SERIEs (2021) 12:281–327
https://doi.org/10.1007/s13209-021-00233-9

SERIEs Journal of the Spanish Economic Association

ORIGINAL ARTICLE

Emission trading systems and the optimal technology mix

Gregor Zoettl[1]

Received: 8 January 2021 / Accepted: 13 April 2021 / Published online: 16 May 2021

Abstract

Cap and trade mechanisms enjoy increasing importance in environmental legislations worldwide. One of the important aspects of designing cap and trade mechanisms is the possibility of authorities to grant emission permits for free. Unlike analyzed in the seminal contributions on cap and trade systems, in reality free allocations are not made lump sum, but are updated contingent on firms' actions, i.e., contingent on production decisions and contingent on facilities entering the market or retiring (see IEA, https://www.oecd-ilibrary.org/content/publication/b7d0842b-en, 2020). As it has already been shown in the literature, such updating yields distorted production decisions of firms (see e.g., Böhringer and Lange in Eur Econ Rev 49(8):2041–2055, 2005, Mackenzie et al. in Environ Resour Econ 39(3):265–282, 2008, or Damon et al. in Rev Environ Econ Policy 13(1):23–42, 2019). The impact of updating on firms' investment and retiring decisions and the resulting technology mix has received much less attention up to now, however. It is the purpose of the present article to shed light on this aspect and to study the impact of a cap and trade mechanism not only on firms' output decisions, but also on their investment incentives in different technologies and to analyze the optimal design of emission trading systems in such an environment.

Keywords Emissions trading · Free allocation · Investment incentives · Technology mix

JEL Classification H21 · H23 · Q55 · Q58 · 033

This research was performed as part of the Energie Campus Nuremberg. The authors thank the Deutsche Forschungsgemeinschaft for their support within Projekt B09 in the Sonderforschungsbereich TRR 154 Mathematical Modelling, Simulation and Optimization using the Example of Gas Networks. We want to thank Veronika Grimm, Karsten Neuhoff and Yves Smeers for many valuable comments.

✉ Gregor Zoettl
gregor.zoettl@fau.de

1 Department of Management and Economics, Friedrich Alexander University Erlangen Nuremberg, Lange Gasse, Nuremberg, Germany

1 Introduction

Cap and trade mechanisms are one of the centrally important policy tools to achieve ambitious greenhouse gas reduction targets required in the context of recent climate agreements (e.g., the 2015 Paris agreement, UNO 2015). In the present article, we analyze the design of cap and trade mechanisms and their impact on firms' investment decisions in different technologies and on their final production decisions. We then derive the optimal design for ideal market conditions, but also for non-ideal situations where firms either exercise market power or competition authorities' decisions are partially constrained by requirements of political or legislative processes.

Cap and trade mechanisms designed to internalize social cost of pollution enjoy increasing importance in environmental legislation worldwide. Well-known examples are given by the European Union Emission Trading System (ETS), the California Cap-and-Trade programme or the Korea ETS. "As of April 2020, there were 23 emissions trading systems covering around 9% of global emissions" (see IEA 2020). A very prominent example of a very large planned system is the China national ETS which is likely to start in 2021.[1] An important aspect when introducing cap and trade mechanisms is the possibility of competition authorities to grant emission permits for free (see, e.g., ICAP 2020). This apparently allows to crucially facilitate the political processes leading to the introduction of cap and trade systems. As Convery (2009) in an early survey on the origins and the development of the EU ETS observes: "The key quid pro quos to secure industry support in Germany and across the EU were agreements that allocation would take place at Member State level [...], and that the allowances would be free." Very similar observations can also be found in many other contributions to that issue.[2]

A one and for all lump sum allocation of permits which is entirely independent of firms' actions has a purely distributive impact in case firms take the market price of emission permits as given. This fundamental insight in principle dates back to Coase (1960). The design of free allocations in currently active cap and trade systems typically does not have such lump sum property, but includes explicit or implicit features of updating (see IEA 2020). Updating of free allocation schemes designed to consistently adapt to an industry's dynamic development has an impact on firms' behavior, however.[3] First, it leads to a distortion of the operation of existing production facilities since for existing cap and trade systems current output and emissions do have an impact on allocations granted to those facilities in the future. Second, it also has an impact on firms' incentives to modify their production facilities through upgrading,

[1] See https://www.reuters.com/article/us-china-climatechange-ets-idUSKBN29G083.

[2] As, for example, Tietenberg (2006) observes: "free distribution of permits (as opposed to auctioning them off) seems to be a key ingredient in the successful implementation of emissions trading programs." Bovenberg et al. (2008) state: "The compensation issue has come to the fore in recent policy discussions. For example, several climate change policy bills recently introduced in the U.S. Congress (for example, one by Senator Jeff Bingaman of New Mexico and another by Senator Dianne Feinstein of California) contain very specific language stating that affected energy companies should receive just enough compensation to prevent their equity values from falling."

[3] As Böhringer and Lange (2005) state: "As a case in point, one major policy concern is that [...] the allocation should account for (major) changes in the activity level of firms. Free allocation schemes must then abstain from lump-sum transfers and revert to output- or emission-based allocation."

retiring and building of new facilities since typically free allocations to some extent are granted contingent on newly installed or retired facilities.[4]

The literature which analyzes the impact of updating on firms' behavior up to now has focused on the first aspect. That is, those contributions provide very rich insight on the impact of updating based on past output or emissions on firms' production and emission decisions, and they abstract from changed free allocations to a firm due to changed installed facilities, however. The most prominent contributions in this context include Moledina et al. (2003), Böhringer and Lange (2005), Rosendahl (2008), Mackenzie et al. (2008), Harstad and Eskeland (2010)), Böhringer et al. (2017), Qiu et al. (2017), Meunier et al. (2017) or Meunier et al. (2018). Reviewing current and planned emission trading schemes worldwide reveals, however, that legislations which provide free emission permits also include elements which update based on newly installed or retired production facilities (Narassimhan et al. 2018; IEA 2020).[5] It is the purpose of the present article to focus on the above-mentioned second aspect and its impact on the technology mix chosen by firms. To the best of our knowledge, this is the first article that formally analyzes the impact of updating on firms' investment incentives in an analytical framework.

Usually, production facilities in most industries allow for production during longer horizons of time. Since demand typically varies over time, it is optimal for firms to invest in a portfolio of different technologies. Technologies with high investment cost which allow for production at low marginal cost are installed to run most of the time, whereas technologies with low investment cost which allow for production at higher marginal cost cover more infrequent demand periods. Prominent examples of capital-intensive and long-lasting investments in production facilities and fluctuating demand due to limited storage possibility are given by, e.g., the production of cement, steel or electricity (with electricity being the most prominent and well-studied example in this context). We thus analyze a framework with fluctuating demand where (strategic) firms first determine their technology mix by investing in different production technologies with different emission intensities. Production facilities allow firms to produce for a continuum of spot markets subject to fluctuating demand. Production at each spot market causes emissions, and total emissions at all spot markets are capped by an emission permit market. We consider the impact of a cap and trade system where different amounts of free allocations are granted for newly built facilities. After having established the market equilibrium both for firms' investment and production decisions and for the emission permit market, we first determine the first best solution as a benchmark. Analogous to most of the previous literature, if distributional concerns do not matter, in an ideal market with perfectly competitive firms updating is not optimal

[4] Numerical simulations by Neuhoff et al. (2006), Pahle et al. (2011) and more recently by Anouliès (2017) and Dardati and Saygili (2020) highlight the importance of the proper design of free allocations in this case.

[5] IEA (2010): "An important detail of systems using grandfathered allocation is the treatment of companies that establish new facilities or close down. Current or proposed schemes generally provide new entrants with the same support as existing facilities. The rationale for this is to avoid investment moving to jurisdictions without carbon pricing. [...] However, there is an obvious political difficulty in continuing to allocate free allowances to facilities that have shut down and almost all emissions trading systems require allowances to be surrendered upon closure."

(i.e., no free allocations should be granted to any investment), and the total emission cap should be set such that the permit price equals to marginal social cost of pollution.

In the main part of the paper, we then analyze the optimal design of a cap and trade system if the market is not ideal. First, we consider the case that firms behave imperfectly competitively when making their investment and their production decisions. In this case, investment and production are inefficiently low. As we show, measures undertaken to stimulate investment incentives (such as capacity payments[6]) have a very close relationship to free allocations made for newly installed production facilities. In this case, it is optimal to either grant capacity payments or free allocations in order to stimulate inefficiently low investment incentives. As we show, however, in a closed system with endogenous permit market, it is not optimal to implement total investment at first best levels since this would imply an inefficiently high permit price which excessively depresses spot market output. Our results do thus have direct implications for the design of capacity markets in the presence of a cap and trade system.

Second, we analyze the case where the design of the cap and trade mechanism is subject to political constraints (compare the second paragraph of this section) and the competition authority has to determine the optimal market design given those constraints.[7] We first analyze how the optimal target on total emissions should be set if free allocations in all technologies are exogenously fixed. As we find, for moderate levels of free allocations the target on total emissions should be set such that the equilibrium permit price is above marginal social cost of pollution. For high levels of free allocation, for example, in case of full allocation where all permits used by a certain technology during a compliance period are freely allocated, the total cap on emissions should be set such that the equilibrium permit price is below marginal social cost of pollution.

We then analyze the case that free allocation only for a specific technology is exogenously fixed and determine the optimal level of free allocation for the remaining technology. In order to avoid excessive distortions of the resulting technology mix, it is typically optimal to grant free allocation for the remaining technology. That is, the insights obtained from the first best benchmark that free allocations are never optimal, are no longer true if allocation to one of the technologies is exogenously fixed. Moreover, if this technology is relatively dirty (as compared to the technology with exogenously fixed allocation), the level of free allocation should remain below the exogenously fixed allocation. If on the contrary the remaining technology is relatively clean, the level of free allocation should even be above the exogenously fixed allocation. Observe that often observed practices of full allocation (see IEA 2020; ICAP 2020) induce a pattern of free allocation which is completely opposed to those findings.

[6] Compare, for example, capacity mechanisms introduced in liberalized electricity markets; for an overview, see Cramton and Stoft (2008), or Cramton and Ockenfels (2012) or more recently Fabra (2018).

[7] Observe that to some extent, this parallels the fundamental approach found in the previous literature: Böhringer and Lange (2005) provide second best rules if (for political reasons) updating has to be based on past output, Harstad and Eskeland (2010) analyze market design if governments cannot commit to full auctioning of permits, and Bovenberg et al. (2005) and Bovenberg et al. (2008) consider the constraint that firms have to be fully compensated for the regulatory burden.

Let us finally mention further related strands of the literature. First, several articles analyze firms' incentives to adopt cleaner technologies. Those include, for example, Requate and Unold (2001), Montero (2002) or Requate and Unold (2003), for a survey of this literature, compare Requate (2005). Whereas all those contribution focus on the comparison of emission taxes and cap and trade systems, our contribution is the first one to consider the phenomenon of updating in the context of technology choice.

Second, from a modeling perspective the present paper also contributes to the literature which analyzes investment decisions in several technologies. For a survey on this literature, see Crew et al. (1995). Further contributions include Zöttl (2010), or more recently Grimm et al. (2017). Our article is (to the best of our knowledge) the first one to introduce an endogenous emission permit market in such a framework, and to derive the optimal design of a cap and trade mechanism with technology-specific free allocations both for ideal and imperfect market conditions.

The remainder of the article is structured as follows: Sect. 2 states the model analyzed throughout this article, and Sect. 3 derives the market equilibrium for a given cap and trade mechanism. We determine the optimal market design for the benchmark case of perfect competition in Sect. 4, for the case of imperfect competition in Sect. 5 and for partially constrained cap and trade mechanisms in Sect. 6. Section 7 concludes.

2 The model

We consider n symmetric firms. Each firm i first has to choose investment in two different types of production facilities prior to producing output $q_i(\theta)$ at many consecutive spot markets with fluctuating demand. We denote total investment of firm i in both technologies by x_{1i} and investment in technology 2 by x_{2i}. Observe that in the framework analyzed, where demand fluctuates over time, it is optimal for firms to invest into a mix of both technologies. We will consider the case that technology 2 allows cheaper production, but exhibits higher investment cost. Those units have to run most of the time in order to recover their high investment cost. (This is typically denoted "baseload–technology.") Technology 1 has relatively low investment cost, but produces at high marginal cost. Those units are built in order to serve during periods of high demand (this is typically denoted "peakload–technology"), but run idle if demand is low. Inverse demand for the quantity $Q(\theta)$ at spot market $\theta \in [\underline{\theta}, \overline{\theta}]$ is given by $P(Q(\theta), \theta)$. Output produced by the different facilities causes emissions, and all emissions have to be covered by emission permits. Permits are issued in the context of a cap and trade mechanism and can be traded at a permit market. The total amount of permits available is limited, and this induces a market clearing price e for permits. Firms receive free allocation of permits based on installed capacities, i.e., each firm i receives the technology-specific amount $(A_2 x_{2i} + A_1\,(x_{1i} - x_{2i}))$ of permits for free, all remaining permits required for production have to be purchased (can be sold) at the permit market.[8] For the overall timing of our setup, see Fig. 1. Subsequently, we provide all definitions and assumptions for our setup in detail.

[8] We provide a discussion of the relationship of our modeling approach and existing cap and trade systems when presenting our results, see the last paragraphs of Sects. 3 to 6.

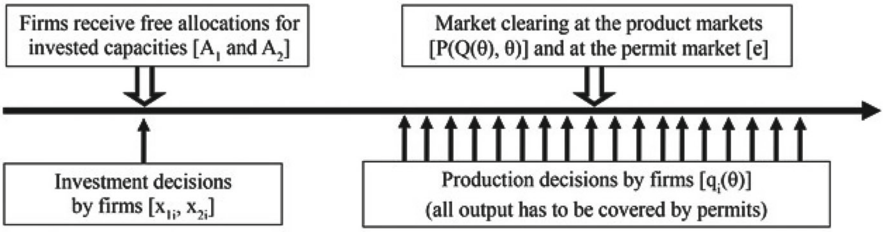

Fig. 1 Illustration of timing

Investment and production decisions Firms have to choose investment in capacities for two different technologies, $t = 1, 2$. Each technology t has constant marginal cost of investment k_t, constant marginal cost of production c_t and an emission factor w_t which measures the amount of the pollutant emitted per unit of output. We denote total investment of firm i in both technologies by x_{1i} and investment in technology 2 by x_{2i}. For aggregate values, we define $X_t = \sum_{i=1}^{n} x_{ti}$.[9] For an illustration of the resulting marginal cost function, see Fig. 2. We denote by $q(\theta) = (q_1(\theta), \ldots, q_n(\theta))$ the vector of spot market outputs of the n firms in demand scenario θ, and by $Q(\theta) = \sum_{i=1}^{n} q_i$ total spot market output in scenario θ.

