Supply flexibility in electricity markets

Claude Crampes* and Jérôme Renault†

23rd July 2018

Abstract

The development of non-dispatchable renewable sources of energy requires more flexible reliable thermal equipment to match residual demand. We analyze the advantages of delaying production decisions to benefit from more precise information on states of the world, at the expense of higher production costs in a two-period framework where two technologies with different flexibility characteristics are available. We determine first-best production levels ex ante and ex post, that is, when demand is still random and is known with certainty respectively. We then show that, under perfect competition, first best can be implemented indifferently either by means of ex post state-contingent markets or by means of a day-ahead market followed by adjustment markets. By contrast, when the industry is imperfectly competitive, the two market designs are not equivalent.

JEL codes: C72, D24, D47, L23, L94

Key words: flexibility, electricity, production costs, day-ahead market, real-time market, perfect competition, imperfect competition

*Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France. E-mail: claude.crampes@tse-fr.eu

†Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France. E-mail: jerome.renault@tse-fr.eu

1 Introduction

1.1 Flexibility in electricity production

In most industries, production processes are at their highest productivity when they are steady, that is, when inputs and outputs are almost invariant in time. By contrast, starting a production process involves warming up periods, changing speed requires ramping up and down costly procedures, and stopping cannot be instantaneous without damages. These unsteady periods are costly but they generally do not occur very often because both the inputs and the outputs can be stored. Then the flexibility of the production process, that is, its capacity to become efficient at different speed levels within a short lag, is not critical. This is not true in the electricity industry because energy storage is not feasible at large scale and any gap between production and demand (the latter being weakly responsive to scarcity signals) provokes costly damages. Moreover, the need for reliable flexible production is increasing with the development of non-dispatchable renewable sources. With operating costs close to zero, electricity made from solar and wind energy is given priority in the merit order for dispatching, but this energy depends on states of nature. Therefore, it is not fully reliable and must be backed by energy from hydro or thermal plants ready to produce in real time above or below the planned level, to compensate for variations in the observed output of green energy.

Overall, the electricity industry is characterized by constant changes in the level of required production, with the consequence that flexibility of reliable equipment is a highly valuable quality for the whole system. Since the opening up to market mechanisms, most countries have accommodated flexibility requirements by designing a two-layer system for energy transactions:

- most transactions[1] are settled in a day-ahead framework where each of the 24 hours of the next day is a market. Typically, there is a deadline (say 12:00) for the submission of demand and supply bids for power that will be delivered the following day. A computer system (a surrogate for the Walras auctioneer) calculates the hourly prices that balance supply and demand. Later (say at 1.00 pm) these equilibrium prices are made public and operations are settled. The next day, power is provided hour for hour according to the agreed contracts.

- in balancing markets, buyers and sellers can trade power close to real time to re-balance demand and supply if something wrong has occurred since the corresponding day-ahead clearing. Balancing markets typically are continuous markets where prices are set based on a pay-as-bid basis, instead of the unique hourly price of the day-ahead system.

[1] We mean "most transactions in the wholesale markets". Actually, the largest portion of electricity transactions is fixed by contract.

Clearly, participating in balancing trade requires more flexible equipment than participating in day-ahead.

1.2 Flexibility degrees

In the electricity industry, recent interest in flexibility problems is due to the challenging development of non-dispatchable renewable energy, mainly solar and wind (see e.g. Bertsch *et al.*, 2012). Flexibility means the ability of a system "*to accommodate increasing levels of uncertainty while maintaining satisfactory levels of performance at minimal additional cost for any timescale*" (Silva, 2010). On practical grounds, most of the analysts consider two groups of technologies among conventional thermal sources, depending on whether or not the output of a source can be ramped up or down at short notice. In this binary classification, there is a clear separation between inflexible nuclear and coal-fired plants on the one hand, and flexible natural gas turbines on the other hand (see for example Eisenack, 2015; Kök *et al.* 2016)[2]. The output from inflexible technologies must be planned before knowing the real state of demand and production from renewables. The adjustment is made ex post by dispatching the flexible gas turbines in real time. It results that the historically dominant inflexible technologies are now challenged by the pair "renewables + flexible turbines" and face the risk of being pushed out of the energy mix.

Actually, nuclear and coal-fired plants are not as inflexible as stated in the quoted papers. Lykidi and Gourdel (2015) for example recognize the flexibility of nuclear plants at least in so far as they accommodate the seasonal component of demand. But seasonal flexibility is just one part of the flexibility problem. Data published on line clearly show that all thermal plants have as some hourly flexibility, less than hydroelectric installations, but sufficient to deserve some attention.[3] In particular, existing thermal power plants can provide more flexibility than is often assumed. Coal fired power plants are already providing large operational flexibility: "*they are adjusting their output on a 15-minute basis (intraday market) and even on a 5-minute basis (balancing market) to variation in renewable generation and demand*" (Pescia 2017).

However, it remains true that for real-time adjustment, in particular for balancing the variability of wind, most systems rely mainly on gas-fired power turbines on top of demand response, pump storage and energy trade with neighboring countries. Some researchers also insist on the need for improving the real-time market design to be sure that thermal plants needed to back up renewables will survive the massive entry of non-dispatchable sources (Ma *et al.*, 2013; Finon,

[2]As stated in PJM (2017) "*Inflexible units are those with declining average costs that are unable to economically produce power within a certain range, or that require an economic minimum output. Inflexible units can be of all fuel types, including coal, nuclear and large gas units, which are inflexible based on either their technology or the way they purchase natural gas.*"

[3]See for example the charts proposed by http://www.rte-france.com/en/eco2mix/eco2mix-mix-energetique-en for the French system.

2015; Bertsch *et al.*, 2016). Finally, note that European regulators (ACER and CEER, 2017) want to favor a so-called "holistic approach" to support market flexibility. They would like the whole electric system to be involved, including all consumers, who should have the opportunity to participate in all relevant markets and other arrangements for valuing flexibility, in particular through aggregation by independent operators. The regulators quote changes in how energy is consumed, e.g. electric vehicles, combined with electricity storage, home automation and progress in Information and Communication Technologies ICT), which may facilitate the provision of flexibility from new sources. From this view, flexibility platforms are to be operated close to consumers, with a pivotal role devoted to distribution system operators.

We rather favour an analytical approach to study the potential flexibility of large plants that already participate in wholesale markets. We mainly have in mind generators using gas, coal, water from reservoirs, or nuclear energy. We assume that every type of production plant has some flexibility potential and we analyze how to combine several installations in an efficient way. We mainly focus on first best and perfect competition outcomes. However, we also address the question of flexibility under imperfect competition.

1.3 Outline of the paper

In section 2 we first recall that most economic models of investment and operation of the electric industry emphasize the production capacity constraint. They therefore consider production plants as fully upward inflexible and fully downward flexible. We then explain how the timing dimension of flexibility can be modeled in a two-period framework where information on the state of the world improves as time passes by, but production costs increase when production decisions must become effective within a short lag. In section 3 we determine the first-best allocation of production in this two-period setting and we show that, under perfect competition, it can be implemented by either a day-ahead market, followed by a state-dependent adjustment market, or only ex post markets. In section 4, we illustrate the trade-off between information gain and adjustment cost by using a multilinear cost function and a quadratic surplus function. We then consider a duopoly model of competition in quantities (section 5) and show that, contrary to perfect competition, market design matters. We conclude in section 6.

2 Tools for the analysis of flexibility

As noted in subsection 2.1, in standard economic analysis of electricity production there is little room for upward flexibility at the level of individual equipment. The overall system is made flexible by stacking heterogeneous technologies characterized by their installation and operation costs. In this approach, time appears

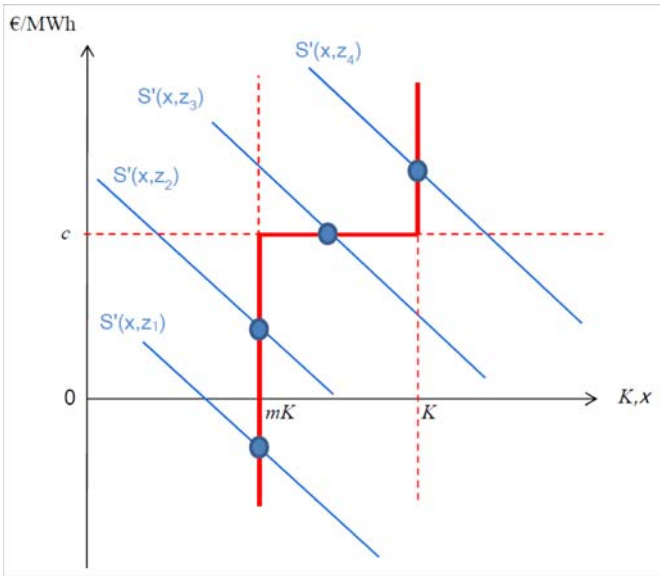

Figure 1: Partially flexible equipment

in the form of use duration, a key determinant for the choice of equipment. Actually, time should also be taken into account in terms of the lag between decision and execution, in particular when information on the state of demand and intermittent renewable output become more precise. We emphasize this in subsection 2.2 to underline the importance of information in the decision process.

