimized with flexible generation and with curtailment, the value of curtailment is still significant. We also find that under medium-to-high intermittent energy penetration levels, when storage is present, the average balancing cost reduction per unit of energy curtailed is actually higher than if the storage is absent. Thus, storage operations may increase the value per unit of curtailed energy. Finally, when the intermediate capacity becomes more flexible, the value of curtailment decreases significantly, but this value increases faster as the intermittent energy penetration level increases.

The rest of the paper is organized as follows. We review the literature in §2 and construct an electricity system model in §3. Structures of the optimal policy are presented in §4 and §5. The value of curtailment is quantified for an electrical system in §6. We conclude the paper in §7.

2. Literature Review

The issue of managing the variabilities introduced by intermittent energy sources was discussed when technologies emerged to use these sources at significant scale. Sørensen (1978) was the first to examine the need for energy storage for hypothetical large wind power generators. Kahn (1979) discusses how intermittency leads to extra cost in the form of added safety capacity. Farmer, Newman, and Ashmole (1980) show that the wind power variations of time-scale ranging from several minutes to a few hours require a significant enhancement of the flexible generation resources and possibly additional storage facilities. These earlier works focus on capacity requirement to accommodate wind power and do not model the effect on balancing cost, especially those costs of the intermediate-load units.

The first series of studies that model the costs of intermediate-load units are by Grubb (1988, 1991a,b). Grubb developed a statistical approach to quantify the impact of intermittency on part-load penalty, cycling cost, and peaking premium. A minimum generation level is also modeled in his work, but no min-gen penalty is imposed because capacity adjustment speed is not modeled. The statistical method is useful in that approximation formulae for these costs can be derived, but the dynamic nature of the system is greatly simplified, as discussed in Grubb (1991b).

ILEX Energy Consulting and Strbac (2002) find that the intermittency of renewables is the single largest driver of system costs, which include the investments in additional flexible generation assets and the system balancing costs. Our study focuses on the system balancing costs. The balancing cost components in their study are similar to ours. Balancing the system in their study requires synchronized reserve provided by part-loaded intermediate-load units and standing reserve provided by peaking units and storage plant. In their study, intermittent energy is curtailed only when inflexible generation exceeds the demand. We also consider such a curtailment practice, and furthermore consider the effect of economic curtailment. Their approach is mainly simulation-based, while we analyze and solve stochastic dynamic programs.

The extra effort in balancing the system due to intermittent generation also leads to environmental concerns. Katzenstein and Apt (2009) estimate the emissions from natural gas generators used to compensate for intermittent wind and solar power. They find that, due to extra emissions from system balancing, the carbon dioxide emission reductions are likely to be 75-80% of those presently assumed by policy makers. Hutzler (2010) summarizes a few recent studies and discusses the impact of cycling on pollution and carbon dioxide emissions.

Our work is also related to the production-inventory literature that considers capacity constraint. Rocklin, Kashper, and Varvaloucas (1984) are among the first to study capacity expansion and contraction under stochastic demand processes. The key tradeoff is between having too much capacity (thus paying unnecessary capital, labor, and maintenance costs) and having too little capacity (thus meeting demand at a higher cost). A target capacity interval policy is shown to be optimal: In each period, there exists an interval, and it is optimal to make the smallest capacity change that brings the capacity into this target interval. Eberly and Van Mieghem (1997) generalize the above problem and study the multidimensional capacity expansion and contraction under uncertainty. The above works assume no inventory can be carried over periods.

Angelus and Porteus (2002) examined the capacity adjustment problem with and without inventory. The model without inventory is similar to Rocklin et al. (1984), and the target capacity interval is solved explicitly when demand distributions are independent. When inventories can be carried across periods, they show that inventory policy has a basestock structure and that inventory and capacity are economic substitutes.

The model setups in Angelus and Porteus (2002) have several similarities to the electrical systems: Demand stochastically rises and falls; capacity expansion resembles the start-up of intermediate-load units; capacity overhead cost resembles the part-load penalty; shortage cost resembles the peaking cost. There are also important differences. First, Angelus and Porteus (2002) consider only one cycle starting with zero capacity, and production is decided before demand realizes. As a result, capacity is always fully used (no part-load penalty). Second, intermediate-load units have minimum production constraints and capacity adjustment delays, which are not modeled in most of the work in the production-inventory literature. Third, for electrical systems, overproduction cannot be easily stored, storage incurs energy conversion losses, and shrinking capacity does not generate a return, whereas in Angelus and Porteus (2002), inventory can be stored without a limit or loss and shrinking capacity generates a return to the firm.

To the best of our knowledge, Davis et al. (1987) is the only work that considers a time delay between a capacity expansion decision and the time when capacity becomes available. The firm controls the rate of investment; when the cumulative investment reaches a random level with known distribution, the capacity is expanded by a given size. Thus, when the capacity will be available is also random. They show that the optimal control is to either invest at maximal speed or do nothing. We model capacity adjustment delay that reflects the characteristics of electrical systems.

3. The Model

Power generation and transmission systems are extremely complicated and different trade-offs exist on tactical and strategic levels. The key elements important for studying the impact of intermittency are on the time-scale of minutes to hours. These key elements include the balancing cost components and time delays in capacity adjustment introduced in Figure 2. In this section, we provide an overview of the model, and then describe each element in more detail.

For a given fleet of generation resources, the objective is to minimize the operating cost while meeting the electricity demand. For modeling purposes, we assume that generation units are characterized by three levels of flexibility. Baseload generation units are inflexible – we assume they are generating energy at a constant level and thus the baseload production cost is sunk. Intermediate-load units have limited flexibility reflected by the costs introduced in Figure 2, which are described in §3.1. Intermediate-load units cannot instantaneously be started up or shut down: During the startup process, capacity gradually becomes dispatchable; similarly for the shutdown process. These time delays in capacity adjustment are modeled in §3.2. The most flexible units are peaking units: Their output can be costlessly and instantaneously adjusted, but they are very expensive. The system cost structure is described in §3.3. The models for electricity demand, intermittent generation, and economic curtailment are presented in §3.4 and §3.5.

The problem is formulated as a Markov decision process in §3.6 and is previewed below. Let $t \in \{0, 1, \ldots, T\}$ index periods, with each period representing the time between two successive dispatch instructions sent by the system operator. The industry standard for this time interval is 15 minutes. The states of the system include the level of intermittent generation, electricity demand, dispatchable capacity, pending-up (initiated to start but not dispatchable yet) and pending-down (initiated to shut down but still dispatchable) capacity of the intermediate-load units. Every period, the system operator decides how much capacity to start up or shut down, how much energy should be produced by each type of flexible resources, and how much intermittent energy should be curtailed if any. The objective is to minimize the system's total cost.

3.1 Flexible Resources: Peaking Units and Intermediate-Load Units

We assume the system has many identical peaking units, which are typically combustion turbines (e.g., single-cycle gas-fired units). Let c^P denote the production cost per unit of energy from the peaking units. Producing Q^P units of energy from the peaking units costs $c^P Q^P$.

The system also has many identical intermediate-load units. Below, we define the cost components for an individual intermediate-load unit and then derive the aggregate cost.

For an individual unit, let κ denote the maximum output per period, or unit's capacity. A typical unit contains a steam turbine that requires the unit to operate above a certain minimum generation level $\alpha\kappa$, with $\alpha \in (0,1)$. Reducing the output rate below $\alpha\kappa$ but not shutting down the unit will cause damage to the unit and may lead to serious consequences. For a natural gas combined-cycle unit, α is typically 40% to 50%. Let $c(q) > 0$ denote the cost of producing $q \in [\alpha\kappa, \kappa]$ per period. When the unit is up, the output rate q can be costlessly adjusted within $[\alpha\kappa, \kappa]$. For analytical convenience, we extend the definition of $c(q)$ to the region $q < \alpha\kappa$ and impose a penalty cost, p, per unit of output below $\alpha\kappa$. This represents the min-gen penalty introduced in Figure 2. Specifically,

$$c(q) \overset{\text{def}}{=} c(\alpha\kappa), \qquad \text{for } q \in [0, \alpha\kappa), \tag{1}$$

and the total production cost per period of an intermediate-load unit is

$$c(q) + (\alpha\kappa - q)^+ p, \qquad \text{for } q \in [0, \kappa]. \tag{2}$$

Assumption 1 *(i) $c(q)$ is non-decreasing and convex in q, for $q \in [0,\kappa]$; (ii) $c(q)/q$ decreases in q, for $q \in (0,\kappa]$; (iii) $c^P > c'(\kappa)$.*

The convexity in part (i) can be verified in practice and is typically assumed in the literature, e.g., Lu and Shahidehpour (2004) use convex quadratic functions to model the cost of combined-cycle units. Part (ii) assumes declining average cost in output, i.e., operating at the full load κ is the most efficient (least average cost), and operating at any load below κ results an increase in the average cost, which is the part-load penalty introduced in Figure 2. For a combined-cycle unit, Boyce (2010) shows that the average cost increases by about 17% when the unit operates at 40% load. Part (iii) says that peaking units have a higher marginal cost than the intermediate-load units (note that, due to convexity, $c'(\kappa)$ is the highest marginal cost of the intermediate-load unit).

Because of the min-gen penalty, operating an intermediate-load unit at very low load is costly and, therefore, during the low-demand periods, it is more efficient to shut down some units and start them up later. Wear and tear costs are incurred during the shutdown and startup processes. The startup process also requires extra fuel to warm up the turbine and bring the unit to the normal working condition. These costs are referred to as the cycling cost of the intermediate-load unit. We denote the cycling cost per unit of capacity per cycle by c^s.

We now derive the aggregate cost of the intermediate-load units. When n identical intermediate-load units are fully started up, we refer to $K = n\kappa$ as the *dispatchable capacity.* To achieve a given output rate $Q^I \leq K$, it is optimal to let all n units be equally loaded, because each individual unit's cost in (2) is convex in q. Therefore, the minimum total production cost per period is:

$$n\, c(Q^I/n) + n(\alpha\kappa - Q^I/n)^+ p \;\equiv\; C(Q^I, K) + (\alpha K - Q^I)^+ p, \tag{3}$$

where we define $C(Q^I, K) \overset{\text{def}}{=} n\, c(Q^I/n)$ for $K = n\kappa$. Following from (1), we have $C(Q^I, K) \equiv C(\alpha K, K)$ for $Q^I < \alpha K$. Thus, the total production cost in (3) decreases first and then increases in Q^I, with the minimum at $Q^I = \alpha K$. For analytical convenience, we allow $n = K/\kappa$ to be a positive real number and generalize the definition for $C(Q^I, K)$:

$$C(Q^I, K) \overset{\text{def}}{=} \frac{K}{\kappa} c\Big(\frac{Q^I}{K}\kappa\Big), \qquad \text{for } Q^I \leq K,\; K > 0. \tag{4}$$

The following lemma describes the properties of $C(Q^I, K)$ with the proof in the online supplement.

Lemma 1 *$C(Q^I, K)$ is increasing in Q^I and K, and jointly convex in (Q^I, K).*

Let $c^I = c(\kappa)/\kappa$ denote the average production cost of the intermediate-load units when they operate at full load. The monotonicity in Lemma 1 implies that $K = Q^I$ minimizes the cost of producing Q^I per period and that $K > Q^I$ leads to inefficiency known as the part-load penalty:

$$\text{Part-load penalty} \;=\; C(Q^I, K) - c^I Q^I, \qquad \text{for } K > Q^I. \tag{5}$$

At the aggregate level, shutting down and starting up intermediate capacity of size Δ incurs a cycling cost of $c^s\Delta$.

3.2 Capacity Adjustment

Adjusting the dispatchable capacity of intermediate-load units not only incurs cycling costs but also takes time. A typical natural gas combined-cycle unit can be fully started up in about 90 minutes. Since each period in our model represents 15 minutes, the startup process takes several periods.

Let K^I denote the total capacity of all intermediate-load units and let $K_t \in [0, K^I]$ denote the dispatchable capacity in period t. We assume that if capacity of size Δ_t^u starts up in period t, then $\gamma^u \Delta_t^u$ becomes dispatchable in period $t+1$, where $\gamma^u \in (0, 1]$ is a constant; the remaining $(1-\gamma^u)\Delta_t^u$ is referred to as *pending-up capacity.* In every following period, fraction γ^u of this pending-up capacity becomes dispatchable. Let R_t^u (R and u for "ramp up" or "remain to be up") denote the pending-up capacity in period t before starting up Δ_t^u. The process R_t^u evolves as follows:

$$R_{t+1}^u = (1-\gamma^u)(R_t^u + \Delta_t^u). \tag{6}$$

Similarly, we assume that if capacity of size Δ_t^d begins to shut down in period t, then $\gamma^d \Delta_t^d$ shuts down in period $t+1$, where $\gamma^d \in (0,1]$; the remaining $(1-\gamma^d)\Delta_t^d$ is referred to as *pending-down capacity*. In every following period, fraction γ^d of this pending-down capacity shuts down. Let R_t^d denote the pending-down capacity in period t before shutting down Δ_t^d. We have:

$$R_{t+1}^d = (1-\gamma^d)(R_t^d + \Delta_t^d). \tag{7}$$

The above approximation of the startup and shutdown processes reflects the time delay in capacity adjustment and that the startup speed is higher when more units are not yet started. Due to the geometric pattern we assumed, the entire capacity K^I will never be fully started up or shut down, but in most applications, this does not affect the insights derived from the model. Such an assumption, however, allows us to capture the underlying dynamics of the entire fleet of intermediate-load units by a few state variables instead of a full vector of history (we do not need to keep track of which stage each unit is in during the startup or shutdown process).