Demand Inverse demand for some quantity Q at spot market $\theta \in [\underline{\theta}, \overline{\theta}]$ is given by $P(Q, \theta)$. Each spot market has weight $f(\theta)$ with corresponding distribution $F(\theta) = \int_{\underline{\theta}}^{\theta} f(\theta)\mathrm{d}\theta$.[10] Inverse demand satisfies $P_q(Q, \theta) < 0$, $P_\theta(Q, \theta) \geq 0$, $P_{q\theta}(Q, \theta) \geq 0$ and $P_q(Q, \theta) + P_{qq}(Q, \theta)\frac{Q}{n} < 0$ for all $Q, \theta \in \mathbb{R}$.[11]

Cap and trade mechanism and social cost of pollution Total emissions produced at all spot markets $\theta \in [\underline{\theta}, \bar{\theta}]$ are given by $\mathcal{T}$. The social cost associated with emissions is denoted by $D(\mathcal{T})$. We assume $D_T(\mathcal{T}) \geq 0$ and $D_{TT}(\mathcal{T}) \geq 0$. A cap and trade mechanism limits total emissions at some level T such that $\mathcal{T} \leq T$. We denote the resulting market price for emission permits by e. We make the following assumptions regarding trade at the permit market, i.e., (i) Emission permit trading is arbitrage-free, and storage of permits is costless, and (ii) firms are price takers at the permit market.[12] Notice, finally, that our formulation allows for technology-specific free allocations (A_1, A_2). Free allocations might, for example, depend on the emission intensities w_1, w_2 and (or) the usage rates of each technology.

[9] Thus, aggregate investment in technology 1 is given by $X_1 - X_2$. This at first sight unusual notation allows for a considerably simplified representation of our results throughout the article.

[10] Formally, we treat the frequencies associated with the realizations of θ by making use of a density and a distribution function. Notice, however, that there is no uncertainty in the framework presented. All realizations of $\theta \in [\underline{\theta}, \overline{\theta}]$ indeed realize, with the corresponding frequency $f(\theta)$.

[11] We denote the derivative of a function $g(x, y)$ with respect to the argument x, by $g_x(x, y)$, the second derivative with respect to that argument by $g_{xx}(x, y)$ and the cross-derivative by $g_{xy}(x, y)$.

[12] Since emission trading systems typically encompass larger regions which include several industries, it seems unlikely that single firms are able to exercise market power on the emission permit market. Within a specific industry, however, the product market might well be subject to the exercise of market power, and we thus explicitly include this case in our analysis.

3 The market equilibrium

In this section, we derive the market equilibrium with a cap and trade mechanism, for the case of perfect and imperfect competition on the product market. To characterize the market equilibrium for a given cap and trade mechanism (T, A_1, A_2), we first determine firms' profits π_i, given investments x_1, x_2 and given spot market output $q(\theta)$.

$$\begin{aligned}\pi_i(x_{1i}, x_{2i}) = &\int_{\underline{\theta}}^{\theta_{\overline{B}}} (P(Q,\theta)-c_2-w_2e)\, q_i(\theta, x) dF(\theta) + \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} (P(X_2,\theta)-c_2-w_2e)\, x_{2i} dF(\theta)\\ &+\int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} (P(Q,\theta)-c_1-w_1e)\, q_i(\theta, x) dF(\theta) + \int_{\theta_{\overline{P}}}^{\overline{\theta}} (P(X_1,\theta)-c_1-w_1e)\, x_{1i} dF(\theta)\\ &-\int_{\theta_{\underline{P}}}^{\overline{\theta}} ((c_1+w_1e)-(c_2+w_2e))\, x_{2i} dF(\theta) - (k_2-A_2e)x_{2i} - (k_1-A_1e)(x_{1i}-x_{2i}). \end{aligned} \quad (1)$$

Note that the permit market affects both the firms' marginal production cost and their investment cost. No matter whether permits have been allocated for free or have to be bought at the permit market, firms face opportunity cost of $w_t e$ when deciding to produce one unit of output with technology $t = 1, 2$. This opportunity cost increases their marginal production cost to $c_t + w_t e$, $t = 1, 2$. Investment cost is affected by the firms' anticipation of a free allocation of permits. A free allocation is equivalent to a subsidy paid upon investment: If each unit of capacity invested is assigned A_t permits, investment cost k_t is reduced by their value, that is by $A_t e$ for $t = 1, 2$.

The critical spot market scenarios[13] $\theta_{\overline{B}}, \theta_{\underline{P}}, \theta_{\overline{P}}$ indicate whether firms produce either at the capacity bounds x_2, x_1 (that is, at the vertical pieces of their marginal cost curves), or on the flat (i.e., unconstrained) parts of their marginal cost curves. They depend on the intensity of competition at the spot market and are illustrated in Fig. 2, both for the case of perfect and imperfect competition. For $\theta \in [\underline{\theta}, \theta_{\overline{B}}]$, firms produce the output at marginal cost c_2. For $\theta \in [\theta_{\overline{B}}, \theta_{\underline{P}}]$, firms are constrained by their investment in the base load technology and produce X_2, still at marginal cost c_2, and prices are driven by the demand function. At those demand levels, using the peak load technology 1 is not yet profitable. Observe that $F(\theta_{\underline{P}}) - F(\theta_{\overline{B}})$ measures the fraction of time where investment in the base load technology is binding which we will refer to as *constrained base duration*. For $\theta \in [\theta_{\underline{P}}, \theta_{\overline{P}}]$, firms produce output at marginal cost c_1, and we denote $1 - F(\theta_{\underline{P}})$ as *peak duration*.[14] Finally, for all realizations above $\theta_{\overline{P}}$, firms are constrained by their total capacity choice X_1, and prices are driven exclusively by the demand function; we denote $1 - F(\theta_{\bar{P}})$ as *constrained peak duration*. In the subsequent lemma, we characterize the market equilibrium when firms invest in the base load and in the peak load technology.

[13] For the precise definition of those critical spot market scenarios, see Appendix A.

[14] The equivalent *base duration* would be given by $1 - F(\underline{\theta}) = 1$, it is not explicitly introduced, however.

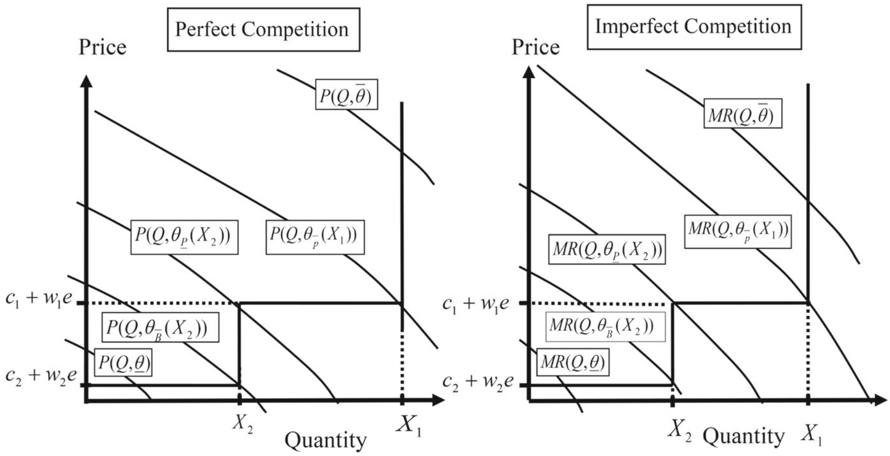

Fig. 2 Illustration of the critical spot market scenarios. Left: The case of a perfectly competitive market, right: the general case with imperfect competition. In the figure, we denote marginal revenue by $MR(Q, \theta) := P(Q, \theta) + P_q(Q, \theta)\frac{Q}{n}$

Lemma 1 *For a given cap and trade mechanism* (T, A_1, A_2)*, define the total investment condition* Ψ_I*, the base investment condition* Ψ_{II} *and the permit pricing condition*[15] Ψ_E *as follows:*

$$\Psi_I := \int_{\theta_{\overline{P}}}^{\overline{\theta}} \left[P(X_1^*, \theta) + P_q(X_1^*, \theta)\frac{X_1^*}{n} - (c_1 + w_1 e^*) \right] \mathrm{d}F(\theta) - \left(k_1 - A_1 e^*\right) \tag{2}$$

$$\Psi_{II} := \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} \left[P(X_2^*, \theta) + P_q(X_2^*, \theta)\frac{X_2^*}{n} - (c_2 + w_2 e^*) \right] \mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\overline{\theta}} (c_1 - c_2) + (w_1 - w_2)e^* \mathrm{d}F(\theta) - (k_2 - k_1) + (A_2 - A_1)e^* \tag{3}$$

$$\Psi_E := \int_{\underline{\theta}}^{\theta_{\overline{B}}} w_2 Q(e^*, \theta)\mathrm{d}F(\theta) + \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} w_2 X_2^* \mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} w_1 Q(e^*, \theta)\mathrm{d}F(\theta) + \int_{\theta_{\overline{P}}}^{\overline{\theta}} w_1 X_1^* \mathrm{d}F(\theta) - \int_{\theta_{\underline{P}}}^{\overline{\theta}} (w_1 - w_2) X_2^* \mathrm{d}F(\theta) - T \tag{4}$$

Equilibrium investment X_1^*, X_2^* *and the equilibrium permit price* e^* *simultaneously solve* $\Psi_I = \Psi_{II} = \Psi_E = 0$.

Proof See Appendix A. □

[15] For a positive permit price, we might also obtain the situation, where production for very low demand realizations is suppressed and positive output is produced only for demand realizations which satisfy θ : $P(0, \theta) - C(0, \theta) - e > 0$. For ease of notation, we disregard this corner solution, which could be easily included in the entire analysis.

In the lemma, (2) is the first-order condition that determines total investment. Firms choose their total investment $\frac{X_1^*}{n}$ as to equal marginal profits generated by their last running unit (running at total marginal cost $c_1 + w_1 e^*$) to the investment cost of that unit (given by $k_1 - A_1 e^*$). As already mentioned above, under a cap and trade mechanism, the value of the permits required for production at the spot market is part of the firms' marginal production cost, and the value of free allocations is of firms' marginal cost of investment.

Now, let us provide some intuition on the determinants of the optimal base load investment. Since total investment X_1 has already been fixed [it is determined by (2)], the firms' decision when choosing X_2 has to be interpreted as a decision of virtually replacing units of technology 1 by technology 2. The cost of such virtual replacement of the marginal unit (given by $k_2 - k_1 - (A_2 - A_1)e^*$) has to equal the extra profits generated by that unit due to lower marginal production cost. Lower production cost of one additional unit has two effects: First, for all demand realizations $\theta \in [\theta_{\overline{B}}, \theta_{\underline{P}}]$ one more unit is produced (that would not have been produced without the replacement); for $\theta \in [\theta_{\underline{P}}, \overline{\theta}]$, one more unit can be produced at lower marginal cost $c_2 + w_2 e^*$ (instead of $c_1 + w_1 e^*$) due to the replacement. (Compare also Fig. 2.)

The market price for permits, e^*, depends on the emission target T set by the market designer as well as the technology mix installed by the firms. At the equilibrium permit price, the market exactly clears, allowing for total emission of T units of the pollutant. Notice that in expression (4), total emissions are determined by multiplying total production at all spot markets with the relevant emission factors of the respective technologies (w_1, w_2).

Finally, observe that Lemma 1 characterizes the market solution when firms decide to invest in both technologies, that is, when indeed $0 < X_2^* < X_1^*$ is obtained. First, whenever the base load technology (k_2, c_2) is very unattractive,[16] then only the peak load technology (k_1, c_1) is active. Second, if the base load technology (k_2, c_2) is always more attractive[17] than the peak load technology (k_1, c_1), then only technology (k_2, c_2) is active in the market equilibrium. Notice that in principle, the case of investment in a single technology is covered by our framework, it is obtained by eliminating the possibility to invest in technology 2, and expression (2) then determines investment in the single technology. To keep the notational burden limited, however, we do not explicitly include those corner solution in the exposition of the paper, but opted to focus on all those cases when firms indeed choose to investment in both technologies.

To conclude the discussion of Lemma 1, let us already at this point mention the relevance of endogenously modeling the emission permit market as compared to the case which assumes an exogenously fixed price for pollution. Observe that for a constant emission price, equilibrium investment under imperfect competition differs from that obtained under perfect competition by the terms $\int_{\theta_{\overline{P}}}^{\overline{\theta}} P_q(X_1, \theta) \frac{X_1}{n}$ and $\int_{\theta_{\overline{B}}}^{\theta_P} P_q(X_2, \theta) \frac{X_2}{n} \mathrm{d}F(\theta)$, respectively. This corresponds to the difference between scarcity prices and marginal scarcity profits. Since those terms are negative (and profits concave given our assumptions), investment incentives under imperfect competition

[16] That is expression (3) yields $X_2^* \leq 0$.

[17] That is expressions (3) and (2) yield $X_2^* \geq X_1^*$.

are lower than under perfect competition. That is, in the absence of an explicit market for emission permits (when pollution is, for example, taxed at some fixed level e^0) subsidies for investment (for example, by granting free tax vouchers $A_1 > 0$ and $A_2 > A_1$, respectively) which exactly compensate for those differences would induce optimal investment incentives. Since the emission price is endogenous in our framework, however, we will obtain a different result (compare Theorem 2).

Before we now discuss existence of the market equilibrium, we introduce the following definitions which will simplify the subsequent analysis and allow for a more intuitive discussion of our results:

Definition 1

(i) We denote the impact of increased total investment on total emissions (for fixed e) by $A_1^E := \frac{\partial \Psi_E}{\partial X_1^*} = \left(1 - F(\theta_{\bar{P}})\right) w_1$, observe $A_1^E > 0$. This allows to state the impact of changed emission price e^* on the equilibrium condition Ψ_I as follows $\frac{\partial \Psi_I}{\partial e^*} = A_1 - A_1^E$.

(ii) We denote the impact of increased base load investment on total emissions (for fixed e) by $A_2^E := \frac{\partial \Psi_E}{\partial X_2^*} = \left(1 - F(\theta_{\bar{B}})\right) w_2 - \left(1 - F(\theta_{\underline{P}})\right) w_1$. This allows to state the impact of changed emission price e^* on the equilibrium condition Ψ_{II} as follows $\frac{\partial \Psi_{II}}{\partial e^*} = A_2 - A_1 - A_2^E$. We furthermore denote $w_2^E := \frac{1-F(\theta_{\underline{P}})}{1-F(\theta_{\bar{B}})} w_1$ (which implies $A_2^E > 0$ if and only if $w_2 > w_2^E$) and $w_2^L := \frac{F(\theta_{\bar{P}})-F(\theta_{\underline{P}})}{1-F(\theta_{\bar{B}})} w_1$ (which implies $A_1^E + A_2^E > 0$ if and only if $w_2 > w_2^L$).