2.1 Quantity and price adjustments

2.1.1 Production limitations

Isolated plant The standard modeling of electricity production is based on the hypothesis that, after installing capacity K at unit cost r, energy x can be produced at unit cost c, as long as it remains below capacity. The cost function is then $C(x, K) = cx + rK$, $x \leq K$. This means that production is totally inflexible beyond K and perfectly downward flexible. To model downward inflexibility, one can add a constraint $x \geq mK$ with m between 0 and 1. If $m = 0$, the plant is fully downward flexible; if $m = 1$, it can work only at level K. As shown in Figure 1, this elementary approach explains some essential features of electricity production. We have represented different levels of marginal surplus, that is, the derivative of the consumer's surplus $S(x, z)$ from the consumption of x when nature is in state z. We assume that, in each state z, the function S is a continuously differentiable concave function of x, increasing up to a saturation point and then decreasing. The resulting marginal surplus $S'(x, z) \stackrel{def}{=} \partial S(x, z)/\partial x$ is decreasing in x and becomes negative when consumption is large.

Figure 1 depicts the optimal (and competitive-equilibrium) short-run dis-

patching of a given capacity K when z, then marginal surplus (= demand) $S'(x,z)$, takes different values.

i) For medium values of demand (in the neighborhood of z_3), production is flexible so that market balancing is done by quantity adjustment. The competitive price is just equal to the unit operation cost c.

ii) For high values of demand (in the neighborhood of z_4), production is inflexible. Market balancing results from price increases above the unit operation cost (or, if prices cannot increase, from demand rationing). The margin $S^{'}(K,z)-c>0$ contributes to the payment of capacity cost r.

iii) For low values of demand (in the neighborhood of z_2), production is inflexible. Market balancing results from price decreases below the unit operation cost (or, if prices cannot decrease, from supply rationing). The margin loss $S^{'}(K,z)-c<0$ impairs the recovery of capacity cost r.

iv) For very low values of demand (in the neighborhood of z_1), production is also downward inflexible. Market balancing results in negative prices, i.e. the producer is ready to pay consumers to get rid of excess production.

The combination of *i*) and *ii*) is the basis of the so-called peak-load pricing theory: the costs of inflexible capacity must be billed to consumers only in states of nature where the capacity constraint is binding (Boiteux, 1949). The recognition of cases *iii*) and *iv*) is more recent in the economic literature on energy markets. It is mainly motivated by the development of solar and wind energy that lowers residual demand to thermal producers and some states of nature. (see e.g. Nicolosi, 2010)

Merit order Overall, the inflexibility of individual plants is alleviated by stacking heterogeneous plants that differ in terms of installation and production costs. Let c_i denote the cost of producing 1MWh and r_i the yearly cost of installing and maintaining 1MW with technology i. Then providing 1MW of power during h hours costs $hc_i + r_i$ per year. It results that the choice between two types of equipment, say 1 and 2 with $c_2 > c_1$ and $r_2 < r_1$, depends on the expected working duration:

$$hc_2 + r_2 \gtrless hc_1 + r_1 \Longleftrightarrow h \gtrless \frac{r_1 - r_2}{c_2 - c_1}$$

For base demand, all along the 8760 hours of the year (minus the periods of maintenance), the least-cost choice is technology 1, the one with high capacity cost r_1 and low operation cost c_1. Technology 2 will be preferred for low duration demand. Given the demand levels expected throughout the year, one can order demands starting from the highest requested capacity to obtain the so-called load-duration curve, a curve that shows how long demand will be below a given capacity level. Then one can adapt investment in K_1 and K_2 to meet these demand requirements. After the installation is done, capacity costs are sunk. Then, only operation cost matters. It results that, when demand is low ($x < K_1$), only plant 1 is dispatched since $c_1 < c_2$. When $x > K_1$, being constrained by

the upward inflexibility of equipment 1, the dispatcher calls type 2 plants to complement the supply from K_1. Then, as long as $K_1 < x < K_1 + K_2$, demand can be supplied by the full capacity of plant 1 and the flexible production $x - K_1$ from plant 2 at the operating cost c_2. This is the merit order, that is, the staircase supply function that gives some flexibility to the entire electric system.[4] In the same vein, the interconnections provide some flexibility since the dispatcher can rely on heterogeneous technologies located in adjacent regions (Bistline, 2017).

Ramping rates The above reference cases contrast capacity and production. Actually, inflexibility is also a matter of variation in production in a given period of time. Denoting by x_t the quantity produced at date t, the producer is constrained by

$$-\underline{\Delta}_\delta \leq x_{t+\delta} - x_t \leq \overline{\Delta}_\delta$$

where $\overline{\Delta}_\delta$ and $\underline{\Delta}_\delta$ are the maximum variations in production that are sustainable between dates t and $t + \delta$, upwards and downwards respectively. Production is dynamically constrained to remain within a tunnel, the size of which increases with time lapse δ.

Flexibility costs All the cases where inflexibility is characterized by physical constraints, like the ones considered in the former subsections, represent engineers' concerns. This does not mean that this modeling does not include a reference to cost. Actually, the cost dimension appears as the dual variable associated to each physical constraint when the dispatching program is launched. However, it remains true that on pure engineering grounds, costs do not matter very much. Physical performance is the main concern.[5]

By contrast, in the following sections, we consider explicit differences in cost levels depending on the time lapse separating decision and operation, namely the difference in costs when production must be planned before the state of (residual) demand is known and when decision can be delayed until the revelation time of demand level.

2.2 Ex ante and ex post decisions

Changing the level of production can be done at different speeds, then at different costs. From now on we focus on the cost feature of flexibility. We do not use a continuous time model where the adaptation duration would be endogenous. Instead we rely on a two period model where production can be dispatched among

[4] See for example the Aggregated Curves in the Market Results at https://www.apxgroup.com//.

[5] Similarly, there is no reference to cost minimization in the following quotation of the European energy regulators: "*Flexibility can be defined as the ability of the electricity system to respond to fluctuations of supply and demand while, at the same time, maintaining system reliability*" (ACER and CEER, 2017).

the periods at different costs. The two periods are distinct by the information on the surplus function, then on demand. In period 1 (ex ante), demand is still random. In period 2 (ex post), the decision maker has perfect information on the surplus function. We keep the notation $S(x,z)$ to represent the surplus derived from the consumption of x when the state of nature is z, and $S'(x,z)$ for marginal utility, that is, the derivative of $S(x,z)$ with respect to its first argument.

We begin by assuming that the cost function is the same in the two periods. This is a critical assumption that we relax later.

2.2.1 Gains from delaying decisions

Let $C(x)$ be the cost of producing $x \geq 0$, whatever the period. This is an increasing, convex and continuously differentiable function. The surplus function $S(x,z)$ has the same property as before. Additionally, we assume $S'(0,z) > C'(0)$ for all z, so that production is always profitable. The random variable z is distributed according to a continuous density dG on a compact interval $[s,t]$.

- If x is to be chosen before knowing the realized value of z, the problem to solve is

$$\underset{x\geq 0}{\text{Max}}\, I\!E\left[S(x,z) - C(x)\right]. \tag{1}$$

The ex ante solution is then the value $\widehat{x}$ that solves $I\!E\left(S'(x,z)\right) = C'(x)$, hence the expected performance $\underline{W} = I\!E\left[S(\widehat{x},z) - C(\widehat{x})\right]$.

- If the decision maker can wait until the state of nature is revealed, the problem is, in each state z :

$$\underset{x\geq 0}{\text{Max}}\, S(x,z) - C(x) \tag{2}$$

The ex post solution is the function $x(z)$ that solves $S'(x,z) = C'(x)$, hence an expected performance $\overline{W} = I\!E\left[S(x(z),z) - C(x(z))\right]$.

Since the expectation of a supremum is at least the supremum of the expectations, we have that

$$\overline{W} > \underline{W}. \tag{3}$$

The intuition is that ex post the decision maker has more accurate information on the state of demand (or technology, or regulation, etc.) than ex ante. He/she will therefore make a more appropriate decision for the circumstances, and the performance will consequently be better.[6] Whenever it is possible, the producer isolated from strategic considerations is better off by delaying decisions until the very last minute.[7]

[6] Sandmo (1971) and Leland (1972) were the first authors to provide an analytical treatment of the advantages for firms of delaying decisions.

[7] The gains from flexibility are emphasized by Goutte and Vassilopoulos (2017). They distinguish between the "immediacy" value as the urgency of the delivery increases (this value is

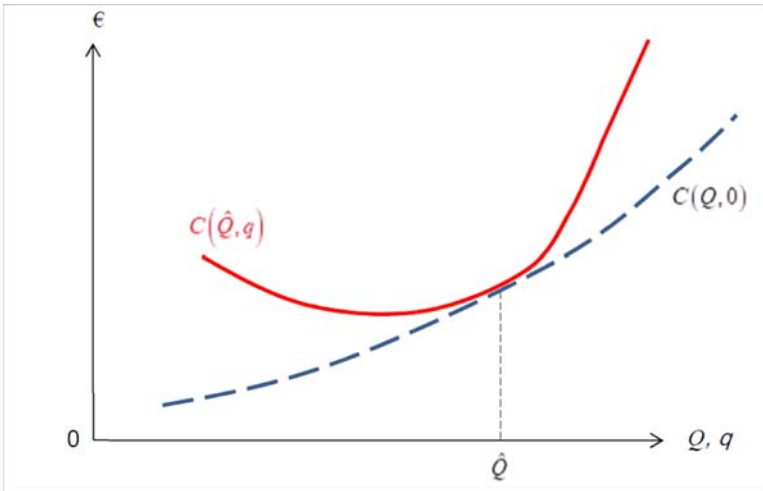

Figure 2: Costs of planned and adjusted production

2.2.2 Costs from delaying decisions

Actually the cost for doing so is different from the cost incurred when planning in advance. This is the essence of flexibility: taking advantage of better information at the expense of higher production costs. If we assume that the production necessary to satisfy demand x can be partly planned ex ante (denoted by Q) and partly produced ex post (denoted by q), we can write the cost function as $C(Q, q)$ with $Q + q = x$, and $C(x, 0) < C(0, x)\ \forall x$. In Figure 2 we have plotted a curve $C(Q, 0)$ representing the low cost of producing "at regular speed" and a curve $C(\widehat{Q}, q)$ representing the cost of adding to or removing from $\widehat{Q}$ the adjustment q.