Following the dynamics for pending capacities described above, the dispatchable capacity K_t evolves as follows:

$$\begin{aligned} K_{t+1} &= K_t + \gamma^u(R_t^u + \Delta_t^u) - \gamma^d(R_t^d + \Delta_t^d) \\ &= K_{t+1}^o + \gamma^u \Delta_t^u - \gamma^d \Delta_t^d, \end{aligned} \tag{8}$$

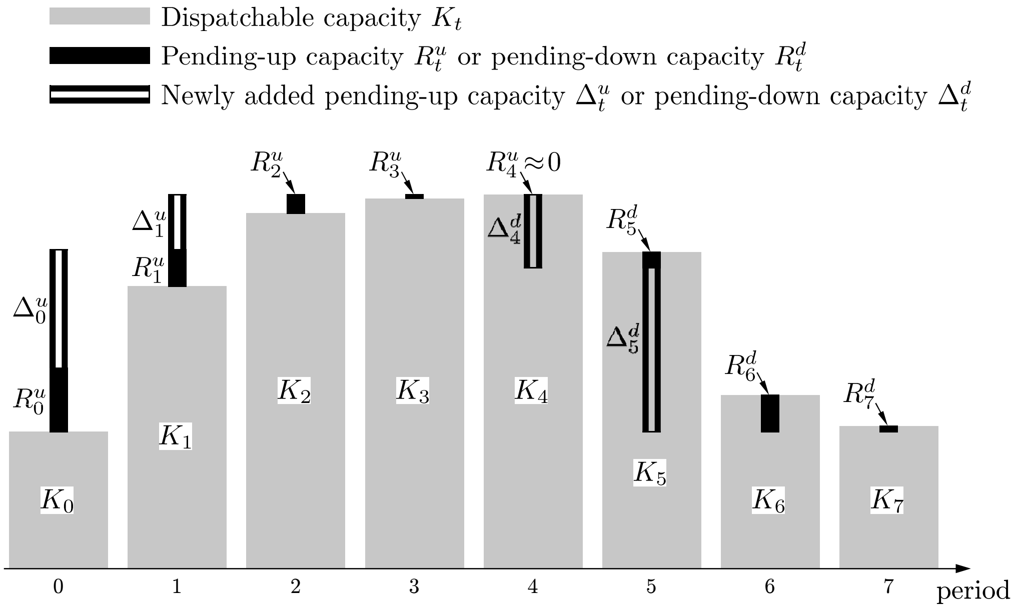
where $K_{t+1}^o \overset{\text{def}}{=} K_t + \gamma^u R_t^u - \gamma^d R_t^d$ is the dispatchable capacity in period $t+1$ if no new pending capacities are added in period t (i.e., $\Delta_t^u = \Delta_t^d = 0$).

Figure 3 illustrates an example of the capacity adjustment process, where the dispatchable capacity increases first and then decreases, as is commonly seen in the daily operations of the intermediate-load units. For ease of illustration, we set γ^u and γ^d very high.

Note that Δ_t^u and Δ_t^d must be non-negative, due to engineering restrictions: Pending-up capacity cannot be shut down, and pending-down capacity cannot be started up. This implies that a startup decision in one period affects the dispatchable capacity in multiple periods. This characteristic is important for electrical systems but cannot be captured by modeling capacity using a single state.

If the system operator needs to achieve the maximum dispatchable capacity for the next period, it can start up all the remaining non-dispatchable capacity that is not already pending-up, $K^I - K_t - R_t^u$,

Figure 3: Capacity Adjustment Process ($\gamma^u = \gamma^d = 0.8$)



and the maximum dispatchable capacity is:

$$K_{t+1}^{\max} \;=\; K_{t+1}^{o} + \gamma^u (K^I - K_t - R_t^u) \;=\; K_t + \gamma^u (K^I - K_t) - \gamma^d R_t^d. \tag{9}$$

The system operator can initiate the shutdown process on all the dispatchable capacity that is not already pending-down, $K_t - R_t^d$, to achieve the minimum dispatchable capacity for the next period:

$$K_{t+1}^{\min} \;=\; K_{t+1}^{o} - \gamma^d (K_t - R_t^d) \;=\; K_t + \gamma^u R_t^u - \gamma^d K_t \;=\; (1-\gamma^d) K_t + \gamma^u R_t^u. \tag{10}$$

Although we have two control variables, Δ_t^u and Δ_t^d, there is no economic reason to initiate startup and shutdown processes at the same time. Therefore, it is equivalent to use K_{t+1} as the single control variable for capacity, and we have relations:

$$\Delta_t^u = (K_{t+1} - K_{t+1}^{o})^+ / \gamma^u, \qquad \Delta_t^d = (K_{t+1}^{o} - K_{t+1})^+ / \gamma^d. \tag{11}$$

Substituting (11) into (6) and (7), we have

$$R_{t+1}^u = (1-\gamma^u)\Big(R_t^u + \frac{(K_{t+1} - K_{t+1}^{o})^+}{\gamma^u}\Big), \qquad R_{t+1}^d = (1-\gamma^d)\Big(R_t^d + \frac{(K_{t+1}^{o} - K_{t+1})^+}{\gamma^d}\Big). \tag{12}$$

In sum, in every period t, the system operator observes three capacity states: dispatchable capacity K_t, pending-up capacity R_t^u, and pending-down capacity R_t^d. Based on the accurate forecast for the demand and intermittent generation in the next period (described in §3.4), the system operator decides the dispatchable capacity $K_{t+1} \in [K_{t+1}^{\min},\ K_{t+1}^{\max}]$, the production, and curtailment of intermittent generation for the next period. The pending capacities evolve according to (12).

3.3 System Cost Structure

Since the production cost of intermittent power is negligible and the baseload production cost is constant, the system's total cost is driven by the costs of the flexible resources, which consist of the capacity adjustment (cycling) cost and the production cost. Unlike most production systems where production decisions are made before demand is realized, in an electrical system, when the system operator decides the production for the next 15-minute period, most of the randomness in the demand for the next period has been resolved so that the demand can be assumed to be known when the production decision is made. Thus, we do not consider the cost related to over- or under-production.

Recall c^s is the cycling cost per unit of intermediate capacity per cycle. For analytical convenience, we charge the cycling cost right after the startup decision is made. Specifically, we charge $\Delta^u_{t-1}c^s = \frac{(K_t - K^o_t)^+}{\gamma^u}c^s$ to period t. Also incurred in period t is the production cost. Let Q_t be the total production of the flexible resources in period t. Recall that K_t is the dispatchable capacity of intermediate-load units in period t. Because peaking units are very flexible and have a higher marginal production cost than the intermediate-load units (Assumption 1(iii)), it is optimal to produce $Q_t \wedge K_t \equiv \min\{Q_t, K_t\}$ from the intermediate-load units and produce $(Q_t - K_t)^+$ from the peaking units. Thus, the production cost of the flexible resources is

$$f(Q_t, K_t) \overset{\text{def}}{=} C(Q_t \wedge K_t, K_t) + (\alpha K_t - Q_t)^+ p + (Q_t - K_t)^+ c^P, \tag{13}$$

where the first two terms are the production cost of intermediate-load units, which follow from (3), and the last term is the peaking units production cost. Following from the discussion after (3), we see that $f(Q_t, K_t)$ decreases first and then increases in Q_t, with the minimum at $Q^I = \alpha K$.

Finally, we assume that curtailing the intermittent generation involves negligible cost (e.g., pitching the blades to curtail the wind power).

3.4 Demand and Intermittent Generation

Let $D_t > 0$ denote the total electricity demand minus the (constant) baseload in period t. We assume D_t is a deterministic function of a vector $\mathbf{D}_t$, which includes weather factors, time of the year, time of the day, etc; $\mathbf{D}_t$ itself is assumed to be a Markovian process. A simple example is $\mathbf{D}_t = \{t, D^r_t\}$ and $D_t = d(t) + D^r_t$, where the deterministic function $d(t)$ models the predictable component of the demand and D^r_t models the random fluctuations of the demand.

Let $W_t \geq 0$ denote the total wind power if not curtailed, which is a deterministic function of a Markovian vector $\mathbf{W}_t$. Elements of $\mathbf{W}_t$ include predictable and random variations in wind power. The random variations may contain multiple components, including regime switching and within-regime variations; see §6 for a specific example.

We refer to $D_t - W_t$ as the *net demand* on the flexible resources, which can be negative.

3.5 Curtailment Policies

We study two curtailment policies: the priority dispatch policy and the economic curtailment policy, denoted respectively by superscript P and E in the following analysis.

Under the priority dispatch policy, the system accommodates intermittent energy whenever it is possible to absorb the fluctuations. That is, intermittent generation is curtailed only when it exceeds the demand. The curtailed energy, denoted as w_t^P (w for 'waste'), is:

$$w_t^P = (W_t - D_t)^+. \tag{14}$$

The production of the flexible resources under the priority dispatch policy is

$$Q_t^P = D_t - W_t + w_t^P = (D_t - W_t)^+. \tag{15}$$

Under the economic curtailment policy, the curtailment decision is made jointly with all other system decisions to minimize the system cost. The production cannot exceed the demand D_t (it equals D_t when all intermittent generation is curtailed). The minimum production is the net demand (if positive) or zero (if the net demand is negative). Thus,

$$Q_t^E \in [(D_t - W_t)^+,\ D_t]. \tag{16}$$

The curtailed energy under the economic curtailment policy is:

$$w_t^E \ = \ Q_t^E + W_t - D_t \ \in \ [(W_t - D_t)^+,\ W_t]. \tag{17}$$

Clearly, less curtailment occurs under the priority dispatch policy: w_t^P defined in (14) is the minimum value of w_t^E in (17); Q_t^P defined in (15) is the minimum value of Q_t^E in (16).

3.6 Problem Formulation

In period $t-1$, the system operator observes the capacity state $\mathbf{K}_{t-1} \overset{\text{def}}{=} (K_{t-1}, R_{t-1}^u, R_{t-1}^d)$ and the accurate forecast for the demand and the intermittent generation in period t, $\mathbf{D}_t$ and $\mathbf{W}_t$. The system operator decides the dispatchable capacity K_t and the flexible generation Q_t, which in turn determine the pending capacities R_t^u, R_t^d, and the amount of curtailment. We assume the forecast for $\mathbf{D}_t$ and $\mathbf{W}_t$ is accurate, because the time interval considered is short (15 minutes).[1]

[1] In practice, the system operator typically solves a mixed-inter program or linear program each period to determine the operations for the next period (demand and intermittent generation for the next period are assumed to be known in the mathematical programs). During a period, the realized demand and intermittent generation can deviate from the forecasts, and the production quantities can deviate from the instructed quantities. But these deviations are typically small, and are absorbed by the so-called regulation services. In our model, we do not consider these deviations.

The transitions of the system state $(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ are specified as follows. We assume $\mathbf{D}_t$ and $\mathbf{W}_t$ follow Markov processes. The dispatchable capacity K_t is part of the decision, and R_t^u and R_t^d follow the dynamics in (12), which are slightly rewritten as follows:

$$R_t^u = (1-\gamma^u)\Big(R_{t-1}^u + \frac{(K_t - K_t^o)^+}{\gamma^u}\Big), \qquad R_t^d = (1-\gamma^d)\Big(R_{t-1}^d + \frac{(K_t^o - K_t)^+}{\gamma^d}\Big). \tag{18}$$

Let ρ be the discount factor. Let V_t^P and V_t^E denote the minimum expected discounted cost from period t onward under the priority dispatch policy and the economic curtailment policy, respectively. The terminal condition is $V_{T+1}^E = V_{T+1}^P = 0$. Then, we have:

$$\begin{aligned} V_t^P(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t} & \big\{ f(Q_t, K_t) + \frac{(K_t - K_t^o)^+}{\gamma^u} c^s + \rho \mathsf{E}_t\big[V_{t+1}^P(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})\big]\big\} \\ & s.t.\ K_t \in [K_t^{\min},\ K_t^{\max}], \quad Q_t = (D_t - W_t)^+, \quad \text{and (18)}, \end{aligned} \tag{19}$$

$$\begin{aligned} V_t^E(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t, Q_t} & \big\{ f(Q_t, K_t) + \frac{(K_t - K_t^o)^+}{\gamma^u} c^s + \rho \mathsf{E}_t\big[V_{t+1}^E(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})\big]\big\} \\ & s.t.\ K_t \in [K_t^{\min},\ K_t^{\max}], \quad Q_t \in [(D_t - W_t)^+,\ D_t], \quad \text{and (18)}. \end{aligned} \tag{20}$$

where $f(Q_t, K_t) = C(Q_t \wedge K_t, K_t) + (\alpha K_t - Q_t)^+ p + (Q_t - K_t)^+ c^P$ is defined in (13).

4. Optimal Capacity Adjustment and Wind Curtailment Policy

This section analyzes the structure of the optimal policy for the problem in (20).

4.1 Production and Curtailment under Given Capacity

We first assume the dispatchable capacity K_t is already decided and we need to choose production Q_t and the amount of curtailment. Proposition 1 states the results with proofs in the online supplement.

Proposition 1 *Under the economic curtailment policy, for given dispatchable capacity K_t of the intermediate-load units, the optimal policy is to produce (intermediate-load and peaking units combined production quantity):*

$$Q_t^*(K_t, D_t, W_t) = (D_t - W_t) \vee (\alpha K_t) \wedge D_t \tag{21}$$

and curtail intermittent generation by $w_t^* = Q_t^*(K_t, D_t, W_t) + W_t - D_t$.