(iii) We denote the impact of changed X_1 on the equilibrium condition Ψ_I by $\Psi_{I1} := \frac{\partial \Psi_I}{\partial X_1^*}$, the impact of changed X_2 on the equilibrium condition Ψ_{II} by $\Psi_{II2} := \frac{\partial \Psi_{II}}{\partial X_2^*}$, and the impact of changed e on the equilibrium condition Ψ_E by $\Psi_{\text{Ee}} := \frac{\partial \Psi_E}{\partial e^*}$. Observe that those three expressions are negative.

Observe[18] that $A_1^E = \left(1 - F(\theta_{\bar{P}})\right) w_1$ determines the total amount of additionally necessary permits resulting from an additionally invested unit of total capacity (formally given by the partial derivative of total emission with respect to X_1, i.e., $\frac{\partial \Psi_E}{\partial X_1}$). An increase of the permit price e now has two opposing effects on total investment incentives: On the one hand, investment incentives are reduced by the amount A_1^E; on the other hand, they increase by A_1 due to the increased value of free allocations.

A similar reasoning is obtained for investment incentives in the base load technology. A_2^E determines the total amount of additionally necessary permits resulting from the replacement of one unit of the peak technology with one unit of the base technology (formally given by the partial derivative of total emission with respect to X_2, i.e., $\frac{\partial \Psi_E}{\partial X_2}$). An increase of the permit price e has two opposing effects on total investment incentives: On the one hand, they are reduced by the amount A_2^E; on the other hand, they increase by $(A_2 - A_1)$ due to the increased value of free allocations.

Notice that $A_1^E \geq 0$, whereas A_2^E can also become negative. That is, an increased level of total investment X_1^* always implies additionally necessary emission permits. An increased level of base investment X_2^* does only imply additionally necessary

[18] Notice that the statements of Definition 1 and the subsequent discussion exclusively refer to partial derivatives. In equilibrium, total emissions do not change since they are capped at T.

emission permits if the base technology is "dirtier" than the peak technology. Interestingly, the cutoff point is obtained for $w_2 = w_2^E < w_1$ since an increased level of X_2^* leads to increased emissions for $\theta \in [\theta_{\underline{P}}, \overline{\theta}]$ if $w_2 > w_2$, but also leads to one unit of additional output for the demand levels $\theta \in [\theta_{\overline{B}}, \theta_{\underline{P}}]$.

As already argued, Lemma 1 only characterizes the market equilibrium by establishing necessary conditions. In the subsequent lemma, we now want to establish second-order conditions for the existence of the market equilibrium.

Lemma 2 (Second-order conditions)

- *Lemma 1 characterizes the market equilibrium if*

$$(a)\ \left(A_1 - A_1^E\right) A_1^E - \Psi_{I1}\Psi_{\mathrm{Ee}} < 0, \quad (b)\ \left(A_2 - A_1 - A_2^E\right) A_2^E - \Psi_{II2}\Psi_{\mathrm{Ee}} < 0$$

$$(c)\ \left(\left(A_1 - A_1^E\right) A_1^E - \Psi_{I1}\Psi_{\mathrm{Ee}}\right)\left(\left(A_2 - A_1 - A_2^E\right) A_2^E - \Psi_{II2}\Psi_{\mathrm{Ee}}\right)$$
$$> \left(A_1 - A_1^E\right) A_1^E \left(A_2 - A_1 - A_2^E\right) A_2^E$$

- *If the levels of free allocation satisfy* $\left(A_1 - A_1^E\right) A_1^E \le 0$ *and* $\left(A_2 - A_1 - A_2^E\right) A_2^E \le 0$*, then condition (i) is satisfied.*
- *Define by* $A_1^{\lim}$ *the highest* A_1 *yielding* $\left(A_1 - A_1^E\right) A_1^E - \Psi_{I1}\Psi_{\mathrm{Ee}} \le 0$*; define by* $A_2^{\lim}$ *the highest* A_2 *yielding* $\left(A_2 - A_1^E - A_2^E\right)\left(A_1^E + A_2^E\right) - (\Psi_{I1} + \Psi_{II2})\Psi_{\mathrm{Ee}} \le 0$*. The second-order conditions (i) cannot be satisfied if either* $A_1 \ge A_1^{\lim}$ *or* $A_2 \ge A_1^{\lim}$.

Proof See Appendix B. □

Part (i) of the lemma establishes the standard second-order conditions which guarantee negative semi-definiteness of the Hessian matrix of firms' optimization problems. It allows the usual application of the implicit function theorem in order to conduct an analysis of comparative statics for the equilibrium characterized in Lemma 1. In part (ii), we establish conditions when those second-order conditions are satisfied and part (iii) provides an upper bound on the levels of free allocation such that higher allocations always violate those second-order conditions.

Notice that for the case of a monopolistic or a perfectly competitive market, the second-order conditions established in Lemma 2(i) guarantee that there exists a unique market equilibrium which is characterized by Lemma 1. For the case of oligopoly, when firms behave strategically, also asymmetric equilibria might arise, however. Let us explicitly mention at this point that in this case, our analysis focuses on symmetric investment decisions only.

After having established the market equilibrium, we now determine the impact of changing the parameters of the cap and trade mechanism (A_1, A_2, T) in an analysis of comparative statics. If the second-order conditions specified in Lemma 2 (i) are satisfied, we obtain the following results:

Lemma 3 (Comparative statics of the market equilibrium)

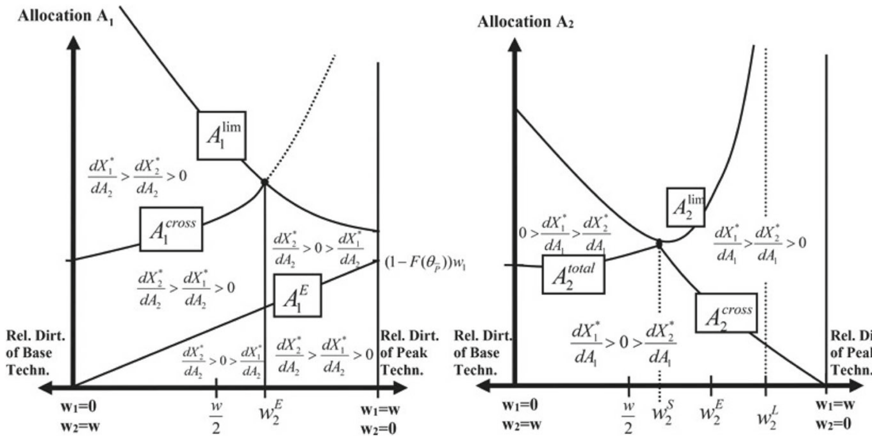

Fig. 3 Results of comparative statics in the degree of free allocation. Left: For the degree of free allocation to the base load technology A_2, right: for the degree of free allocation to the peak load technology A_1. For the case of linear demand, we obtain, left: $A_1^{\text{cross}}(w_1{=}0, w_2{=}w) = \frac{1-F(\theta_{\overline{P}})}{1-F(\theta_{\overline{B}})} F(\theta_{\overline{B}})w$, $A_1^{\text{cross}}(w_1{=}w - w_2^E, w_2{=}w_2^E) = w_2^E$, $A_1^{\text{lim}}(w_1{=}w, w_2{=}0) = \left(1 - F(\theta_{\underline{P}})\right) w$ and right: $A_2^{\text{total}}(w_1{=}0, w_2{=}w) = \left(1 - \frac{1-F(\theta_{\overline{P}})}{1-F(\theta_{\overline{B}})} F(\theta_{\overline{B}})\right) w$, $A_2^{\text{cross}}(w_1{=}w - w_2^S, w_2{=}w_2^S) = w_2^E$

(i) Higher free allocation for the base load technology A_2 always yields higher investment in the base load technology (i.e., $\frac{\mathrm{d}X_2}{\mathrm{d}A_2} > 0$). We furthermore obtain $\frac{\mathrm{d}X_1}{\mathrm{d}A_2} < 0$ if and only if $\left(A_1 - A_1^E\right) A_2^E < 0$. Define A_1^{cross} as the highest A_1 yielding $\left(A_1 - A_1^E\right) A_2^E \leq \Psi_{\text{Ee}}\Psi_{I1} - \left(A_1 - A_1^E\right) A_1^E$, we obtain $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} < \frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}$ if and only if $\left(w_2 > w_2^E\right)$ and $A_1 \in (A_1^{\text{cross}}, A_1^{\text{lim}})$.

(ii) Higher free allocation for the peak load technology A_1 always yields higher investment in the peak load technology (i.e., $\frac{\mathrm{d}X_1}{\mathrm{d}A_1} > \frac{\mathrm{d}X_2}{\mathrm{d}A_1}$). Define by A_2^{total} the highest A_2 which yields $\left(A_2 - A_1^E - A_2^E\right) A_2^E - \Psi_{I12}\Psi_{\text{Ee}} \leq 0$ and by A_2^{cross} the highest A_2 which yields $\left(A_2 - A_1^E - A_2^E\right) A_1^E - \Psi_{I1}\Psi_{\text{Ee}} \leq 0$. There exists a unique w_2^S with $w_2^E < w_2^S \leq w_1$ such that $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} < 0$ if and only if $w_2 > w_2^S$ and $A_2 \in (A_2^{\text{total}}, A_2^{\text{lim}})$. Furthermore, we obtain $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} > 0$ if and only if $w_2 < w_2^S$ and $A_2 \in (A_2^{\text{cross}}, A_2^{\text{lim}})$.

(iii) For a change of the total emission cap T, we obtain $\frac{\mathrm{d}X_1^*}{\mathrm{d}T} > 0$ if and only if $\left(A_1 < A_1^E\right)$, and we furthermore obtain $\frac{\mathrm{d}X_2^*}{\mathrm{d}T} > 0$ if and only if $(A_2 - A_1 < A_2^E)$.

Proof See Appendix C. □

As we establish in the lemma, an increase of the free allocation A_2 in the base load technology always leads to increased base load investment [i.e., $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} > 0$, see point (i)], and an increase of the free allocation A_1 in the peak load technology always leads to increased investment in the peak load technology [i.e., $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} > \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$, see point (ii)]. The impact of such changes on the remaining investment decisions is more ambiguous. In the subsequent paragraphs, we briefly sketch the central trade-offs, and a complete proof is only provided in the appendix, however. First, consider a variation of the free allocation A_2 and determine its impact on the system of equilibrium conditions established in Lemma 1. The total differential yields:[19]

$$\frac{\mathrm{d}\Psi_I}{\mathrm{d}A_2} = \Psi_{I1}\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} + \Psi_{\mathrm{Ie}}\frac{\mathrm{d}e^*}{\mathrm{d}A_2} = 0 \tag{5}$$

$$\frac{\mathrm{d}\Psi_{II}}{\mathrm{d}A_2} = \Psi_{II2}\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} + \Psi_{\mathrm{IIe}}\frac{\mathrm{d}e^*}{\mathrm{d}A_2} + \frac{\partial\Psi_{II}}{\partial A_2} = 0 \tag{6}$$

$$\frac{\mathrm{d}\Psi_E}{\mathrm{d}A_2} = \Psi_{E1}\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} + \Psi_{E2}\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} + \Psi_{\mathrm{Ee}}\frac{\mathrm{d}e^*}{\mathrm{d}A_2} = 0 \tag{7}$$

In order to directly evaluate the impact of the changed emission price $\frac{\mathrm{d}e^*}{\mathrm{d}A_2}$ on the equilibrium conditions for total investment and investment in the base load technology, we solve expression (7) for $\frac{\mathrm{d}e^*}{\mathrm{d}A_2} = \frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} + \frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}$ and plug into expression (5) which yields:

$$\begin{aligned}\frac{\mathrm{d}\Psi_I}{\mathrm{d}A_2} &= \left(\Psi_{I1} + \Psi_{\mathrm{Ie}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right)\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} + \left(\Psi_{\mathrm{Ie}}\frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\right)\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} = 0\\ &\Leftrightarrow \left(-\Psi_{\mathrm{Ee}}\Psi_{I1} + \left(A_1 - A_1^E\right)A_1^E\right)\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} + \left(\left(A_1 - A_1^E\right)A_2^E\right)\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} = 0\end{aligned} \tag{8}$$

Observe that the coefficient on the expression $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}$ determines the total impact of changed X_1^* on the equilibrium condition Ψ_I. This is given by the direct impact (i.e., Ψ_{I1}) and the indirect impact which takes into account the impact of changed X_1^* on the emission price and its feedback on the equilibrium condition Ψ_I (i.e., $\Psi_{\mathrm{Ie}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}} = \frac{1}{-\Psi_{\mathrm{Ee}}}\left(A_1 - A_1^E\right)A_1^E$). Observe that the total impact of changed X_1^* on the equilibrium condition Ψ_I is negative if the second-order conditions established in Lemma 2(i) are to be satisfied. This directly illustrates why $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}$ cannot drop to zero.

Furthermore, observe that the total impact of changed X_2^* on the equilibrium condition Ψ_I is only indirect, since Ψ_I does not directly depend on X_2. That is, we only have to take into account the impact of increased X_2^* on the emission price e^* and its feedback on the equilibrium condition Ψ_I. According to Definition 1, an increase of X_2 leads to an increased equilibrium emission price if $A_2^E > 0$ (i.e., for $w_2 > w_2^E$, we obtain a decreased equilibrium emission price if $A_2^E < 0$, i.e., for $w_2 < w_2^E$). The impact of an increased emission price on the equilibrium condition Ψ_I depends on the degree of free allocation A_1. Whenever $A_1 < A_1^E$ (i.e., $\frac{\partial\Psi_I}{\partial e^*} < 0$, compare Definition 1), an increased emission price leads to a decrease of firms' total investment activity X_1^*. In this case, the reduction of scarcity rents (obtained when total capacity is binding) caused by the increased emission price dominates the increased value associated with the permits granted for free. The reverse holds true for a high level of free allocation, i.e., $A_1 > A_1^E$ where an increased emission price leads to increased total investment

[19] For a better traceability of our computations, we denote the partial derivatives $\frac{\partial\Psi_I}{\partial e^*} = \Psi_{\mathrm{Ie}}$, $\frac{\partial\Psi_{II}}{\partial e^*} = \Psi_{\mathrm{IIe}}$, $\frac{\partial\Psi_E}{\partial X_1^*} = \Psi_{E1}$ and $\frac{\partial\Psi_E}{\partial X_2^*} = \Psi_{E2}$; in a second step, we make use of A_1^E and A_2^E introduced in Definition 1.

X_1^*. Whenever the impact of increased investment X_2^* yields a decreased emission price which is obtained for cleaner base load technologies (for $A_2^E < 0$, i.e., $w_2 < w_2^E$), we obtain the opposite results. In sum, $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} > 0$ if and only if $\left(A_1 - A_1^E\right) A_2^E > 0$, as stated in the theorem.