In this more realistic framework, the decision maker solves:

$$\underset{Q}{\text{Max}}\ I\!E_z \left(\max_q\ S(Q+q, z) - C(Q, q) \right). \tag{4}$$

Ex post, knowing Q and z, the entrepreneur determines $q(Q, z)$ like in (2), except that a volume Q is already available. Ex ante, anticipating this adjustment, he decides upon Q like in (1) . We then face a trade-off between the informational advantages of a delayed decision and the extra costs due to unplanned adjustment. As shown in the following, it results that with $Q > 0$ and $q \neq 0$, the performance is generically higher than if either $Q = 0$ or $q = 0$.

revealed during the intraday process and is closely linked to risk) from the "flexibility" as a resource can capture variations of shorter granularity (more related to asset optimization and already priced a day ahead).

3 Optimal planning and optimal adjustment

In this section, we analyze the trade-off between information gain and adjustment cost in a general framework. We successively determine the first-best allocation and the competition allocation under two market designs: first when there are both day-ahead and adjustment markets, then when there are only adjustment markets.

3.1 First best

We assume a non-specified surplus function $S(Q+q,z)$ continuously differentiable concave in each state z and a non-specified cost function $C(Q,q)$ continuously differentiable convex. The problem to solve is (4) . We solve it backwards.

- Maximizing the net surplus in state z, the first order condition is

$$S^{'}(Q+q,z) = C^{'}_{q}(Q,q) \tag{5}$$

with $C^{'}_{q} \overset{def}{=} \partial C(Q,q)/\partial q > 0$. From (5) we obtain the optimal adjustment $q(Q,z)$. Differentiating the condition, we have that

$$\frac{\partial q(Q,z)}{\partial Q} = \frac{S^{"} - C^{"}_{qQ}}{C^{"}_{qq} - S^{"}}$$

where $S^{"} \overset{def}{=} \partial^2 S(x,z)/\partial x^2 < 0$, and $C^{"}_{qq} \overset{def}{=} \partial^2 C(Q,q)/\partial q^2 > 0$. Then the denominator is positive. If the cross second derivative $C^{"}_{qQ}$ is nil or positive (or negative but small in absolute value) then $\frac{\partial q}{\partial Q} < 0$. Indeed, when z necessitates an upward (downward) adjustment, the adjustment will be small (large) if Q has been fixed at a large value.

- Ex-ante, the problem to solve is

$$\max_{Q} I\!E_z W(Q, q(Q,z); z)$$

where $W(Q,q;z) = S(Q+q,z) - C(Q,q)$.

Given the anticipated ex post adjustments (5), the first order condition that determines the optimal planned output Q^* and the resulting optimal adjustment in each state of nature $q(Q^*,z)$ is

$$I\!E_z \left[S^{'}(Q^* + q(Q^*,z), z) - C^{'}_{Q}(Q^*, q(Q^*,z)) \right] = 0 \tag{6}$$

where $S^{'}$ and $C^{'}_{Q}$ denote partial derivatives wrt Q.

3.2 Perfect competition

We now analyze how first best can be implemented by market mechanisms. In this section, we consider competitive markets were all agents are price-takers. Consumers are supposed to be price reactive, either directly or through retailers and energy service providers.

3.2.1 Day-ahead and adjustment markets

The standard organization for electricity trade in liberalized economies is a set of two successive markets: i) a day-ahead market where competitive producers and consumers trade quantity Q at price P, followed by ii) a balancing market (given state of nature z) where they trade quantity q at price p.

- In the adjustment market determined by z, the supply function is the solution to $\max_q pq - C(Q,q)$ and demand is the solution to $\max_q \ S(Q+q,z) - pq$. From the first order conditions $S^{'}(Q+q,z) - p = 0$ and $C_q^{'}(Q,q) - p = 0$, we have the inverse demand function $p^d(q,Q,z) \stackrel{def}{=} S^{'}(Q+q,z)$ and the inverse supply function $p^s(q,Q) \stackrel{def}{=} C_q^{'}(Q,q)$ respectively.

Matching demand and supply,

$$p^d(q,Q,z) = p^s(q,Q) \Longrightarrow S^{'}(Q+q,z) = C_q^{'}(Q,q) \tag{7}$$

we obtain the equilibrium quantity $q(Q,z)$ and price $p(Q,z)$.

- In the day-ahead market, demand is the solution to

$$\max_Q \ I\!E_z \left[S\left(Q + q\left(Q,z\right),z\right) - p(Q,z) q\left(Q,z\right) \right] - PQ \tag{8}$$

where $\partial p(Q,z)/\partial Q \equiv 0$ since consumers are price-takers and $S^{'}(Q+q\left(Q,z\right),z) - p\left(Q,z\right) = 0$ by the anticipated behavior on the adjustment market. Consequently, from the first order condition we obtain the inverse demand function

$$P^d(Q) \stackrel{def}{=} I\!E_z \left(S^{'}(Q + q\left(Q,z\right),z) \right). \tag{9}$$

Producers solve

$$\max_Q \ PQ + I\!E_z \left[p(Q,z) q\left(Q,z\right) - C\left(Q, q\left(Q,z\right)\right) \right] \tag{10}$$

. Again, $\partial p(Q,z)/\partial Q \equiv 0$ since they are price-takers and $C_q^{'}(Q,q) - p\left(Q,z\right) = 0$ by the ex post adjustment. Consequently, the inverse supply funtion is

$$P^s(Q) \stackrel{def}{=} I\!E_z \left(C_Q^{'}\left(Q, q\left(Q,z\right)\right)\right). \tag{11}$$

The result is the day-ahead equilibrium $P^s(Q) = P^d(Q) \Longrightarrow I\!E_z \left(S^{'}(Q + q\left(Q,z\right),z) \right) = I\!E_z \left(C_Q^{'}\left(Q, q\left(Q,z\right)\right)\right)$, which is the same relationship as (6).

We then conclude the following:

Proposition 3.1. *A market framework made up of a day-ahead competitive market followed by a competitive adjustment market contingent to the observed state of nature allows one to decentralize first best.*

Note that this requires that consumers, or the intermediaries who represent them, be allowed to sell contracted energy on the adjustment market. For low values of z, the adjustment q can be negative, so that the actual consumption $Q+q$ is lower than the planned consumption Q. Then the market mechanism is efficient only if flexibility also comes from demand response.[8]

3.2.2 No day-ahead energy market

In all countries where the two-stage architecture we have just analyzed has been set up, most energy is traded in the day-ahead market.[9] What if there were only ex post (state contingent) markets,[10] i.e. if all quantities, whether planned or adjusted, were sold at the same contingent price?

Ex-post, consumers solve $\max_x \ S(x,z) - px$. Then the demand function in state z is $p^d(q+Q,z) \overset{def}{=} S'(Q+q,z)$. Producers solve $\max_q \ p \times (Q+q) - C(Q,q)$, from which we derive the supply function $p^s(q+Q,z) \overset{def}{=} C'_q(Q,q)$. At equilibrium,

$$p^s(q+Q,z) = p^d(q+Q,z) \Longrightarrow C'_q(Q,q) = S'(Q+q,z) \tag{12}$$

determines the quantity and price equilibrium $q(Q,z)$, $p(Q,z)$.

At the ex ante stage there is no trade. Only competitive producers have to determine the output Q to maximize their expected profit

$$\max_Q \ I\!E_z \left[p(Q,z)(Q + q(Q,z)) - C(Q, q(Q,z)) \right]$$

with $\partial p(Q,z)/\partial Q \equiv 0$ since they are price takers. From the first order condition and using the ex post adjustment condition, the ex ante production plan is the solution to

$$I\!E_z \left[p(Q,z) - C'_Q(Q, q(Q,z)) \right] = 0 \tag{13}$$

Now, observe that at ex post equilibrium in state z, we have $S'(Q+q(Q,z),z) = p(Q,z)$. Combining this with (13), we obtain the same condition as (6).

We then conclude the following:

Proposition 3.2. *Under perfect competition, first best can be reached with ex post state-contingent markets and no day-ahead market. Even though producers are only partially flexible, the day-ahead market is not necessary.*

[8] On demand response, see Crampes and Léautier (2012).

[9] Actually, the largest fraction is traded through bilateral contracts. Here, we refer only to energy traded in a pool.

[10] We do not consider the case where only day-ahead trade is organized and there is no exchange for ex post adjustement. That would entail energy outage or energy waste depending on the ex post demand value.