Proposition 1 leads to the following optimal production and curtailment policy:

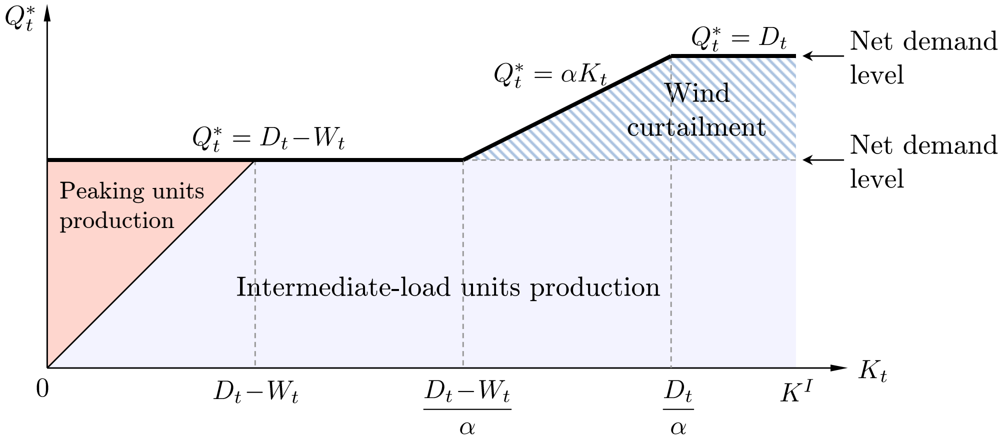
(i) If the net demand is above the min-gen level, $D_t - W_t \geq \alpha K_t$, then produce the net demand $Q_t^* = D_t - W_t$ and no intermittent generation is curtailed, $w_t^* = 0$.

(ii) If the min-gen level is above the net demand but below the total demand, $D_t - W_t < \alpha K_t < D_t$, then produce the minimum output $Q_t^* = \alpha K_t$ and partially curtail intermittent generation: $w_t^* = \alpha K_t + W_t - D_t \in (0, W_t)$.

(iii) If the min-gen level is at or above the total demand level, $\alpha K_t \geq D_t$, then produce the demand $Q_t^* = D_t$ (paying min-gen penalty $(\alpha K_t - D_t)p$) and curtail all intermittent generation: $w_t^* = W_t$.

Figure 4 illustrates the optimal policy under various levels of the dispatchable capacity K_t.

Figure 4: Optimal production as a function of K_t for given D_t and W_t



4.2 Capacity Adjustment

Using the optimal production quantity in (21), we write the optimal production cost of the flexible resources as

$$f(K_t; D_t, W_t) \overset{\text{def}}{=} f\big((D_t - W_t) \vee (\alpha K_t) \wedge D_t\big), K_t\big). \tag{22}$$

Then, the problem in (20) becomes:

$$V_t^E(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t} \Big\{ f(K_t; D_t, W_t) + \frac{(K_t - K_t^o)^+}{\gamma^u} c^s + \rho \mathsf{E}_t \big[V_{t+1}^E(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1}) \big] \Big\} \tag{23}$$
$$s.t.\ K_t \in [K_t^{\min},\ K_t^{\max}],\ \text{and (18)}.$$

The last term in the objective in (23) can be written as:

$$W_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) \overset{\text{def}}{=} \rho \mathsf{E}_t \big[V_{t+1}^E(K_t,\ R_t^u(K_t, \mathbf{K}_{t-1}),\ R_t^d(K_t, \mathbf{K}_{t-1}), \mathbf{D}_{t+1}, \mathbf{W}_{t+1}) \big],$$

where R_t^u and R_t^d relate to K_t and $\mathbf{K}_{t-1}$ according to (18).

Lemma 2 *(i) $V_t^E(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is jointly convex in $\mathbf{K}_{t-1} = (K_{t-1}, R_{t-1}^u, R_{t-1}^d)$ for any $\mathbf{D}_t$ and $\mathbf{W}_t$. (ii) The objective function in (23) is convex in K_t. In particular, $f(K_t; D_t, W_t)$ is convex in K_t, and $\frac{(K_t - K_t^o)^+}{\gamma^u} c^s + W_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is convex in K_t.*

Note that the value function is generally not monotone in $\mathbf{K}_t$ and, therefore, the convexity of the objective (part (ii) of the lemma) is not derived from the composition of convex functions. In fact, because R_t^u and R_t^d in (18) are piece-wise linear in K_t, we see $W_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is convex in K_t

for $K_t \geq K_t^o$ and $K_t \leq K_t^o$, but may have a kink at $K_t = K_t^o$. However, we prove that the sum of $\frac{(K_t - K_t^o)^+}{\gamma^u} c^s$ and $W_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is convex in K_t.

The convexity leads to the following optimal policy structure:

Proposition 2 *The optimal capacity adjustment policy is characterized by two pending capacity targets, $y^u(K_{t-1}, R_{t-1}^d, \mathbf{D}_t, \mathbf{W}_t)$ and $y^d(K_{t-1}, R_{t-1}^u, \mathbf{D}_t, \mathbf{W}_t)$:*
(i) If the pending-up capacity R_{t-1}^u is below the target $y^u(K_{t-1}, R_{t-1}^d, \mathbf{D}_t, \mathbf{W}_t)$, then bring the pending-up capacity up to the target by starting up $y^u(K_{t-1}, R_{t-1}^d, \mathbf{D}_t, \mathbf{W}_t) - R_{t-1}^u$ units of capacity;
(ii) If the pending-down capacity R_{t-1}^d is below the target $y^d(K_{t-1}, R_{t-1}^u, \mathbf{D}_t, \mathbf{W}_t)$, then bring the pending-down capacity up to the target by initiating the shutdown process on $y^d(K_{t-1}, R_{t-1}^u, \mathbf{D}_t, \mathbf{W}_t) - R_{t-1}^d$ units of capacity.
(iii) Parts (i) and (ii) cannot occur at the same time.
(iv) When neither (i) nor (ii) occurs, it is optimal not to adjust capacity.

We discuss intuitively some features of the optimal capacity adjustment and the role of economic curtailment. When the net demand exhibits daily cycles, the dispatchable capacity of the intermediate-load units goes through four phases every day: An expansion phase in the morning hours, a constant phase in the middle of the day, a downsizing phase in the evening hours, and another constant phase at night (similar to Angelus and Porteus 2002). Economic curtailment reduces the min-gen penalty at night, allowing more intermediate units to stay dispatchable throughout the night. Thus, cycling cost is also reduced, while the part-load penalty increases.

Capacity adjustment speed also affects the value of economic curtailment. In many electrical systems, the intermediate capacity is not flexible enough to ramp up to the increasing demand in the morning and thus significant peaking premium is incurred to meet the rising demand. Furthermore, starting up more intermediate-load units before the morning hours is undesirable due to the min-gen and part-load penalties. Curtailing intermittent generation in the very early morning can help increase the net demand, allowing more intermediate-load units to start up early without violating the min-gen restriction. Thus, economic curtailment also reduces the peaking premium.

5. Capacity Adjustment and Wind Curtailment Policy with Storage Operations

5.1 Model for Energy Storage

Energy storage is expensive due to both the initial investment and conversion losses. Electricity must be converted to other forms of energy to store (referred to as storing operation) and converted back to electricity when needed (referred to as releasing operation). Unlike other physical goods for which storage cost is mainly inventory holding cost, most energy losses occur during the conversions. The storage efficiency, denoted by η, measures the proportion of energy recovered after storing and releasing operations. For a hydroelectric pumped storage, η is typically 70-80%.

In this paper, "storage level" or "inventory level" refer to the amount of energy that the storage can release until empty. Let $\overline{S}$ denote the maximum storage level and S_t denote the inventory level in period t. Raising the inventory level by one unit requires $1/\eta$ units of energy from the system.

Storage operations also have speed limits. Let $\underline{\lambda}$ denote the maximum amount of energy that can be released per period, and $\overline{\lambda}$ denote the maximum amount of energy demanded by the storage per period. We assume that the storage can absorb $\overline{\lambda}$ units of energy even when it is full (extra energy is wasted). For example, a hydroelectric pumped storage can take energy while releasing water at the same time. Wasting energy via storage may be needed for relieving min-gen penalty.

To model inventory dynamics, let $x_t < 0$ be the amount of energy released from the storage in period t, and $x_t > 0$ be the amount of energy demanded by the storage. The range of x_t is

$$x_t \in [-\min\{\underline{\lambda}, S_{t-1}\},\, \overline{\lambda}].$$

The inventory dynamics are described as follows:

$$S_t = \begin{cases} S_{t-1} + x_t, & \text{if } x_t \leq 0, \\ \min[\overline{S},\, S_{t-1} + \eta x_t], & \text{if } x_t > 0. \end{cases} \tag{24}$$

5.2 Problem Formulation with Storage

With storage operations, let V_t^{PS} and V_t^{ES} denote the minimum expected discounted cost from period t onward under the priority dispatch policy and the economic curtailment policy, respectively. The terminal condition is $V_{T+1}^{ES} = V_{T+1}^{PS} = 0$. Then, we have:

$$\begin{aligned} V_t^{PS}(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t, x_t} \big\{ f(Q_t, K_t) + \frac{(K_t - K_t^o)^+}{\gamma^u} c^s + \rho \mathsf{E}_t \big[V_{t+1}^{PS}(\mathbf{K}_t, S_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1}) \big] \big\} \\ s.t.\ K_t \in [K_t^{\min},\ K_t^{\max}], \quad x_t \in [-\min\{\underline{\lambda}, S_{t-1}\},\, \overline{\lambda}], \\ Q_t = (D_t + x_t - W_t)^+,\ (24),\ \text{and } (18). \end{aligned} \tag{25}$$

$$\begin{aligned} V_t^{ES}(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t, Q_t, x_t} \big\{ f(Q_t, K_t) + \frac{(K_t - K_t^o)^+}{\gamma^u} c^s + \rho \mathsf{E}_t \big[V_{t+1}^{ES}(\mathbf{K}_t, S_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1}) \big] \big\} \\ s.t.\ K_t \in [K_t^{\min},\ K_t^{\max}], \quad x_t \in [-\min\{\underline{\lambda}, S_{t-1}\},\, \overline{\lambda}], \\ Q_t \in [(D_t + x_t - W_t)^+,\, D_t + x_t\,],\ (24),\ \text{and } (18). \end{aligned} \tag{26}$$

Note that when $\overline{\lambda} = \underline{\lambda} = 0$ (no storage case), (25) and (26) reduce to (19) and (20), respectively.

5.3 Optimal Policy under Economic Curtailment

For given storage energy flow x_t, the demand on the rest of the electrical system becomes $D_t + x_t$. Following from Proposition 1, the optimal production under given dispatchable capacity K_t and storage energy flow x_t is:

$$Q_t^*(K_t, D_t + x_t, W_t) = (D_t + x_t - W_t) \vee (\alpha K_t) \wedge (D_t + x_t), \tag{27}$$

and the optimal production cost of the flexible resources is

$$f(K_t; D_t + x_t, W_t) = f\big(Q_t^*(K_t, D_t + x_t, W_t)\big), K_t\big).$$

Thus, the problem in (26) becomes:

$$V_t^{ES}(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t, x_t} \big\{ f(K_t; D_t + x_t, W_t) + \frac{(K_t - K_t^o)^+}{\gamma^u} c^s + \rho \mathsf{E}_t \big[V_{t+1}^{ES}(\mathbf{K}_t, S_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1}) \big] \big\}$$

$$s.t.\ K_t \in [K_t^{\min},\ K_t^{\max}], \quad x_t \in [-\min\{\underline{\lambda}, S_{t-1}\},\ \overline{\lambda}],\ (24),\ \text{and } (18).$$

Proposition 3 *(i)* $V_t^{ES}(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ *is decreasing in* S_{t-1}.
(ii) $V_t^{ES}(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ *is jointly convex in* $(\mathbf{K}_{t-1}, S_{t-1})$ *for any* $\mathbf{D}_t$ *and* $\mathbf{W}_t$.

With storage operations, the optimal capacity adjustment policy has a similar structure to that in Proposition 2: If the pending-up capacity is below a target, it is raised to the target; furthermore, the target is independent of the pending-up capacity. A similar structure holds for the pending-down capacity. However, the optimal storage operations cannot be characterized by an inventory-independent target.

If the intermediate-load units were as flexible as the peaking units, the sole role of the storage would be to move the energy produced by the intermediate-load units during the off-peak periods to the peak periods for consumption, thereby reducing the total cost of energy production. In reality, with the costly capacity adjustment and the limited adjustment speed, the storage operations become more strategic. When the net demand is very low, some intermediate-load units may have to shut down to reduce the min-gen penalty. Storage can be used to keep some of these units stay dispatchable by storing the energy they produce. This lowers the cycling cost and min-gen penalty.

Storage also helps reduce the cost incurred during the transition from the off-peak to the peak hours. Without storage, during typical morning hours, peaking units have to be used when the demand grows faster than the upward capacity adjustment speed. The stored energy could be used to decrease the peaking premium. However, because storage is limited, using the stored energy in the morning hours results in less stored energy for use during the rest of the day. Storage allows for a usually better alternative: It can allow more intermediate-load units to get started before the morning hours by storing the energy they produce. With more intermediate-load units running, demand for peaking units is decreased during the morning hours.

6. Numerical Analysis

The goal of our numerical analysis is to explore the implications of our theoretical model and derive insights that may be useful for electricity system operators and policy makers. We will address the series of research questions regarding the value of curtailing intermittent generation raised in §1. To set our model parameters close to a real electrical system, we estimate the parameters based on data from Midwest Independent System Operator (MISO), as discussed in §6.1. (We do not intend to assess the performance of MISO.) The extra balancing cost due to intermittent generation is studied in §6.2. The impact of curtailment without and with the storage is analyzed in §6.3 and §6.4.