Finally, expression (8) also provides the intuition when $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} \geq \frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}$ is obtained (i.e., also investment in the peak load technology increases). To this end, observe that $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} = \frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}$ if and only if in expression (8), the total impact of changed X_1^* is precisely of the same size as the total impact of changed X_2^*, but of opposite sign. As shown in the theorem, this is only obtained if the increase of investment in the base technology leads to an increase of the emission price (for $A_2^E > 0$, i.e., $w_2 > w_2^E$) and if this increase has a sufficiently positive impact on the equilibrium condition Ψ_I, i.e., for allocation A_1 sufficiently big (for $A_1 > A_1^{\text{cross}} > A_1^E$). All those results are illustrated in the left graph of Fig. 3.

Likewise, we can analyze the impact of changing A_1, as established in Theorem 3(ii). Analogous to expressions (19) (20) and (21), we can determine the total derivative and solve for $\frac{de^*}{\mathrm{d}A_1}$. After plugging in, we obtain for $\frac{\mathrm{d}\Psi_I}{\mathrm{d}A_1} + \frac{\mathrm{d}\Psi_{II}}{\mathrm{d}A_1}$ (observe $\frac{\partial \Psi_I}{\partial A_1} = -\frac{\partial \Psi_{II}}{\partial A_1} = e$):

$$\begin{aligned}
\frac{\mathrm{d}\Psi_I}{\mathrm{d}A_1} + \frac{\mathrm{d}\Psi_{II}}{\mathrm{d}A_1} &= \left(\Psi_{I1} + \Psi_{\mathrm{Ie}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}} + \Psi_{\mathrm{IIe}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right)\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} \\
&\quad + \left(\Psi_{II2} + \Psi_{\mathrm{IIe}}\frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}} + \Psi_{\mathrm{Ie}}\frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\right)\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} = 0 \\
&\Leftrightarrow \left(-\Psi_{\mathrm{Ee}}\Psi_{I1} + \left(A_2 - A_1^E - A_2^E\right)A_1^E\right)\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} \\
&\quad + \left(-\Psi_{\mathrm{Ee}}\Psi_{II2} + \left(A_2 - A_1^E - A_2^E\right)A_2^E\right)\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} = 0 \qquad (9)
\end{aligned}$$

Analogous to above the coefficients on the expressions $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ and $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$, determine the impact of changed investment X_1^* or X_2^* on both equilibrium conditions. The sum of both coefficients is strictly negative if second-order conditions are not to be violated [compare Lemma 2(iii)]. This directly illustrates why $\frac{\mathrm{d}X_2}{\mathrm{d}A_1}$ cannot reach the level of $\frac{\mathrm{d}X_1}{\mathrm{d}A_1}$. (In other words, increased free allocation A_1 cannot leave investment in the peak load technology unchanged.)

Furthermore, as we show, for small A_2 both coefficients are negative (thus, $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$ and $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ have opposite signs); since $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} > \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$, this implies $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} < 0$. Observe that the coefficient of the expression $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$ is increasing in A_2, and the coefficient of expression $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$ is increasing in A_2 if $A_2^E > 0$ (i.e., $w_2 > w_2^E$). That is, for A_2 high enough the coefficients become nonnegative, leading to altered monotonicity behavior. As we show in the theorem, we can establish a relative level of dirtiness w_2^S (with $w_2^S \geq w_2^E$ and $w_2^S = w_2^E$ in the case of linear demand) which separates the cases when either of the coefficients becomes zero for higher levels of A_2. (Remember the sum of both coefficients has to be negative in order to satisfy the second-order conditions, see above.) Whenever the coefficient of $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$ is equal to zero, expression (9) directly implies $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} = 0$ and vice versa, as stated in the theorem.

Finally, in Theorem 3(iii) we provide the results of comparative statics with respect to the parameter T. For an intuition of those results, observe first of all that an increase of the total emission cap T leads to a reduction of the equilibrium permit price. This in turn induces increased total investment X_1^* if [similar to the intuition for part (i)] the increase of scarcity rents (which is obtained due to lower emission price) dominates the decreased value of the emission permits granted for free, i.e., $A_1 < A_1^E$. The opposite result is obtained for $A_1 > A_1^E$. Similarly, the reduced emission price induces increased investment in the base load technology X_2^* if the total impact of reduced emission price on the base load investment condition is negative, i.e., if and only if $A_2 < A_1 + A_2^E$ (i.e., $\Psi_{\mathrm{IIe}} < 0$). If we denote total emissions which obtain in the absence of any environmental policy by $\bar{T}$, then lowering the cap on total emissions T below $\bar{T}$ corresponds to the introduction of a cap and trade mechanism.

Discussion and policy implications: Let us conclude this section by briefly discussing policy implications of our results. Many existing and planned cap and trade systems worldwide do grant free allocations; in most cases, we observe updating which also considers market entry of new capacities or retirements (IEA 2020). In those cases, investment in production facilities receives free allocations based on their expected technology-specific needs (see IEA 2020; ICAP 2020). Moreover, however, updating in many cases additionally includes components which grant free allocations based on past output choices. This induces additional distortions, both on investment and output choices, since firms' decisions anticipate free allocations received in later compliance periods. Unlike previous contributions (compare, for example, Böhringer and Lange 2005; Mackenzie et al. 2008, or Damon et al. 2019), we are able to explicitly analyze the impact of free allocations on investment incentives. To keep our analysis tractable, however, we refrain from considering several compliance periods in a dynamic setup. For an interpretation of our analytical results in the context of cap and trade systems which grant free allocations based on past output and additionally grant free allocations to newly built facilities, a careful calibration of the parameters (A_1, A_2) of our setup is required to reasonably approximate real-world incentives for our model.[20] An interesting reference case is obtained when free allocations are granted such as to cover all emissions caused by a certain installation. In our framework, we define this by A_1^{full} and A_2^{full}. Observe $A_1^{\mathrm{full}} \in [1 - (F(\theta_{\overline{P}}))w_1, 1 - (F(\theta_{\underline{P}}))w_1]$ and $A_2^{\mathrm{full}} \in [1 - (F(\theta_{\overline{B}}))w_2, w_2]$.

In the context of our formal results, for an allocation scheme $(A_1^{\mathrm{full}}, A_2^{\mathrm{full}})$ we obtain increased total investment and increased base load investment when introducing the

[20] An immediate calibration of the parameters (A_1, A_2) and $(A_1^{\mathrm{full}}, A_2^{\mathrm{full}})$ of our setup is possible if free allocations are explicitly granted contingent on installed capacities. An example for such allocation rule is given, for example, by the two initial phases of the EU ETS (2005–2012) for the electricity sector. In this case, free allocations have been determined by a technology-specific emission factor which measures average emissions per unit of electricity produced (0.365 tCO2/MWh for gaseous fuels and 0.750 tCO2/MWh for solid and liquid fuels) multiplied by a pre-established technology-specific average usage. For open cycle gas turbines, e.g., in Germany, the average usage has been established at 0.11 (i.e., 1000 h per year), for coal and combined cycle gas turbines, it was given by 0.86 (i.e., 7500 h per year), and for lignite plants, it was given by 0.94 (i.e., 8250 h per year), see Appendices 3 and 4 of German-Parliament (2007).

trading system (i.e., the emission cap is lowered below $\bar{T}$ in our framework). To see this, first, observe that $A_1^{\text{full}} > A_1^E$ which according to Lemma 3 (iii) leads to an increase of X_1^*. Second, observe that $A_2^{\text{full}} - A_1^{\text{full}} > A_2^E$ (since $A_2^{\text{full}} > (1 - F(\theta_{\overline{B}}))w_2$ and $A_1^{\text{full}} < (1 - F(\theta_{\underline{P}}))w_1$) which according to Lemma 3 (iii) leads to increased investment in the base load technology.

To apply the findings of Lemma 3(i), consider the example of an electricity market with lignite or coal-fired plants as a representative base load technology and open cycle gas turbines as a representative peak load technology. Since open cycle gas turbines have lower emission factors, we obtain $w_2 > w_1$, which directly implies $w_2 > w_2^E$ (compare Definition 1). Since furthermore $A_1^{\text{full}} > A_1^E$, as established above, we can directly conclude that an increase (decrease) of the free allocation A_2 not only would yield increased base load investment but also an increased (decreased) emission price and increased (decreased) total investment.

After having analyzed the market equilibrium which is obtained in the presence of an emission trading system and derived its properties of comparative statics, we now proceed to the main part of this article and analyze the optimal design of a cap and trade mechanism.

4 Optimal market design under perfect competition

In this section, we determine the optimal cap and trade mechanism. We first determine the first best solution as a benchmark, which is obtained for the case of a perfectly competitive market when a regulator can freely choose all parameters (A_1, A_2, T) of the cap and trade mechanism (see Theorem 1). We then analyze several market imperfections and solve for the corresponding second best solutions. We first determine the optimal cap and trade mechanism which should be chosen for an imperfectly competitive market (see Sect. 5). We then analyze the case when competition authorities cannot freely choose all parameters (A_1, A_2, T) of the cap and trade mechanism but only a subset of them (see Sect. 6). In order to answer all those questions, we first determine total welfare generated in a market with some cap and trade mechanism (A_1, A_2, T):

$$
\begin{aligned}
W(A_1, A_2, T) = & \int_{\underline{\theta}}^{\theta_{\overline{B}}} \left[\int_0^{Q^*} (P(Y,\theta) - c_2)\, Y \mathrm{d}Y \right] dF(\theta) + \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} \left[\int_0^{X_2^*} \big(P(X_2^*,\theta) - c_2\big)\, Y \mathrm{d}Y \right] dF(\theta) \\
& \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \left[\int_0^{Q^*} (P(Y,\theta) - c_1)\, Y \mathrm{d}Y \right] dF(\theta) + \int_{\theta_{\overline{P}}}^{\overline{\theta}} \left[\int_0^{X_1^*} \big(P(X_1^*,\theta) - c_1\big)\, Y \mathrm{d}Y \right] dF(\theta) \\
& - \int_{\theta_{\underline{P}}}^{\overline{\theta}} (c_1 - c_2) X_2^* dF(\theta) - k_2 X_2^* - k_1 (X_1^* - X_2^*) - D(T). \qquad (10)
\end{aligned}
$$

Observe that welfare does not directly depend on the parameters (A_1, A_2, T) chosen for the cap and trade mechanism, but only indirectly through the implied investment and production decisions X_1^*, X_2^* and Q^*. In order to maintain presentability of the results, we relegated all computations to appendix and directly characterize the optimal cap and trade mechanism in the subsequent lemma.

Lemma 4 *The optimal cap and trade mechanism solves the following conditions:*

$$(i)\;\; W_{A_1} := \frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\Omega_I + \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\Omega_{II} = 0$$
$$(ii)\;\; W_{A_2} := \frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}\Omega_I + \frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}\Omega_{II} = 0$$
$$(iii)\;\; W_T := \frac{\mathrm{d}X_1^*}{\mathrm{d}T}\Omega_I + \frac{\mathrm{d}X_2^*}{\mathrm{d}T}\Omega_{II} - D_T(T) + e^* + \frac{\Delta}{n} = 0.$$

The expressions Ω_I and Ω_{II} determine the total impact of changed X_1^* and changed X_2^*, respectively, on total welfare. They are defined as follows:

$$\Omega_I := \int_{\theta_{\overline{P}}}^{\overline{\theta}} \frac{-P_q X_1^*}{n}\mathrm{d}F(\theta) - A_1 e^* - \frac{\Delta}{n}A_1^E$$
$$\Omega_{II} := \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} \frac{-P_q X_2^*}{n}\mathrm{d}F(\theta) - (A_2 - A_1)e^* - \frac{\Delta}{n}A_2^E.$$

The term $\frac{\Delta}{n} := \frac{\int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{\mathrm{d}Q^*}{de^*}\left(-P_q\frac{Q^*}{n}\right)\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{\mathrm{d}Q^*}{de^*}\left(-P_q\frac{Q^*}{n}\right)\mathrm{d}F(\theta)}{\int_{\underline{\theta}}^{\theta_{\overline{B}}}\left(\frac{\mathrm{d}Q^*}{de^*}w_2\right)\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}}\left(\frac{\mathrm{d}Q^*}{de^*}w_1\right)\mathrm{d}F(\theta)} > 0$ *determines the impact of changed emissions on welfare for those spot markets where investment is not binding.*

Proof See Appendix D. □

We now provide some intuition for the conditions which characterize an optimal cap and trade mechanism. We first consider the optimal choice of the free allocation to the peak load technology given by A_1. Observe that the optimality conditions (i) and (ii) express the impact of changed free allocation on total welfare exclusively through the channel of changed investment in the base load technology X_2^* and changed total investment X_1^*. The total impact of changed investment on total welfare is denoted by Ω_I and Ω_{II}, and this total impact can be broken down into three components corresponding to the three summands of Ω_I and Ω_{II}, respectively.

First, observe that at all those spot markets where total investment is binding (i.e., for $\theta \in [\theta_{\overline{B}}, \theta_{\underline{P}}]$ and $\theta \in [\theta_{\overline{P}}, \overline{\theta}]$, respectively), imperfectly competitive investment behavior induces too low investment incentives, and an increase of investment X_2^* or X_1^* leads to increased welfare given by the markup $-P_q\frac{X}{n}$. Second, free allocation $A_1 > 0$ or $(A_2 - A_1) > 0$ induces too high investment incentives, and thus, an increase of investment would lead to a reduction of welfare given by the monetary value of the free allocation (i.e., $A_1 e^*$ and $(A_2 - A_1)e^*$). Notice that in a world with exogenously fixed emission price e^*, the optimal level of free allocation should be chosen such as to balance those two effects.[21] Since the emission price is endogenous in our analysis, an additional term is obtained. An increase of investment $\mathrm{d}X_1^*$ or $\mathrm{d}X_2^*$ leads to increased emissions of $\mathrm{d}X_1^* A_1^E$ and $\mathrm{d}X_2^* A_2^E$ at those spot markets where investment is binding. Since total emissions are capped by T, however, this necessarily has to imply an equivalent reduction of emissions at those spot markets where investment is not binding (i.e., for $\theta \in [\underline{\theta}, \theta_{\overline{B}}]$ or $\theta \in [\theta_{\underline{P}}, \theta_{\overline{P}}]$). Since production decisions are also imperfectly competitive, a reduction of output leads to reduced welfare generated at those spot markets. This impact is quantified by the term $\frac{\Delta}{n}$ defined in the lemma. That is, taking into account the endogenous nature of the emission price leads to a lower degree of optimal free allocation A_1 than suggested by an analysis with exogenously fixed emission price.

The impact of a changed emission cap T on total welfare has a similar structure than the impact of changed free allocations. Analogous to above, a changed emission cap leads to changed investment incentives, and the impact of changed investment incentives on welfare is given by the terms Ω_I and Ω_{II} which have already been discussed above. As we will see later on in Theorems 1 and 2, if the levels of free allocation are chosen optimally such as to obtain $\Omega_I = \Omega_{II} = 0$, those terms will not be relevant for the optimal choice of the emission cap. If the levels of free allocation are not chosen optimally, however, they have to be considered when determining the optimal level of the emission cap T (compare Theorems 3, 4, 5 and 6).

