Incentive properties of coincident peak pricing

Ross Baldick

Journal of Regulatory Economics

ISSN 0922-680X

J Regul Econ
DOI 10.1007/s11149-018-9367-9

Springer

Journal of Regulatory Economics
https://doi.org/10.1007/s11149-018-9367-9

ORIGINAL ARTICLE

Incentive properties of coincident peak pricing

Ross Baldick[1]

Abstract

Coincident peak pricing is used in several electricity markets to recover the embedded cost of infrastructure, such as transmission. In this approach, measured consumption at the time of the peak is used to set charges for that pricing period or a subsequent period. If transmission costs are truly sunk, then such a recovery is unlikely to be efficient. However, in the context of growing peak demand, new additions must be built. We consider the incentive properties of coincident peak pricing when related investments are not considered to be sunk, finding that it can reproduce the incentive properties of an ideal time-varying price. We also consider several variations on this assumption.

Keywords Coincident peak pricing · Transmission systems · Electric power system expansion

JEL Classification D45 · D47 · L94 · Q41

List of symbols

Y	Number of hours in a year
$t \in [0, Y]$	Time over year
q_t	Power at time t
p_t	Energy price at time t
C_G, C_{Gi}	Generator operating cost functions per unit time
a_G	Linear coefficient of generator operating cost function
D, D_j	Demand functions
W, W_i	Inverse demand functions
N	Load-duration characteristic
T	Transmission capacity
C_T	Annualized transmission cost function
a_T	Linear coefficient of annualized transmission cost function

✉ Ross Baldick
baldick@ece.utexas.edu

[1] Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712, USA

Published online: 22 September 2018

P_G, P_{Gt}	Energy price functions
P_T	Transmission price function
p_T	Volumetric adder per unit energy for transmission

1 Introduction and literature review

Retail electric customers in restructured electricity markets pay for their consumption under tariffs that are designed to recover the energy, transmission and distribution, and per customer components of costs of delivered electricity. Typically, the energy is purchased by retail customers under a volumetric, that is, per kWh, tariff component that directly or indirectly reflects wholesale energy market prices. On the other hand, the charges for transmission and distribution and per customer costs are recovered in a variety of ways. At one extreme, these charges might be included in a volumetric charge based on the ratio of the total transmission, distribution and per customer costs divided by total energy sales. We will refer to this type of approach as a "volumetric adder per unit energy." At the other extreme, the charges for transmission and distribution could be unbundled into specific tariff components. For recent discussions of such tariffs, see Abdelmotteleb et al. (2018), Borenstein (2012), Faruqui and Aydin (2017), Haro et al. (2017), Passey et al. (2017).

In this context, coincident peak pricing is used in several electricity markets to recover annualized capital costs of electricity assets from end-use consumers. The charge to the end-use consumers is in proportion to their measured consumption at the times of overall system peaks or at the times of the peaks of a particular subsystem. The assets could, in principle, be generation, transmission, or distribution related; however, coincident peak pricing is most typically used for transmission and distribution assets. In the context of transmission, measured consumption at the time of the system peak in one year may be used to set the required payment for transmission that year or for the following year. The total payment is designed to recover the total annual "revenue requirements" of the transmission system to finance the existing capacity and potentially to finance the building of necessary new capacity to meet the peak demand forecast for the following year or years.

For example, coincident peak pricing is used for recovering some transmission system costs in the Electric Reliability Council of Texas (ERCOT) for, among others, large commercial and industrial consumers (Zarnikau 2017b). In ERCOT, the transmission charges for each large consumer over a particular calendar year are assessed on the average of its 15-min electrical demands occurring at the time of the four monthly 15-min peak demands of the entire ERCOT system for the months of June, July, August, and September during the prior year (Electric Reliability Council of Texas 2017, Section 9.17.1). The resulting charges are referred to as "four coincident peak" (4CP) charges, and the intervals when these peaks occur are called "4CP events."

In other regions of the United States, several other tariffs are also based on coincident peak consumption, differing in detail but broadly similar to the ERCOT 4CP charges. (See openei.org/wiki/Utility_Rate_Database for a database of tariffs in the United States.) For example, other regions in the United States use a different number of intervals to assess the coincident peak.

Coincident peak pricing is also used in the United Kingdom, where the transmission network use of system charges for each large consumer are based on the average of its 30-min electrical demands occurring at the time of the three highest 30-min peak demands of the entire system that are separated by at least 10 days (National Grid 2015). The resulting charges are referred to as "Triad" charges. To summarize, coincident peak pricing is utilized in a variety of electricity markets to recover transmission (and, in some cases, distribution) system costs.

In some jurisdictions, customer non-coincident peak demand may be used as a proxy to evaluating the contribution to coincident peak. For a discussion of the relationship between the number of intervals used in evaluating the customer non-coincident peak and the contribution to the coincident peak, see Passey et al. (2017). In this paper, we will assume that the contribution to the coincident peak of each end-user can be directly measured without error.

Several researchers have criticized the use of coincident peak pricing and similar charging methods to recover sunk costs. For example, see Borenstein (2012) for a general criticism of such approaches and (Hogan and Pope 2017) for specific criticism in the context of ERCOT. For a numerical simulation of this effect in conjunction with options for self-generation, see Schittekatte et al. (2018).

While not disputing the observations in Borenstein (2012), Hogan and Pope (2017), Schittekatte et al. (2018) that capacity-based charges are not appropriate for recovering investments that do not depend on the level of consumption, it should be understood that annual load growth together with necessary replacement and maintenance costs imply that at least some transmission and distribution investment is not sunk and that the necessary annual investments therefore depend on the (forecast) demand levels. That is, continuing investments must be made each year to increase capacity or even to maintain capacity at a fixed level. In these circumstances, some of the investment going forward in transmission is potentially "avoidable" in the sense that the amount of expenditure depends on the forecast peak demand level.

For example, in regions with population growth and growth in peak demand, such as in ERCOT, there are ongoing investments in transmission and distribution that could be avoided in the absence of that peak demand growth. A recent case in point is the "Houston Import Project" built for anticipated load growth in the Houston area and to compensate for retirements of generation within Houston (Electric Reliability Council of Texas 2014). The Houston Import Project involves approximately half a billion dollars in transmission investment.

The collective value of such incremental investments can add significantly to the annual "revenue requirements" for transmission. This is particularly the case in United States, European, and Australian networks where much of the existing system was built decades ago and the costs of these prior investments have been mostly recovered in the past, so that costs associated with existing transmission are mostly due to maintenance costs, whereas costs associated with new transmission include land and construction costs. Moreover, the real cost of building transmission and distribution has increased significantly compared to past epochs of construction due to the cost of land and environmental restrictions.

An important motivation for this paper is therefore that in several markets the transmission and distribution charges are becoming a significant fraction of the total

retail bill, prompting concerns that resulting high retail prices may result in inefficient reduction in consumption at the time of the peak or inefficiently high levels of self-generation at the time of the peak. See Hogan and Pope (2017, p. 80) for a discussion of this in the context of ERCOT and Agency for the Cooperation of Energy Regulators (2017, p. 13) for a discussion in the European context. This is particularly of concern in systems where there is significant investment or potential investment in renewables (Schittekatte et al. 2018). A large "wedge" between wholesale and retail prices due to charges for sunk transmission investments embedded in volumetric retail tariffs implies that the avoided wholesale production cost at peak times due to self-generation may be lower than the cost of self-generation, even if this self-generation cost is itself lower than the effective retail price. This situation will tend to result in over-investment in self-generation resources.

Such over-investment in self-generation could also potentially occur with coincident peak pricing if used to recover sunk costs. For example in ERCOT, the anticipation of a 4CP event, which typically occurs due to high air-conditioning load during high temperature days, and the implications for the transmission charge in the following year, results in significant reduction in net consumption (Zarnikau 2017b). As discussed in Zarnikau (2017a, b), large market participants in ERCOT are actively and successfully forecasting the times of 4CP events and are able to modulate their demand during these periods to reduce 4CP charges the following year. Effectively, the peak of net consumption is "clipped" (Zarnikau 2017a). At least some of this apparent reduction is likely due to self-generation. To the extent that the 4CP prices send incorrect incentives for net consumption by reflecting sunk costs, they are also distorting the demand side of the market and thereby resulting in poor overall investment decisions, including in self-generation.

Although there is a clear concern that 4CP and similar charges will result in inefficient decisions when they are used to recover sunk costs, this paper seeks to understand under what circumstances 4CP and similar charges would actually lead to efficient consumption decisions. A key methodological issue is that the time-varying demand and non-storability of electricity implies that efficiency of capital decisions must be assessed on the basis of the benefits accruing over time, not just at the time of the peak demand.

Building on the foregoing, in this paper we will primarily consider the case that is polar opposite to sunk costs and model the transmission capacity as being effectively "rented" on an annual basis. While this assumption, and the opposite assumption of fully sunk costs, is not perfectly realistic, we explore the situation for rentable transmission capacity as applied to coincident peak pricing. In a practical context, the assumed rental cost in this paper could be construed as being based on an estimate of the annualized average incremental cost of building additional transmission capacity to meet forecast demand that is growing over time.

In Sect. 5, we return to the consideration of sunk costs in the light of the formulation in this paper, observing that a hybrid approach could be used to recover some of the costs using coincident peak pricing based on the incremental costs and with recovery of the rest of the costs using a non-distorting tariff component. A recent example of this approach is described in Abdelmotteleb et al. (2018).

As well as analyzing coincident peak pricing, we will also compare it to other charging methodologies and variations. For example, we will consider the case where some of the transmission costs are recovered on the basis of a volumetric adder per unit energy. Such a volumetric adder tends to increase the on-peak consumption compared to the welfare optimal level, all else equal, which acts in the opposite direction to the effect of 4CP charges.