6.1 Data and Setup

6.1.1 Seasonal Decomposition of Load and Wind Data

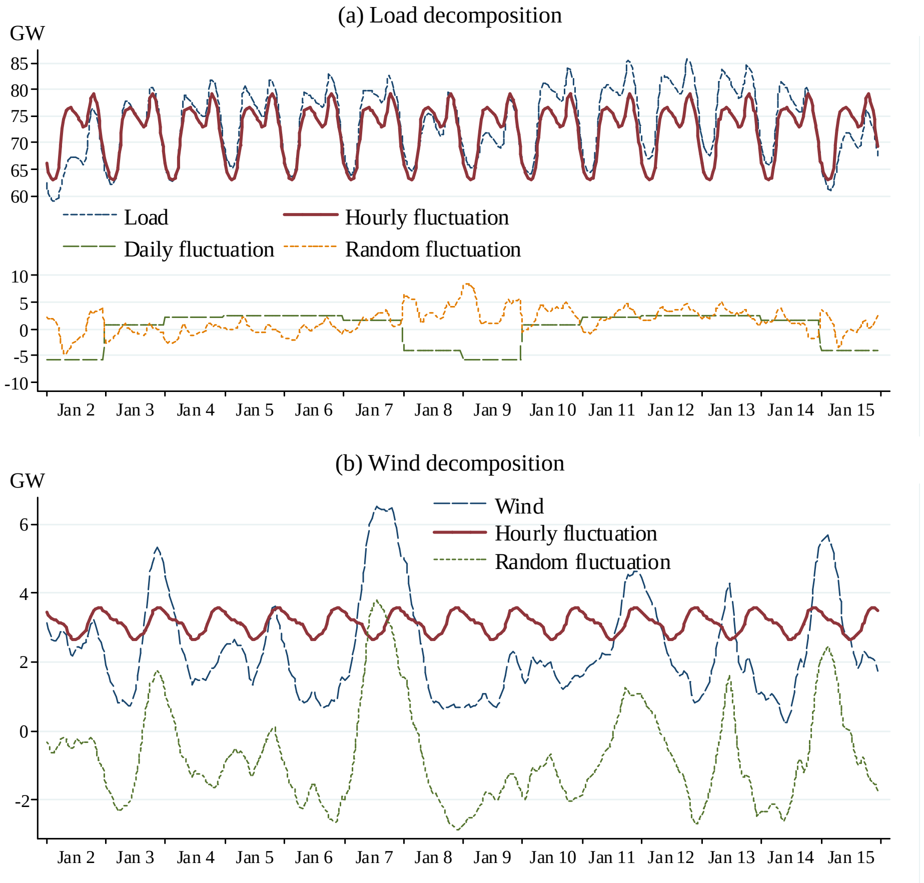
The data for the hourly electrical load and wind power in the Midwest electrical system are from MISO's website: https://www.midwestiso.org/Library/MarketReports/Pages/MarketReports.aspx.

We focus on the load and wind power variations on the time-scale of minutes to hours, which result in the balancing cost described in Figure 2 in §1. Following Ailliot and Monbet (2010), we remove inter-annual and inter-seasonal variations by limiting the data to short time periods and assume that the changes due to the season can be neglected. We analyze an eight-week (1,344-hour) period from January 2, 2011 to February 26, 2011. The data in this period do not exhibit a noticeable seasonal trend. Analysis on other time periods results in qualitatively similar results.

Figure 5 shows the data for the first two weeks. The average load is 72.21 GW and the average wind power is 3.12 GW. The wind energy penetration is about 4.3% (= 3.12/72.21) for this period of study. We decompose the variations into daily, hourly, and random components (daily component captures the weekday effect and is needed only for the load). These components are obtained from an ordinary least-squared regression with day-of-week and hour-of-day dummies. The daily and random components are set to have zero means; the hourly component has the same mean as the original data. We observe two known features of the wind power: The wind power is higher in the evenings than in the day; the variability in wind power is mostly contributed by its random component.

For both load and wind power, we use the hourly component (the thick curve in Figure 5) to model the predictable variability. Because each period in our analysis is a 15-minute interval, we

Figure 5: Load and wind power in MISO's footprint: January 2-15, 2011



(a) Load decomposition

(b) Wind decomposition

linearly interpolate three values between adjacent hourly values. Thus, the predictable variability is represented by 96 values, repeating daily. For the random component, we conduct a structural estimation of the parameters as detailed in the next subsection.

6.1.2 Models for the Random Variations in Load and Wind

The random component of the load, denoted as X_t, is assumed to be a mean-reverting process: $dX_t = -\eta X_t + \varepsilon_t$, where η represents the mean-reversion speed, and ε_t is white noise. Using the hourly data and the maximum likelihood method for autoregressive processes, we estimate that $\eta = 0.041$ and the standard deviation of the noise is 0.944 GW. We discretize the space for random load into seven levels: $-2, -\frac{4}{3}, -\frac{2}{3}$, 0, $\frac{2}{3}$, $\frac{4}{3}$, 2 GWh per 15 minutes. From each level, the random load either stays at the current level for the next 15-minute period or transits to another level with a maximum change of $\pm\frac{4}{3}$ GWh (i.e., move at most two levels above or below). The transition probabilities are set to match the mean and variance of the mean-reverting process.

Various models for wind speed and/or wind power exist; see a survey provided by Monbet, Ailliot, and Prevosto (2007). In recent years, regime switching models have been promoted by many researchers (Monbet et al. 2007, Pinson et al. 2008, Barber et al. 2010) to capture the existence of weather regimes. We model the random component of the wind power, denoted as Y_t, as the sum of two mean-reverting processes: $dY_{it} = -\eta_i Y_{it} + \varepsilon_{it}$, $i = 1, 2$, with $\eta_1 < \eta_2$ and $Y_t = Y_{1t} + Y_{2t}$. The slower mean-reverting process Y_{1t} represents the regime-switching process; the faster mean-reverting process Y_{2t} presents variations within regimes. Using the maximum likelihood method, we estimate that $\eta_1 = 0.064$ and $\eta_2 = 0.279$, both statistically significant at all conventional levels. The estimates for the standard deviations of ε_{1t} and ε_{2t} are statistically insignificant, because the combined variations are difficult to separate.

In discretization for Y_{1t}, we use three levels to represent high, medium, and low wind power regimes. For Y_{2t}, we use five levels. Thus, we have a total of 15 states for the random component of wind power. We let these 15 states be evenly spaced between -0.65 and 0.65 GWh per 15 minutes, with the top/middle/bottom five levels in the high/medium/low wind power regime. We set the transition probabilities such that the variability of the discrete process matches the data.

To study the effect of increase in wind penetration level, we need to specify how the extra wind power is correlated with the existing wind power. Let σ^2 be the variance of the random component of the existing wind power, and assume that wind power increases k times. In an extreme case, where the added wind power is perfectly correlated with existing wind power, the variance of the random component increases to $k^2\sigma^2$. In another extreme case, where the added and existing wind power are independent, the variance of the random component increases to $k\sigma^2$. The realistic case is mostly likely in between and we assume that the variance increases to $k^{1.5}\sigma^2$ in our computation. In discretization, as in the base case, we let the 15 states be evenly spaced between $-0.65k^{0.75}$ and $0.65k^{0.75}$ GWh per 15 minutes; the transition probabilities are set to reflect the increased variance.

6.1.3 Operating Parameters of the Resources

In our base case, wind power averages 3.12 GW, the generation fleet consists of 50 GW of baseload capacity, 15 GW of intermediate capacity, and enough peaking capacity to cover the maximum load.

To study the impact of the flexibility of intermediate-load units, we consider two types of intermediate-load units: natural gas combined-cycle (NGCC) units and coal units designed for cycling (not for baseload). We assume that the 15-GW intermediate capacity is of the same type. NGCC units are more flexible than coal units. The operating cost parameters are listed below.

- *Cycling cost.* At natural gas price \$5 per MBtu, starting up a NGCC unit costs \$50 per MW per start; at coal price \$3 per MBtu, starting up a coal unit costs \$235 per MW per start (based on the startup fuel requirement in Wu and Bennett (2010)). Lew et al. (2011) shows that wear and tear cost is comparable to the startup fuel cost. We assume that the total cycling cost is, for NGCC units, $c^s = \$100$ per MW per cycle, and for coal units, $c^s = \$470$ per MW per cycle.

- *Min-gen penalty.* The minimum generation level for a NGCC unit is typically 40-50% of the full load; the minimum generation level for a coal unit is higher. We set $\alpha = 0.5$ for NGCC units, and $\alpha = 0.7$ for coal units. The min-gen penalty is very high and assumed to be $p = \$1000$ per MWh below the minimum generation level for NGCC units and $p = \$2000$ per MWh below the minimum generation level for coal units.
- *Part-load penalty.* At natural gas price \$5 per MBtu, the average production cost of a NGCC unit is \$35 per MWh at full load, and \$42 per MWh at 50% load (Wu and Bennett 2010). For a typical coal unit serving intermediate load, the average production cost is \$30 per MWh at full load. We assume the aggregate production cost functions are:

$$\begin{aligned} &\text{NGCC units:} \quad C(Q,K) = 10K + 19Q + 6Q^2/K, \\ &\text{Coal units:} \quad\;\; C(Q,K) = \;6K + 18Q + 6Q^2/K. \end{aligned}$$

  These cost functions are consistent with the data. For example, the average cost for NGCC unit is $\frac{C(Q,K)}{Q} = 10\frac{K}{Q} + 19 + 6\frac{Q}{K}$, which equals \$35 when fully loaded ($Q = K$) and equals \$42 when part-loaded ($Q = K/2$).
- *Capacity adjustment speed.* For NGCC units, we assume $\gamma^u = 0.2$ and $\gamma^d = 0.4$. That is, from the down state, 20% of the capacity can become dispatchable in the first 15 minutes; from the up state, 40% of the capacity can be shut down in the first 15 minutes. Coal units have slower startup and shutdown processes: we assume $\gamma^u = 0.1$ and $\gamma^d = 0.2$.
- *Peaking cost.* Most peaking units are single-cycle natural gas-fired units. At natural gas price \$5 per MBtu, peaking cost is set to be $c^P = \$50$ per MWh.

The storage considered is a large hydroelectric pumped storage with the maximum water level equivalent to 12 GWh. The maximum water level change is 0.5 GWh per 15-minute period. The storing efficiency is 80%, and thus the maximum demand the storage can create is $\overline{\lambda} = 0.5/0.8 = 0.625$ GWh per period. The releasing efficiency is 94%, and thus the maximum supply the storage can provide is $\underline{\lambda} = 0.47$ GWh per period. The round-trip efficiency is $\eta = 0.8 \times 0.94 = 75.2\%$. These parameters are approximately close to those for the Ludington pumped storage in Michigan, one of the largest pumped storage plants in the world.

6.1.4 Average Cost Objective and Problem Size

We use the long-run average cost criterion in the numerical analysis. The finite-horizon discounted objective is used in problem formulations for analytical convenience. Because each period in our model is 15 minutes, the appropriate discount rate is very close to one, which effectively leads to the average cost criterion. Another benefit of using the long-run average cost criterion is that it does not depend on the initial state of the system.

The intermediate capacity, 3.75 GWh per period (or 15 GWh per hour), is discretized into n^I levels: $0,\ \delta,\ 2\delta, \ldots, (n^I - 1)\delta = 3.75$. When the dispatchable capacity $K_t = i\delta$, the pending-down capacity R_t^d can take $i + 1$ possible levels, and the pending-up capacity R_t^u can take $n^I - i$ possible levels. It can be shown that the capacity vector $\mathbf{K}_t = (K_t, R_t^u, R_t^d)$ can have $\frac{1}{6}n^I(n^I + 1)(n^I + 2)$ possible states. In our analysis, we set $n^I = 19$ and we have a total of 1,330 capacity states. We discretize the 12 GWh of storage space into 25 levels; each step is 0.5 GWh. As discussed earlier, we have 7 random load levels, 15 random wind power levels, and 96 periods each day.

In total, we have $1,330 \times 25 \times 7 \times 15 \times 96 = 335.16$ million states. The algorithm we implemented is the value iteration for unichain models with average cost criteria (Puterman 1994, §8.5). On a workstation with 2.40 GHz CPU, a typical problem instance can be solved within 2 hours.

6.2 Extra Balancing Cost for Wind Power

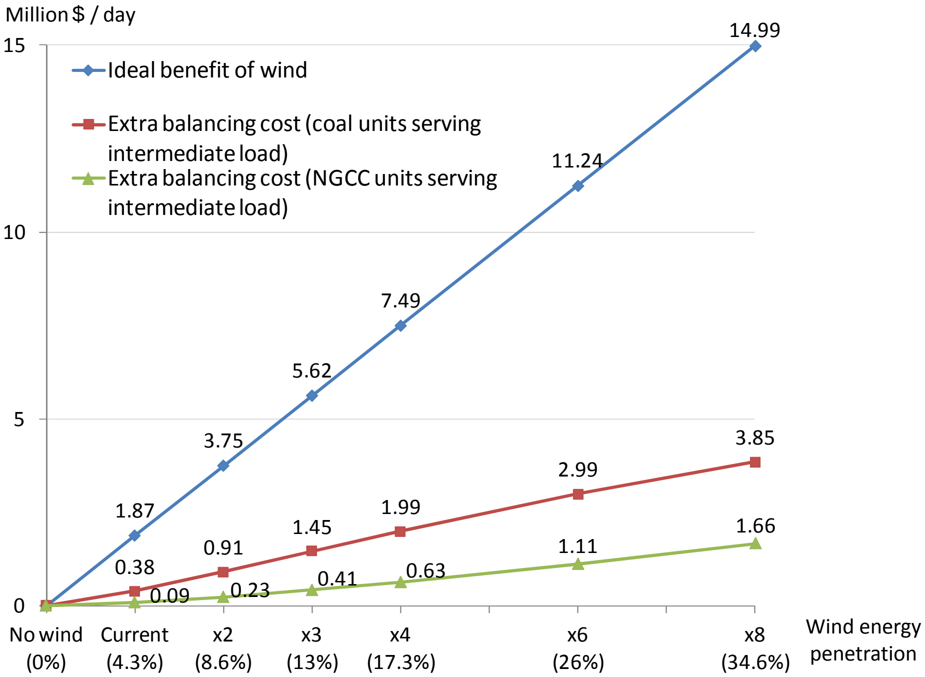
If wind power had no variability, it would reduce the net demand by a constant, which in turn reduces the baseload production by an amount equal to the wind power. During the period of our study, the average wind power is 74,930 MWh per day. Assuming the baseload production cost at \$25 per MWh, the wind power would ideally save \$1.87 million per day. In Figure 6, the top straight line shows the ideal benefit of the wind power as the wind penetration level increases.