Apart from having an impact on investment incentives, a changed emission cap T leads to changed welfare also through several other channels. First, most apparently an increased emission cap leads to increased emissions which reduce welfare by the marginal social cost of pollution D_T. Second, observe that on the other hand, an increased emission cap leads to a welfare increase since it implies a reduced emission price which allows for increased output. The welfare increase at each spot market is given by the changed output multiplied by the difference between marginal cost as perceived by the firms and true marginal cost, i.e., $dQ(w_i e^*)$, for $i = 1, 2$. Put differently, however, this corresponds to the changed pollution at each spot market multiplied by the emission price e^*, and the change in welfare at all spot markets then is simply given by the total change of emissions multiplied by the emission price, i.e., $\mathrm{d}Te^*$. As we will see in the subsequent Theorem 1, for a perfectly competitive market the optimal cap and trade mechanism only balances those two effects and equates the marginal social cost of pollution to the emission price (i.e., $e^* = D_T$).[22] Third, observe that an increased emission cap T leads to a reduced emission price. This allows to reduce the welfare loss obtained due to imperfect competition at those spot markets where investment is not binding and output too low. Notice that the impact of changed emissions on welfare at those spot markets where investment is not binding has already been discussed above, it is given by $\frac{\Delta}{n}$.

[21] That is, the monetary subsidy $A_1 e^*$, for example, should then equate to the integral of the markups over all relevant spot markets. The intuition for this result in some sense parallels the quite well known insight obtained for a simple static model where a monopolist can be induced to produce first best output if he obtains a subsidy corresponding to his markup.

[22] This parallels the fundamental trade-off obtained in a simple static model where a Pigou tax should just equal to the marginal social damage of pollution.

Based on the findings of Lemma 4 as the first best benchmark, we can now directly establish the optimal cap and trade mechanism which is obtained for a perfectly competitive market

Theorem 1 (Optimal market design, first best benchmark) *Under perfect competition, the optimal market design satisfies*

$$(i)\;\; A_1^* = 0 \qquad (ii)\;\; A_2^* = 0 \qquad (iii)\;\; T^*: \quad e^* = D_T(T).$$

Proof See Appendix E. □

Discussion and policy implications: The theorem demonstrates that in a competitive market (i.e., $n \to \infty$), where all aspects of the cap and trade system (here: parameters A_1, A_2 and T) can be freely chosen by authorities, full auctioning is unambiguously optimal (i.e., no free allocations should be granted) and the emission target T should be set such that the equilibrium permit price equals marginal social cost of environmental damage. That is, as already discussed above, the optimal cap and trade mechanism balances welfare losses due to foregone production at all spot markets given by e^* with the marginal social cost of pollution given by D_T. In case of the EU-ETS, for example, in the early phases I and II (2005–2012) free allocations have been granted to all sectors including the electricity sector. One important reason has apparently been of political nature allowing for industry support when introducing the cap and trade system (see, e.g., Convery 2009). For the electricity sector, this has been changed since the start of phase III (2013), since then no free allocations have granted to electricity producers[23]. This seems to be perfectly in line with the results obtained in Theorem 1. Observe, however, that the theorem requires all aspects of the cap and trade mechanism to be set optimally, including the total amount T of permits issued by authorities. Likely, also this total amount T of emissions is subject to political constraints which require the emission price e to remain within certain limits (likely below the socially ideal level). Let us mention already at this point, however, that we are able to derive the optimal design also in case of an exogenously fixed overall emissions cap T which is not optimally chosen. As we show later on in Theorem 3, also in this case it is optimal to fully auction permits without granting any free allocations. Thus, the elimination of free allocations in the electricity sector in the EU-ETS since 2013 is perfectly in line with our results, provided we can assume that producing firms interacts in a perfectly competitive way (for the relaxation of this assumption, see the subsequent section).

In the subsequent two sections, we now consider market imperfections which make an attainment of the first best outcome impossible. First, we analyze the case of imperfectly competitive markets (see Sect. 5). Apart from imperfect competition, another source of market imperfection arises when authorities cannot freely choose all parameters (A_1, A_2, T) of the cap and trade mechanism, but only a subset. Such situations arise, for example, when the level of free allocation for (some of) the different technologies or the total emission cap is exogenously fixed due to political arrangements or lobbing of firms and the competition authority can only determine the remaining parameters (see Sect. 6).

[23] There are exceptions for those member states with a GDP below 60% of the average, see IEA (2020).

5 Optimal market design under imperfect competition

After having determined the first best benchmark (Theorem 1), we now analyze the case of an imperfectly competitive market. Under imperfect competition, investment incentives are generally too low (compare Sect. 3). One of the usually proposed measures to overcome problems of under-investment is given by capacity payments which are assigned in capacity markets.[24] Notice that the fundamental impact of a capacity payment is equivalent to that of free allocations in our framework. That is, in our framework it is equivalent if a monetary payment S_t or free allocations of value $A_t e^*$ for $t = 1, 2$ are granted to a firm per unit of investment. In total, those measures lead to a reduction of marginal cost of investment by $A_t e^* + S_t$. Thus, the optimal design of both instruments is closely related. In the present section, we want to shed light on this interdependence and determine the optimal design of capacity payments and of the cap and trade system.

Theorem 2 (Optimal Market Design under Imperfect Competition) *Under imperfect competition, the optimal market design satisfies*

$$(i) \quad A_1^* e^* + S_1^* = \int_{\theta_{\overline{P}}}^{\overline{\theta}} \left(\frac{-P_q X_1^*}{n} \right) \mathrm{d}F(\theta) - \frac{\Delta}{n} A_1^E$$

$$(ii) \quad A_2^* e^* + S_2^* = \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} \left(\frac{-P_q X_2^*}{n} \right) \mathrm{d}F(\theta) + \int_{\theta_{\overline{P}}}^{\overline{\theta}} \left(\frac{-P_q X_1^*}{n} \right) \mathrm{d}F(\theta) - \frac{\Delta}{n} \left(A_1^E + A_2^E \right)$$

$$(iii) \quad T^* : \quad e^* = D_T(T) - \frac{\Delta}{n}.$$

Now, assume that $P_{q\theta} = 0$. We then obtain $A_1^* e^* + S_1^* > 0$. For $w_2 \leq w_2^E$, we obtain $A_2^* e^* + S_2^* > A_1^* e^* + S_1^*$; for $w_2 > w_2^E$, we can obtain $A_2^* e^* + S_2^* = 0$.

Proof See Appendix E. □

The optimal levels of total free allocations and capacity payments $A_t^* e^* + S_t^*$ under imperfect competition are thus typically different from zero, a striking difference to the result obtained under perfect competition (see Theorem 1). The fundamental reason why this is the case follows directly from the insights provided by Lemma 1 and the subsequent discussion of the results: Imperfectly competitive firms not only exercise market power at the spot markets, but also choose their capacity to optimally benefit from scarcity prices, implying reduced investment incentives.

As already discussed in the text following Lemma 1 (compare the last paragraph which discusses Lemma 1), for an exogenously fixed price for pollution (e.g., a Pigouvian tax at some fixed level e^*) optimal investment incentives are obtained by subsidizing investment such as to precisely compensate for the difference between scarcity rents and marginal scarcity profits. To stick as close as possible to our notation, such subsidy could be made by assigning the amounts A_1 and A_2 of free tax

[24] For a survey on those capacity markets, compare, for example, Cramton and Stoft (2008), Cramton and Ockenfels (2012), or Fabra (2018).

vouchers to each unit invested in either of the technologies. Equivalently, such subsidy could also be realized by appropriately chosen levels of capacity payments S_t, for $t = 1, 2$. Formally, for an exogenous emission price the total amount of subsidies $A_t e^* + S_t$ is given by expressions (i) and (ii) of Theorem 2 by setting $\Delta = 0$. (Notice that for exogenously fixed permit price, we have $\Delta = 0$.) Those would be the levels of capacity payments granted by an optimal capacity market which disregards the endogenous nature of the emission permit price.

However, if the emission permit price is not exogenous but changes with firms' investment incentives (indeed $\Delta > 0$), those levels of capacity payments are not optimal, as we show. Positive free allocation leads to increased investment incentives, which (through an increased emission price) can lead to reduced output (and thus pollution) at those spot market where investment is not binding. The terms including the expression $\frac{\Delta}{n}$ take this welfare loss into account. This leads to a reduced level of the optimal degree of free allocation. As we show in the theorem, under imperfect competition the degree of free allocation for the peak load technology is always positive, i.e., $A_1^* e^* + S_1^* > 0$. For the optimal allocation for the base load technology, ambiguous results are obtained. If the base load technology is less emission intensive than the peak load technology (i.e., $w_2 \leq w_1$), increased investment in the base load technology leads to reduced emissions and thus allows for more output at spot markets where investment is not binding. As we show, this always implies that the base load technology should receive more total subsidies than the peak load technology, i.e., $A_2^* e^* + S_2^* > A_1^* e^* + S_1^*$. On the other hand, if the base load technology is more emission intensive than the peak load technology ($w_2 > w_1$, i.e., an increase of base load investment leads to increased emission price), then it might be optimal to grant less, i.e., $A_2^* e^* + S_2^* < A_1^* e^* + S_1^*$ or even $A_2^* e^* + S_2^* < 0$, as we show.

Finally, consider the optimal choice of the total emission cap T for the case of imperfect competition. A brief look at the optimality condition (iii) established in Lemma 4 reveals that the impact of a changed emission cap on investment decisions can be neglected since the levels of free allocation are determined optimally (such as to obtain $\Omega_I = \Omega_{II} = 0$). What matters, however, is the fact that an increased emission cap leads to a reduced emission price which in turn allows to reduce the welfare loss induced by imperfectly competitive production decisions at those spot markets where investment is not binding (given by $\frac{\Delta}{n}$). As a result, the optimal cap on total emissions is chosen such as to yield an emission price below the marginal social cost of pollution.

Discussion and policy implications: Based on our results, we can draw several important conclusions for the debate on the design of cap and trade mechanisms. First of all, policy makers should be very aware of the close relationship of capacity markets and cap and trade systems which update based on installed production facilities. Most importantly, as we show, the optimal capacity payments have to be fundamentally different for markets with exogenous emission price (e.g., for an emission tax) than for markets with endogenous price (as for a cap and trade system). Disregarding those aspects when designing capacity payments for a market with a cap and trade system will lead to flawed market outcomes.

Furthermore, policy makers might be able to exploit this strongly interdependent nature of capacity payments and free allocations, especially if there are rigidities present in either of the instruments. On the one hand, those rigidities might be present for the cap and trade system when free allocations have to be granted to enhance political support (compare our discussion in Sect. 1). In this case, optimal capacity payments should be designed such as to grant a total level of subsidies specified by our optimality conditions. On the other hand, the introduction of capacity markets to address problems of under-investment might be desirable, but politically infeasible for certain markets or industries. In such case, appropriately designed levels of free allocations might at least partially overcome this inefficiency. Our results established in Theorem 2 characterize the optimal policy choice for all those cases.

6 Optimal design of a partially constrained cap and trade mechanism

In Theorems 1 and 2, we determined the optimal design of a cap and trade mechanism when all its parameters (A_1, A_2, T) can be freely chosen by the competition authority. We first analyzed the case of a perfectly competitive market, which yields the first best benchmark (Theorem 1) and then the case of imperfect competition (Theorem 2). Another source of market imperfection, apart from imperfect competition, arises when the competition authorities cannot freely choose all parameters (A_1, A_2, T) of the cap and trade mechanism. Such rigidities might be due to political constraints and arrangements or due to lobbing of firms. As already discussed extensively in the introduction of this article, free allocations have been key to guarantee the political support necessary to introduce cap and trade systems, compare Convery (2009), Tietenberg (2006), Bovenberg et al. (2008) or, for example, Grubb and Neuhoff (2006)[25]. It is the purpose of the present section to analyze how a competition authority should optimally design a cap and trade mechanism if it can determine only a subset of the parameters of the cap and trade mechanism, whereas the remaining parameters are exogenously fixed due to the above discussed problems.

Theorem 3 determines the optimal degree of free allocations for the case of exogenously fixed level of the total emission cap T. In Theorems 4, 5 and 6, we determine the optimal degree of free allocation to the remaining technologies and the corresponding level of the optimal total emission cap T. Observe that our results obtained in Lemma 4 in principle would allow for a detailed analysis of those questions both for the cases of perfect and imperfect competition. In order to limit the notational burden in the present paper, we restrict ourselves to the case of perfect competition, however. In this case, the optimality conditions determined in Lemma 4 read as follows

$$W_{A_1} := \frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}(-A_1)\,e^* + \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}(A_1 - A_2)\,e^* = 0 \tag{11}$$

$$W_{A_2} := \frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}(-A_1)\,e^* + \frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}(A_1 - A_2)\,e^* = 0 \tag{12}$$

[25] "Due in part to the sheer scale of the EU ETS, governments are subject to intense lobbying relating to the distributional impact of the scheme, and are constrained by this and by concerns about the impact of the system on industrial competitiveness. Few academics understand the real difficulties that policy-makers face when confronted with economically important industries claiming that government policy risks putting them at a disadvantage relative to competitors."

$$W_T := \frac{\mathrm{d}X_1^*}{\mathrm{d}T}\,(-A_1)\,e^* + \frac{\mathrm{d}X_2^*}{\mathrm{d}T}\,(A_1 - A_2)\,e^* - D_T(T) + e^* = 0. \tag{13}$$

We first analyze the case of an exogenously fixed level of the cap on total emissions T, and an example might be a situation where politicians are willing to introduce a cap and trade mechanism, but are reluctant to induce too severe (even though optimal from an overall welfare point of view) distortions on the economy. The above optimality conditions directly reveal that in a perfectly competitive market, no free allocations should be granted to firms, independently of the level of the emission cap.[26] This is summarized in Theorem 3.

Theorem 3 (Optimal design for fixed emission cap T) *For any exogenously fixed total emission cap T, it is optimal to choose the levels of free allocation $A_1^* = A_2^* = 0$.*

That is, the result obtained in the first best benchmark (Theorem 1) where no free allocation has been found to be optimal is also obtained if the total emission cap is not set at an optimal level. This seems to be perfectly in line with the abandonment of free allocations for the electricity sector in most member states in the context of the EU-ETS since 2013. For a more detailed discussion of the policy implications, see the discussion following Theorem 2. Observe that the reverse does not hold as we show in the subsequent theorem, however.

Theorem 4 (Optimal design for fixed allocations A_1 and A_2) *Suppose the initial allocations A_1 and A_2 are fixed exogenously. Define*

$$\Gamma_0(A_1, A_2) := (A_1 - A_1^E)A_1\Psi_{I1} + (A_2 - A_1 - A_2^E)(A_2 - A_1)\Psi_{II2}. \tag{14}$$

The optimal emission cap T^* has to be set such as to satisfy $e^* = D_T(T^*)$ for $\Gamma_0(A_1, A_2) = 0$, $e^* > D_T(T^*)$ for $\Gamma_0(A_1, A_2) > 0$, and $e^* < D_T(T^*)$ for $\Gamma_0(A_1, A_2) < 0$.