As mentioned above, an important aspect of this work is the consideration of an extended time horizon considering temporal variation of demand to assess welfare. This type of analysis is necessary because considering just a snapshot of time, for example at peak demand conditions, cannot reveal the welfare optimal level of transmission capacity. While models that consider welfare over an extended horizon have antecedents such as Stoft (2003), Klemperer and Meyer (1989), Green and Newbery (1992), Green (1996, 1999), the application to coincident peak pricing is novel.

Throughout, the setting will abstract from reality along several dimensions. First, we will consider a single capacitated transmission line, whereas practical transmission systems are meshed and there are multiple limiting elements in the system whose constraints are binding at different times of the day or during different seasons. Second, in meshed systems, locational marginal prices translate transmission limitations and generator marginal costs into time-varying prices to demand that increase when transmission limits are reached, reflecting increasing marginal delivered costs. However, we abstract from this to consider peak levels of demand where, because of transmission limitations, there is no available additional generation capacity that can be delivered, and no storage capacity available at the demand to mitigate peak consumption. Third, both transmission and distribution capacity is required to deliver electricity to consumers, but we conflate all such constraints into a single capacitated transmission line. Fourth, although we will consider variation of demand over time, the analysis will be deterministic. In practice, transmission and distribution capacity is typically built to meet extreme conditions that may occur relatively rarely, requiring some adjustment to our analysis to be utilized in practice. Fifth, we will evaluate equilibria in prices, consumption, and capacities assuming that these are simultaneously determined, although markets such as ERCOT set transmission prices for a given year based on consumption in the prior year. Sixth, we ignore strategic behavior, assuming that the demand-side is a price-taker and that the entity designing and building transmission has welfare optimization as its goal.

The contributions of this paper are as follows. We formulate a planning problem in welfare-optimal transmission capacity that considers the variability of underlying demand over time enabling an annualized assessment of welfare contributions in order to balance capital and operating costs with willingness-to-pay. As an ideal for comparison, we first consider energy-only prices that would induce the welfare optimizing consumption over time, assuming price-taking behavior. We highlight the informational difficulty in setting such prices. We then consider how coincident peak pricing for transmission could provide the coordinating signal in a more decentralized fashion. To the best of the author's knowledge, this is the first demonstration of economic efficiency of coincident peak pricing of capacity with non-sunk costs. A numerical example highlights the issues and comparisons are made to other approaches to charging for capacity under the same underlying formulation.

The rest of this paper is organized as follows. Section 2 formulates the basic problem and Sect. 3 considers ideal energy-only prices that provide incentives for welfare optimal behavior and then discusses energy and transmission prices as a model of coincident peak pricing that reproduces the incentives for welfare optimal behavior in a more decentralized fashion. An extended example in Sect. 4 is used to illustrate the results. Section 5 considers several variations, including a comparison to the case where network costs are recovered by an adder to energy prices, as is customary for most small consumers such as residential retail customers in the United States and Europe.[1] Section 6 concludes.

2 Formulation

There are several possible approaches to analyzing the effect of tariffs on consumption and efficiency. For example, an analytical approach is taken in Li (2007), simulation based on aggregate demand and supply models is used in Abdelmotteleb et al. (2018), and agent-based simulation is used in Manuel de Villena et al. (2017). We will take an analytical approach based on aggregate demand and supply models, using a model that is similar to that used in the supply function equilibrium literature (Klemperer and Meyer 1989; Green and Newbery 1992; Green 1996, 1999), but which is also similar to other related literature on the economics of transmission expansion (Léautier 2000). We do not, however, consider strategic behavior on the part of market participants.

In the rest of this section, we discuss the representation of time, the demand side, the supply (that is, generation) side, the transmission line, the combined model of the industry, welfare, and welfare optimization. The formulation is deliberately abstracted to enable a focus on transmission pricing.

2.1 Time

Because of the non-storability of electricity and the significant variation in demand over time due to human rhythms and seasons, we must explicitly consider the representation of temporal issues into the model. We primarily consider variation over a year, with time measured in hours, and with time represented by $t \in [0, Y]$, where $Y = 8760$ h. For convenience in the examples in Sect. 4, we will assume that time has been re-ordered so that highest demand corresponds to $t = 0$, and lowest to $t = Y$ in the format of a "load-duration curve." However, for the initial theoretical analysis no such assumption is required and we will tacitly assume that t simply represents time measured in hours over a year. We ignore the fact that electricity markets are cleared on an hourly, 15-minutely, or 5-minutely basis and, instead, view t as a continuous variable.

2.2 Demand side

We initially consider an undifferentiated demand side with aggregate demand function, $D : \mathbb{R}_+ \times [0, Y] \to \mathbb{R}$, such that for each price $p \in \mathbb{R}_+$ and each $t \in [0, Y]$, $D(p, t)$

[1] In some regions, the prevalence of rooftop solar in residential areas is prompting a move to peak charges for residential customers. See, for example Ergon Energy (2018).

evaluates the net electrical demand power (measured in GW) at time t given price p (in \$/MWh). We assume a corresponding inverse demand function or willingness-to-pay, $W : \mathbb{R} \times [0, Y] \to \mathbb{R}_+$, that evaluates the marginal benefit of consumption per hour and satisfies:

$$\forall p \in \mathbb{R}_+, \forall t \in [0, Y], W(D(p, t), t) = p.$$

For convenience, we assume that W is continuous and that it is non-increasing in the quantity argument.

2.3 Supply side

Abstracting from the reality of electric generation, we model aggregate supply as a single generation resource. We assume that the aggregate capacity is larger than the largest possible demand level $\max_{t\in[0,Y]}\{D(0, t)\}$, and that the aggregate operating cost is specified by $C_G : \mathbb{R}_+ \to \mathbb{R}_+$ with interpretation that the operating cost per hour of generating at power level q is $C_G(q)$. We assume that C_G is differentiable, non-decreasing, and convex. We ignore unit commitment issues.

2.4 Transmission

As discussed in Sect. 1, we assume that transmission capacity T is effectively available for annual rental. In particular, we assume that the annualized rental cost of transmission is described by a function $C_T : \mathbb{R}_+ \to \mathbb{R}_+$ so that the annual rental cost for transmission capacity T in a particular year is $C_T(T)$.[2] For analytic convenience, we assume that C_T is differentiable, non-decreasing, and convex, although we acknowledge that this may not hold in practice due to lumpiness of transmission expansion and economies of scale in transmission capital costs (Baldick and Kahn 1993; Léautier 2000, Section 3; Dixit and Baldick 2003). We return to this briefly in Sect. 3.3.

We assume that to deliver power generation level q to the demand side, there must be transmission capacity of at least q. That is, with capacity equal to T, the level of generation q (and therefore the level of consumption net of any self-generation) would be limited by the constraint $q \leq T$.

2.5 Industry model

Figure 1 shows the conceptual configuration where a generator located at the left injects power into the capacitated transmission line at the generator electrical bus symbolized by the thick vertical line. The transmission line allows delivery to the bus at the right, also symbolized by another thick vertical line. Electricity is locally reticulated on a distribution system that is not explicitly modeled, and the total demand D, modeled as

[2] We do not address the appropriate regulatory regime to induce a transmission company to build efficiently. See, for example, Bushnell and Stoft (1997), Léautier (2000), Kristiansen and Rosellón (2006), Léautier and Thelen (2009), Rosellón and Weigt (2011) for discussion of such mechanisms. We also do not discuss the technical issues involved with efficient planning. See, for example, Majidi-Qadikolai and Baldick (2016a, b, 2018) for a discussion of optimization-based approaches to transmission expansion planning.

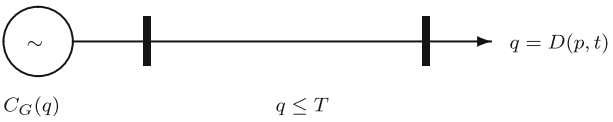

Fig. 1 Industry model

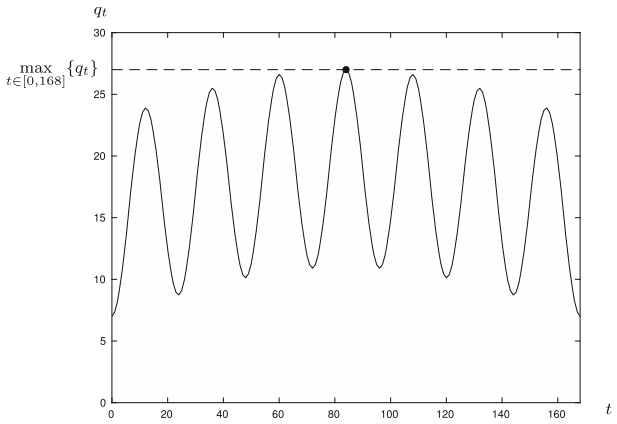

Fig. 2 Figure showing maximum demand over a week

the sum of the distribution system loads, is equal to generation q, ignoring electrical losses. The transmission line capacity implicitly limits the production by the generator and the delivery to the demand to be no more than the transmission capacity T.[3]

Conversely, the transmission capacity must be at least as large as the largest value of generation q_t over the time interval of interest. Figure 2 shows an example where the time interval is a week, of duration 168 h. The solid curve shows the realized supply, or equivalently, demand, q_t, over this period. The value of q_t varies over the week and the transmission capacity must be at least as large as $\max_{t\in[0,168]}\{q_t\}$, which is indicated by the height of the dashed line, with the maximum occurring at the time indicated by the bullet.