In reality, wind power introduces additional variability into the system, which needs to be balanced by flexible resources (intermediate-load units and peaking units). The extra balancing cost is expected to increase as more wind power is brought into the system.

As shown in Figure 6, at the current wind energy penetration level (4.3%), the extra balancing cost is on average \$0.38 million per day (or 20.5% of the ideal cost savings) when coal units provide the intermediate capacity. When wind power doubles, the extra balancing cost increases to \$0.91 million per day, which offsets 24.2% of the ideal cost savings. This percentage stabilizes at about 26% when the wind energy penetration level is between 17% and 35%.

When NGCC units serve intermediate load, the extra balancing costs are significantly lower. At the current wind penetration level, the extra balancing cost is 4.6% of the ideal cost savings of the wind power. However, this percentage increases steadily as the wind penetration increases: The extra balancing cost is 6.2%, 8.4%, 9.9%, and 11.1% of the ideal cost savings when the wind penetration level increases to 2, 4, 6, and 8 times, respectively.

Figure 6: Ideal benefit vs. extra balancing cost of wind power



In the above computation, we assumed that the baseload production is reduced by an amount equal to the wind power (i.e., the same as the ideal situation where wind has no variability). In other words, we keep the average net demand on flexible resources constant. This allows us to focus on the effect of intermittency on the system balancing cost. (Without holding the average net demand on flexible resources constant, the change in the system balancing cost will be attributed to not only intermittency but also the amount of net demand to balance, making the comparison less intuitive for readers.)

6.3 Impact of Economic Curtailment

To prepare for the analysis, we first discuss a flow balance constraint. When comparing two systems with different policies or setups, let $\Delta\overline{Q}^P$ and $\Delta\overline{Q}^I$ be the changes in the long-run average output rate of the peaking units and the intermediate-load units, respectively. The demand is assumed to be the same for both systems and it must be met. Thus, the sum of the changes, $\Delta\overline{Q}^P + \Delta\overline{Q}^I$, must equal the change in the long-run average wind power curtailment, denoted as $\Delta\overline{w}$, plus the change in the long-run average conversion loss of the storage per period, denoted as $\Delta\bar{l}$:

$$\Delta\overline{Q}^I + \Delta\overline{Q}^P \;=\; \Delta\overline{w} + \Delta\bar{l}. \tag{28}$$

Recall that c^I is the average production cost of the intermediate-load units when they operate at the full load and that the part-load penalty in (5) is incurred when they operate at a lower load. The total balancing cost change can be decomposed into six components: cycling cost change, part-load penalty change, min-gen penalty change, and three additional components (corresponding to the changes in peaking cost and intermediate-load production cost evaluated at c^I):

$$c^P \Delta\overline{Q}^P + c^I \Delta\overline{Q}^I = \overbrace{(c^P - c^I)\Delta\overline{Q}^P}^{\text{Peaking premium change}} + \overbrace{c^I \Delta\overline{w}}^{\text{Wind curtailment cost change}} + \overbrace{c^I \Delta\overline{l}}^{\text{Storage loss change}} ,$$

where the equality is due to (28). We will quantify these six components in the following analysis.

For each type of intermediate capacity (coal and NGCC units) and each wind penetration level, we compute the balancing cost components under economic curtailment, and compare them with those under priority dispatch. Figure 7 shows the cost component changes (stacked bars) and the net balancing cost reduction (curve) brought by the economic curtailment policy.

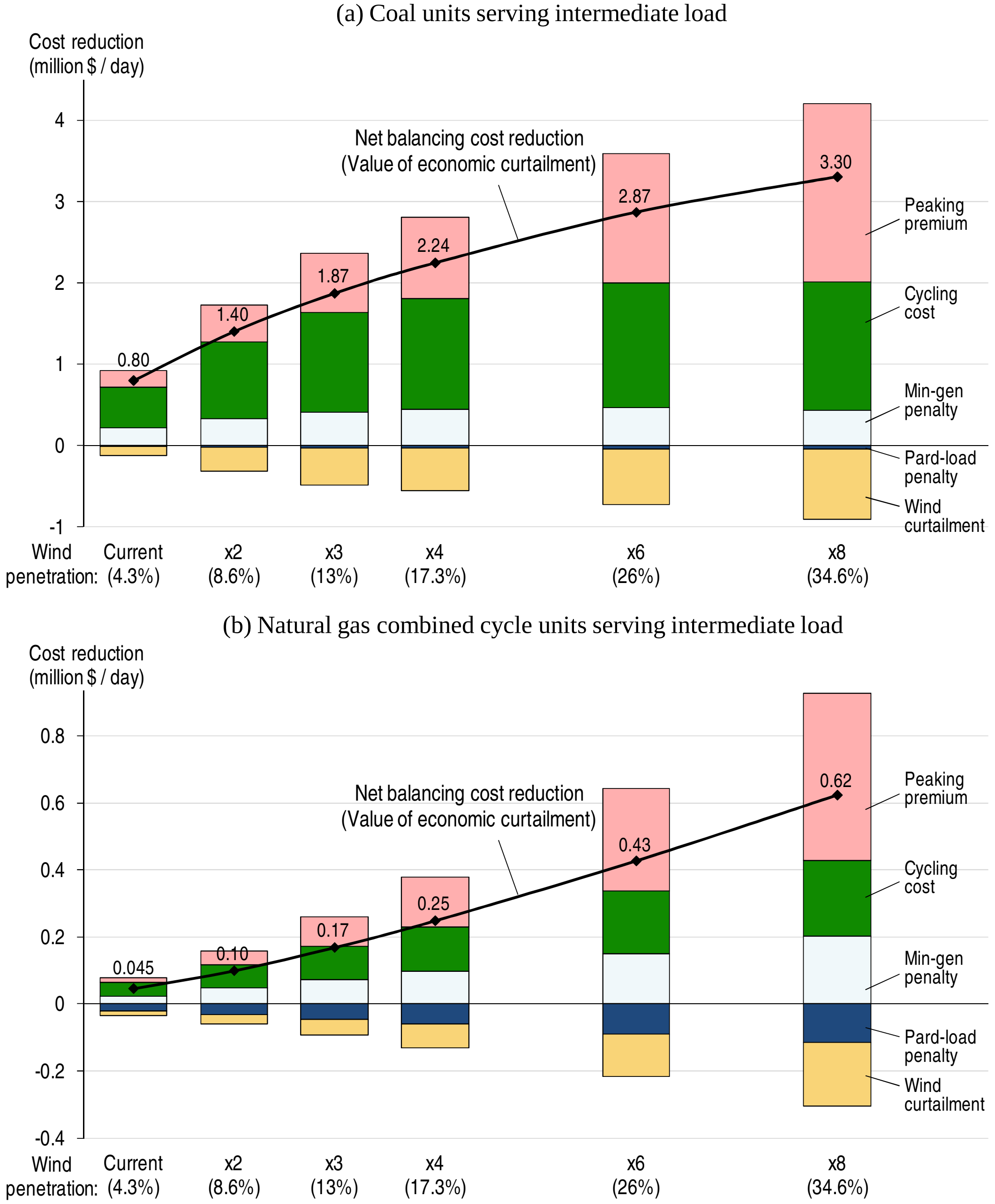
Three cost components decrease: peaking premium, cycling cost, and min-gen penalty, represented by the three stacked bars above the horizontal axis. Two cost components increase: part-load penalty and wind curtailment cost, represented by the stacked bars under the axis. Among these five components, the two dominant contributors to the balancing cost reduction are cycling cost and peaking premium. For both types of intermediate capacity, the cycling cost reduction plays a major role when the wind penetration level is relatively low, whereas the contribution from the peaking premium reduction becomes more prominent as the wind penetration increases.

When coal units serve intermediate load and the wind penetration is relatively low, the balancing cost reduction brought by economic curtailment can offset the entire extra balancing cost due to intermittency and bring the extra balancing cost to negative, as shown in Table 1. This implies that, with economic curtailment of wind power, the total system balancing cost may be lower than the balancing cost of a system without wind power. Hence, intermittency of the wind power coupled with economic curtailment does not necessarily increase the system balancing cost, and can benefit the system under low wind penetration levels.

To investigate how the balancing cost under economic curtailment can fall below the balancing cost without wind power, we examined the changes in the cost components. The major contributors are the cycling cost and min-gen penalty. Intuitively, with no wind power in the system, during the low-demand period, intermediate-load units need to be cycled down to reduce the min-gen penalty. When wind power is introduced into the system and allowed to be curtailed, the curtailment effectively allows the system to raise the net demand and shut down fewer units, which reduces the min-gen penalty and the cycling cost.

Table 1 also shows the percentage of wind power curtailed. Under priority dispatch, the amount

Figure 7: Impact of economic curtailment on balancing cost



(a) Coal units serving intermediate load

(b) Natural gas combined cycle units serving intermediate load

of wind curtailed is the same for both types of intermediate capacity, because it is curtailed only when net demand is negative. Economic curtailment results in about 6% curtailment of wind power when coal units serve intermediate load and much less is curtailed when NGCC units are used. In terms of balancing cost reduction, economic curtailment appears to be less valuable when the intermediate-load units are more flexible. However, when NGCC units serve intermediate load, the balancing cost savings increase in wind penetration at a faster rate than if coal units are used: The balancing cost reduction curve in Figure 7(a) is concave, whereas the curve in Figure 7(b) is convex.

Table 1: Impact of economic curtailment

PD = Priority dispatch EC = Economic curtailment

	Current (4.3%)	×2 (8.6%)	×3 (13%)	×4 (17.3%)	×6 (26%)	×8 (34.6%)
Coal units serve intermediate load:						
Extra balancing cost under PD (mil $/day)	0.38	0.91	1.45	1.99	2.99	3.85
Extra balancing cost under EC (mil $/day)	−0.42	−0.49	−0.42	−0.25	0.12	0.55
Balancing cost reduction (mil $/day)	0.80	1.40	1.87	2.24	2.87	3.30
Curtailment under PD	0.02%	0.08%	0.15%	0.22%	0.39%	0.58%
Curtailment under EC	4.88%	6.65%	6.89%	6.04%	5.46%	5.34%
Avg. cost reduction of curtailment ($/MWh)	219.0	142.2	123.4	128.7	125.7	115.7
NGCC units serve intermediate load:						
Extra balancing cost under PD (mil $/day)	0.087	0.23	0.41	0.63	1.11	1.66
Extra balancing cost under EC (mil $/day)	0.041	0.13	0.25	0.38	0.69	1.04
Balancing cost reduction (mil $/day)	0.045	0.10	0.17	0.25	0.43	0.62
Curtailment under PD	0.02%	0.08%	0.15%	0.22%	0.39%	0.58%
Curtailment under EC	0.47%	0.59%	0.74%	0.89%	1.20%	1.49%
Avg. cost reduction of curtailment ($/MWh)	134.5	128.5	126.0	122.7	117.5	114.3

Within the current range of penetration levels, the intuitions are as follows. At low wind penetration, the flexibility offered by the NGCC units well accommodates wind power variability and, therefore, economic curtailment does not benefit the system as much as when coal units serve intermediate load. At a higher wind penetration, the flexibility of NGCC units becomes inadequate, making the economic curtailment policy more valuable, resulting in a fast increase in the percentage of wind power curtailed, shown in Table 1. With coal units, the percentage of curtailment increases first and then decreases at higher wind penetration, due to some degree of the pooling of wind power (see §6.1.2). Note that reaching wind penetration levels as high as 35% may take decades; during that time, the capacity of the generation fleet will also change. Our purpose here is to understand what the balancing cost would be if capacity remained unchanged.

Table 1 also shows the average contribution to the balancing cost reduction per MWh of wind that is curtailed. The average contributions exceed $110 per MWh in all cases. Note that the peaking cost is assumed to be $50 per MWh. Thus, optimally curtailing wind power is, on average, more valuable than using the wind power to offset the fossil generation. Furthermore, although the absolute effect of curtailment in the case of NGCC units is much weaker than for coal units, the per-MWh contribution from the curtailed wind power is comparable for both types of intermediate capacities, especially at the medium-to-high wind penetration levels.

6.4 Impact of Economic Curtailment with Storage Available

The storage described in §6.1.3 can create a maximum demand of 0.625 GWh per period for 24 periods (6 hours). The wind power modeled in §6.1.2 can increase at most by 0.65 GWh per period (one regime change and two-level change in variation within regimes). Hence, the storage is able to store most of the wind power that would otherwise be curtailed. In other words, the storage may play nearly the same role as economic curtailment.

Dedicating the storage to reduce curtailment, however, may not be its best use, and may underestimate the value of economic curtailment. In the following analysis, we optimize the storage use under priority dispatch and economic curtailment, respectively, and compute the balancing cost reduction due to economic curtailment.