We can explain this result as follows. When there is a day-ahead market, the demand function used for planned production (9) is based on the expectation of ex post prices. Then in the only-expost market scenario firms sell zero kWh ex ante but determine their output targeting the expectation of ex post prices while when day-ahead trade is possible, producers do sell their ex ante production at a price equal to the expectation of ex post prices. Consequently, their expected profits are the same in the two scenarios, which entails that they take the same decisions. This also explains why, in the two-market scenario, consumers buy on the day-ahead market whereas they could wait until the state of nature is revealed. Since they can resell excess contracted quantity (a crucial hypothesis), with an ex ante equilibrium price equal to the expectation of the ex post equilibrium prices they are indifferent between buying ex ante or ex post. The equality between the ex ante price and the expectation of ex post prices plays the role of a no-arbitrage condition.

With an ex post market for each state of nature, the day-ahead market is redundant to implement first best. Then, why do day-ahead markets exist in energy exchanges such as epexspot?[11] Several reasons can be mentionned: for example the lack of ex post markets in actual design, last-minute transaction costs, the time required to match demand and supply in uniform-price auctions, the time to check that the dispatching is feasible given the grid constraints, and the lack of demand-response. Maybe more important is agents' risk aversion. In our model, both the producer and the consumer are risk neutral. That is obvious for the producer as it maximizes its expected profit. The consumer maximizes their expected net surplus which is measured in monetary terms. With such a quasi-linearity structure, the model fails to bring out the reluctance of agents to risky outcomes, then the attractiveness of day-ahead transactions.

4 Linear specification

To illustrate the flexibility problem, we use a multilinear cost function and a quadratic surplus function. After a presentation of the notations and properties of the model, in subsection 4.2 we determine the optimal dispatching and in subsection 4.3 we consider the specific case of a uniform distribution of the demand parameter. Finally, in subsection 4.4 we address the case of heterogeneous technologies.

[11] See https://www.epexspot.com/en/

4.1 Cost and surplus

4.1.1 Cost function

Assume that the total cost for producing $Q + q \geq 0$ is given by:

$$C(Q,q) = \begin{cases} \alpha Q + aq & \text{if} \quad q \geq 0 \\ \alpha Q & \text{if} \quad q \leq 0 \end{cases},$$

where $Q \geq 0$ is to be chosen ex ante and the decision on q can be delayed.[12] The parameters satisfy $a \geq \alpha > 0$. This is a particular case of the curves in Figure 2, with a linear cost of planned production, $C(Q,0) = \alpha Q$, a linear cost of adjusted production $C(\widehat{Q},q) = \alpha\widehat{Q} + aq$ on the right of $\widehat{Q}$, and an horizontal line on the left of $\widehat{Q}$. Note that in this simplified modeling the adjustment cost has the same shape whatever the value of the planned production since $\partial^2 C/\partial q \partial Q \equiv 0$.

The above cost function means that all planned costs are sunk: there is no gain from decreasing production below the planned volume. This modeling is similar to that of Dixit (1980). However the Dixit's model is an approach of entry deterrence by an incumbent firm that invests in sunk capacity before producing and selling on a deterministic market.

The asymmetry between costly upward and free downward adjustments is a useful simplifying hypothesis when the objective is just to illustrate the general results of the former sections. Actually, thermal plants can save at least on fuel and emission permits when producing below the planned level. In this first approach, we want to emphasize the cost of increasing production in the very short term since the main fear of private and public decision-makers in the electricity industry remains the risk of black-outs. As noted in subsection 2.1, downward inflexibility is becoming a new concern because of the development of non-reliable renewables.

The cost function $C(Q,q)$ is convex in q. If $q \geq 0$, we have $C(Q,q) - C(Q+q,0) = (a-\alpha)q \geq 0$, which is the extra cost for producing $Q+q$ after planning only Q. Similarly if $q \leq 0$, we have $C(Q,q) - C(Q+q,0) = -\alpha q \geq 0$, which is the extra cost for producing $Q+q$ after planning Q, instead of producing directly $Q+q \leq Q$ in the first stage.

4.1.2 Maximum and minimum net surplus

Assume that the gross surplus from consuming x in state z is given by :

$$S(x,z) = (z - \frac{x}{2})x,$$

[12]Because we focus on short-term and very short-term decisions, we discard the capacity constraints and costs. These costs are partially financed through capacity mechanisms. For an analysis of the complex relationship between markets for energy, availability and flexibility on the one hand, and capacity mechanisms on the other hand, see Höschle et al. (2017).

where z is distributed according to the c.d.f. $F(z)$ with a continuous density $dG(z)$ on the interval $[s,t]$, where we assume $s \geq a$ (so that, for each $z \geq s$, $S'_x(0,z) = z \geq a \geq \alpha = C'_Q(0,q)$).

If it was possible to wait until the realization of the event to decide on production, since $a \geq \alpha$ it would be optimal to set $q = 0$ and to choose Q so as to maximize $(z - \frac{Q}{2})Q - \alpha Q$, which gives the state-contingent output $Q = z - \alpha$. The expected performance would be the maximum maximorum $\overline{W} = \frac{1}{2}\mathbb{E}\ (z-\alpha)^2$. By contrast, if all production had to be chosen ex ante, production would be based on the expected value $\mathbb{E}(z)$. With the non-contingent output $Q = \mathbb{E}(z) - \alpha$, and no ex-post adjustment $q = 0$, we would obtain a low performance equal to $\underline{W} = \frac{1}{2}(\mathbb{E}(z) - \alpha)^2$. Given (3), the actual performance W^*, that is the value of W when we solve (4) , will be between these two extreme values:

$$\frac{1}{2}(\mathbb{E}(z) - \alpha)^2 \leq W^* \leq \frac{1}{2}\mathbb{E}(z-\alpha)^2,$$

Notice that $\mathbb{E}\,(z-\alpha)^2 = \mathbb{E}\,(z - \mathbb{E}(z) + \mathbb{E}(z) - \alpha)^2 = \mathbb{E}(z - \mathbb{E}(z))^2 + (\mathbb{E}(z) - \alpha)^2$. Consequently the gap $\frac{1}{2}\mathbb{E}(z-\alpha)^2 - \frac{1}{2}(\mathbb{E}(z) - \alpha)^2 = \frac{1}{2}\mathbb{E}(z - \mathbb{E}(z))^2$ (which equals half the variance of z) is independent of α. Then the larger the demand uncertainty, measured by its variance, the higher the potential benefit from a technology that allows ex-post adjustment.

4.2 Optimal dispatching

4.2.1 Adjustment

Given the quantity Q planned at stage 1, for each realization z of demand observed at stage 2, the adjusted production q should be set so as to maximize $W(Q,q;z) = S(Q+q,z) - C(Q,q)$. Since $W(Q,q;z)$ is concave in q, the first order condition is sufficient to determine the unique solution.

- For $q > 0$, we have $W(Q,q;z)) = (z - \frac{Q+q}{2})(Q+q) - \alpha Q - aq$. Therefore:

$$W'_q(q;Q,z) \geq 0 \Longleftrightarrow q \leq z - (Q+a).$$

- For $q < 0$, we have $W(Q,q;z) = (z - \frac{Q+q}{2})(Q+q) - \alpha Q$. Therefore:

$$W'_q(q;Q,z) \geq 0 \Longleftrightarrow q \leq z - Q.$$

Consequently, if $z > a + Q$, the expost surplus $W(Q,q;z))$ is uniquely maximized for $q = z - (Q+a)$. If $z < Q$, $W(Q,q;z)$ is uniquely maximized for $q = z - Q$. And if $z \in [Q, a+Q]$, the ex-post surplus is uniquely maximized for $q = 0$. We then obtain:

Lemma 4.1. ***Optimal adjustment:***

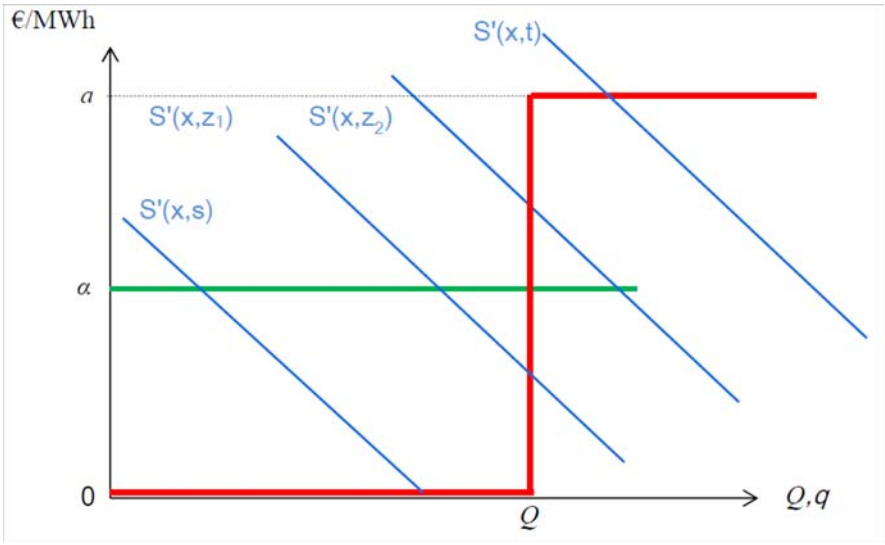

Figure 3: Unit costs and demand span

If $z < Q$, the optimal adjustment is $q = z - Q < 0$, and $W(Q,z) = \frac{1}{2}z^2 - \alpha Q$.