As mentioned in Sect. 1, this model does not capture the complications of an actual power system for several reasons. For example, while self-generation at a demand location can be represented in terms of the net demand function, the main results will not immediately generalize to the presence of economically separated generation offering into the market at the receiving end of the transmission line. The setting should be seen as characterizing the case where all such local generation is generating at its maximum capacity and there are no other sources that can deliver power to the demand. Moreover, the notion of loading on the line is *much* more complicated in a realistic meshed transmission system, where there are many thousands of transmission

[3] Given the simple configuration of the industry, the capacity constraint could also represent a constraint on generation, which in turn could be used to endogenously represent generation expansion instead of or as well as transmission expansion. However, we have assumed that the aggregate generation supply is exogenous with large capacity in order to focus on transmission and transmission costs and expansion. Extension to a model with multiple potentially binding capacity constraints is mentioned in the Conclusion.

lines and the specific location of both demand and supply will affect the loading on any particular transmission line (Bergen and Vittal 2000; Wood and Wollenberg 1996), and where the cost to increase transmission capacity of individual lines may exhibit lumpiness and economies of scale.

2.6 Welfare

The welfare is, by definition, the benefits of consumption minus the costs. The costs are assumed to be due both to the operating costs of the generator and to the rental of the transmission. Assume that generation and consumption at time t is q_t, while the transmission capacity is T. Then the annual welfare is:

$$A = \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t) \right] dt - C_T(T).$$

Welfare optimization is, by definition:

$$A^\star = \max_{\substack{T \geq 0, \\ q_t \geq 0, \forall t \in [0,Y]}} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t) \right] dt - C_T(T) \middle| q_t \leq T, \forall t \in [0,Y] \right\}. \tag{1}$$

A transformation of problem (1) is more convenient for characterizing welfare optimality. In particular, we observe that problem (1) is equivalent to optimizing the objective over choices of T in an outer problem, with an inner problem representing the optimum over choices of $q_t \leq T, \forall t \in [0, Y]$. That is, problem (1) is equivalent to:

$$\max_{T \geq 0} \left\{ \max_{q_t \geq 0, \forall t \in [0,Y]} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t) \right] dt \middle| q_t \leq T, \forall t \in [0,Y] \right\} - C_T(T) \right\}. \tag{2}$$

Moreover, we note that because of the continuity of the integrand in the objective of (2), and since there are no coupling constraints between q_t and q_s for $t \neq s$, we can swap the order of the maximization over q_t with the integration over t to obtain another equivalent problem:

$$\max_{T \geq 0} \left\{ \int_{t=0}^{t=Y} \max_{q_t \geq 0} \left\{ \left[\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t) \right] \middle| q_t \leq T \right\} dt - C_T(T) \right\}. \tag{3}$$

We have now expressed welfare maximization as the integral of a collection of inner problems each of which is a pointwise maximum over $0 \leq q_t \leq T$. The form of problem (3) facilitates characterization of conditions for welfare optimality.

Consider the objective of the inner problem for a given t in problem (3):

$$\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t). \tag{4}$$

By assumption, W is continuous and non-decreasing and C_G is differentiable and convex. Therefore, for each t, the *unconstrained* maximizer of this inner objective is obtained by differentiating the inner objective with respect to q_t and setting equal to zero, resulting in the condition:

$$W(q_t,t) - \frac{\partial C_G}{\partial q}(q_t) = 0, \tag{5}$$

Consider particular values of time t and of transmission capacity T. If the solution of (5) is no larger than T then the corresponding *constrained* maximizer of the inner problem for t and the given T is the same as the solution of (5). On the other hand, if the solution of (5) is larger than T then the corresponding constrained maximizer of the inner problem for t and the given T is T itself. We can summarize this observation by defining $\hat{q}_t : \mathbb{R}_+ \to \mathbb{R}_+, \forall t \in [0, Y]$ to be:

$$\forall T \in \mathbb{R}_+, \hat{q}_t(T) = \max\{0, \min\{T, \text{ solution of (5)}\}\}. \tag{6}$$

This expression characterizes the optimal production decisions, given a transmission level T.

Substituting the optimal production decisions into the welfare maximization formulation (3), we obtain:

$$\max_{T \geq 0} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=\hat{q}_t(T)} W(q,t)dq - C_G(\hat{q}_t(T)) \right] dt - C_T(T) \right\}.$$

Making the reasonable assumption that the constraint $T \geq 0$ is not binding, then differentiating this expression with respect to T yields the conditions for the optimal transmission level $T^\star$:

$$\int_{t=0}^{t=Y} \left[W(\hat{q}_t(T^\star), t) - \frac{\partial C_G}{\partial q}(\hat{q}_t(T^\star)) \right] \frac{\partial \hat{q}_t}{\partial T}(T^\star)dt - \frac{\partial C_T}{\partial T}(T^\star) = 0, \tag{7}$$

where we note that, strictly speaking, there will be values of t for which $\hat{q}_t$ is not differentiable, but so long as the set of such times is of measure zero this will not affect the integral. The optimal production decisions are then given by $q_t^\star = \hat{q}_t(T^\star), \forall t \in [0, Y]$.

Summarizing, note that if $(T^\star; q_t^\star, \forall t \in [0, Y])$ is a solution of (1), then for t such that $q_t^\star < T^\star$, the value $q_t^\star$ is the corresponding unconstrained maximizer of $\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t)$ over $q_t \geq 0$. Furthermore, again because of the assumption that W is continuous and non-decreasing and the assumption that C_G is differentiable and convex, and further assuming that the non-negativity constraint on q_t is not binding,

then $q_t^\star$ satisfies $W(q_t^\star, t) = \dfrac{\partial C_G}{\partial q}(q_t^\star)$ if $0 < q_t^\star < T^\star$. Note that since W is non-increasing, we have that $W(q_t^\star, t) \geq W(T^\star, t)$ in this case, so we could also write the condition on $q_t^\star$ as $W(q_t^\star, t) = \max\left\{\dfrac{\partial C_G}{\partial q}(q_t^\star), W(T^\star, t)\right\}$ if $0 < q_t^\star < T^\star$.

On the other hand, if the unconstrained maximizer of $\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t)$ equals or exceeds $T^\star$ then we have that $q_t^\star = T^\star$, so that $W(q_t^\star, t) = W(T^\star, t)$. Moreover, we have that $\dfrac{\partial C_G}{\partial q}(q_t^\star) < W(q_t^\star, t)$ since, otherwise, $q_t^\star$ would be the unconstrained maximizer of $\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t)$. That is, in this case, we again have that $q_t^\star$ satisfies $W(q_t^\star, t) = \max\left\{\dfrac{\partial C_G}{\partial q}(q_t^\star), W(T^\star, t)\right\}$.

To summarize, given the welfare optimal value of transmission, $T^\star$, then the welfare optimal consumption, $q_t^\star$, is characterized by:

$$W(q_t^\star, t) - P_{Gt}(q_t^\star) = 0, \tag{8}$$

where $P_{Gt} : \mathbb{R}_+ \to \mathbb{R}_+$ is defined for each $t \in [0, Y]$ by:

$$\forall t \in [0, Y], \forall q_t \in \mathbb{R}_+, P_{Gt}(q_t) = \max\left\{\frac{\partial C_G}{\partial q}(q_t), W(T^\star, t)\right\}. \tag{9}$$

and with welfare optimal consumption:

$$\forall t \in [0, Y], q_t^\star = \min\{T^\star, \text{ solution of (8)}\}.$$

As we will see in the next section, at each time t, $P_{Gt}(q_t^\star)$ is the welfare inducing price.

3 Welfare inducing prices

In this section, we first consider energy prices that will induce behavior consistent with welfare maximization, given a known solution $T^\star$ of the welfare-optimizing transmission capacity. Unsurprisingly, the prices will involve the function P_{Gt} defined in (9). It should be emphasized that to the best of the author's knowledge, no jurisdiction uses such prices to recover transmsission costs. This analysis is used simply to exhibit how energy prices could induce welfare maximizing consumption and to provide a point of comparison for coincident peak pricing, which we will argue is more practically implementable.

3.1 Energy-only prices

We claim that energy prices of the following form will induce welfare optimality:

$$\forall t \in [0, Y], p_t = \max\left\{\frac{\partial C_G}{\partial q}(q_t^\star), W(T^\star, t)\right\}.$$

Moreover, if both the demand side and the supply side cannot affect energy prices, that is, if they are "energy price takers," then we claim that the collection of energy price functions $P_{Gt} : \mathbb{R}_+ \to \mathbb{R}_+, \forall t \in [0, Y]$, defined in (9), induce an equilibrium that maximizes welfare. Although the price function P_{Gt} is defined for all $q_t \in \mathbb{R}_+$, only values of q_t that are less than or equal to $T^\star$ will be realized at equilibrium. The energy price function P_{Gt} requires explicit knowledge of the marginal cost of production of the generator and the aggregate willingness-to-pay of the demand. We consider the supply-side and the demand-side in turn.

Since the price functions P_{Gt} depend explicitly on the marginal cost of production of the generator, implementation of (9) as a mechanism for setting prices implicitly assumes that marginal generating cost information is available to evaluate the price function. In principle, an electricity system "independent system operator" (ISO) could implement the price function as the outcome of so-called offer-based economic dispatch. In general, such mechanisms involve:

- the supply-side specifying offers to the ISO,
- the ISO making short-term forecasts of the demand (for a "real-time market"),
- the ISO finding the dispatch levels for the generator to meet the short-term forecast, and
- the ISO then setting the energy prices.

Modeling the supply side as consisting of a larger number of small (in principle, infinitesimal) generators, then we have the standard result that each generator will maximize its profits by setting its offer equal to its marginal costs. This then validates the assumption that the marginal cost can be evaluated by the ISO to form the price functions: the marginal cost function is equal to the aggregation of the generation offer functions under competitive conditions and the aggregation of offers is a proxy to the marginal cost under more practical conditions of imperfect competition.