Proof See Appendix F. □

That is, for levels of free allocation A_1, A_2 which are not set optimally the optimal cap on emissions T typically does not implement an emission price e^* equal to the social cost of pollution D_T. To get an intuition for the result, note first that the cap T on total emissions governs the price for emission certificates e^* which in turn influences both investment decisions and unconstrained production decisions at those spot markets where investment is not binding. Optimal production decisions are induced by an emission price equal to the social cost of pollution. This is only overall optimal in case of optimal investment incentives.

Now, first observe that in case of positive free allocations (as considered in the theorem), investment incentives are distorted, however. That is, for $A_1 > 0$ investment incentives in the peak load technology are too high; for $A_2 > A_1$ ($A_2 < A_1$), investment incentives in the base load technology are too high (low). A distortion of the emission price can then be suited to at least partially adjust investment incentives.

[26] The results of Theorem 3 for the case of imperfect competition are obtained analogously, and the optimal levels of free allocation are given by conditions (i) and (ii) established in Theorem 2.

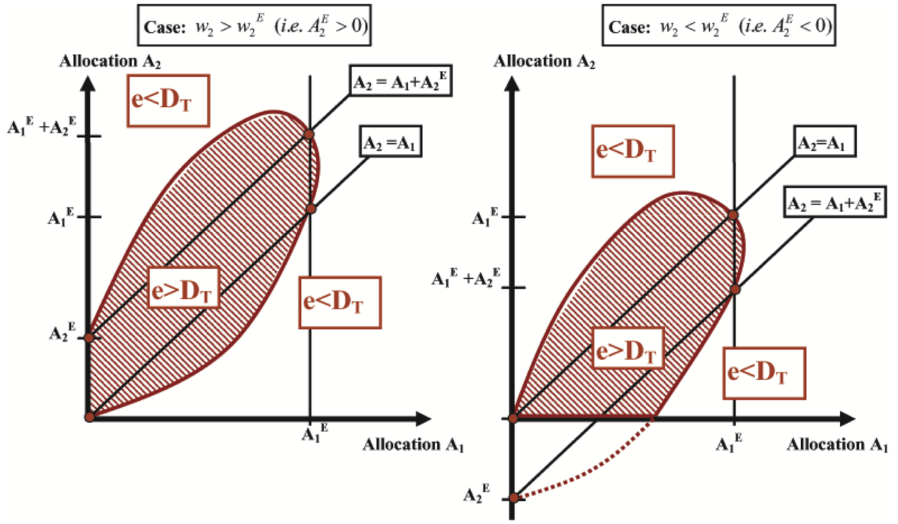

Fig. 4 Choosing the optimal T^* for exogenously fixed initial allocations A_1 and A_2. Left: for relatively dirty base technology, i.e., $w_2 > w_2^E$, right: for relatively clean base technology, i.e., $w_2 < w_2^E$

Second, observe that the impact of a changed emission cap T on investment incentives already has been derived in Lemma 3 (iii) and was discussed in the subsequent text. As established there, a higher emission cap T (implying a lower emission price e^*) leads to increased investment in the peak load technology X_1^* if and only if $A_1 < A_1^E$, and it leads to increased investment in the base load technology X_2^* if and only if $A_2 < A_1 + A_2^E$.

Intuitively, Theorem 4 formally joins those two effects; that is, whenever the levels of free allocation A_1, A_2 are such as to induce over investment, the total cap on emissions should be set such as to induce an emission price which leads to a reduction of investment incentives and vice versa. All those findings are illustrated graphically in Fig. 4.

Consider the case $A_2 = A_1 > 0$, where all technologies get the same amount of free allocations (the 45-degree line of Fig. 4). In the light of the above discussion, this implies first of all that investment incentives in the base load technology are undistorted (since $A_2 = A_1$) and investment incentives in the peak load technology are too high. For $A_1 < A_1^E$, investment incentives are reduced for a higher emission price; for $A_1 > A_1^E$, they are reduced for a lower emission price. Next, consider the case $A_2 = A_1 + A_2^E$. In this case, a changed emission price e^* has no impact on investment in the base load technology, and analogous to above, the optimal cap T is thus designed exclusively such as to reduce the too high investment incentives in the peak load technology (i.e., for $A_1 < A_1^E$, we have $e^* > D_T$ and vice versa).[27]

[27] Observe that an analogous reasoning is obtained for the case $A_1 = 0$ when only investment incentives in the base load technology are distorted and the case $A_1 = A_1^E$ when a changed emission price has no impact on investment in the peak load technology and only distortions of base load investment are to be adjusted by the total emission cap.

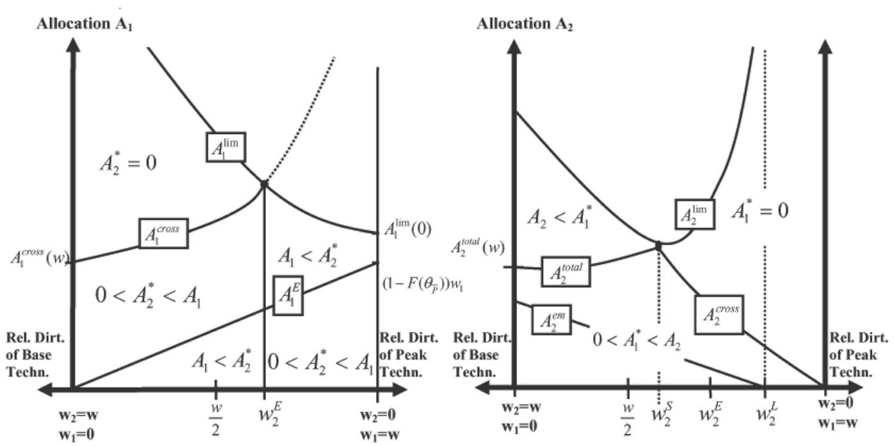

Fig. 5 Left: Choosing the optimal A_2^* for exogenously fixed initial allocation A_1. Right: Choosing the optimal A_1^* for exogenously fixed initial allocation A_2

Discussion and policy implications: We conclude the discussion of Theorem 4 by applying our findings to the above discussed policy of full allocation $(A_1^{\text{full}}, A_2^{\text{full}})$ as already introduced at the end of Sect. 3. Remember that we derived the following properties for the levels of full allocation: $A_1^{\text{full}} > A_1^E$ and $A_2^{\text{full}} > A_1^{\text{full}} + A_2^E$. As already discussed, if we consider either lignite or coal-fired plants as the representative base load technology and open cycle gas turbines as the representative peak load technology, we also obtain $A_2^{\text{full}} > A_1^{\text{full}}$. For our framework, we thus obtain that the optimal cap on total emissions has to be set such that the equilibrium permit price is lower than the social cost of pollution, i.e., $e^* < D_T$.

In the subsequent Theorem 5, we consider the case that only the allocation for the peak load technology A_1 is exogenously fixed, and allocation for the base load technology A_2 and the total emission cap T can be determined optimally, however.

Theorem 5 (Optimal design for fixed allocation A_1) *Suppose the allocation for the peak technology A_1 is exogenously fixed. The optimal allocation for the base technology then solves $A_2^* = \frac{\mathrm{d}X_2^*/\mathrm{d}A_2 - \mathrm{d}X_1^*/\mathrm{d}A_2}{\mathrm{d}X_2^*/\mathrm{d}A_2} A_1$. More specifically, we obtain (see left graph of Fig. 5)*

$$\begin{cases} A_2^* = 0 & if \;\; (A_1^{\text{cross}} \leq A_1 < A_1^{\lim}) \\ 0 < A_2^* < A_1 & if \;\; \left((0 < A_1 < A_1^E)\,\&\,(w_2 < w_2^E)\right) OR \left((A_1^E < A_1 < A_1^{\text{cross}})\,\&\,(w_2 > w_2^E)\right) \\ A_1 < A_2^* & if \;\; \left((A_1^E < A_1 < A_1^{\text{cross}})\,\&\,(w_2 < w_2^E)\right) OR \left((0 < A_1 < A_1^E)\,\&\,(w_2 > w_2^E)\right). \end{cases}$$

The optimal cap T^* is such that $e^* > D_T$ if $A_1 < A_1^E$ and $e^* < D_T$ if $A_1 > A_1^E$.

Proof See Appendix G. □

Observe that the optimality condition for A_2 as stated in the theorem is obtained directly by rearranging expression (12). To derive the properties of the optimal degree of allocation for the base load technology A_2^* as stated in the theorem, we can now make use of the properties of comparative statics derived in Lemma 3(i). Most importantly, as established there, we always obtain $\frac{dX_2^*}{dA_2} > 0$. We thus obtain $A_2^* > A_1$ if and only if $\frac{dX_1^*}{dA_2} < 0$. Since we only consider nonnegative levels of free allocation, we furthermore obtain $A_2^* = 0$ whenever $0 < \frac{dX_2^*}{dA_2} < \frac{dX_1^*}{dA_2}$. All those results of comparative statics have been derived in Lemma 3 and have been discussed subsequently in Sect. 3. Figures 3 and 5 do thus in principle look identical; observe, however, that Fig. 3 exclusively states results of comparative statics, whereas Fig. 5 illustrates the properties of the optimal allocation A_2^* by making use of the previously obtained findings.

The intuition for the optimal level of free allocation T^* in principle goes along the same lines as the one provided for the findings of Theorem F. As compared to the first best benchmark, for positive allocation A_1 investment incentives in the peak load technology are too high. Whenever $A_1 < A_1^E$, we obtain $\Psi_{\text{Ie}} < 0$ which implies that a higher emission price e^* allows to reduce investment incentives in the peak load technology. Observe furthermore that investment incentives in the base load technology as induced by A_2^* are either too high or too low (i.e., $A_2^* > A_1$ for $w_2 > w_2^E$ and vice versa). As we show in the theorem, it is optimal to set the total cap such as to obtain an emission price $e^* > D_T$ which induce reduced investment incentives in the peak load technology whenever $A_1 < A_1^C$. Observe that for $A_1 = A_1^C$, we obtain $A_2^* = A_1 = A_1^C$ which implies undistorted investment incentives in the base load technology; in this case, we thus obtain $e^* = D_T$. The reverse holds true for the case $A_1 > A_1^C$ where the optimal emission cap T^* has to be set to obtain $e^* < D_T$ which induces reduced investment incentives in the peak load technology.

In the subsequent Theorem 6, we consider the case that allocation for the peak load technology A_2 is exogenously fixed, and allocation for the base load technology A_2 and the total emission cap T is determined optimally.

Theorem 6 (Optimal design for fixed allocation A_2) *Suppose the allocation for the base technology A_2 is exogenously fixed. The optimal allocation for the peak technology then solves $A_1^* = \frac{dX_2^*/dA_1}{dX_2^*/dA_1 - dX_1^*/dA_1} A_2$. More specifically (see right graph of Fig. 5)*

$$\begin{cases} A_1^* = 0 & if \quad (A_2^{\text{cross}} \leq A_2 < A_2^{\text{lim}}) \\ 0 < A_1^* < A_2 & if \quad \left((0 < A_2 < A_2^{\text{cross}}) \;\&\; (w_2 < w_2^S)\right) OR \left((0 < A_2 < A_2^{\text{total}}) \;\&\; (w_2 > w_2^S)\right) \\ A_2 < A_1^* & if \quad (A_2^{\text{total}} < A_2 < A_2^{\text{lim}}). \end{cases}$$

Define $A_2^{\text{em}} := A_1^E + A_2^E$. We obtain $A_2^{\text{em}} < A_2^{\text{total}}$ and $A_2^{\text{em}} < A_2^{\text{cross}}$. The optimal cap T^* is such that $e^* < D_T$ if $A_2 < A_2^{\text{em}}$ and $e^* > D_T$ if $A_2 > A_2^{\text{em}}$.

Proof See Appendix H. □

Observe that the optimality condition for A_1 as stated in the theorem is obtained directly by rearranging expression (11). To derive the properties of the optimal degree of allocation for the peak load technology A_1^* as stated in the theorem, we can now make use of the properties of comparative statics derived in Lemma 3(ii). Most importantly, as established there, we always obtain $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} > \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$. Since we only consider nonnegative levels of free allocation, we obtain $A_1^* > 0$ whenever $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} < 0$. Furthermore, we obtain $A_1^* > A_2$ whenever $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} < \frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} < 0$.

Let us finally provide some intuition for the optimal cap on total emissions T. First of all, observe that for $A_2 < A_2^{\text{total}}$ and $A_2 < A_2^{\text{cross}}$, we always obtain $0 < A_1^* < A_2$ which implies that investment incentives both in the base load and the peak load technology are too high as compared to the first best benchmark. For low levels of allocation to the base load technology (i.e., $A_2 < A_1^*(A_2) + A_2^E$), we obtain $\Psi_{\text{IIe}} < 0$ which makes it optimal to induce an emission price $e^* > D_T$ to lower investment incentives for both technologies.[28] Observe that for $A_2 = A_1^*(A_2) + A_2^E$, we obtain $\Psi_{\text{IIe}} = 0$, and a distortion of the emission price above (or below) social cost of pollution has no impact on investment incentives in the base load technology. However, investment incentives in the peak load technology are too high (since $A_1^* > 0$). Since $\Psi_{\text{Ie}} < 0$, the distortion of the emission price above social cost of pollution is thus still suited to reduce investment incentives in the peak load technology. In total, the theorem thus balances increased investment incentives in the base load technology with reduced investment incentives in the peak load technology. The cutoff is reached where $\Psi_{\text{IIe}} + \Psi_{\text{Ie}} = 0$ which implies $A_2 = A_1^E + A_2^E = A_2^{\text{em}}$. That is for $A_2 < A_2^{\text{em}}$, it is optimal to set an emission cap T^* which induces $e^* > D_T$ and for $A_2 > A_2^{\text{em}}$, the optimal cap T^* induces $e^* < D_T$.[29] All those results are illustrated in Fig. 5.

Discussion and policy implications: We conclude the discussion of Theorems 5 and 6 by applying our findings to the above discussed policy of full allocation $(A_1^{\text{full}}, A_2^{\text{full}})$ which served as the main illustrating example throughout this article:

First, consider the case of Theorem 5, where the allocation A_1 for the peak technology is exogenously fixed. As already shown, under full allocation we obtain $A_1^{\text{full}} \in [A_1^E, (1 - F(\underline{\theta_P}))w_1]$. For the case of a completely clean base load technology (that is $w_2 = 0$, in the context of electricity markets this would be the case for nuclear power plants, for example) under the current rules such technology would not obtain any permits. As our results directly show, however, such technology should be granted more free allocations than the peak technology, i.e., $A_2^* > A_1^{\text{full}}$. Moreover, the total emission cap T^* should be chosen such as to implement $e^* < D_T$ in order to dampen excessive investment incentives induced by those levels of free allocation.