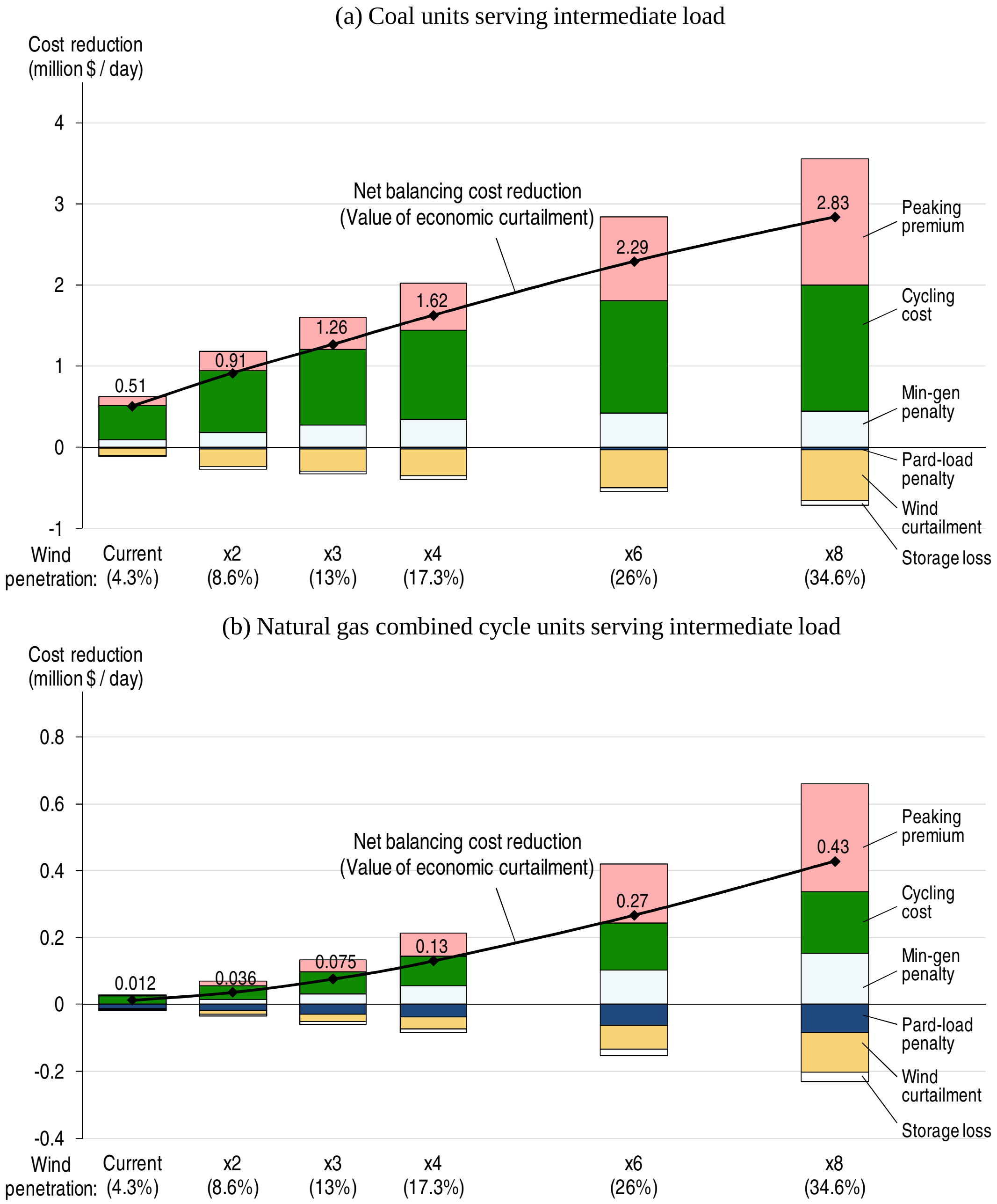
Figure 8 shows the cost reduction due to the economic curtailment with storage available. Compared to Figure 7, an additional component of the balancing cost is the conversion loss of the storage. Table 2 provides details of the balancing cost and amount of wind power curtailed.

Table 2: Impact of economic curtailment with storage available

PD = Priority dispatch EC = Economic curtailment

	Current	×2	×3	×4	×6	×8
	(4.3%)	(8.6%)	(13%)	(17.3%)	(26%)	(34.6%)
Coal units serve intermediate load:						
Extra balancing cost under PD (mil $/day)	0.25	0.65	1.13	1.63	2.62	3.55
Extra balancing cost under EC (mil $/day)	−0.25	−0.26	−0.14	0.003	0.34	0.72
Balancing cost reduction (mil $/day)	0.51	0.91	1.26	1.62	2.29	2.83
Curtailment under PD	0.0003%	0.003%	0.02%	0.05%	0.15%	0.27%
Curtailment under EC	3.80%	4.84%	4.01%	3.70%	3.60%	3.73%
Avg. cost reduction of curtailment ($/MWh)	177.9	126.0	141.0	148.5	147.4	136.5
$\frac{\text{Cost reduction by EC with storage}}{\text{Cost reduction by EC without storage}}$	63.5%	65.1%	67.6%	72.3%	79.7%	85.8%
NGCC units serve intermediate load:						
Extra balancing cost under PD (mil $/day)	0.055	0.15	0.28	0.44	0.84	1.30
Extra balancing cost under EC (mil $/day)	0.043	0.12	0.21	0.31	0.57	0.87
Balancing cost reduction (mil $/day)	0.012	0.036	0.075	0.13	0.27	0.43
Curtailment under PD	0.0002%	0.006%	0.03%	0.06%	0.16%	0.28%
Curtailment under EC	0.16%	0.22%	0.30%	0.40%	0.62%	0.85%
Avg. cost reduction of curtailment ($/MWh)	96.3	112.3	122.3	128.3	128.8	124.8
$\frac{\text{Cost reduction by EC with storage}}{\text{Cost reduction by EC without storage}}$	25.8%	35.9%	44.9%	52.3%	62.4%	68.7%

Figure 8: Impact of economic curtailment on balancing cost with storage available


(a) Coal units serving intermediate load

(b) Natural gas combined cycle units serving intermediate load

Compared to the results without storage, when storage is available, economic curtailment has a weaker effect on the reductions of all cost components. However, the cost reductions are still significant, as indicated by the ratio of cost reduction with storage over the cost reduction without storage in Table 2. This ratio increases with the wind penetration, indicating that the need rises for balancing the extra variability introduced by the wind; economic curtailment plays a more significant role if no new storage is added to balance this extra variability. The ratio is larger when coal units serve intermediate load, because the coal units are less flexible than the NGCC units. However, even for the case of NGCC units serving intermediate load, this ratio increases dramatically when wind penetration increases.

A surprising result revealed in Table 2 is that curtailing wind power may have a higher average contribution to the cost reduction when storage is present than if storage is absent. This is true for medium-to-high wind penetration levels and for both types of intermediate capacities. Intuitively, because storage operations reduce the need for curtailment, the system will choose to curtail wind power in situations when storage is unable to further reduce the balancing cost. Thus, the storage actually increases the value of per MWh of energy curtailed, rather than eliminating the value of curtailment.

7. Concluding Remarks

This paper analyzes the effects of curtailing intermittent generation on the balancing cost of electrical systems. Curtailing intermittent generation during the low-demand hours helps avoid the min-gen penalty and reduces the need to shut down intermediate-load units, thereby reducing the cycling cost. Curtailment also allows earlier startups of intermediate-load units, increasing the dispatchable capacity in the morning hours and reducing the peaking premium. Our numerical analysis shows that the value of curtailment is significant whether storage is present or absent. The presence of storage reduces the curtailment, but does not diminish its value; storage may actually increase the value per unit of curtailed energy. We also find that when intermediate capacity is more flexible, the value of curtailment is lower, but increases faster in the intermittent energy penetration level.

There are several limitations of this study. We do not consider the value of curtailment in a transmission network with possible transmission congestions. Ela (2009) shows the value of economic curtailment in a simple transmission network. Because transmission congestion may prevent the use of flexible generation resources and energy storage in one location to lower the balancing cost in other locations, economic curtailment may become more valuable in network settings.

We assume in our analysis that peaking capacity is unlimited. In practice, increased intermittent generation requires more safety capacity known as operating reserves to maintain grid reliability. Safety capacity can be provided by flexible generation units and storage facilities. Our results show that economic curtailment allows more intermediate-load units to operate at part-load, which effectively increases their capability to provide safety capacity. Thus, economic curtailment also helps reduce the investment needed for additional safety capacity.

This study focuses on system cost minimization, but also sheds light on the implementation of economic curtailment to benefit all parties in the system. Renewable energy producers need compensation for the curtailment, otherwise such a policy is unlikely to be well implemented. Our results show that the value per unit of energy curtailed significantly exceeds the marginal cost of energy in the system, which is determined by the most expensive operating unit in the system. Hence, our results reveal that fully compensating renewable energy producers for the energy curtailed is a feasible policy.

References

Ailliot, P., V. Monbet 2010. Markov-switching autoregressive models for wind time series. *Preprint.* Available at http://pagesperso.univ-brest.fr/∼ailliot/doc/wind.pdf.

Angelus, A., E. L. Porteus 2002. Simultaneous capacity and production management of short-life-cycle, produce-to-stock goods under stochastic demand. *Management Science* **48**(3) 399–413.

Barber, C., J. Bockhorst, P. Roebber 2010. Auto-regressive HMM inference with incomplete data for short-horizonwind forecasting. *Proceedings of the 24th Annual Conference on Neural Information Processing Systems.*

Boyce, M. P. 2010. *Handbook for Cogeneration and Combined Cycle Power Plants.* 2nd edn. ASME Press. New York, NY.

Davis, M. H. A., M. A. H. Dempster, S. P. Sethi, D. Vermes 1987. Optimal capacity expansion under uncertainty. *Advances in Applied Probability* **19**(1) 156–176.

Eberly, J. C., J. A. Van Mieghem 1997. Multi-factor dynamic investment under uncertainty. *Journal of Economic Theory* **75**(2) 345–387.

Ela, E. 2009. Using economics to determine the efficient curtailment of wind energy. National Renewable Energy Laboratory, NREL/TP-550-45071.

Energy Information Administration 2010. Electric power annual with data for 2009. Released November 2010, available at http://www.eia.gov/cneaf/electricity/epa/epates.html.

European Union 2009. Directive 2009/28/EC of the European Parliament and of the Council on the promotion of the use of energy from renewable sources.

Farmer, E., V. Newman, P. Ashmole 1980. Economic and operational implications of a complex of wind-driven generators on a power system. *IEE Proceedings A* **127**(5) 289–295.

Gross, R., P. Heptonstall, D. Anderson, T. Green, M. Leach, J. Skea 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. Report by the UK Energy Research Centre's Technology and Policy Assessment function.

Grubb, M. J. 1988. The economic value of wind energy at high power system penetrations: An analysis of models, sensitivities and assumptions. *Wind Engineering* **12**(1) 1–26.

Grubb, M. J. 1991a. The integration of renewable electricity sources. *Energy Policy* **19**(7) 670–688.

Grubb, M. J. 1991b. Value of variable sources on power systems. *IEE Proceedings C* **138**(2) 149–165.

Hutzler, M. 2010. Wind integration: Does it reduce pollution and greenhouse gas emissions? Institute for Energy Research. Available at http://www.instituteforenergyresearch.org.

ILEX Energy Consulting, G. Strbac 2002. Quantifying the system costs of additional renewables in 2020. Report to the British Department of Trade and Industry.

Kahn, E. 1979. The compatibility of wind and solar technology with conventional energy systems. *Annual Review of Energy* **4** 313–352.

Katzenstein, W., J. Apt 2009. Air emissions due to wind and solar power. *Environmental Science and Technology* **43**(2) 253–258.

Lew, D., G. Brinkman, M. Milligan, S. Lefton, D. Piwko 2011. How does wind affect coal? Cycling, emissions, and costs. Presentation by the National Renewable Energy Laboratory, NREL/PR-5500-51579.

Lu, B., M. Shahidehpour 2004. Short-term scheduling of combined cycle units. *Power Systems, IEEE Transactions on* **19**(3) 1616–1625.

Monbet, V., P. Ailliot, M. Prevosto 2007. Survey of stochastic models for wind and sea state time series. *Probabilistic Engineering Mechanics* **22** 113–126.

Monsen, W. A., L. Norin 2011. Proposed market rules may increase wind curtailments. *North American Windpower* **8**(8) pp.

National Renewable Energy Laboratory 2008. Energy storage and wind power. Available from http://www.nrel.gov/wind/systemsintegration/energy_storage.html, accessed on Sept 18, 2011.

Pinson, P., L. E. A. Christensen, H. Madsen, P. E. Sørensen, M. H. Donovan, L. E. Jensen 2008. Regime-switching modelling of the fluctuations of offshore wind generation. *Journal of Wind Engineering and Industrial Aerodynamics* (96) 2327–2347.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.. New York, NY.

Rocklin, S. M., A. Kashper, G. C. Varvaloucas 1984. Capacity expansion/contraction of a facility with demand augmentation dynamics. *Operations Research* **32**(1) 133–147.

Rogers, J., S. Fink, K. Porter 2010. Examples of wind energy curtailment practices. National Renewable Energy Laboratory, NREL/SR-550-48737.

Sørensen, B. 1978. On the fluctuating power generation of large wind energy converters, with and without storage facilities. *Solar Energy* **20** 321–331.

Swider, D. J., C. Weber 2007. The costs of wind's intermittency in germany: application of a stochastic electricity market model. *Eurpean Transactions on Electrical Power* **17** 151–172.

Wu, O. Q., N. D. Bennett 2010. Sustainability at Detroit Edison: Using natural gas as a transition fuel. Case 1-429-143, William Davidson Institute, University of Michigan.

Online Supplement

Proof of Lemma 1. By (4) and Assumption 1(i), $C(Q^I, K)$ is increasing in Q^I. Rewrite the cost function as $C(Q^I, K) = \dfrac{c(Q^I\kappa/K)}{Q^I\kappa/K}Q^I$. Because the average cost $\dfrac{c(q)}{q}$ decreases in q by Assumption 1(ii), $C(Q^I, K)$ increases in K.