We now turn to the demand-side. First, to see that the price function results in welfare maximizing consumption for a price taking demand-side, note that the demand-side would choose consumption $q_t^\star$ that maximizes demand-side benefits minus energy charges. That is, $q_t^\star$ maximizes, over choices q_t, the following objective:

$$\int_{q=0}^{q=q_t} W(q, t)dq - P_{Gt}(q_t^\star)q_t, \tag{10}$$

where we have applied the price taking assumption that the energy price at each time is not affected by consumption decisions at that time, by substituting the welfare-optimizing value $q_t^\star$ into the price function P_{Gt} for each time $t \in [0, Y]$.

By assumption on W, the objective (10) is concave and differentiable. Taking the derivative of (10) with respect to q_t, we obtain:

$$W(q_t, t) - P_{Gt}(q_t^\star).$$

Setting this derivative equal to zero to maximize the demand-side benefits minus energy charges, we obtain precisely the same condition as (8) for welfare optimality. We observe that if W has a region of zero derivative about $q_t = T^\star$ then the maximizer of (10) may not be unique and there will be a need for the ISO to specify the

welfare optimal demand to be $q_t^\star = T^\star$ so that the transmission flow is within the transmission limit $T^\star$. That is, by analogy with a similar situation for dispatch of generation, we might say that the prices "support" welfare optimal consumption in that there is a maximizer of demand-side benefits minus energy charges that is consistent with welfare maximization. However, in the case that there is a non-unique maximizer of demand-side benefits minus energy charges, then the prices do not "strictly support" the welfare optimal solution in that the ISO must select the value of demand $q_t^\star = T^\star$ among those maximizers. To summarize, given the optimal level of transmission capacity $T^\star$, energy prices based on the energy price functions P_{Gt} will support welfare optimal generation and consumption decisions. Note that the result is similar to generic results on peak load pricing, except that we have explicitly considered variation in willingness-to-pay over the horizon.

Although these energy-only prices would induce welfare optimal behavior, they suffer from several drawbacks. For example, in practical electricity markets, even if marginal costs of generation are revealed by offers, the function W representing the demand side may only be approximately known, and may not be explicitly exhibited to the market. Consequently, there may be insufficient information to perform the welfare maximization in detail for each time t to determine the price functions. That is, the ISO would require specific, time varying information about the willingness-to-pay of the demand side. Moreover, the resulting price functions vary with time, even though the generator cost function is assumed here to be independent of time.[4] Indeed, as mentioned above, there do not appear to be any jurisdictions that use energy prices based on willingness-to-pay to induce welfare-optimizing behavior in this manner, although locational marginal prices use time and location varying prices to reflect variations in marginal generation costs as they relate to transmission limitations. In the next section, we will consider an alternative approach to inducing welfare-optimal behavior.

3.2 Energy and transmission prices

In this section, we explore an alternative approach that partially mitigates the problems of energy-only prices by providing a more decentralizable solution through coincident peak pricing of transmission. The implications for information needed by the demand-side is discussed in Sect. 3.3.

Consider an annual rental price for using transmission levied on the maximum power level over the year and based on a transmission price function $P_T : \mathbb{R}_+ \to \mathbb{R}_+$ of the form $P_T = \dfrac{\partial C_T}{\partial T}$, together with an energy price function $P_G : \mathbb{R}_+ \to \mathbb{R}_+$ of the form $P_G = \dfrac{\partial C_G}{\partial q}$. That is, there is a charge for energy based on energy consumption integrated over time, together with a charge for transmission based on peak power consumption. The latter implements the coincident peak pricing.

[4] A time-varying energy price could be used to reflect time varying energy prices, but the point here is that energy and transmission prices must vary over time to represent the effect of transmission capacity even when marginal costs for energy production are constant.

Note that, as distinct from P_{Gt} defined in (9), the energy price function P_G is independent of time. However, in addition to paying for energy based on the energy price P_G, there is an additional charge for transmission capacity. The charge for transmission capacity is based on the transmission price function P_T evaluated at the welfare optimal level of transmission and is applied to the maximum consumption over the year. That is, over a year, the total payment by the demand is:

$$\int_{t=0}^{t=Y} P_G(q_t^\star) q_t dt + P_T(T^\star) \max_{t\in[0,Y]} \{q_t\},$$

where we have maintained the energy price taker assumption on the demand-side by substituting $q_t^\star$ into the energy price function P_G and additionally assumed that the demand-side is a "transmission price taker" by also substituting $T^\star$ into the transmission price function P_T. That is, the charge for transmission capacity uses coincident peak pricing with price defined by the marginal cost of transmission.

To see that these prices for energy and transmission result in welfare maximizing consumption for an energy and transmission price taking demand side, note that the demand side would choose consumption $q_t^\star$ that maximizes demand-side benefits minus energy and transmission charges. That is, $q_t^\star$ maximizes, over choices q_t, the following problem:

$$\begin{aligned}
&\max_{q_t\geq 0, \forall t\in[0,Y]} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=q_t} W(q,t)dq - P_G(q_t^\star) q_t \right] dt - P_T(T^\star) \max_{t\in[0,Y]} \{q_t\} \right\} \\
&= \max_{q_t\geq 0, \forall t\in[0,Y]} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=q_t} W(q,t)dq - \frac{\partial C_G}{\partial q}(q_t^\star) q_t \right] dt \right. \\
&\qquad \left. - \frac{\partial C_T}{\partial T}(T^\star) \max_{t\in[0,Y]} \{q_t\} \right\}, \\
&\qquad \text{by definition of } P_G \text{ and } P_T, \\
&= \max_{T'\geq 0} \left\{ \max_{q_t\geq 0, \forall t\in[0,Y]} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=q_t} W(q,t)dq - \frac{\partial C_G}{\partial q}(q_t^\star) q_t \right] \right. \right. \\
&\qquad \left. \left. dt - \frac{\partial C_T}{\partial T}(T^\star) T' \,\middle|\, q_t \leq T', \forall t \in [0,Y] \right\} \right\}, \\
&\qquad \text{introducing a new variable } T' \text{that represents the maximum demand,} \\
&\qquad \text{recognizing that the realized transmission capacity will be no larger than} \\
&\qquad \text{the demand at the optimal solution,} \\
&= \max_{T'\geq 0} \left\{ \int_{t=0}^{t=Y} \left[\int_{q=0}^{q=\hat{q}_t(T')} W(q,t)dq - \frac{\partial C_G}{\partial q}(q_t^\star) \hat{q}_t(T') \right] dt - \frac{\partial C_T}{\partial T}(T^\star) T' \right\},
\end{aligned}$$

where the function $\hat{q}_t$ was defined in (6). Differentiating the objective of the last problem with respect to T' results in the same conditions on T' as for optimal transmission in (7). Therefore, again assuming that there is some non-zero consumption and that the optimality conditions have a unique solution, then the optimal value of the maximum

demand T' will match the welfare optimal transmission capacity $T^\star$ and the demand side will consume at the welfare optimal levels $q_t^\star = \hat{q}_t(T^\star), \forall t \in [0, Y]$.

To summarize, we have shown that if demand is an energy price taker and a transmission price taker then the resulting consumption choices are consistent with welfare maximization. That is, we have shown conditions under which coincident peak pricing induces welfare optimal consumption decisions.

3.3 Discussion

As shown in the previous section, energy and transmission prices can induce welfare maximizing consumption decisions without the ISO needing to know as much detailed information about the willingness-to-pay as required for the energy-only prices. In practice, even with energy and transmission prices, the marginal cost of transmission, and the times and level of maximum demand may not be known with certainty *ex ante*. Consequently, efficiency in practice should be understood in an expected sense, relying on rational beliefs about transmission cost functions and the timing and level of coincident peak demand. Transmission cost prediction can be obtained from historical data, although this is subject to variation based on the specifics of particular lines (Dixit and Baldick 2003). However, even if average cost per unit capacity is well-characterized, designed transmission capacities are subject to lumpiness. Moreover, there are inherent uncertainties in levels of peak demand due to, for example, ambient conditions at the time of peak demand. Nevertheless, as mentioned in Sect. 1, the timing of coincident peak demand events seem to be well forecasted by market participants in ERCOT, particularly since they are primarily driven by residential air-conditioning load and therefore are highly correlated with weather phenomena, which are themselves predicted fairly accurately (Zarnikau 2017a, b). Consequently, coincident peak pricing using, for example, average incremental costs of transmission capacity and forecasted levels of peak demand, could provide a workable approximation to inducing close to welfare optimal behavior.

4 Example

In this section, we develop an example, based on Green (1996), Day and Bunn (2001) and further explored in Baldick and Hogan (2002). Although we are not focused on competition in the generation sector in this paper, similar models are typically used in the supply function equilibrium literature (Klemperer and Meyer 1989; Green and Newbery 1992; Green 1996, 1999). Following the supply function equilibrium literature, and as mentioned in the introduction, rather than considering demand chronologically as illustrated in Fig. 2, we will consider that time has been re-ordered so that D represents a load-duration curve. The example in Green (1996), Day and Bunn (2001), Baldick and Hogan (2002) assumed quadratic generation costs, but to simplify the example and implicitly enforce the energy and transmission price taker assumptions, we will consider linear costs; that is, constant marginal costs for both energy and transmission.

4.1 Undifferentiated demand side

Following Green (1996), Day and Bunn (2001), Baldick and Hogan (2002), we assume that $D : \mathbb{R}_+ \times [0, Y] \to \mathbb{R}_+$ has the specific form:

$$\forall p \in \mathbb{R}_+, \forall t \in [0, Y], D(p, t) = N(t) - \gamma p, \tag{11}$$

where:

- the function $N : [0, Y] \to \mathbb{R}_+$ is the non-increasing *load-duration* characteristic, and
- the parameter $\gamma \in \mathbb{R}_+$ is minus the slope of the demand curve.