Next, consider the case of Theorem 6 where the allocation A_2 for the base technology is exogenously fixed. The optimal level of free allocation for the peak technology has to be strictly positive if the peak technology is less emission intense than the base technology. In particular, if the peak technology is completely clean (that is $w_1 = 0$, in the context of electricity markets this would be the case for small biogas-fired engines or turbines for example), under full free allocation this technology would not receive any free permits. As our results directly show, however, such technology should be granted a positive amount of free permits. Unlike in the case discussed in the pre-

[28] Observe that for $A_2^E < 0$, this range is degenerated at zero.

[29] For $A_2 > A_2^{\text{cross}}$, only investment incentives in the base load technology are distorted; since $A_2^{\text{cross}} > A_2^{\text{em}}$, the optimal cap then clearly has to implement $e^* > D_T$. For $A_2 > A_2^{\text{total}}$, the optimal $A_1^*(A_2) > A_2$ is so large that the optimal cap also has to implement $e^* > D_T$, as we show.

ceding paragraph, however, the level of free allocation for the peak load technology should remain below the exogenously fixed level of free allocation for the base load technology. The reason for this difference lies in the fact that free allocations A_2 for the base load technology only have an impact on the resulting technology mix, and free allocations A_1 to the peak load technology have an impact on the resulting technology mix and on firms' total investment activity. For exogenously fixed A_2, the optimal level of A_1^* is thus more moderate since it also leads to distorted total investment decisions. Finally, notice for the optimal cap on total emissions: Since $A_2^{\text{em}} < (1 - F(\theta_{\overline{B}}))w_2 < A_2^{\text{full}}$ (compare Definition 1 and the last paragraph of Sect. 3), we can directly conclude that the total cap on emission has to be chosen such as to implement $e^* < D_T$ in order to dampen excessive investment incentives induced by those levels of free allocation.

In sum, if one of the technologies is granted an exogenously fixed level of free allocation (e.g., due to lobbying), then the optimal pattern of allocations to the remaining technology is completely different from the one that is obtained under full allocation. Furthermore, for high levels of free allocation, the cap on total emissions should be chosen such as to induce an emission permit price which is below marginal social cost of pollution in order to reduce the distortions on the resulting technology mix.

7 Conclusion

Tradeable pollution permits are an increasingly important policy tool in environmental legislation worldwide. The possibility to freely allocate permits provides an important possibility to share the regulatory burden. This seems to significantly enhance the political support for recently introduced legislations (see, for example, Tietenberg 2006; Bovenberg et al. 2008, or Convery (2009)). Since free allocations typically are subject to implicit or explicit updating, the allocation of permits has an impact on firms' decisions: First, updating of free allocations on the one hand has an impact on firms' operation of existing production facilities if they believe that current output or emissions do have an impact on allowances granted in future periods. Second, updating will also have an impact on firms' incentives to modify their production facilities. The first aspect has already been intensively analyzed in the literature (compare, for example, Böhringer and Lange 2005; Mackenzie et al. 2008; Böhringer et al. 2017, or Meunier et al. 2018).

All those contributions focus on the impact of free allocations with updating on output decisions and abstract from investment decisions. However, free allocations with updating also have a direct impact on firms' investment incentives which determine production capacities and the technology mix in the long run. It has been the purpose of the present article to analyze its impact on firms' investment decisions and to derive the optimal cap and trade mechanism in such an environment.

In the present article, we have thus analyzed an analytical framework with tradeable permits and a cap on total emissions. Potentially strategically acting firms have been able to invest into production facilities (with different emission intensities) which allow for production for a longer horizon of time. After establishing the market equilibrium and the resulting technology mix, we have analyzed the optimal design of the cap and trade mechanism.

As a benchmark, we established the first best solution which is obtained for an ideal market. We then have derived the optimal design of the cap and trade system for a series of market imperfections. First, we have analyzed the case of strategic investment and production decisions in an imperfectly competitive market. This allowed to highlight the close interdependency of mechanisms to overcome low investment incentives (such as, for example, capacity markets, compare Fabra 2018) and cap and trade systems: If the endogenous nature of emission prices in the presence of a cap and trade system is disregarded, too high investment incentives are induced by such mechanisms.

We then have analyzed the case that the competition authority cannot freely choose all parameters of the cap and trade system due to restrictions imposed by the political processes. The optimal choice of the remaining parameters differs substantially from that observed for the first best benchmark. Our result showed, for example, that if a certain technology receives free allocations, it is typically optimal to grant free allocations also to the other technology. Interestingly, those free allocations granted to the other technology should be higher in case this technology is less emission intensive.

A Proof of Lemma 1

Note that given our assumptions on demand and cost, existence of spot market equilibrium at each demand scenario θ is ensured for the case of perfect and imperfect competition. We denote by $q_i^*(\theta, x)$ spot market output of firm i in scenario θ, given investments $x = (x_{11}, \ldots, x_{1n}, x_{21}, \ldots, x_{2n})$. Remember, X_1 and X_2 denote industry investment in either technology and $Q^*(\theta)$ industry output at each spot market θ, and it is given as follows:

$$Q^* = \begin{cases} Q : P(Q,\theta) + P_q(Q,\theta)\frac{Q}{n} - c_2 - w_2 e^* = 0 & if \;\; \theta \;\; \in [\underline{\theta}, \theta_{\overline{B}}] \\ X_2 & if \;\; \theta \;\; \in [\theta_{\overline{B}}, \theta_{\underline{P}}] \\ Q : P(Q,\theta) + P_q(Q,\theta)\frac{Q}{n} - c_1 - w_1 e^* = 0 & if \;\; \theta \;\; \in [\theta_{\underline{P}}, \theta_{\overline{P}}] \\ X_1 & if \;\; \theta \;\; \in [\theta_{\overline{P}}, \overline{\theta}] \end{cases} \tag{15}$$

The critical spot market scenarios are defined as follows:

$$\theta_{\overline{B}} : \;\; P(X_2, \theta_{\overline{B}}) + P_q(X_2, \theta_{\overline{B}})\frac{1}{n} - c_2 - w_2 e^* = 0$$

$$\theta_{\underline{P}}: \; P(X_2,\theta_{\underline{P}}) + P_q(X_2,\theta_{\underline{P}})\frac{1}{n} - c_1 - w_1 e^* = 0$$
$$\theta_{\overline{P}}: \; P(X_1,\theta_{\overline{P}}) + P_q(X_1,\theta_{\overline{P}})\frac{1}{n} - c_1 - w_1 e^* = 0$$

That is, at spot market $\theta_{\overline{B}}$ investment X_2 in the base load technology (c_2, k_2) starts to be binding, at $\theta_{\underline{P}}$, firms start to produce with the peak load technology (c_1, k_1) and at $\theta_{\overline{P}}$, the total capacity bound X_1 is met. The first-order conditions stated in Lemma (1) are obtained when equating expressions (16), (17) and (18) to zero. Notice that the case solution for the case perfect competition is obtained as the special case where $n \to \infty$.

We first derive the first-order conditions for optimal investment decisions. Note that although in equilibrium at different demand realizations θ firm i might sometimes produce an unconstrained equilibrium quantity and sometimes is constrained by its choice x_{1i} or x_{2i}, equilibrium profit of firm i is continuous in θ. Thus, by Leibniz' rule, the first derivatives of the profit function are given as follows:

$$\frac{\mathrm{d}\pi_i}{dx_{1i}} = \int_{\theta_{\overline{P}}}^{\overline{\theta}} \left[P(X_1,\theta) + P_q(X_1,\theta)\frac{X_1}{n} - (c_1 + w_1 e) \right] \mathrm{d}F(\theta) - (k_1 - A_1 e) \tag{16}$$

$$\frac{\mathrm{d}\pi_i}{dx_{2i}} = \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} \left[P(X_2,\theta) + P_q(X_2,\theta)\frac{X_2}{n} - (c_2 + w_2 e) \right] \mathrm{d}F(\theta)$$
$$+ \int_{\theta_{\underline{P}}}^{\overline{\theta}} (c_1 - c_2) + (w_1 - w_2) e \mathrm{d}F(\theta) - (k_2 - A_2 e) + (k_1 - A_1 e) \tag{17}$$

In the market solution, the emission price e has to be such as to equate the following expression to zero:

$$\int_{\underline{\theta}}^{\theta_{\overline{B}}} w_2 Q^* \mathrm{d}F(\theta) + \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} w_2 X_2 \mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} w_1 Q^* \mathrm{d}F(\theta)$$
$$+ \int_{\theta_{\overline{P}}}^{\overline{\theta}} w_1 X_1 \mathrm{d}F(\theta) - \int_{\theta_{\underline{P}}}^{\overline{\theta}} (w_1 - w_2) X_2 \mathrm{d}F(\theta) - T \tag{18}$$

which are the conditions Ψ_I, Ψ_{II} and Ψ_E as given in the lemma.

Let us directly at this point determine all partial derivatives of the equilibrium system characterized in the lemma. The partial derivatives of Ψ_I (expression (16)), Ψ_{II} (expression (17)) and Ψ_E (expression (18)) read as follows:

$$\frac{\partial \Psi_I}{\partial X_1^*} = \Psi_{I1} = \int_{\theta_{\bar{P}}}^{\bar{\theta}} P_q\left(X_1^*, \theta\right) \frac{n+1}{n} + P_{qq}\left(X_1^*, \theta\right) \frac{X_1^*}{n} \mathrm{d}F(\theta) < 0$$
$$\frac{\partial \Psi_{II}}{\partial X_2^*} = \Psi_{II2} = \int_{\theta_{\bar{B}}}^{\theta_{\underline{P}}} P_q\left(X_2^*, \theta\right) \frac{n+1}{n} + P_{qq}\left(X_2^*, \theta\right) \frac{X_2^*}{n} dF(\theta) < 0$$

$$\frac{\partial\Psi_E}{\partial e}=\Psi_{\text{Ee}}=\int_{\underline{\theta}}^{\theta_{\bar{B}}}\frac{w_2^2}{P_q(Q^*,\theta)\frac{n+1}{n}+P_{qq}(Q^*,\theta)\frac{Q^*}{n}}\,\mathrm{d}F(\theta)$$
$$+\int_{\theta_{\underline{P}}}^{\theta_{\bar{P}}}\frac{w_1^2}{P_q(Q^*,\theta)\frac{n+1}{n}+P_{qq}(Q^*,\theta)\frac{Q^*}{n}}\,\mathrm{d}F(\theta)<0$$
$$\frac{\partial\Psi_E}{\partial X_1^*}=\Psi_{E1}=\left(1-F(\theta_{\overline{P}})\right)w_1=A_1^E>0\,\frac{\partial\Psi_I}{\partial e}=\Psi_{\text{Ie}}=A_1-(1-F(\theta_{\bar{P}}))w_1=A_1-A_1^E$$
$$\frac{\partial\Psi_E}{\partial X_2^*}=\Psi_{E2}=(1-F(\theta_{\bar{B}}))w_2-(1-F(\theta_{\underline{P}}))w_1=A_2^E$$
$$\frac{\partial\Psi_{II}}{\partial e}=\Psi_{\text{IIe}}=A_2-A_1-(1-F(\theta_{\bar{B}}))w_2+(1-F(\theta_{\underline{P}}))w_1=A_2-A_1-A_2^E$$
$$\frac{\partial\Psi_I}{\partial A_1}=e\qquad\frac{\partial\Psi_{II}}{\partial A_1}=-e\qquad\frac{\partial\Psi_{II}}{\partial A_2}=e\qquad\frac{\partial\Psi_E}{\partial T}=-1.$$

B Proof of Lemma 2

Part (i) To derive the second-order conditions established in Lemma 2, first observe that differentiation of the permit pricing condition Ψ_E with respect to X_1 and slight rearranging yields $\frac{\mathrm{d}e^*}{\mathrm{d}X_1}=\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}$. Plugging into the derivatives of Ψ_I and Ψ_{II} and replacing for A_1^E and A_2^E as introduced in Definition 1 yield:

$$\frac{\mathrm{d}\Psi_I}{\mathrm{d}X_1}=\Psi_{I1}+\Psi_{\text{Ie}}\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}=\Psi_{I1}+(A_1-A_1^E)\frac{A_1^E}{-\Psi_{\text{Ee}}}$$
$$\frac{\mathrm{d}\Psi_{II}}{\mathrm{d}X_1}=\Psi_{\text{IIe}}\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}=(A_2-A_1-A_2^E)\frac{A_1^E}{-\Psi_{\text{Ee}}}$$

Likewise, we obtain

$$\frac{\mathrm{d}\Psi_I}{\mathrm{d}X_2}=\Psi_{\text{Ie}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}=(A_1-A_1^E)\frac{A_2^E}{-\Psi_{\text{Ee}}}$$
$$\frac{\mathrm{d}\Psi_{II}}{\mathrm{d}X_1}=\Psi_{II2}+\Psi_{\text{IIe}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}=\Psi_{II2}+(A_2-A_1-A_2^E)\frac{A_2^E}{-\Psi_{\text{Ee}}}$$

The matrix $H=\begin{pmatrix}d\Psi_I/\mathrm{d}X_1 & d\Psi_I/\mathrm{d}X_2\\ d\Psi_{II}/\mathrm{d}X_1 & d\Psi_{II}/\mathrm{d}X_2\end{pmatrix}$ is negative definite if and only if conditions (a), (b) and (c) established in Lemma 2 (i) are satisfied. To save on notation, we introduce $C:=det(H)$; observe that $C\geq 0$ if H is negative definite [compare (c) in Lemma 2 (i)].

Part (ii) Since $\Psi_{I1}<0$, $\Psi_{II2}<0$ and $\Psi_{\text{Ee}}<0$ (see Appendix A), the conditions provided in Lemma 2 (ii) are sufficient to guarantee negative definiteness of the matrix H.

Part (iii) To see why this is true, just observe that for $A_1 > A_1^{\text{lim}}$, condition (a) established in Lemma 2(i) will be violated. The condition defining A_2^{lim} is given by the sum of conditions (a) and (b) of Lemma 2(i); for $A_1 > A_1^{\text{lim}}$, at least one of those two conditions will be violated.