To see that $C(Q^I, K)$ is jointly convex in (Q^I, K), pick two arbitrary points (Q_1, K_1) and (Q_2, K_2), and we have:

$$\begin{aligned} C(Q_1, K_1) + C(Q_2, K_2) &= \tfrac{K_1}{\kappa}c\big(\tfrac{Q_1}{K_1}\kappa\big) + \tfrac{K_2}{\kappa}c\big(\tfrac{Q_2}{K_2}\kappa\big) \\ &= \tfrac{K_1+K_2}{\kappa}\big[\tfrac{K_1}{K_1+K_2}c\big(\tfrac{Q_1}{K_1}\kappa\big) + \tfrac{K_2}{K_1+K_2}c\big(\tfrac{Q_2}{K_2}\kappa\big)\big] \\ &\geq \tfrac{K_1+K_2}{\kappa}c\big(\tfrac{K_1}{K_1+K_2}\tfrac{Q_1}{K_1}\kappa + \tfrac{K_2}{K_1+K_2}\tfrac{Q_2}{K_2}\kappa\big) \\ &= \tfrac{K_1+K_2}{\kappa}c\big(\tfrac{Q_1+Q_2}{K_1+K_2}\kappa\big) \\ &= 2C\big(\tfrac{Q_1+Q_2}{2}, \tfrac{K_1+K_2}{2}\big) \end{aligned}$$

where the inequality is due to the convexity of $c(q)$. ∎

Proof of Proposition 1. The problem in (20) can be solved sequentially. For any given capacity level K_t, we first solve the optimal production by solving:

$$\min_{Q_t}\big\{f(Q_t, K_t) \ : \ Q_t \in [(D_t - W_t)^+, \ D_t]\big\}.$$

Because $f(Q_t, K_t)$ is decreasing in Q_t for $Q_t \leq \alpha K_t$, and then increasing in Q_t for $Q_t > \alpha K_t$, $f(Q_t, K_t)$ is minimized at αK_t, and the optimal production is to produce αK_t or as close as possible. Therefore, we have:

$$Q_t^*(K_t, D_t, W_t) = (D_t - W_t) \vee (\alpha K_t) \wedge D_t.$$

The amount of wind power curtailed is given by (17) with Q_t replaced by $Q_t^*(K_t, D_t, W_t)$. ∎

Proof of Lemma 2. First, we prove that the production cost $f(K_t; D_t, W_t)$ is convex in K_t for any D_t and W_t. The function $f(K_t; D_t, W_t)$ has at most four segments:

$$f(K_t; D_t, W_t) = \begin{cases} C(K_t, K_t) + (D_t - W_t - K_t)c^P, & K_t \in [0, D_t - W_t) \\ C(D_t - W_t, K_t), & K_t \in [D_t - W_t, \frac{D_t - W_t}{\alpha}) \\ C(\alpha K_t, K_t), & K_t \in [\frac{D_t - W_t}{\alpha}, \frac{D_t}{\alpha}) \\ C(\alpha K_t, K_t) + (\alpha K_t - D_t)p, & K_t \in [\frac{D_t}{\alpha}, K^I] \end{cases}$$

When $K^I < D_t/\alpha$, there will be fewer segments of the function.

The convexity of $C(Q, K)$ in Lemma 1 suggests that each of the four segments of $f(K_t; D_t, W_t)$ is convex in K_t. We next compare the slope at the three connection points. Note that $\frac{\partial C(Q,K)}{\partial Q} = c'\big(\frac{Q}{K}\kappa\big)$

follows from the definition of $C(Q, K)$ in (4). Then,

$$\left(\frac{\partial C(K_t, K_t)}{\partial Q} + \frac{\partial C(K_t, K_t)}{\partial K} - c^P - \frac{\partial C(D_t - W_t, K_t)}{\partial K}\right)\Bigg|_{K_t = D_t - W_t} = c'(\kappa) - c^P < 0$$

$$\left(\frac{\partial C(D_t - W_t, K_t)}{\partial K} - \frac{\partial C(\alpha K_t, K_t)}{\partial Q}\alpha - \frac{\partial C(\alpha K_t, K_t)}{\partial K}\right)\Bigg|_{K_t = \frac{D_t - W_t}{\alpha}} = -c'(\alpha\kappa)\alpha \leq 0$$

$$\left(\frac{\partial C(\alpha K_t, K_t)}{\partial Q}\alpha + \frac{\partial C(\alpha K_t, K_t)}{\partial K} - \frac{\partial C(D_t, K_t)}{\partial K} - \alpha p\right)\Bigg|_{K_t = \frac{D_t}{\alpha}} = \alpha(c'(\alpha\kappa) - p) < 0$$

The above three inequalities follow from Assumption 1 (iii), (i), respectively.

Second, we prove $V_t(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ in (23) can be found using alternative decision variables Δ^u_{t-1} and Δ^d_{t-1} without imposing the property that $\Delta^u_{t-1} \cdot \Delta^d_{t-1} = 0$:

$$V_t(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{\Delta^u_{t-1}, \Delta^d_{t-1}} \left\{ f(K_t; D_t, W_t) + \Delta^u_{t-1} c^s + \rho \mathsf{E}_t[V_{t+1}(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})] \right\} \tag{A.1}$$

$$\Delta^u_{t-1} \in [0, K^I - K_{t-1} - R^u_{t-1}], \qquad \Delta^d_{t-1} \in [0, K_{t-1} - R^d_{t-1}] \tag{A.2}$$

$$K_t = K_{t-1} + \gamma^u(R^u_{t-1} + \Delta^u_{t-1}) - \gamma^d(R^d_{t-1} + \Delta^d_{t-1}) \tag{A.3}$$

$$R^u_t = (1 - \gamma^u)(R^u_{t-1} + \Delta^u_{t-1}) \tag{A.4}$$

$$R^d_t = (1 - \gamma^d)(R^d_{t-1} + \Delta^d_{t-1}) \tag{A.5}$$

To see the equivalence between the original formulation and the above alternative, we need to prove $\Delta^u_{t-1} \cdot \Delta^d_{t-1} = 0$, which implies $\Delta^u_{t-1} = (K_t - K^o_t)^+/\gamma^u$ and $\Delta^d_{t-1} = (K^o_t - K_t)^+/\gamma^d$ and, consequently, the formulation in (A.1)-(A.5) reduces to that in (23).

To prove $\Delta^u_t \Delta^d_t = 0$, we consider any feasible solution with $\Delta^u_t > 0$ and $\Delta^d_t > 0$. Consider a revised policy:

$$\begin{aligned}
&\widetilde{\Delta}^u_t = \Delta^u_t - \varepsilon/\gamma^u, \qquad \widetilde{\Delta}^d_t = \Delta^d_t - \varepsilon/\gamma^d, \qquad \text{with } \varepsilon \in (0, \min\{\gamma^u \Delta^u_t, \gamma^d \Delta^d_t\}) \\
&\widetilde{\Delta}^u_{t+1} = \Delta^u_{t+1} + \frac{1-\gamma^u}{\gamma^u}\varepsilon, \qquad \widetilde{\Delta}^d_{t+1} = \Delta^d_{t+1} + \frac{1-\gamma^d}{\gamma^d}\varepsilon \\
&\widetilde{\Delta}^d_s = \Delta^d_s, \qquad \widetilde{\Delta}^u_s = \Delta^u_s, \qquad \text{for } s > t+1.
\end{aligned}$$

Under the new policy, in period $t+1$, the dispatchable capacity does not change, while the pending capacities decline:

$$\begin{aligned}
\widetilde{K}_{t+1} &= K_t + \gamma^u(R^u_t + \Delta^u_t) - \varepsilon - \gamma^d(R^d_t + \Delta^d_t) + \varepsilon \;=\; K_{t+1} \\
\widetilde{R}^u_{t+1} &= (1-\gamma^u)(R^u_t + \widetilde{\Delta}^u_t) = R^u_{t+1} - \tfrac{1-\gamma^u}{\gamma^u}\varepsilon \\
\widetilde{R}^d_{t+1} &= (1-\gamma^d)(R^d_t + \widetilde{\Delta}^d_t) = R^d_{t+1} - \tfrac{1-\gamma^d}{\gamma^d}\varepsilon
\end{aligned}$$

and in period $t+2$, we have $\widetilde{\mathbf{K}}_{t+2} = \mathbf{K}_{t+2}$ because after taking the actions $\widetilde{\Delta}^u_{t+1}$ and $\widetilde{\Delta}^d_{t+1}$, the pending capacities are the same as in the original policy:

$$\widetilde{R}^u_{t+1} + \widetilde{\Delta}^u_{t+1} = R^u_{t+1} + \Delta^u_{t+1}, \qquad \widetilde{R}^d_{t+1} + \widetilde{\Delta}^d_{t+1} = R^d_{t+1} + \Delta^d_{t+1}$$

The rest of the capacity dynamics of the new policy is the same as the original policy.

Because the dispatchable capacity process under the two policies are the same, the production cost are the same as well. The capacity adjustment costs are different: The new policy saves $c^s\varepsilon/\gamma^u$ in period t while spending $c^s\frac{1-\gamma^u}{\gamma^u}\varepsilon$ in period $t+1$. The net saving is $c^s\varepsilon(\frac{1}{\gamma^u} - \rho\frac{1-\gamma^u}{\gamma^u}) > 0$. Thus, the new policy is strictly better than the original policy. Thus, as long as $\Delta^u_t\Delta^d_t > 0$, we can make strict improvement and, therefore, the optimal policy must have $\Delta^u_t\Delta^d_t = 0$.

The equivalence established above ensures that the properties derived from the alternative formulation of the problem also hold for the original problem.

Third, we prove the convexity of the value function by induction. The terminal value function V_{T+1} is assumed to be zero. Suppose that $V_{t+1}(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})$ is convex in $\mathbf{K}_t$ for any given $\mathbf{D}_{t+1}$ and $\mathbf{W}_{t+1}$. Then, the objective function in (A.1) is convex in $\mathbf{K}_t$. Equations (A.3)-(A.5) show that $\mathbf{K}_t$ is a linear function of $(\mathbf{K}_{t-1}, \Delta^u_{t-1}, \Delta^d_{t-1})$. Hence, the objective function in (A.1) is jointly convex in $(\mathbf{K}_{t-1}, \Delta^u_{t-1}, \Delta^d_{t-1})$ on a closed convex set defined as

$$\big\{(\mathbf{K}_{t-1}, \Delta^u_{t-1}, \Delta^d_{t-1}) : K_{t-1} \in [0, K^I],\ R^u_{t-1} \in [0, K^I - K_{t-1}],\ R^d_{t-1} \in [0, K_{t-1}],\ \text{and (A.2)}\big\}.$$

By the theorem on convexity preservation under minimization (Heyman and Sobel 1984, p. 525), we conclude that $V_t(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is convex in $\mathbf{K}_{t-1}$.

Fourth, we prove the convexity of the objective in (23) in K_t. We write

$$V_t(\mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{K_t}\big\{f(K_t; D_t, W_t) + F_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) : K_t \in [K^{\min}_t,\ K^{\max}_t]\big\} \tag{A.6}$$

and $F_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ can be written as follows, without imposing $\Delta^u_{t-1} \cdot \Delta^d_{t-1} = 0$:

$$F_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{\Delta^u_{t-1}}\Big\{\Delta^u_{t-1}c^s + \rho\mathsf{E}_t[V_{t+1}(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})]\Big\} \tag{A.7}$$

$$\Delta^u_{t-1} \in [0, K^I - K_{t-1} - R^u_{t-1}], \tag{A.8}$$

$$\Delta^u_{t-1} \geq \frac{K_t - K^o_t}{\gamma^u}, \qquad \Delta^u_{t-1} \leq \frac{K_t - K^{\min}_t}{\gamma^u} \tag{A.9}$$

$$R^u_t = (1-\gamma^u)(R^u_{t-1} + \Delta^u_{t-1}) \tag{A.10}$$

$$R^d_t = (1-\gamma^d)\frac{1}{\gamma^d}(K_{t-1} + \gamma^u(R^u_{t-1} + \Delta^u_{t-1}) - K_t) \tag{A.11}$$

where the inequalities in (A.9) are derived from the constraint $\Delta^d_{t-1} \in [0, K_{t-1} - R^d_{t-1}]$, and (A.11) follows from (A.3) and (A.5).

Because $V_{t+1}(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})$ is convex in $\mathbf{K}_t$ for any given $\mathbf{D}_{t+1}$ and $\mathbf{W}_{t+1}$, and because of the linear relations in (A.10)-(A.11), the objective function in (A.7) is jointly convex in (K_t, Δ^u_{t-1}) on a closed convex set defined by $K_t \in [K^{\min}_t,\ K^{\max}_t]$, (A.8) and (A.9). Therefore, $F_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is convex in K_t.