We will use a value of $\gamma = 0.125$ GW/(\$/MWh). A slight deviation from Green (1996), Day and Bunn (2001), Baldick and Hogan (2002) is that we model time in hours over a year, so that $t \in [0, Y]$ in the argument of the load-duration characteristic N, instead of a normalized time range $[0, 1]$ as in Green (1996), Day and Bunn (2001), Baldick and Hogan (2002). This involves only cosmetic changes to the model.

Although the highest value of N occurs for $t = 0$, we will not *a priori* assume that the highest demand actually occurs at $t = 0$; this will depend on the outcome of prices, but will indeed turn out to be true. The specific form of N that we will consider is:

$$\begin{aligned} \forall t \in [0, Y], N(t) &= 7 + 20(1 - t/Y), \\ &= N(0) - \delta t, \end{aligned}$$

where $N(0) = 27$ and $\delta = (20/Y)$ GW/h, with quantities measured in GW. That is, N varies linearly from 27 GW at $t = 0$ to 7 GW at $t = Y$. This load-duration characteristic is illustrated in Fig. 3.

The corresponding willingness-to-pay W is the inverse of D. That is,

$$\forall q \in \mathbb{R}_+, \forall t \in [0, Y], W(q, t) = (N(t) - q)/\gamma = (N(0) - \delta t - q)/\gamma.$$

4.2 Undifferentiated supply side

For simplicity, we assume that the marginal cost for energy is constant, with:

$$\forall q \in \mathbb{R}_+, C_G(q) = a_G q,$$

with $a_G = 10$ \$/MWh.

4.3 Transmission cost

Again for simplicity, and to implicitly enforce that transmission price taker assumption, we will assume that the marginal cost for transmission is constant, with:

$$\forall T \in \mathbb{R}_+, C_T(T) = a_T T,$$

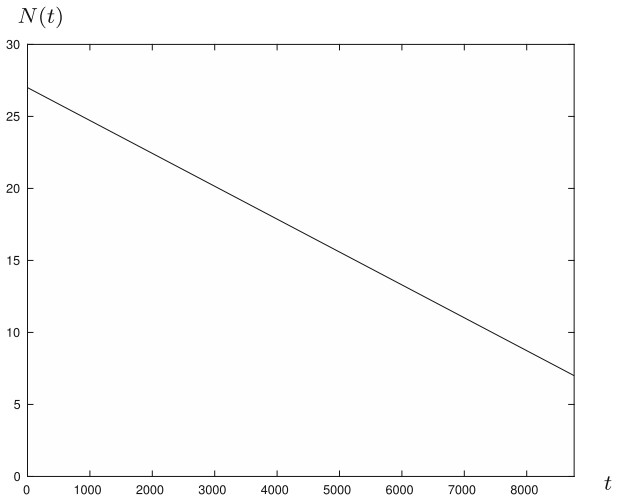

Fig. 3 Load-duration characteristic for example

with $a_T = 10^6$ \$/GW year $= 10^3$ \$/MW year. The transmission cost was chosen so that the annualized cost was a relatively small fraction of the total costs; however, the actual cost depends on the particulars of the transmission investment requirements to access generation resources.

4.4 Welfare maximization

The Appendix derives the conditions for welfare maximization in detail. It follows the analysis in Sect. 2.6, first considering the objective (4) of the inner problem in problem (3) for each t and finding the conditions for welfare maximization. The welfare maximizing level of transmission, $T^\star$, and the maximum welfare, $A^\star$, are given, respectively, by:

$$T^\star = N(0) - \gamma a_G - \sqrt{2\delta\gamma a_T}, \tag{12}$$

$$A^\star = [(N(0) - a_G\gamma)^3 - (N(0) - a_G\gamma - \delta Y)^3]/(6\gamma\delta) - (\sqrt{2\delta\gamma a_T})^3/(6\gamma\delta) - a_T T^\star, \tag{13}$$

assuming that:

$$\sqrt{2\gamma a_T/\delta} < Y. \tag{14}$$

The first term in the expression (13) for optimal welfare is independent of transmission costs and represents the optimal welfare if the transmission were costless. The last two terms represent the reduction in welfare due to the non-zero costs of transmission, with the last term being the direct cost of the transmission and the second-last term being the reduction in welfare due to the transmission limit.

4.5 Numerical evaluation

For the assumed values $N(0) = 27$, $\delta = (20/Y)$ GW/h, $\gamma = 0.125$ GW/(\$/MWh), $a_G = 10$ \$/MWh, $a_T = 10^6$\$/GW year $= 10^3$ \$/MW year, we have that $\sqrt{2\gamma a_T/\delta} \approx 330\,\text{h} < 8760\,\text{h} = Y$, satisfying (14), so that indeed we have that the optimal level of transmission is given by (12): $T^\star = N(0) - \gamma a_G - \sqrt{2\delta\gamma a_T} \approx 25$ GW. The optimal welfare is, according to (13):

$$\begin{aligned} A^\star &= [(N(0)-a_G\gamma)^3-(N(0)-a_G\gamma-\delta Y)^3]/(6\gamma\delta)-(\sqrt{2\delta\gamma a_T})^3/(6\gamma\delta) - a_T T^\star \\ &= 9.86\times 10^9 - 2.5\times 10^{-1} - 2.50\times 10^7, \\ &= 9.835\times 10^9\ \$/\text{year}. \end{aligned}$$

Note that for this example the direct cost of the transmission is much larger than the reduction in welfare due to the transmission limit.

4.6 Prices and quantities

In this section, we evaluate prices that induce welfare-optimality, and the corresponding consumption quantities.

4.6.1 Energy-only prices

From the analysis in Sect. 3.1, the energy-only prices are defined in (9):

$$\begin{aligned} \forall t \in [0,Y], \forall q_t \in \mathbb{R}_+, P_{Gt}(q_t) &= \max\left\{\frac{\partial C_G}{\partial q}(q_t), W(T^\star,t)\right\}, \\ &= \max\{a_G, a_G + (\sqrt{2\delta\gamma a_T} - \delta t)/\gamma\}. \end{aligned}$$

by assumption on C_G and (18). The resulting equilibrium consumptions $q_t^\star$ and prices $p_t^\star$ are:

$$\begin{aligned} q_t^\star &= \min\{T^\star, N(0) - \delta t - \gamma a_G\}, \\ p_t^\star &= \max\{a_G, a_G + (\sqrt{2\delta\gamma a_T} - \delta t)/\gamma\}. \end{aligned} \tag{15}$$

We note that the energy-only price is given by the marginal cost of generation whenever the transmission constraint is not binding, but when the transmission constraint is binding, the prices rise to constrain consumption to within the transmission limit.

4.6.2 Equilibrium quantities

The equilibrium quantities are illustrated in Fig. 4. Note that the demand is constant during the highest 330 h of demand: the peak of the load-duration characteristic has been clipped.

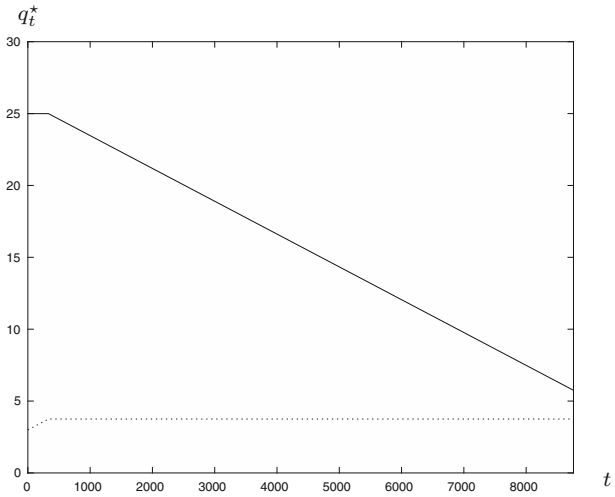

Fig. 4 Equilibrium consumption for example

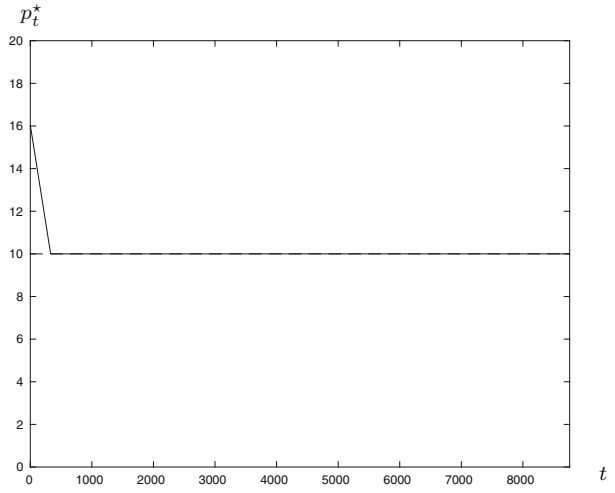

Fig. 5 Equilibrium prices for example

4.6.3 Equilibrium energy-only prices

The equilibrium energy-only prices are illustrated as the solid line in Fig. 5. They are high at the time of the peak of the load-duration characteristic in order to clip the demand to the level of the optimal transmission capacity. Note that these prices depend on explicit knowledge of the demand willingness-to-pay, W. We have argued that in typical electricity market arrangements this information may not be easily accessible to the ISO.

4.6.4 Energy and transmission prices

Because of the assumed linear form of the generation and transmission cost functions, the energy and transmission prices are, respectively, equal to $a_G = 10$ \$/MWh and $a_T = 10^3$ \$/MW year. The equilibrium consumption is as specified in (15) and the same as illustrated in Fig. 4. The equilibrium energy prices are constant over all time and do not display the peak that is observed in the energy-only price case illustrated in Fig. 5. As observed in Zarnikau (2017a), the coincident peak prices clip the peak of the demand, but without an explicit increase in the energy price.