If $z \in [Q, a+Q]$, no adjustment is needed: $q = 0$, and $W(Q,z) = Q(z - \alpha - \frac{Q}{2})$.

If $z > a + Q$, the optimal adjustment is $q = z - a - Q > 0$, and $W(Q,z) = \frac{1}{2}(z-a)^2 + (a-\alpha)Q$.

These adjustments are shown in Figure 3, where $S'_x(x,z) = z - x$ is the marginal surplus in state z. In particular for $Q \in [s, t-a]$, if z is at the lower bound of the support ($z = s$) there is a downward adjustment $q = s - Q < 0$ whereas if it is at its higher bound ($z = t$), there is an upward adjustment $q = (t-a) - Q > 0$. For 'medium' values such as z_1 and z_2, the marginal surplus at Q is $0 < z - Q < a$ so that no adjustment is profitable.

4.2.2 Planned production

Using the optimal adjustment rule of Lemma 4.1, problem (4) consists in $\text{Max}_{Q \geq 0} W(Q)$, with:

$$W(Q) = -\alpha Q + \int_s^Q \frac{z^2}{2} dG(z) + \int_Q^{a+Q} \left[Q\left(z - \frac{Q}{2}\right) \right] dG(z) + \int_{a+Q}^t \left[aQ + \frac{(z-a)^2}{2} \right] dG(z) \tag{14}$$

where the three integrals correspond, in sequence, to $q < 0$, $q = 0$, and $q > 0$.

- Note first that if we set $Q < s - a$, all adjustment must be upwards and the welfare function reduces to $W(Q) = -\alpha Q + \int_s^t \left[aQ + \frac{(z-a)^2}{2} \right] dG(z)$. Consequently, $W'(Q) = a - \alpha > 0$, which means that Q has to be set at least equal to $s - a$.

The simple case where $a = \alpha$ results in $W'(Q) = 0$ for all $Q \leq s - \alpha$. Setting $Q \leq s - \alpha$ and adapting at the second stage up to $z - \alpha$ gives the best possible social benefit $\overline{W} = \frac{1}{2}\mathbb{E}\left((z-\alpha)^2\right)$. This is because increasing production *a posteriori* costs no more than if planned from scratch. Since initial costs are sunk, it is optimal to choose in the first stage any output level no larger than to the lowest value of demand $s - \alpha$, and later to systematically adapt production upwards

- Symmetrically, fixing $Q > t$ cannot be optimal since it would give $W(Q) = -\alpha Q + \int_s^t \frac{z^2}{2} dG(z)$, and then $W'(Q) = -\alpha < 0$. Consequently Q must be set at most equal to t to avoid wasteful downwards adjustment.

- Then $W(Q)$ is given by (14). Its first derivative with respect to Q is

$$W^{'}(Q) = -\alpha + \int_Q^{a+Q} (z - Q)\, dG(z) + a \int_{a+Q}^{t} dG(z). \tag{15}$$

  A slight increase in Q costs α. It has no effect on the gains corresponding to low values of z since Q is already two large in the corresponding states of demand. By contrast, it has the benefit of decreasing the set of events where a costly upward adjustment will be necessary to match demand, as shown by the last term on the right-hand side in (15). For intermediary levels of demand (such as those characterized by z_1 and z_2 in Figure 3) the gains represented by the second term in the right-hand side of (15) are the marginal surpluses provided at zero additional cost thanks to a larger planned production.

The optimal production is then the solution to $W^{'}(Q) = 0$. The explicit value of Q is obviously dependent in the shape of the c.d.f of the demand parameter z. In the following, we consider the case of a uniform distribution.

4.3 Uniform distribution of the demand parameter

The explicit solution to $W^{'}(Q) = 0$ (where $W^{'}(Q)$ is given by (15)) is determined by the shape of the distribution function $F(z)$. To obtain a better understanding of the flexibility effect, assume now that z follows a uniform distribution function with density $D = (t-s)^{-1}\mathbf{1}_{[s,t]}$, so that $dG(z) = Ddz$. We also assume that: $t - s > a$. The adjustment rules of Lemma 4.1 obviously remain unchanged. Given these rules, the solution to max$W(Q)$ for $Q \geq 0$ is given by Proposition 4.2. The proof is in the Appendix.

Proposition 4.2. *The function W is concave and continuously differentiable. It has a unique maximizer Q^*.*

Fixing a, α and the mean $\frac{1}{2}(s+t)$, the maximum $W(Q^*)$ is increasing in the difference $t - s$ (or equivalently in the variance of z).

There are three possible regimes for the optimal first stage production Q^* and the achieved maximum:

A) If $\alpha \geq a\left(1-\frac{a}{2(t-s)}\right)$,

$$Q^* = s - a + \sqrt{2(t-s)(a-\alpha)} \in [s-a, s],$$

$$W^* = \frac{2}{3}(a-\alpha)\sqrt{2(t-s)(a-\alpha)} + (s-a)(a-\alpha) + \frac{1}{6}(t^2+st+s^2+3a^2(t-s)-3a(t+s)).$$

B) If $\frac{a^2}{2(t-s)} \leq \alpha \leq a\left(1-\frac{a}{2(t-s)}\right)$,

$$Q^* = t - \frac{1}{2}a - \frac{\alpha}{a}(t-s) \in [s, t-a]$$

$$W^* = \frac{\alpha^2}{2a}(t-s) - \alpha t + \frac{1}{2}a\alpha + \frac{t^3-s^3}{6(t-s)} - \frac{a^3}{24(t-s)}.$$

C) If $\alpha \leq \frac{a^2}{2(t-s)}$,

$$Q^* = t - \sqrt{2\alpha(t-s)} \in [t-a, t]$$

$$W^* = -\alpha t + \frac{2\sqrt{2}}{3}\alpha^{3/2}\sqrt{(t-s)} + \frac{t^3-s^3}{6(t-s)}.$$

Notice that when $a \leq 2\alpha$, case C) is not possible. So we are left with two possible regimes: A) and B).

Proposition 4.2 can be interpreted as follows. Fix s, t and a such that $t - s > a$. Then let α go from a to 0 so that it becomes more and more costly to adapt production upwards compared to the cost of phase 1. In regime A) where $\alpha \geq a\left(1-\frac{a}{2(t-s)}\right)$, we obtain $Q \in [s-a, s]$. Regime B) occurs when $\frac{a^2}{2(t-s)} \leq \alpha \leq a\left(1-\frac{a}{2(t-s)}\right)$ so that $Q \in [s, t-a]$. Finally, regime C) occurs when $\alpha \leq \frac{a^2}{2(t-s)}$, and production is $Q \in [t-a, t]$. Overall, we see that when α decreases, the planned production Q increases as it becomes less profitable to adapt ex post production upwards.

An immediate consequence of Proposition 4.2 and Lemma 4.1 is the following.

Corollary 4.3. *Given the optimal first stage production Q^*, the optimal adjustment is*

$q^*(z) = z - Q^* < 0$ if $z < Q^*$. This only happens in regimes B) and C).

$q^*(z) = 0$ if $z \in [Q^*, a+Q^*]$. This may happen in all regimes A), B), C).

$q^*(z) = z - a - Q^* > 0$ if $z > a + Q^*$. This only happens in regimes A) and B).

In zone A), a is close to α. It is therefore profitable to fix a low Q during the first stage and a high $q > 0$ during the second, rather than a higher Q at high cost αq and an adjustment q that may be negative. By contrast, in zone C), α is so small that it is optimal to produce most of the output during the first stage.

As shown in section 3.2, perfect competition allows one to implement the outcome of Proposition 4.2 and its corollary. With the current specification, adjustment prices are $p^*(z) = a$ when $q^*(z) > 0$, $p^*(z) = 0$ when $q^*(z) < 0$ and $p^*(z) = z - Q^*$ when $q^*(z) = 0$. The ex ante price (if there is a day-ahead market) is the weighted sum of these ex post prices. Regarding the fact that $W(Q^*)$ is increasing with the variance of z, it is the consequence of the risk neutrality of all agents.[13]

4.4 Heterogeneous technologies

In all electricity markets, heterogeneous technologies coexist, in particular because they have different degrees of flexibility. Ma *et al* (2013) consider the provision of flexibility using a portfolio of generating units with different dynamic characteristics. Their model analyzes the balancing of short term (operation) costs and long term (building) costs of flexibility provision. By contrast we keep on thinking in terms of day-ahead and real-time dispatching, neglecting capacity costs that are already sunk.

Assume there are two technologies ($i = 1, 2$) with costs

$$C_i(Q, q) = \begin{cases} \alpha_i Q + a_i q & \text{if} \quad q \geq 0 \\ \alpha_i Q & \text{if} \quad q \leq 0 \end{cases} \tag{16}$$

where $a_1 > a_2 > \alpha_2 > \alpha_1$.

Clearly, by $\alpha_2 > \alpha_1$, it is sub-optimal to dispatch technology 2 for the planned production, and, by $a_1 > a_2$, to use technology 1 for the adjustment production. Therefore the first-best solution is the same as with a single technology characterized by costs a_2 for basic production and α_1 for upward adjustment.