Using the property of $\Delta^u_{t-1}\Delta^d_{t-1} = 0$, we know that the minimizer to (A.7) is $\Delta^u_{t-1} = \frac{(K_t - K^o_t)^+}{\gamma^u}$. Thus, $F_t(K_t; \mathbf{K}_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \frac{(K_t - K^o_t)^+}{\gamma^u} c^s + \rho \mathsf{E}_t\big[V_{t+1}(\mathbf{D}_{t+1}, \mathbf{W}_{t+1}, K_t, R^u_t(K_t), R^d_t(K_t))\big]$ is convex in K_t. ∎

Proof of Proposition 2. Consider minimizing the objective in (23) over two separate regions: $K_t \in [K^{\min}_t, K^o_t]$ and $K_t \in [K^o_t,\ K^{\max}_t]$. The corresponding problems are:

$$\min_{K_t \in [K^{\min}_t,\, K^o_t]} f(K_t; D_t, W_t) + \rho \mathsf{E}_t\big[V^E_{t+1}(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})\big] \tag{A.12}$$

$$s.t. \quad R^u_t = (1-\gamma^u)R^u_{t-1}, \qquad R^d_t = (1-\gamma^d)\Big(R^d_{t-1} + \frac{K^o_t - K_t}{\gamma^d}\Big).$$

$$\min_{K_t \in [K^o_t,\ K^{\max}_t]} f(K_t; D_t, W_t) + \frac{K_t - K^o_t}{\gamma^u} c^s + \rho \mathsf{E}_t\big[V^E_{t+1}(\mathbf{K}_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})\big] \tag{A.13}$$

$$s.t. \quad R^d_t = (1-\gamma^d)R^d_{t-1}, \qquad R^u_t = (1-\gamma^u)\Big(R^u_{t-1} + \frac{K_t - K^o_t}{\gamma^u}\Big).$$

For the problem in (A.12), we change the decision variable to the pending-down capacity after shutting down Δ^d_{t-1},

$$y^d = R^d_{t-1} + \Delta^d_{t-1} = R^d_{t-1} + \frac{K^o_t - K_t}{\gamma^d} = \frac{K_{t-1} + \gamma^u R^u_{t-1} - K_t}{\gamma^d}.$$

For the problem in (A.13), we change the decision variable to the pending-up capacity after starting up Δ^u_{t-1},

$$y^u = R^u_{t-1} + \Delta^u_{t-1} = R^u_{t-1} + \frac{K_t - K^o_t}{\gamma^u} = \frac{K_t - K_{t-1} + \gamma^d R^d_{t-1}}{\gamma^u}.$$

Then, $K_t \in [K^{\min}_t,\ K^o_t]$ is equivalent to $y^d \in [R^d_{t-1}, K_{t-1}]$, and $K_t \in [K^o_t,\ K^{\max}_t]$ is equivalent to $y^u \in [R^u_{t-1}, K^I - K_{t-1}]$. The problems in (A.12) and (A.13) are equivalent to:

$$\min_{y^d \in [R^d_{t-1}, K_{t-1}]} \quad F(K_{t-1}, R^u_{t-1}, y^d, \mathbf{D}_t, \mathbf{W}_t) \tag{A.14}$$

$$\min_{y^u \in [R^u_{t-1}, K^I - K_{t-1}]} F(K_{t-1}, y^u, R^d_{t-1}, \mathbf{D}_t, \mathbf{W}_t) + (y^u - R^u_{t-1})c^s \tag{A.15}$$

where,

$$F(K_{t-1}, y^u, y^d, \mathbf{D}_t, \mathbf{W}_t) \overset{\text{def}}{=} f\big(K_{t-1} + \gamma^u y^u - \gamma^d y^d;\ D_t, W_t\big) \tag{A.16}$$

$$+ \rho \mathsf{E}_t\Big[V^E_{t+1}\big(K_{t-1} + \gamma^u y^u - \gamma^d y^d,\ (1-\gamma^u)y^u,\ (1-\gamma^d)y^d, \mathbf{D}_{t+1}, \mathbf{W}_{t+1}\big)\Big].$$

The problems in (A.14) and (A.15) imply that the optimal policy has the following structures:

If the pending-down capacity R^d_{t-1} is below a target level defined as

$$y^d(K_{t-1}, R^u_{t-1}, \mathbf{D}_t, \mathbf{W}_t) \overset{\text{def}}{=} \inf \underset{y^d \in [0, K_{t-1}]}{\arg\min}\ F(K_{t-1}, R^u_{t-1}, y^d, \mathbf{D}_t, \mathbf{W}_t),$$

then it is optimal to bring the pending-down capacity up to the target.

If the pending-up capacity R^u_{t-1} is below a target level defined as

$$y^u(K_{t-1}, R^d_{t-1}, \mathbf{D}_t, \mathbf{W}_t) \overset{\text{def}}{=} \inf \underset{y^u \in [0, K^I - K_{t-1}]}{\arg\min}\ F(K_{t-1}, y^u, R^d_{t-1}, \mathbf{D}_t, \mathbf{W}_t) + y^u c^s,$$

then it is optimal to bring the pending-up capacity up to the target.

Because the objective in (23) is convex in K_t, as shown in Lemma 2 (ii), if the problem in (A.12) has a minimizer $K^*_t < K^o_t$, then K^*_t minimizes the objective in (23) over the entire feasible region $K_t \in [K^{\min}_t, K^{\max}_t]$. Similarly, when solving (A.13), if a minimizer $K^*_t > K^o_t$ exists, it is also the global minimizer. If neither of the above two situations occur, then $K^*_t = K^o_t$ is the optimal solution.

■

Proof of Proposition 3. (i) We prove by induction that $V^{ES}_t(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ decreases in S_{t-1}. The terminal value function V_{T+1} is assumed to be zero. Suppose that $V^{ES}_{t+1}(\mathbf{K}_t, S_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})$ decreases in S_t. Consider $S^a_{t-1} < S^b_{t-1}$. Starting from state $(\mathbf{K}_{t-1}, S^a_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$, denote the optimal decision as (K^*_t, x^*_t) and the resulting inventory as S^a_t. If the system starts from $(\mathbf{K}_{t-1}, S^b_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$, the decision (K^*_t, x^*_t) remains feasible because $x^*_t \in [-\min\{\underline{\lambda}, S^a_{t-1}\}, \overline{\lambda}] \subseteq [-\min\{\underline{\lambda}, S^b_{t-1}\}, \overline{\lambda}]$, and $S^a_t \leq S^b_t$ according to (24).

$$\begin{aligned}
V^{ES}_t(\mathbf{K}_{t-1}, S^a_{t-1}, \mathbf{D}_t, \mathbf{W}_t) &= f(K^*_t; D_t + x^*_t, W_t) + \frac{(K^*_t - K^{*o}_t)^+}{\gamma^u} c^s + \rho \mathsf{E}_t\big[V^{ES}_{t+1}(\mathbf{K}^*_t, S^a_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})\big]\big\} \\
&\geq f(K^*_t; D_t + x^*_t, W_t) + \frac{(K^*_t - K^{*o}_t)^+}{\gamma^u} c^s + \rho \mathsf{E}_t\big[V^{ES}_{t+1}(\mathbf{K}^*_t, S^b_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})\big]\big\} \\
&\geq V^{ES}_t(\mathbf{K}_{t-1}, S^b_{t-1}, \mathbf{D}_t, \mathbf{W}_t),
\end{aligned}$$

where the first inequality is due to the induction assumption and $S^a_t \leq S^b_t$, and the last inequality follows from the feasibility of (K^*_t, x^*_t).

(ii) We formulate the problem using alternative decision variables Δ^u_{t-1} and Δ^d_{t-1} as in the proof of Lemma 2. Furthermore, we use alternative decision variables x^u_t, x^d_t, and x^w_t, which respectively represent the energy flows for raising inventory, lowering inventory, and being wasted. We reformulate the problem in (26) as follows:

$$
\begin{aligned}
V_t^{ES}(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t) = \min_{\Delta_{t-1}^u, \Delta_{t-1}^d, x_t^u, x_t^d, x_t^w} & \Big\{ f(K_t; D_t + x_t, W_t) + \Delta_{t-1}^u c^s \\
& + \rho \mathsf{E}_t[V_{t+1}^{ES}(\mathbf{K}_t, S_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})] \Big\}
\end{aligned} \tag{A.17}
$$

$$
\begin{aligned}
&\Delta_{t-1}^u \in [0, K^I - K_{t-1} - R_{t-1}^u], && \Delta_{t-1}^d \in [0, K_{t-1} - R_{t-1}^d] \\
&x_t^u \in [0, \min\{\overline{\lambda},\ (\overline{S} - S_{t-1})/\eta\}], && x_t^d \in [0, \min\{\underline{\lambda}, S_{t-1}\}], \\
&x_t^w \in [0, (\overline{\lambda} - (\overline{S} - S_{t-1})/\eta)^+], && \\
&x_t = x_t^u - x_t^d + x_t^w, && S_t = S_{t-1} + \eta x_t^u - x_t^d, \\
&R_t^u = (1-\gamma^u)(R_{t-1}^u + \Delta_{t-1}^u), && R_t^d = (1-\gamma^d)(R_{t-1}^d + \Delta_{t-1}^d) \\
&K_t = K_{t-1} + \gamma^u(R_{t-1}^u + \Delta_{t-1}^u) - \gamma^d(R_{t-1}^d + \Delta_{t-1}^d) &&
\end{aligned}
$$

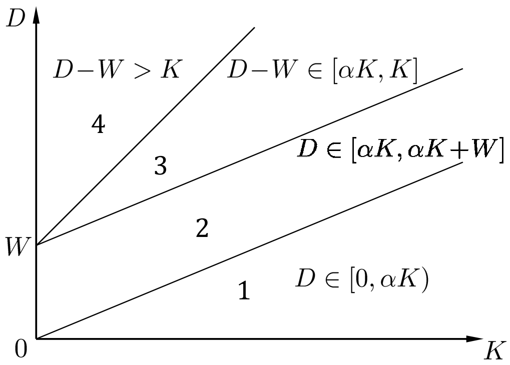
The optimal solution has the property that $x_t^u \cdot x_t^d = 0$, because if $x_t^u \cdot x_t^d > 0$, reducing x_t^u and x_t^d by the same amount does not change x_t, but raises S_t, thereby improving the objective due to the monotonicity in part (i). Similarly, the optimal solution also has the property that $x_t^d \cdot x_t^w = 0$. These two properties, together with $\Delta_{t-1}^u \cdot \Delta_{t-1}^d = 0$, indicate that the above alternative formulation is equivalent to the original formulation in (26).

Next, we prove that $f(K; D, W)$ is jointly convex in (K, D). Expanding the definition in (22), we have

$$
f(K; D, W) = \begin{cases}
C(\alpha K, K) + (\alpha K - D)p, & \text{Region 1: } K \in [\frac{D}{\alpha}, K^I] \\
C(\alpha K, K), & \text{Region 2: } K \in [\frac{D-W}{\alpha}, \frac{D}{\alpha}) \\
C(D - W, K), & \text{Region 3: } K \in [D - W, \frac{D-W}{\alpha}) \\
C(K, K) + (D - W - K)c^P, & \text{Region 4: } K \in [0, D - W)
\end{cases} \tag{A.18}
$$

The four regions are illustrated in Figure A.1.

Figure A.1: Regions for $f(K, D, W)$



Note that $f(K; D, W)$ is jointly convex in (K, D) within each of the four regions, because $C(Q, K)$

is jointly convex in (Q, K). Hence, we need to prove $f(K; D, W)$ is convex across the region boundaries.

Over regions 1 and 2, we can write $f(K; D, W) = C(\alpha K, K) + (\alpha K - D)^+ p$, which is jointly convex in (K, D) because $(\alpha K - D)^+$ is jointly convex in (K, D).

Over regions 2 and 3 but excluding the area with $D < W$, we can write

$$f(K; D, W) = \max\{C(\alpha K, K),\ C(D - W, K)\},$$

which is jointly convex in (K, D) because the maximum of two convex functions is convex.

To prove $f(K; D, W)$ is convex across regions 3 and 4, we define an auxiliary function:

$$\widetilde{c}(q) = \begin{cases} c(q), & q \in [0, \kappa] \\ c(\kappa) + (q - \kappa)c^P, & q > \kappa \end{cases}$$

Note that $\widetilde{c}(q)$ is convex in q due to Assumption 1. We define $\widetilde{C}(Q, K) \overset{\text{def}}{=} \dfrac{K}{\kappa}\widetilde{c}\Big(\dfrac{Q}{K}\kappa\Big)$. Using (4), we can related $\widetilde{C}(Q, K)$ with $C(Q, K)$:

$$\widetilde{C}(Q, K) = \begin{cases} C(Q, K), & Q \in [0, K] \\ C(K, K) + (Q - K)c^P, & Q > K \end{cases}$$

In Lemma 1, we proved that $C(Q, K)$ is jointly convex due to the convexity of $c(q)$. Following the same lines of proof and the convexity of $\widetilde{c}(q)$, we see that $\widetilde{C}(Q, K)$ is jointly convex in (Q, K). Because $f(K; D, W) = \widetilde{C}(D - W, K)$, we conclude that $f(K; D, W)$ is jointly convex in (K, D) across regions 3 and 4.

Now, we prove the convexity of the value function by induction. The terminal value function V_{T+1} is assumed to be zero. Suppose that $V_{t+1}^{ES}(\mathbf{K}_t, S_t, \mathbf{D}_{t+1}, \mathbf{W}_{t+1})$ is jointly convex in $(\mathbf{K}_t, S_t)$ for any given $\mathbf{D}_{t+1}$ and $\mathbf{W}_{t+1}$. Then, because all the constraints of the problem in (A.17) are linear in $(\mathbf{K}_{t-1}, S_{t-1}, \Delta_{t-1}^u, \Delta_{t-1}^d, x_t^u, x_t^d, x_t^w)$, the objective function in (A.17) is jointly convex in $(\mathbf{K}_{t-1}, S_{t-1}, \Delta_{t-1}^u, \Delta_{t-1}^d, x_t^u, x_t^d, x_t^w)$ on a closed convex set. Therefore, $V_t(\mathbf{K}_{t-1}, S_{t-1}, \mathbf{D}_t, \mathbf{W}_t)$ is convex in $(\mathbf{K}_{t-1}, S_{t-1})$. ∎

Reference

Heyman, D. P., M. J. Sobel 1984. *Stochastic Models in Operations Research.* Vol. 2. McGraw-Hill, New York.