4.6.5 Heterogenous demand-side

Although we have modeled a single, undifferentiated demand side, different consumers have different characteristics. For example, residential customers may be inelastic to wholesale prices, and have demand that varies significantly on a seasonal basis, while industrial customers may be elastic, but with underlying demand that is less dependent on seasons. This is intended to stylistically represent characteristics of the demand in ERCOT.

For example, suppose that the demand function in (11) is actually due to the sum of an elastic sector with time-invariant demand function $D_E : \mathbb{R}_+ \times [0, Y] \to \mathbb{R}_+$ and an inelastic sector with time-varying demand function $D_I : \mathbb{R}_+ \times [0, Y] \to \mathbb{R}_+$, defined as:

$$\forall p \in \mathbb{R}_+, \forall t \in [0, Y], D_E(p, t) = 5 - \gamma p,$$
$$\forall p \in \mathbb{R}_+, \forall t \in [0, Y], D_I(p, t) = 22 - \delta t,$$

where as before, $\gamma = 0.125$ GW/(\$/MWh) and $\delta = (20/Y)$ GW/h. That is, we are considering an extreme example where the elastic demand has a demand characteristic that is invariant over time and where the inelastic demand exhibits all of the overall variation in the load-duration characteristic.

The welfare optimal decisions are similar to the previous case, with overall demand the same. The welfare optimal level of demand of the elastic sector is illustrated by the dotted line in Fig. 4. The demand of the inelastic sector is the difference between the solid and the dotted line in Fig. 4. As in the previous cases of undifferentiated demand-side, optimal energy-only and optimal energy and transmission prices induce welfare-optimizing behavior.

5 Variations

In this section we consider several variations on the basic model, making qualitative observations. In some cases, we also analyze quantitatively based on the example. The first two cases continue to assume that transmission capacity is being built in response to exhibited needs and is not sunk. The third case will consider where capacity is not directly related to maximum consumption. The fourth considers the hybrid situation where there are both sunk costs and avoidable costs of transmission. We return to

analyzing the undifferentiated demand function (11) and do not distinguish different demand sectors in the following.

5.1 Monopsony market power

5.1.1 Assumptions

We continue to assume that transmission capacity must be built to accommodate maximum demand and that the supply side is an energy price taker. However, we consider the case that the demand side is a monopsonist, and that the transmission costs are strictly convex in the required transmission.

5.1.2 Qualitative observations

With a concentrated demand-side and with transmission costs strictly convex then, by withholding some demand at the peak, a monopsonist could improve its profits compared to the competitive solution. That is, by reducing consumption it reduces the clearing price for transmission. In general, this would result in an equilibrium with too little transmission capacity compared to welfare optimal.

5.2 Allocation based on energy mark-up

5.2.1 Assumptions

We continue to assume that transmission capacity must be built to accommodate maximum demand. It may be the case that not all of the cost of transmission is allocated through coincident peak charges but instead is "spread" across energy consumption in a volumetric adder per unit energy. For example, in ERCOT, although large industrial and commercial consumers are exposed to 4CP charges, the situation for residential retail consumption is different. In particular, the cost allocated to residential consumers for transmission is generally charged as a fixed volumetric adder per unit energy, rather than a peak demand charge. For simplicity, we will assume that all transmission costs are recovered using a volumetric adder per unit energy, p_T.

5.2.2 Qualitative observations

In contrast to consumers exposed to 4CP charges, consumers that are charged on the basis of a volumetric adder to the energy price will therefore typically consume higher than efficient levels of energy on peak, necessitating more transmission capacity. That is, charging on the basis of energy results in too much transmission capacity compared to welfare optimal. They will also consume lower than efficient levels at off-peak times.

5.2.3 Quantitative analysis

Given an assumed adder to the energy price, p_T, the price to demand for energy is changed from a_G to $a_G + p_T$, resulting in a quantity consumed q_t^{ea}. Equating the price to the inverse demand function at each time results in:

$$\begin{aligned} \forall t \in [0, Y], a_G + p_T &= W(q_t^{ea}, t), \\ &= (N(0) - \delta t - q_t^{ea})/\gamma. \end{aligned}$$

Re-arranging this expression results in:

$$\forall t \in [0, Y], q_t^{ea} = N(0) - \gamma(a_G + p_T) - \delta t.$$

The largest value of consumption occurs at $t = 0$, and this determines the required transmission capacity:

$$T^{ea} = N(0) - \gamma(a_G + p_T),$$

The annual cost of this capacity is $a_T T^{ea} = a_T(N(0) - \gamma(a_G + p_T))$.

The total energy in the year is given by:

$$\begin{aligned} \int_{t=0}^{t=Y} q_t^{ea} dt &= \int_{t=0}^{t=Y} [N(0) - \gamma(a_G + p_T) - \delta t] dt, \\ &= [N(0) - \gamma(a_G + p_T)]Y - \delta(Y)^2/2, \end{aligned}$$

with units of GWh, and resulting the annual payment for transmission $1000 p_T([N(0) - \gamma(a_G + p_T)]Y - \delta(Y)^2/2)$, for p_T in \$/MWh. This payment must recover the annual cost $a_T(N(0) - \gamma(a_G + p_T))$. Equating the payment and annual cost results in a quadratic equation that must be satisfied by p_T:

$$1000 p_T([N(0) - \gamma(a_G + p_T)]Y - \delta(Y)^2/2) = a_T(N(0) - \gamma(a_G + p_T)).$$

Re-arranging, this results in:

$$\begin{aligned} &1000\gamma Y(p_T)^2 + [1000(\gamma a_G Y - N(0)Y + \delta(Y)^2/2) - \gamma a_T]p_T \\ &+ a_T(N(0) - \gamma a_G) = 0. \end{aligned}$$

Solving this quadratic equation in p_T yields $p_T = 0.187$\$/MWh, and $T^{ea} = 25.7$GW, slightly more transmission capacity than welfare optimal. The price charged for all consumption is $a_G + p_T = 10.187$ \$/MWh and the resulting consumption is shown in Fig. 6. The main qualitative difference compared to welfare optimal consumption is that the peak is not clipped, but instead is linearly related to t throughout the load-duration curve. The consumption is above the welfare optimal value in the vicinity of the peak, but below welfare optimal for most of the time. The main qualitative difference compared to the welfare optimal energy-only prices is that the on-peak

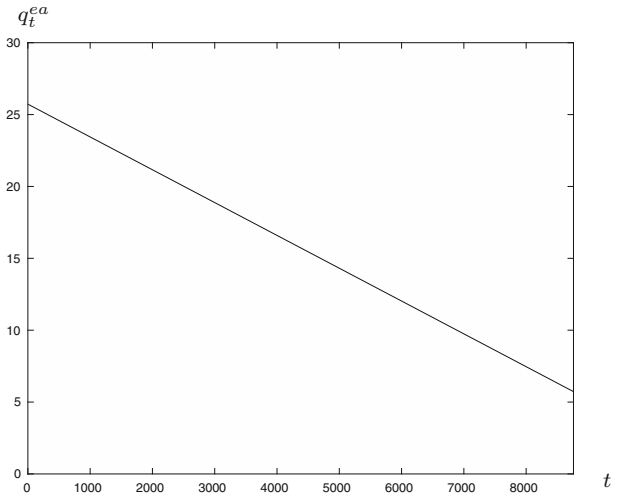

Fig. 6 Equilibrium consumption for example when transmission is charged volumetrically through an energy adder

prices are lower than needed to induce welfare optimality, and the off-peak prices are higher than needed to induce welfare optimality.

This example exhibits only slightly more transmission than optimal, reflecting the relatively small distortion on-peak, and the relatively small contribution to overall energy price of the volumetric transmission cost adder for the assumed cost parameters. The welfare is given by:

$$\begin{aligned}
&\int_{t=0}^{t=Y}\left[\int_{q=0}^{q=q_t^{ea}} W(q,t)dq - a_G q_t^{ea})\right]dt - a_T T^{ea}\\
&= Y\left[(N(0))^2 - 2N(0)\gamma a_G - (\gamma)^2((p_T)^2 - (a_G)^2) - (N(0) - \gamma a_g)\delta Y + (\delta Y)^2/3\right]/(2\gamma) - a_T T^{ea},\\
&= 9.86 \times 10^9 - 2.57 \times 10^7,\\
&= 9.834 \times 10^9 \,\$/\text{year},
\end{aligned}$$

which is almost as high as the optimal welfare.

With higher incremental transmission costs, the distortion of recovering transmission costs with an energy adder becomes more evident. However, with the linear model used, the reduction in welfare is still not extremely substantial. For example, even with annual transmission costs increased fifty-fold to \$50,000/MW, the optimal welfare becomes $\$8.84 \times 10^9$ per year, and this is only reduced by just under 3% to $\$8.58 \times 10^9$ per year using an energy adder. For these assumed values, the energy adder is $p_T = \$9.63$/MWh, which is nearly as large as the energy price alone. This suggests that the policy choice for cost recovery of transmission costs may not be extremely critical if inefficient bypass is not feasible. On the other hand, with high transmission costs and higher elasticity of demand, the welfare losses can be much larger (Schittekatte et al. 2018). Since rooftop solar photovoltaics allow bypass by residential retail customers, the risk of inefficient bypass will be much more significant

as the costs of rooftop solar decrease but nevertheless remain above the underlying wholesale costs of energy.[5]

5.3 "Reliability" investment

5.3.1 Assumptions

In some cases, new transmission construction may not be driven by an equilibrium level of exhibited peak demand, but may be built to accommodate extreme, and unlikely, conditions. In an extreme case of such "reliability" investment, a transmission capacity of $\overline{T}$ is built that exceeds essentially any possible realizable demand.