The only interplay between the two technologies is that a higher a_2 pushes the use of technology 1 at the initial stage up and, with a higher α_1, one can expect a more intensive use of technology 2 at the adjustment stage. Clearly, the result is the same under perfect competition. Only producers endowed with technology 1 start their production day-ahead and only those endowed with technology 2 participate in the real-time market if upward adjustment is required.

A mix of technologies at both stages instead of a full specialization could be justified by relaxing our technological hypotheses, for example in case of capacity constraints, non-(piecewise) linear cost, adjustment cost varying with duration, regulatory requirements, correlated risks of failure, etc.

[13] Oi (1961) proved that competitive producers benefit from price uncertainty because their profit function is convex in p so that they are risk-lovers when the risk comes from price variability.

5 Imperfect competition

The objective of the section is to contrast the one market design and the two-market design when firms have market power. Indeed, in the former section, we have considered market mechanisms only in the case of perfect competition. Hereafter, we assume a duopoly and, given a two-period game setting, we determine the conditions for a subgame-perfect equilibrium when firms compete in quantities. We show that, contrary to the result of Proposition 3.2, equilibrium quantities now differ depending on the number of markets.

5.1 No day-ahead market

Suppose that there is no day-ahead market. Firms sell all their production (made of the planned quantity chosen ex ante and the adjusted quantity chosen ex post) on ex post state-contingent markets.

The duopoly game is as follows:

Stage 1: Firms choose their planned production $Q_i \geq 0$, $i = 1, 2$ simultaneously and independently while demand is still random.

Stage 2: The random parameter z is revealed.

2a: The consumer's demand function is determined and observable by the two firms.

2b: Given Q_1 and Q_2, upon observing demand firms choose their adjusted production q_i, $i = 1, 2$ upwards or downwards, simultaneously and independently.

2c: The entire production $Q_1 + q_1 + Q_2 + q_2$ is sold to meet demand, which determines the equilibrium price and profits.

Let $C^i(Q_i, q_i)$ represent the cost function of firm i.

The profit of firm i is $\Pi_i = p(x, z)(Q_i + q_i) - C_i(Q_i, q_i)$ where $p(x, z)$ is the inverse demand function in state z and $Q_1 + q_1 + Q_2 + q_2 = x$ at demand-supply equilibrium. We solve the model backwards.

- Ex-post
  - Knowing z, the representative consumer solves

    $$\max_x S(x, z) - pz \Longrightarrow S'(x, z) - p = 0$$

    Then the demand function is $p(x, z) = S'(x, z)$
  - Knowing z, Q_1, and Q_2, producer i solves

    $$\max_{q_i} p(Q_i + q_i + Q_{-i} + q_{-i}, z)(q_i + Q_i) - C^i(Q_i, q_i)$$

    The first order condition is (for $i = 1, 2$)

    $$p(Q_i+q_i+Q_{-i}+q_{-i}, z)+p'(Q_i+q_i+Q_{-i}+q_{-i}, z)(q_i + Q_i)-C_q^{i'}(Q_i, q_i) = 0 \quad (17)$$

where $C_q^{i'}\left(Q_i, q_i\right) \stackrel{def}{=} \partial C^i\left(Q_i, q_i\right)/\partial q_i$ and $p^{'}(x,z) = \frac{\partial p(x,z)}{\partial x} < 0$ since firms do have market power.

- From conditions (17) we deduce the best response function of firm i in state z : $q_i^{br} = f_i(q_{-i}; z, Q_i, Q_{-i})$. Combining the two best-response functions, we obtain the Cournot adjustment outputs:

$$q_i^C = f_i(q_{-i}^C; z, Q_i, Q_{-i}) \qquad i = 1, 2$$

and the Cournot price

$$p^{^C}\left(x^{^C}, z\right) = S^{'}(Q_i + Q_{-i} + q_i^C + q_{-i}^C, z)$$

in state z, where $x^{^C} = Q_i + Q_{-i} + q_i^C + q_{-i}^C$.

- Ex-ante, there is no demand. Firm i just launches a fraction Q_i of its total future supply $Q_i + q_i$.

  - firm i solves

$$\max_{Q_i} I\!E_z\left[p^{^C}\left(x^{^C}, z\right)\left(Q_i + q_i^C\right) - C^i\left(Q_i, q_i^C\right)\right]$$

The FOC is

$$I\!E_z[p^{^C}\left(x^{^C}, z\right)\left(1 + \partial q_i^C/\partial Q_i\right) + \frac{dp^{^C}\left(x^{^C}, z\right)}{dQ_i}\left(Q_i + q_i^C\right) - \frac{dC^i\left(Q_i, q_i^C\right)}{dQ_i}] = 0 \quad (18)$$

where

$$\frac{dp^{^C}\left(x^{^C}, z\right)}{dQ_i} = S^{"}\left(Q_i + Q_{-i} + q_i^C + q_{-i}^C, z\right)\left(1 + \partial q_i^C/\partial Q_i + \partial q_{-i}^C/\partial Q_i\right)$$

since a change in Q_i has a direct effect not only on future contingent prices but also on the quantities supplied by i and $-i$ in the adjustment market with indirect consequences on future prices, and

$$\frac{dC^i\left(Q_i, q_i^C\right)}{dQ_i} = C_Q^{i'}\left(Q_i, q_i^C\right) + C_q^{i'}\left(Q_i, q_i^C\right)\partial q_i^C/\partial Q_i.$$

Since $p'(x,z) \equiv S^{"}\left(Q_i + Q_{-i} + q_i^C + q_{-i}^C, z\right)$, after injecting (17) into (18) and simplifying we obtain

$$I\!E_z[p(x^C, z) + p^{'}(x^C, z)\left(Q_i + q_i^C\right)\left(1 + \frac{\partial q_{-i}^C}{\partial Q_i}\right) - C_Q^{i'}\left(Q_i, q_i^C\right)] = 0. \qquad (19)$$

- - from conditions (19) we deduce the best response function of firm i in stage 1: $Q_i^{br} = F_i(Q_{-i})$. Combining the two best-response functions, we obtain the Cournot planned outputs:

$$Q_i^C = F_i(Q_{-i}^C) \qquad i = 1, 2$$

5.2 Day-ahead and adjustment markets

Assume now a framework made of i) a day-ahead market, followed by ii) adjustment markets. Contrary to the former framework, at stage 1 firms sell $Q = Q_1 + Q_2$ on the day-ahead market at price P, and at stage 2 they sell $q = q_1 + q_2$ at price p in the adjustment market.

- Ex-post,
  - knowing z, Q_1, and Q_2, producer i solves

$$\max_{q_i} p(Q+q,z)q_i - C^i(Q_i,q_i)$$

    where the demand function still is $p(Q+q,z) = S^{'}(Q_i+q_i+Q_{-i}+q_{-i},z)$

The first order condition is

$$p(Q_i+q_i+Q_{-i}+q_{-i},z)+p^{'}(Q_i+Q_{-i}+q_i+q_{-i},z)q_i - C_q^{i'}(Q_i,q_i) = 0 \qquad i=1,2. \tag{20}$$

- From conditions (17) we deduce the best response function of firm i in state z : $\widetilde{q}_i^{br} = \widetilde{f}_i(q_{-i};z,Q_i,Q_{-i})$. Combining the two best-response functions, we obtain the Cournot adjustment outputs:

$$\widetilde{q}_i^C = \widetilde{f}_i(q_{-i}^C;z,Q_i,Q_{-i}) \qquad i=1,2$$

  and the Cournot price

$$\widetilde{p}^C\left(\widetilde{x}^C,z\right) = S^{'}(Q_i+Q_{-i}+\widetilde{q}_i^C+\widetilde{q}_{-i}^C,z)$$

  in state z, where $\widetilde{x}^C = Q_i+Q_{-i}+\widetilde{q}_i^C+\widetilde{q}_{-i}^C$.

- Ex-ante, firm i solves

$$\max_{Q_i} P^d(Q_1,Q_2)Q_i + I\!E_z\left[\widetilde{p}^C\left(\widetilde{x}^C,z\right)\widetilde{q}_i^C - C^i\left(Q_i,\widetilde{q}_i^C\right)\right]$$

where the ex-ante demand is $P^d(Q_1,Q_2) = I\!E_z\left(S^{'}(Q+q(z),z)\right)$ with $q(z) = \widetilde{q}_1^C+\widetilde{q}_2^C$.

- Given the expected adjustments, the FOC is

$$P^d(Q_1,Q_2)+\frac{\partial P^d}{\partial Q_i}Q_i + I\!E_z\left[\frac{d\widetilde{p}^C\left(\widetilde{x}^C,z\right)}{dQ_i}\widetilde{q}_i^C + \widetilde{p}^C\left(\widetilde{x}^C,z\right)\frac{\partial\widetilde{q}_i^C}{\partial Q_i} - \frac{dC^i\left(Q_i,\widetilde{q}_i^C\right)}{dQ_i}\right] = 0$$

where

$$\frac{d\widetilde{p}^{C}\left(\tilde{x}^{C},z\right)}{dQ_i}=p'(x,z)\left(1+\partial\widetilde{q}_i^{C}/\partial Q_i+\partial\widetilde{q}_{-i}^{C}/\partial Q_i\right)$$

and $\frac{dC^i\left(Q_i,\widetilde{q}_i^{C}\right)}{dQ_i}=C_Q^{i'}\left(Q_i,\widetilde{q}_i^{C}\right)+C_q^{i'}\left(Q_i,\widetilde{q}_i^{C}\right)\partial\widetilde{q}_i^{C}/\partial Q_i$.

After substituting these two derivatives, we obtain

$$P^d+\frac{\partial P^d}{\partial Q_i}Q_i+I\!E_z\left[\left(\widetilde{p}^{C}-C_q^{i'}\left(Q_i,\widetilde{q}_i^{C}\right)+\widetilde{q}_i^{C}\widetilde{p}'\right)\frac{\partial\widetilde{q}_i^{C}}{\partial Q_i}+\widetilde{p}'\widetilde{q}_i^{C}\left(1+\frac{\partial\widetilde{q}_{-i}^{C}}{\partial Q_i}\right)-C_Q^{i'}\left(Q_i,\widetilde{q}_i^{C}\right)\right]=0$$

and finally, using (20)

$$P^d+\frac{\partial P^d}{\partial Q_i}Q_i+I\!E_z\left[\widetilde{q}_i^{C}\widetilde{p}'\left(1+\frac{\partial\widetilde{q}_{-i}^{C}}{\partial Q_i}\right)-C_Q^{i'}\left(Q_i,\widetilde{q}_i^{C}\right)\right]=0. \tag{21}$$

- – From these conditions we deduce the best response function of firm i in stage 1: $\widetilde{Q}_i^{br}=\widetilde{F}_i(\widetilde{Q}_{-i})$. Combining the two best-response functions, we obtain the Cournot planned outputs:

$$\widetilde{Q}_i^{C}=\widetilde{F}_i(\widetilde{Q}_{-i}^{C})\qquad i=1,2$$

5.3 Which market design?

Given the shape of the FOCs (19) and (21), outputs are generically different in the two market designs, contrary to what we obtained in the case of perfect competition. To facilitate the comparison, let us rewrite the conditions in a simplified format:

$$\text{Without day-ahead market}\qquad I\!E_z[p+p'\left(Q_i^{C}+q_i^{C}\right)(1+\frac{\partial q_{-i}^{C}}{\partial Q_i})-C_Q^{i'}]=0 \tag{22}$$

$$\text{With day-ahead market}\qquad P+P'\widetilde{Q}_i^{C}+I\!E_z\left[\widetilde{p}'\widetilde{q}_i^{C}\left(1+\frac{\partial\widetilde{q}_{-i}^{C}}{\partial Q_i}\right)-C_Q^{i'}\right]=0 \tag{23}$$

Suppose first that there is only one firm. Clearly, in that case $\frac{\partial q_{-i}}{\partial Q_i}\equiv 0$ and the two conditions are the same if $P=E_z\left[p\right]$. In words, a monopoly produces the same quantity with or without a day-ahead market if there is no possibility of arbitrage by consumers. As usual, comparing the resulting condition $I\!E_z\left[p\left(z\right)+p'\left(Q+q\left(z\right)\right)-C_Q'\right]=0$ to the first-best condition given in (6), that is $I\!E_z\left[S'\left(Q+q\left(z\right),z\right)-C_Q'\right]=0$, with or without a day-ahead market the monopoly produces below the optimal level. In both market designs, it has an incentive to launch its production in advance to benefit from the low costs of planned procedure and to cost adjust ex post to maximise profit in the revealed state of nature.

We now switch back to the duopoly case where $\frac{\partial q_{-i}}{\partial Q_i}$ is not nil. We first show that this term is negative. Let $\pi^{i'}_{q_i}(Q_i, q_i, Q_{-i}, q_{-i}, z) = 0$ represent the ex-post first-order condition of firm i, that is (17) or (20). Total differentiation gives $\pi^{i''}_{q_i q_i} dq_i + \pi^{i''}_{q_i Q_{-i}} dQ_{-i} = 0$ where $\pi^{i''}_{q_i Q_{-i}} = \partial \pi^{i'}_{q_i} / \partial Q_{-i}$ and $\pi^{i''}_{q_i q_i} = \partial \pi^{i'}_{q_i} / \partial q_i$. By the second-order condition of profit maximization, we have $\pi^{i''}_{q_i q_i} < 0$ and, from (17) or (20), $\pi^{i''}_{q_i Q_{-i}} = p^{'} + p^{"} q_i$, which is negative for most demand functions used in industrial organization.[14] Then we obtain

$$\frac{\partial q_i}{\partial Q_{-i}} = -\frac{\pi^{i''}_{q_i Q_{-i}}}{\pi^{i"}_{q_i q_i}} < 0$$

and the same obviously applies to $\frac{\partial q_{-i}}{\partial Q_i}$. This term is an incentive à la Stackelberg to produce more than in a standard Cournot framework. As a matter of fact, by increasing its ex ante output, firm i not only benefits from lower costs but also pre-empts a larger future market share by pushing its competitor aside. This term then represents an incentive to produce more. It alleviates the market power of firms in a way similar to forward contracts in the Allaz and Vila's model (1993). It appears in both (22) and (23) .

However, the term $E_z \left[p^{'} Q_i \frac{\partial q^C_{-i}}{\partial Q_i} \right]$ only appears in (22) . Since it is positive, it shifts the expected marginal revenue of firm i upward. Then expected marginal revenue intersects expected marginal cost for a higher level of production when there is no day-ahead market. The result is that the best response function of both firms is shifted outwards when there is no day-ahead market. At first, we could conclude that the day-ahead market is potentially bad for efficiency. However, things are not that simple. First, because shifting the two best-response functions upwards does not necessarily lead to a larger **total** production. If the two firms are asymmetric, the new equilibrium can be such that one firm is closer to the monopoly outcome leaving the other with a tiny production. Overall, the total ex ante output may be larger when there is a day-ahead market. Second, the final net production depends on ex post adjustment, the sign and size of which depend on the flexibility cost. With a higher ex-ante production, one can expect more downward and less upward adjustment. It is hard then to predict the net result without precise specifications of the cost functions and the distribution of probability of demand.

6 Conclusion

In most industries, there is one single efficient technology in operation at a given date. Innovation makes the installed machines obsolete so that active firms must adapt or die. The main reason is that products can be stored and demand is both

[14] As shown by Novshek (1985), it is a necessary condition for the existence of a Cournot equilibrium.

price-responsive and moderately volatile. Therefore the exact date of production is not crucial. This is not true in the electricity industry where *i*) storage can still accommodate only a tiny fraction of total production, *ii*) demand is permanently changing, both cyclically and randomly, and is weakly price-responsive, and *iii*) producers are politically and socially obliged to satisfy demand at all dates (at least in developed countries). As a result, from the very beginning of the electricity industry, several technologies have coexisted, some with low-variable/high-fixed costs to meet stable demand that can be predicted at least one day in advance, others with high-variable/low-fixed costs to meet short-term variations in demand. Actually, all technologies embed some degree of flexibility. They can accelerate and decelerate but at a cost higher than when they produce steadily. The paper analyzes optimal dispatching and market outcomes for technologies that can be operated at low cost day-ahead when demand is only known in probability, and at high cost for last-minute adjustment to observed actual demand.

In the last twenty years or so, developed countries have liberalized the electricity industry, in particular by opening wholesale markets. The design of these has been very similar in all countries, with a day-ahead market followed by intra-day adjustment markets. Our paper shows that when all agents are price-takers and risk-neutral making competition in the wholesale market efficient given demand uncertainty does not necessitate a day-ahead market. By contrast, when producers have some market power, trading only on ex-post markets or on a combination of ex-ante and ex-post markets is not the same. Determining which market design is more socially efficient necessitates the specification of demand, cost and uncertainty, and the use of simulations.

References

[1] ACER and CEER (2017), "Facilitating flexibility", White Paper # 3 relevant to European Commission's Clean Energy Proposals, 22 May

[2] Allaz B. and J. L. Vila (1993), "Cournot Competition, Forward Markets and Efficiency", Journal of Economic Theory, Volume 59, Issue 1, February, Pages 1-16

[3] Bertsch J., C. Growitsch, S. Lorenczik and S. Nagl (2012), "Flexibility options in European electricity markets in high RES-E scenarios", Institute of Energy Economics at the University of Cologne, October.

[4] Bertsch J., C. Growitsch, S. Lorenczik and S. Nagl (2016), "Flexibility in Europe's power sector — An additional requirement or an automatic complement?", Energy Economics 53, 118–131

[5] Bistline, J.E. (2017), "Economic and technical challenges of flexible operations under large-scale variable renewable deployment", Energy Economics 64, 363–372.

[6] Boiteux, M. (1949), « La tarification des demandes de pointe : Application de la théorie de la vente au coût marginal », Revue générale de l'électricité. English translation « Peak Load Pricing », The Journal of Business, 1960, 33(2), 157-179.

[7] Crampes C. and T.-O. Léautier (2012), « Distributed Load-Shedding in the Balancing of Electricity Markets », Loyola de Palacio Programme on Energy Policy, http://idei.fr/sites/default/files/medias/doc/by/crampes/distributed.pdf

[8] Dixit, A. (1980), "The Role of Investment in Entry-Deterrence". The Economic Journal, 90, 95-106

[9] Eisenack K. (2015), "Peak-load pricing with different types of dispatchability: there can only be one", WP, January 19, Carl von Ossietzky University Oldenburg.