5.3.2 Qualitative observations

If the costs $C_T(\overline{T})$ of the transmission investment are allocated on the basis of the share of the coincident peak then, as with other examples of recovering sunk costs on the basis of peak consumption levels, this allocation will tend to depress elastic consumption compared to the optimal consumption for any given level of transmission, which in turn means that the actual consumption may be significantly depressed compared to the built capability of the system. Alternatively, if the cost of the investment is allocated on the basis of energy, it is indeterminate as to whether these combined effects result in more or less transmission than welfare optimal.

5.4 Hybrid case

5.4.1 Assumptions

In practice, there are likely to be costs associated with the transmission system that are sunk, and other costs that are avoidable going forward depending on the level of (forecast) peak demand. We assume that coincident peak pricing is used for the avoidable costs, and another tariff mechanism is used for the sunk costs.

5.4.2 Qualitative analysis

The main results in this paper suggest that coincident peak pricing is appropriate for the avoidable costs that depend on the level of demand going forward. In principle, a non-distorting complementary charge, such as a per customer charge, possibly differentiated by customer class, could be used to recover the remaining sunk costs. A recent example of this approach is described in Abdelmotteleb et al. (2018).

[5] I am indebted to Tim Schittekatte of the Florence School of Regulation for this observation.

6 Conclusion

Coincident peak pricing has been analyzed in this paper. In the case of avoidable transmission capacity represented as having a cost that depends on the peak of net demand, coincident peak pricing can, in principle, be efficient. Other approaches, such as a volumetric adder per unit energy, will typically result in inefficient consumption and transmission levels, although the distortion for the numerical cases examined was not large. In cases where high energy prices may result in inefficient bypass, however, the effect of recovering costs with an energy adder may be more deleterious (Schittekatte et al. 2018). In the practical case where some of the costs are sunk and some are avoidable, a hybrid approach may be workably efficient.

In reality, transmission costs exhibit a much more complicated structure than modeled in this paper, with the required capacity in a meshed system driven by so-called $N-1$ security criteria that require deliverability under any single transmission element failure. Moreover, in addition to transmission costs, there are also electric distribution costs. Future work includes extending the model to consider costs of meshed transmission systems under $N-1$ security, distribution systems, and capacitated generation to investigate if, for example, a collection of prices associated with peak utilization of various transmission, distribution, and generation assets could be efficient. For example, a tariff could consist of a transmission charge based on consumption at the time of peaks on the transmission system plus a distribution charge based on consumption at the time of peak utilization of the local distribution feeder, together with a time-varying energy price and a complementary fixed charge to recover remaining costs. Collectively, this four-part tariff may be able to capture the main drivers of transmission, distribution, energy, and per customer costs to provide efficient incentives for consumption.

Acknowledgements This work was accomplished during a visit in Fall 2017 to the Florence School of Regulation, European University Institute, Fiesole, Italy. Support from the University of Texas Faculty Development Program and Professor Jean-Michel Glachant of the Florence School of Regulation is gratefully acknowledged. The author would like to thank Professor Glachant and Tim Schittekatte of the Florence School of Regulation, Florence, Italy, and Nicolas Astier of the Commission de Régulation de l'Énergie, France, for comments on this work.

Appendix: Welfare maximization

This Appendix presents the detailed derivation of welfare maximization for Sect. 4.4. Repeating the analysis in Sect. 2.6, we again consider the objective (4) of the inner problem in problem (3) for each t. Noting that C_G is linear in the case of the example, differentiating the objective (4) and setting equal to zero yields:

$$\begin{aligned} 0 &= W(q_t, t) - a_G, \\ &= (N(t) - q_t)/\gamma - a_G, \\ &= (N(0) - \delta t - q_t)/\gamma - a_G, \end{aligned}$$

using the assumed functional form of D and the resulting form of W. Re-arranging, we obtain the unconstrained maximizer $\hat{q}_t$ at time t of:

$$\hat{q}_t = N(t) - \gamma a_G = N(0) - \delta t - \gamma a_G.$$

We now consider the constraint $q_t \leq T, \forall t \in [0, Y]$, in the inner problem. Following the same argument as in Sect. 2.6, for a given t, if $\hat{q}_t \leq T$, then the maximizing choice of q_t in the inner problem for t is the same as the unconstrained minimizer $\hat{q}_t$. On the other hand, if $\hat{q}_t > T$ then, because the objective is concave, we have that the maximizing choice of q_t for the inner problem is T. Combining these observations, we obtain that the maximizer of the inner problem is:

$$q_t^\star = \min\{T, N(0) - \delta t - \gamma a_G\}.$$

Consider the condition on t where, given T, the transmission constraint is just binding on the maximizing q_t in the inner problem. If such a time $\hat{t}$ exists for which the constraint is just binding, it satisfies $T = N(0) - \delta\hat{t} - \gamma a_G$, so that $\hat{t} = (N(0) - \gamma a_G - T)/\delta$. Abusing notation, we now think of $\hat{t} : \mathbb{R}_+ \to [0, Y]$ as a function of T and define it by:

$$\forall T \in \mathbb{R}, \hat{t}(T) = \max\{0, \min\{Y, (N(0) - \gamma a_G - T)/\delta\}\},$$

and we note that:

$$q_t^\star = \begin{cases} T, & \text{if} \quad 0 \leq t \leq \hat{t}(T), \\ N(0) - \delta t - \gamma a_G, & \text{if} \quad \hat{t}(T) \leq t \leq Y, \end{cases}$$

so that indeed the highest value of demand occurs at $t = 0$ (and, in the typical case that $0 < \hat{t}(T)$, this highest value occurs throughout the interval $[0, \hat{t}(T)]$.) Moreover, we have that, for each t and for a given T, the maximum value of the objective (4) of the inner problem in problem (3) is:

$$\begin{aligned} &\int_{q=0}^{q=q_t^\star} W(q,t)dq - C_G(q_t^\star) \\ &= \int_{q=0}^{q=q_t^\star} ((N(t) - q)/\gamma)dq - C_G(q_t^\star), \\ &= ((N(t)q_t^\star - (q_t^\star)^2/2)/\gamma) - a_G q_t^\star, \text{ on integrating and using the assumed} \\ &\quad \text{functional form for } C_G, \\ &= \begin{cases} T(N(t) - a_G\gamma - T/2)/\gamma, & \text{if } 0 \leq t \leq \hat{t}(T), \\ (N(t) - \gamma a_G)^2/(2\gamma), & \text{if } \hat{t}(T) \leq t \leq Y, \end{cases} \end{aligned} \tag{16}$$

substituting for the optimal value $q_t^\star$. We can now re-write problem (3) as:

$$\max_{T \geq 0} \left\{ \int_{t=0}^{t=Y} \max_{q_t \geq 0} \left\{ \left[\int_{q=0}^{q=q_t} W(q,t)dq - C_G(q_t) \right] \middle| \, q_t \leq T \right\} dt - C_T(T) \right\}$$

$$
\begin{aligned}
&= \max_{T\geq 0}\left\{\int_{t=0}^{t=Y}\left[\int_{q=0}^{q=q_t^\star} W(q,t)dq - C_G(q_t^\star)\right]dt - C_T(T)\right\},\\
&= \max_{T\geq 0}\left\{\int_{t=0}^{t=\hat{t}(T)}\left[\int_{q=0}^{q=q_t^\star} W(q,t)dq - C_G(q_t^\star)\right]dt\right.\\
&\quad\left.+\int_{t=\hat{t}(T)}^{t=Y}\left[\int_{q=0}^{q=q_t^\star} W(q,t)dq - C_G(q_t^\star)\right]dt - C_T(T)\right\},\\
&= \max_{T\geq 0}\left\{\int_{t=0}^{t=\hat{t}(T)}\left[T(N(t)-a_G\gamma - T/2)/\gamma\right]dt\right.\\
&\quad\left.+\int_{t=\hat{t}(T)}^{t=Y}\left[(N(t)-\gamma a_G)^2/(2\gamma)\right]dt - a_T T\right\},
\end{aligned}
$$

using the assumed functional forms for W, C_G, and C_T, and using (16) to evaluate the inner integrals. We now observe that the objective of the outer problem is of the form:

$$
\int_{t=0}^{t=\hat{t}(T)} f(T,t)dt + \int_{t=\hat{t}(T)}^{t=Y} g(t)dt - a_T T, \tag{17}
$$

where $f : \mathbb{R}_+ \times [0,Y] \to \mathbb{R}$ and $g : [0,Y] \to \mathbb{R}$ are of the form:

$$
\begin{aligned}
\forall T \in \mathbb{R}_+, \forall t \in [0,Y],\ f(T,t) &= T(N(t) - a_G\gamma - T/2)/\gamma,\\
\forall t \in [0,Y],\ g(t) &= (N(t)-\gamma a_G)^2/(2\gamma),
\end{aligned}
$$

and we note that $f(T,\hat{t}(T)) = g(\hat{t}(T)) = (T)^2/(2\gamma)$. The expression in (17) represents the optimal welfare *given* a transmission capacity T. To find the maximizer of welfare as expressed in (17) over T, we differentiate it with respect to T and set the derivative equal to zero:

$$
\begin{aligned}
0 &= \left[f(T,\hat{t}(T)) - g(\hat{t}(T))\right]\frac{\partial \hat{t}(T)}{\partial T} + \int_{t=0}^{t=\hat{t}(T)} \frac{\partial f}{\partial T}(T,t)dt - a_T,\\
&= \int_{t=0}^{t=\hat{t}(T)} [(N(t)-T)/\gamma - a_G]dt - a_T,\\
&\qquad \text{since } f(T,\hat{t}(T)) = g(\hat{t}(T)),\ \text{and by differentiation of } f,\\
&= (N(0) - \gamma a_G - T)^2/(2\gamma\delta) - a_T,
\end{aligned}
$$

on integrating and simplifying, assuming that $0 < \hat{t}(T) < Y$, so that we can simplify the evaluation of $\hat{t}(T)$ to $\hat{t}(T) = (N(0) - \gamma a_G - T)/\delta$. Setting this derivative equal to zero and noting that $N(0) - \gamma a_G - T \geq 0$ yields the optimizer:

$$
T^\star = N(0) - \gamma a_G - \sqrt{2\delta\gamma a_T},
$$

again assuming that $0 < \hat{t}(T)^\star < Y$. Noting that $(N(0) - \gamma a_G - T^\star)/\delta = \sqrt{2\gamma a_T/\delta}$, we observe that $0 < \hat{t}(T^\star) < Y$ if:

$$\sqrt{2\gamma a_T/\delta} < Y.$$

For future convenience, note that:

$$\begin{aligned} W(T^\star, t) &= (N(0) - \delta t - T^\star)/\gamma, \\ &= a_G + (\sqrt{2\delta\gamma a_T} - \delta t)/\gamma. \end{aligned} \tag{18}$$