[10] Finon D. (2015), "Le besoin de marchés de la flexibilité: l'adaptation du design des marchés électriques aux productions d'énergies renouvelables", CEEM Working Paper 2015-13

[11] Goutte S. and P. Vassilopoulos (2017), "The value of flexibility in power markets", Working Paper #26, Dauphine University.

[12] Höschle H., C. De Jonghe, H. Le Cadre, and R. Belmans (2017), Electricity markets for energy, flexibility and availability — Impact of capacity mechanisms on the remuneration of generation technologies, Energy Economics 66, 372–383.

[13] Kök A.G, K. Shang and S. Yücel (2016), "Investments in Renewable and Conventional Energy: The Role of Operational Flexibility", Georgetown McDonough School of Business Research Paper No. 2856013, October

[14] Leland H.E. (1972) "Theory of the Firm Facing Uncertain Demand", American Economic Review 62, June, 278-291.

[15] Lykidi M. and P. Gourdel (2015), "How to manage flexible nuclear power plants in a deregulated electricity market from the point of view of social welfare", Energy, p. 1-14

[16] Ma J., V Silva, R Belhomme, D S Kirschen, L F Ochoa (2013), "Evaluating and Planning Flexibility in Sustainable Power Systems", IEEE Transactions on Sustainable Energy, Vol. 4, No. 1, January, pp. 200-209

[17] Nicolosi M. (2010), "Wind power integration and power system flexibility– An empirical analysis of extreme events in Germany under the new negative price regime", Energy Policy, vol 38, Issue 11, November, p. 7257-7268

[18] Novshek, W., (1985), "On the Existence of Cournot Equilibrium", Review of Economic Studies L(II), pp. 85-98.

[19] Oi W. (1961), "The desirability of price instability under perfect competition", Econometrica, January, Vol. 29, 1, p. 58-64.

[20] Pescia D. (2017), "Flexibility in thermal power plants", Agora Energiewende, June

[21] PJM (2017), "Proposed Enhancements to Energy Price Formation", PJM Interconnection, November 15, http://www.pjm.com/-/media/library/reports-notices/special-reports/20171115-proposed-enhancements-to-energy-price-formation.ashx

[22] Sandmo A. (1971), "On the Theory of the Competitive Firm under Price Uncertainty," American Economic Review, March , 61, 65-73

[23] Silva V. (2010), "Value of flexibility in systems with large wind penetration", Imperial College London, October; https://tel.archives-ouvertes.fr/tel-00724358/document

7 Appendix: Proof of proposition 4.2:

In the maximization of $W(Q)$ for $Q \geq 0$, there are several cases to consider depending on the relative positions of Q and $a+Q$ compared to s and t.

• Zone 1: $Q \leq s-a$.

Then $W(Q) = -\alpha Q + aQ + \frac{1}{2}\mathbb{E}((z-a)^2)$. This expression is strictly increasing in Q, so the maximum in zone 1 is achieved for $Q = s-a$.

• Zone 2: $s-a \leq Q \leq s$. So $a+Q \leq s+a \leq t$. Here,

$$W(Q) = \mathbb{P}(z \in [s, a+Q])\, Q\, \mathbb{E}((z-\alpha-\frac{Q}{2})|z \in [s, a+Q])$$

$$+\mathbb{P}(z > a+Q)\left[(a-\alpha)Q + \frac{1}{2}\mathbb{E}((z-a)^2|z > a+Q)\right]$$

Computations give[15]

$$(t-s)W(Q) = -\frac{1}{6}Q^3 + \frac{1}{2}Q^2(s-a) + Q(-\alpha(t-s) - \frac{1}{2}s^2 - \frac{1}{2}a^2 + at) + \frac{1}{6}(t-a)^3.$$

[15] If X is a random variable uniformly distributed on an interval $[s,t]$, then $\mathbb{E}(X^2) = \frac{1}{3}(s^2 + st + t^2)$.

$$(t-s)W'(Q) = -\frac{1}{2}Q^2 + Q(s-a) + -\alpha(t-s) - \frac{1}{2}s^2 - \frac{1}{2}a^2 + at.$$

The equation $W'(Q) = 0$ has 2 roots $Q_1 = s - a - \sqrt{\Delta}$ and $Q_2 = s - a + \sqrt{\Delta}$, where $\Delta = 2(t-s)(a-\alpha) > 0$. Notice that $Q_1 < s - a < Q_2$, and we have $Q_2 > s$ if and only if $2(t-s)(1-\alpha/a) \geq a$. There are two possible cases:

If $2(t-s)(1-\alpha/a) \leq a$, the maximum of W in zone 2 is achieved at $Q_2 = s - a + \sqrt{2(t-s)(a-\alpha)}$.

If $2(t-s)(1-\alpha/a) \geq a$, W is increasing in zone 2, and the maximum of W in this zone is achieved at s.

• Zone 3: $s \leq Q \leq t - a$.

Here, $(t-s)W(Q)$ reads:

$$-\alpha Q(t-s) + (Q-s)\frac{1}{6}(s^2+Q^2+sQ) + \frac{1}{2}aQ(a+Q) + (t-a-Q)(aQ + \frac{1}{6}(Q^2+(t-a)^2+Q(t-a))).$$

Hence

$$(t-s)W'(Q) = -aQ - \alpha(t-s) + ta - \frac{1}{2}a^2.$$

So we have $W'(Q) > 0$ if and only if $Q < Q_3$, where

$$Q_3 := t - \frac{1}{2}a - \frac{\alpha}{a}(t-s).$$

We have:

$$Q_3 \geq s \iff (1 - \frac{\alpha}{a})(t-s) \geq \frac{a}{2},$$
$$Q_3 \leq t - a \iff \frac{\alpha}{a}(t-s) \geq \frac{a}{2}.$$

And we have three possible cases for zone 3:

If $Q_3 \leq s$, then W is decreasing in zone 3 and the maximum of W in this zone is achieved for $Q = s$. If $Q_3 \geq t - a$, then W is increasing in zone 3 and the maximum of W in this zone is achieved for $Q = t - a$. If $s \leq Q_3 \leq t - a$, the maximum of W in this zone is achieved for $Q = Q_3$.

• Zone 4: $t - a \leq Q \leq t$. Here,

$$W(Q) = I\!P(z \in [s,Q]) \left[-\alpha Q + \frac{1}{2}I\!E(z^2 | z \in [s,Q])\right] + I\!P(z \in [Q,t])\, Q\, I\!E((z - \alpha - \frac{Q}{2}) | z \in [Q,t]).$$

Computations give:

$$(t-s)W(Q) = -\alpha Q(t-s) - \frac{1}{6}(t-Q)^3 + \frac{1}{6}(t^3 - s^3).$$

So that $(t-s)W'(Q) = -\alpha(t-s) + \frac{1}{2}(t-Q)^2$, and we have $W'(Q) > 0$ if and only if $Q < Q_4$, where

$$Q_4 := t - \sqrt{2\alpha(t-s)}.$$

$$Q_4 \geq t - a \Longleftrightarrow \alpha(t-s) \geq \frac{a^2}{2}.$$

And we have two possible cases for zone 4:

If $Q_4 \leq t - a$, then W is decreasing in zone 4 and the maximum of W in this zone is achieved for $Q = t - a$. If $Q_4 \geq t - a$, then the maximum of W on this zone is achieved for $Q = Q_4$.

• Zone 5: $t \leq Q$. It is easily seen that W is decreasing on this zone, so the maximum of W on this zone is achieved for $Q = t$.

Notice that:

$$(1 - \frac{\alpha}{a})(t-s) \leq \frac{a}{2} \Longleftrightarrow Q_2 \leq s \Longleftrightarrow Q_3 \leq s.$$

and if $(1 - \frac{\alpha}{a})(t-s) \geq \frac{a}{2}$, we have:

$$\frac{\alpha}{a}(t-s) \geq \frac{a}{2} \Longleftrightarrow Q_3 \leq t - a \Longleftrightarrow Q_4 \leq t - a.$$

One can check that W is C^1 and concave on $\mathbb{R}_+$, and we obtain the solution to the planification problem as announced in proposition 4.2.

Fix finally a, α and the mean $m = \frac{1}{2}(s+t)$, and write $l = t - s$.

In zone A,

$$\frac{\partial W^*}{\partial l} = \frac{\sqrt{2}(a-\alpha)^{3/2}}{3\sqrt{l}} - \frac{(a-\alpha)}{2} + \frac{l}{12} + \frac{a^2}{2}.$$

This function is mimimized at $l = 2(a-\alpha)$, and $\frac{\partial W^*}{\partial l}(2(a-\alpha)) = \frac{a^2}{2} > 0$.

In zone B,

$$\frac{\partial W^*}{\partial l} = \frac{\alpha^2}{2a} - \frac{\alpha}{2} + \frac{l}{12} + \frac{a^3}{24l^2}.$$

This function is mimimized at $l = a$, and $\frac{\partial W^*}{\partial l}(a) = \frac{a}{2}(\frac{\alpha}{a} - \frac{1}{2})^2 \geq 0$.

Finally, in zone C,

$$\frac{\partial W^*}{\partial l} = -\frac{\alpha}{2} + \frac{l}{12} + \frac{\sqrt{2}\alpha^{3/2}}{3\sqrt{l}}.$$

This function is mimimized at $l = 2\alpha$, and $\frac{\partial W^*}{\partial l}(2\alpha) = 0$. This concludes the proof of proposition 4.2.