We now evaluate the optimal welfare; that is, we evaluate (17) at the welfare optimizing transmission level $T^\star$. We initially assume that the condition $0 < \hat{t}(T) < Y$ is satisfied, but numerically verify that this condition is satisfied. From (17), we have that the optimal welfare is (re-arranging (17)):

$$\begin{aligned} &\int_{t=\hat{t}(T^\star)}^{t=Y} g(t)dt + \int_{t=0}^{t=\hat{t}(T^\star)} f(T^\star, t)dt - a_T T^\star \\ &= \int_{t=0}^{t=Y} g(t)dt + \int_{t=0}^{t=\hat{t}(T^\star)} (f(T^\star, t) - g(t))dt - a_T T^\star, \\ &\quad \text{where we note that the first integral does not depend on } T^\star, \\ &\quad \text{whereas the second does depend on } T^\star, \\ &= \int_{t=0}^{t=Y} \left[(N(0) - \gamma a_G - \delta t)^2/(2\gamma)\right] dt \\ &\quad + \int_{t=0}^{t=\hat{t}(T^\star)} \left[(N(0) - a_G\gamma - T^\star - \delta t)/(-2\gamma)\right] dt - a_T T^\star, \\ &\quad \text{on evaluating the terms,} \\ &= [(N(0) - a_G\gamma)^3 - (N(0) - a_G\gamma - \delta Y)^3]/(6\gamma\delta) - (\sqrt{2\delta\gamma a_T})^3/(6\gamma\delta) - a_T T^\star, \end{aligned}$$

on integrating and simplifying. As noted in Sect. 4.4, the first term in the expression for optimal welfare is independent of transmission costs and represents the optimal welfare if the transmission were costless. The last two terms represent the reduction in welfare due to the non-zero costs of transmission, with the last term being the direct cost of the transmission and the second-last term being the reduction in welfare due to the transmission limit.[6]

References

Abdelmotteleb, I., Gómez, T., Ávila, J. P. C., & Reneses, J. (2018). Designing efficient distribution network charges in the context of active customers. *Applied Energy*, *210*, 815–826.

[6] Because the system is radial, there is no "indirect" term relating to the change in the admittance associated with upgrading the line (Caramanis et al. 1982; Léautier 2000, Section 3). This issue would be necessary to consider in meshed systems.

Agency for the Cooperation of Energy Regulators. (2017). Annual report on the results of monitoring the internal electricity and gas markets in 2016: Electricity and gas retail markets volume. Technical report, 2017. https://acer.europa.edu. Accessed December 2017.

Baldick, R., & Hogan, W. (2002). Capacity constrained supply function equilibrium models of electricity markets: Stability, non-decreasing constraints, and function space iterations. University of California Energy Institute POWER Paper PWP-089. https://ei.haas.berkeley.edu/research/papers/PWP/pwp089.pdf. Accessed September 2018.

Baldick, R., & Kahn, E. (1993). Network costs and the regulation of wholesale competition in electric power. *Journal of Regulatory Economics*, *5*(4), 367–384.

Bergen, A. R., & Vittal, V. (2000). *Power systems analysis* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Borenstein, S. (2012). The economics of fixed cost recovery by utilities. *The Electricity Journal*, *29*, 5–12.

Bushnell, J. B., & Stoft, S. E. (1997). Improving private incentives for electric grid investment. *Resource and Energy Economics*, *19*, 85–108.

Caramanis, M. C., Bohn, R. E., & Schweppe, F. C. (1982). Optimal spot pricing: Practice and theory. *IEEE Transactions on Power Apparatus and Systems, PAS*, *101*(9), 3234–3245.

Day, C. J., & Bunn, D. W. (2001). Divestiture of generation assets in the electricity pool of England and Wales: A computational approach to analyzing market power. *Journal of Regulatory Economics*, *19*(2), 123–141.

Dixit, K., & Baldick, R. (2003). An empirical study of the economies of scale in AC transmission line construction costs. http://www.ece.utexas.edu/~baldick/papers/economiesofscale.pdf. Accessed September 2018

Electric Reliability Council of Texas. (2014). New transmission project to deliver more power to Houston area, April 2014. http://www.ercot.com/news/releases/show/26616. Accessed October 2017.

Electric Reliability Council of Texas. (2017). ERCOT nodal protocols. http://nodal.ercot.com/protocols/index.html. Accessed December 4, 2017.

Ergon Energy. (2018). Demand tariffs. Technical report, 2018. https://www.ergon.com.au/retail/residential/tariffs-and-prices/demand-tariffs. Accessed July 2018.

Faruqui, A., & Aydin, M. G. (2017). Moving forward with electric tariff reform. *Regulation*, *40*(3), 42–48.

Green, R. (1996). Increasing competition in the British electricity spot market. *The Journal of Industrial Economics*, *XLIV*(2), 205–216.

Green, R. (1999). The electricity contract market in England and Wales. *The Journal of Industrial Economics*, *XLVII*(1), 107–124.

Green, R., & Newbery, D. M. (1992). Competition in the British electricity spot market. *Journal of Political Economy*, *100*(5), 929–953.

Haro, S., Aragonés, V., Martínez, M., Moreda, E., Morata, A., Arbós, E., et al. (2017). Toward dynamic network tariffs: A proposal for spain. In F. P. Sioshansi (Ed.), *Competitive electricity markets, chapter 12* (pp. 221–241). Amsterdam: Elsevier Science Ltd.

Hogan, W. W., & Pope, S. L. (2017). Priorities for the evolution of an energy-only market design in ERCOT. Technical report, FTI Consulting, Washington, DC, 2017. https://sites.hks.harvard.edu/fs/whogan/Hogan_Pope_ERCOT_050917.pdf. Accessed October 2017.

Klemperer, P. D., & Meyer, M. A. (1989). Supply function equilibria in oligopoly under uncertainty. *Econometrica*, *57*(6), 1243–1277.

Kristiansen, T., & Rosellón, J. (2006). A merchant mechanism for electricity transmission expansion. *Journal of Regulatory Economics*, *29*(2), 167–193.

Léautier, T.-O. (2000). Regulation of an electric power transmission company. *The Energy Journal*, *21*(4), 61–92.

Léautier, T.-O., & Thelen, V. (2009). Optimal expansion of the power transmission grid: Why not? *Journal of Regulatory Economics*, *36*, 127–153.

Li, F. (2007). Long-run marginal cost pricing based on network spare capacity. *IEEE Transactions on Power Systems*, *22*(2), 885–886.

Majidi-Qadikolai, M., & Baldick, R. (2016a). Integration of $n-1$ contingency analysis with systematic transmission capacity expansion planning: Ercot case study. *IEEE Transactions on Power Systems*, *31*(3), 2234–2245.

Majidi-Qadikolai, M., & Baldick, R. (2016b). Stochastic transmission capacity expansion planning with special scenario selection for integrating $n-1$ contingency analysis. *IEEE Transactions on Power Systems*, *31*(6), 4901–4912.

Majidi-Qadikolai, M., & Baldick, R. (2018). A generalized decomposition framework for large-scale transmission expansion planning. *IEEE Transactions on Power Systems*, *33*(2), 1635–1649.

Manuel de Villena, M., Gautier, A., Fonteneau, R., & Ernst, D. (2017). A simulator to explore tarification designs for distribution networks. Unpublished manuscript.

National Grid. (2015). Introduction to triads. Technical report, 2015. https://www.nationalgrid.com/sites/default/files/documents/44940-Triads%20Information.pdf. Accessed July 2018.

Passey, R., Haghdadib, N., Bruce, A., & MacGill, I. (2017). Designing more cost reflective electricity network tariffs with demand. *Energy Policy*, *109*, 642–649.

Rosellón, J., & Weigt, H. (2011). A dynamic incentive mechanism for transmission expansion in electricity networks: Theory, modeling, and application. *The Energy Journal*, *32*(1), 119–148.

Schittekatte, T., Momber, I., & Meeus, L. (2018). Future-proof tariff design: Recovering sunk grid costs in a world where consumers are pushing back. *Energy Economics*, *70*, 484–498.

Stoft, S. (2003). The demand for operating reserves: Key to price spikes and investment. *IEEE Transactions on Power Systems*, *18*(2), 470–477.

Wood, A. J., & Wollenberg, B. F. (1996). *Power generation, operation, and control* (second ed.). New York: Wiley.

Zarnikau, J. (2017a). Three simple steps to clip the peak in the texas (ERCOT) electricity market. Technical report, USAEE, September 2013. Working Paper No. 13-143. https://ssrn.com/abstract=2334001. Accessed January 1, 2017.

Zarnikau, J. (2017b). Results from the 2013 survey of LSEs to obtain retail DR and dynamic pricing information. Technical report, Frontier Associates and ERCOT, Austin and Taylor, TX, 2014. http://www.ercot.com/calendar/2014/6/25/32352-DSWG. Accessed October 2017.