C Proof of Lemma 3

C.1 Preliminaries: comparative statics

The differentials for $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$, $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}$, $\frac{\mathrm{d}X_1^*}{\mathrm{d}\mathcal{T}}$ and $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$, $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}$, $\frac{\mathrm{d}X_2^*}{\mathrm{d}\mathcal{T}}$ are obtained by applying the implicit function theorem to the equilibrium conditions established in Lemma 1. The total derivative of this equations system with respect to the parameter A_1 yields:

$$\Psi_I : \quad \Psi_{I1}\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} + \Psi_{\text{Ie}}\frac{\mathrm{d}e^*}{\mathrm{d}A_1} + \frac{\partial \Psi_I}{\partial A_1} \equiv 0 \tag{19}$$

$$\Psi_{II} : \quad \Psi_{II2}\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} + \Psi_{\text{IIe}}\frac{\mathrm{d}e^*}{\mathrm{d}A_1} + \frac{\partial \Psi_{II}}{\partial A_1} \equiv 0 \tag{20}$$

$$\Psi_E : \quad \Psi_{E1}\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} + \Psi_{E2}\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} + \Psi_{\text{Ee}}\frac{\mathrm{d}e^*}{\mathrm{d}A_1} + \frac{\partial \Psi_E}{\partial A_1} \equiv 0 \tag{21}$$

In order to derive an explicit formulation for $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$, we solve expression (21) for $\frac{\mathrm{d}e^*}{\mathrm{d}A_1}$ and expression (20) for $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$. Plugging into expression (19) yields:

$$\left(\left(\Psi_{I1} + \Psi_{\text{Ie}}\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right) - \left(\Psi_{\text{Ie}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right)\frac{\left(\Psi_{\text{IIe}}\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right)}{\left(\Psi_{II2} + \Psi_{\text{IIe}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right)}\right)\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$$
$$+\frac{\partial \Psi_I}{\partial A_1} - \frac{\left(\Psi_{\text{Ie}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right)\frac{\partial \Psi_{II}}{\partial A_1}}{\left(\Psi_{II2} + \Psi_{\text{IIe}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right)} = 0$$

By making use of the definition of the variable C (compare Appendix B), we can rearrange this expression and obtain:

$$\begin{aligned}\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} &= (-1)\frac{\frac{\partial \Psi_I}{\partial A_1}\left(\Psi_{II2} + \Psi_{\text{IIe}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right) - \frac{\partial \Psi_{II}}{\partial A_1}\left(\Psi_{\text{Ie}}\frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right)}{C} \\ &= \left((A_2 - A_1^E - A_2^E)A_2^E - \Psi_{II2}\Psi_{\text{Ee}}\right)\frac{-e}{-\Psi_{\text{Ee}}C}\end{aligned} \tag{22}$$

Likewise, to obtain an explicit formulation for $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$, we analogously solve expression (21) for $\frac{\mathrm{d}e^*}{\mathrm{d}A_1}$ and expression (19) for $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$. Plugging into expression (20) and solving for $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ yield:

$$\begin{aligned}\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} &= (-1)\frac{\frac{\partial \Psi_{II}}{\partial A_1}\left(\Psi_{I1} + \Psi_{\mathrm{Ie}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right) - \frac{\partial \Psi_I}{\partial A_1}\left(\Psi_{\mathrm{IIe}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right)}{C} \\ &= \left((A_2 - A_1^E - A_2^E)A_1^E - \Psi_{I1}\Psi_{\mathrm{Ee}}\right)\frac{e}{-\Psi_{\mathrm{Ee}}C}\end{aligned} \tag{23}$$

Analogously, we obtain:

$$\begin{aligned}&\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} = \left((A_1 - A_1^E)A_2^E\right)\frac{e}{-\Psi_{\mathrm{Ee}}C} \qquad \frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} = \left((A_1 - A_1^E)A_1^E - \Psi_{I1}\Psi_{\mathrm{Ee}}\right)\frac{-e}{-\Psi_{\mathrm{Ee}}C} \\ &\frac{\mathrm{d}X_1^*}{\mathrm{d}T} = \left(A_1 - A_1^E\right)\frac{\Psi_{II2}}{-\Psi_{\mathrm{Ee}}C} \qquad \frac{\mathrm{d}X_2^*}{\mathrm{d}T} = \left(A_2 - A_1 - A_2^E\right)\frac{\Psi_{I1}}{-\Psi_{\mathrm{Ee}}C}\end{aligned} \tag{24}$$

C.2 Proof of Lemma 3(i)

First, we define

$$\begin{aligned}\Gamma_1^{\lim}(A_1) &:= \Psi_{I1} + \Psi_{\mathrm{Ie}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}} \\ &= \int_{\theta_{\overline{P}}}^{\overline{\theta}} P_q(X_1^*, \theta)\mathrm{d}F(\theta) \\ &\quad + \frac{(A_1 - A_1^E)A_1^E}{\int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{w_2^2}{-P_q(Q^*,\theta)}\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{w_1^2}{-P_q(Q^*,\theta)}\mathrm{d}F(\theta)}.\end{aligned} \tag{25}$$

Observe that the second-order sufficient conditions for existence of the market equilibrium specified in 2(i) require $\Gamma_1^{\lim}(A_1) < 0$. Since $\Gamma_1^{\lim}(A_1)$ is increasing in A_1, we can define a unique $A_1^{\lim}$ which solves $\Gamma_1^{\lim}(A_1^{\lim}) = 0$ and conclude that $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} > 0$ for all $A_1 < A_1^{\lim}$.
Second, we define

$$\Gamma_1^{\mathrm{total}}(A_1) := \Psi_{\mathrm{Ie}}\frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}} = \frac{\left(A_1 - A_1^E\right)A_2^E}{\int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{w_2^2}{-P_q(Q^*,\theta)}\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{w_1^2}{-P_q(Q^*,\theta)}\mathrm{d}F(\theta)} \tag{26}$$

This allows us to rewrite $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}$ as established in expression (24) as follows:

$$\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} = \Gamma_1^{\mathrm{total}}(A_1)\frac{e^*}{C} \tag{27}$$

We finally show that $A_1^E < A_1^{\lim}$. To see this, observe that $\Gamma_1^{\lim}(A_1^E) = \Psi_{I1} < 0$. Since $\Gamma_1^{\lim}(A_1)$ is increasing in A_1, we necessarily obtain $A_1^{\lim} > A_1^E$.
Third, we define

$$\Gamma_1^{\text{cross}}(A_1) := \Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}} + \Psi_{\text{Ie}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}$$
$$= \int_{\theta_{\overline{P}}}^{\overline{\theta}} P_q(X_1^*, \theta) \mathrm{d}F(\theta) + \frac{\left(A_1 - A_1^E\right)\left(A_1^E + A_2^E\right)}{\int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{w_2^2}{-P_q(Q^*,\theta)} \mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{w_1^2}{-P_q(Q^*,\theta)} \mathrm{d}F(\theta)}$$

Observe that $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} < \frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}$ if and only if $\Gamma_1^{\text{cross}}(A_1) > 0$ (compare expression (24)). We define the locus where $\Gamma_1^{\text{cross}}(A_1) = 0$ by A_1^{cross}.

We now compare the critical allocation A_1^{cross} relative to the critical values $A_1^{\lim}$ and A_1^E:

- For $w_2 > w_2^E$, we can establish the following ranking: $A_1^E < A_1^{\text{cross}} < A_1^{\lim}$.
  To show the first inequality, observe that for all $A_1 \le A_1^E$, we obtain $\Gamma_1^{\text{total}}(A_1) \le 0$. As shown above, we also obtain $\Gamma_1^{\lim}(A_1) \le 0$, which implies $\Gamma_1^{\text{cross}}(A_1) = \Gamma_1^{\lim}(A_1) + \Gamma_1^{\text{total}}(A_1) \le 0$. This directly implies, however, that A_1^{cross} cannot be in the interval $[0, A_1^E]$.
  To show the second inequality, observe that for $A_1 > A_1^E$, we have $\Gamma_1^{\text{total}}(A_1) > 0$. Whenever $\Gamma_1^{\text{cross}}(A_1) = \Gamma_1^{\lim}(A_1) + \Gamma_1^{\text{total}}(A_1) = 0$, we must thus have $\Gamma_1^{\lim}(A_1) < 0$. Since $\Gamma_1^{\lim}(A_1)$ is strictly increasing in A_1, this implies $A_1^{\lim} > A_1^{\text{cross}}$.
- For $w_2 < w_2^E$, we establish that $\Gamma_1^{\text{cross}}(A_1) < 0$ (i.e., $A_1^* > 0$) for all $A_1 \in (0, A_1^{\lim}]$.
  First, observe that for $A_1 > A_1^E$ we obtain $\Gamma_1^{\text{total}}(A_1) < 0$ which implies that $\Gamma_1^{\text{cross}}(A_1) = \Gamma_1^{\lim}(A_1) + \Gamma_1^{\text{total}}(A_1) < 0$ for $A_1 \in [A_1^E, A_1^{\lim}]$.
  Second, observe that for $A_1 \le A_1^E$ and $\left((1 - F(\theta_{\overline{B}}))w_2 - (F(\theta_{\overline{P}}) - F(\theta_{\underline{P}}))w_1\right) > 0$, we obtain $\Gamma_1^{\text{cross}}(A_1) < 0$.
  Third, observe that for $A_1 \le A_1^E$ and $\left((1 - F(\theta_{\overline{B}}))w_2 - (F(\theta_{\overline{P}}) - F(\theta_{\underline{P}}))w_1\right) < 0$, $\Gamma_1^{\text{cross}}(A_1)$ is maximized for $A_1 = 0$. Expression (28) then reads as follows:

$$\Gamma_1^{\text{cross}}(0) = \int_{\theta_{\overline{P}}}^{\overline{\theta}} P_q(X_1^*, \theta) \mathrm{d}F(\theta)$$
$$+ \frac{-\left(1 - F(\theta_{\overline{P}})\right) w_1 \left((1 - F(\theta_{\overline{B}}))w_2 - (F(\theta_{\overline{P}}) - F(\theta_{\underline{P}}))w_1\right)}{\int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{w_2^2}{-P_q(Q^*,\theta)} \mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{w_1^2}{-P_q(Q^*,\theta)} \mathrm{d}F(\theta)}$$

  which can be guaranteed to be negative if $P_{qq} \le 0$ and $P_{q\theta} \ge 0$. Without those additional assumptions, it might happen that $A_1^* = 0$ in the region where $A_1 \le A_1^E$ and $\left((1 - F(\theta_{\overline{B}}))w_2 - (F(\theta_{\overline{P}}) - F(\theta_{\underline{P}}))w_1\right) < 0$.

C.3 Proof of Lemma 3(ii)

First, we define

$$\Gamma_2^{\lim}(A_2) := \left(\Psi_{II2} + \Psi_{I1} + (\Psi_{\text{IIe}} + \Psi_{\text{Ie}}) \frac{\Psi_{E2} + \Psi_{E1}}{-\Psi_{\text{Ee}}}\right). \tag{28}$$

In order to precisely define $A_2^{\lim}$, first observe that $\Psi_{E2} + \Psi_{E1} = A_1^E + A_2^E$ (compare Definition 1), we thus have to consider the following two cases:

- For the case $w_2 > w_2^L$ (i.e., $A_1^E + A_2^E > 0$), $\Gamma_2^{\lim}$ is strictly increasing in A_2, and we can define a unique $A_2^{\lim}$ which solves $\Gamma_2^{\lim}(A_2^{\lim}) = 0$
- For the case $w_2 \leq w_2^L$ (i.e., $A_1^E + A_2^E \leq 0$), $\Gamma_2^{\lim}(A_2)$ is non-increasing in A_2, and it is thus minimized for $A_2 = 0$ which yields

$$\Gamma_2^{\lim} = \Psi_{II2} + \Psi_{I1} + \left(A_2 - (A_1^E + A_2^E)\right) \frac{A_1^E + A_2^E}{-\Psi_{\text{Ee}}} < 0.$$

That is, $\Gamma_2^{\lim}(A_1) < 0$ for all $A_2 \geq 0$. For ease of notation, we thus define $A_2^{\lim} = \infty$ in this case.

We now determine $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ as given by expressions (22) and (22). After plugging in for $\frac{\partial \Psi_I}{\partial A_1} = e^*$ and $\frac{\partial \Psi_{II}}{\partial A_1} = -e^*$ (compare Appendix A), we obtain for $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$:

$$\begin{aligned}&\left(\left(\Psi_{II2} + \Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right) + \left(\Psi_{\text{Ie}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right) + \left(\Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right) + \left(\Psi_{\text{IIe}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right)\right) \frac{-e^*}{C}\\ &= \frac{-e^*}{C} \Gamma_2^{\lim}(A_2)\end{aligned}$$

Since $\Gamma_2^{\lim}(A_2) < 0$ as established above, we conclude that $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} > \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ for all $A_2 \in [0, A_2^{\lim}]$.
Second, we define

$$\begin{aligned}\Gamma_2^{\text{total}}(A_2) &:= \left(\left(\Psi_{II2} + \Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right) + \left(\Psi_{\text{Ie}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right)\right)\\ &= \left(\Psi_{II2} + \left(A_2 - A_1^E - A_2^E\right) \frac{A_2^E}{-\Psi_{\text{Ee}}}\right)\end{aligned} \tag{29}$$

- For $w_2 > w_2^E$ (i.e., $A_2^E > 0$, see Definition 1), $\Gamma_2^{\text{total}}(A_2)$ is strictly increasing in A_2. Since $\Gamma_2^{\text{total}}(0) < 0$, we can thus define a unique $A_2^{\text{total}} > 0$ which satisfies $\Gamma_2^{\text{total}}(A_2^{\text{total}}) = 0$.
- For the case $w_2 \leq w_2^E$ (i.e., $A_2^E \leq 0$, see Definition 1), $\Gamma_2^{\text{total}}(A_2)$ is non-increasing in A_2. Observe that in this case, Γ_2^{total} is maximized for $(A_2 = 0, w_2 = 0)$ which

yields:

$$\Gamma_2^{\text{total}}(0) = \left(\Psi_{II2} + \left(A_2 - A_1^E - A_2^E\right)\frac{A_2^E}{-\Psi_{\text{Ee}}}\right) < 0. \tag{30}$$

That is, for $A_2 \geq 0$ we obtain $\Gamma_2^{\text{total}}(A_2) < 0$. For ease of notation we thus define $A_2^{\text{total}} := \infty$ whenever $w_2 \leq w_2^E$.

Observe now that we can rewrite $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}$ [established in expression (23)] in terms of Γ_2^{total} which yields:

$$\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} = \Gamma_2^{\text{total}}(A_2)\frac{-e^*}{C} \tag{31}$$

Third, we define

$$\Gamma_2^{\text{cross}}(A_2) := \left(\Psi_{I1} + \Psi_{\text{Ie}}\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right) + \left(\Psi_{\text{IIe}}\frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right) = \Psi_{I1} + \left(A_2 - A_1^E - A_2^E\right)\frac{A_1^E}{-\Psi_{\text{Ee}}} \tag{32}$$

We now rewrite $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ as established in expression (23) in terms of Γ_2^{cross} which yields:

$$\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} = \Gamma_2^{\text{cross}}(A_2)\frac{e^*}{C} \tag{33}$$

Observe that Γ_2^{cross} is strictly increasing in A_2 (see Appendix A). Since $\Gamma_2^{\text{cross}}(0) < 0$, we can thus define a unique $A_2^{\text{cross}} > 0$ which satisfies $\Gamma_2^{\text{cross}}(A_2^{\text{cross}}) = 0$:

$$A_2^{\text{cross}} = A_1^E + A_2^E + \frac{\Psi_{I1}\Psi_{\text{Ee}}}{A_1^E} \tag{34}$$

Finally, we compare the different critical values: A_2^{lim}, A_2^{total}, and A_2^{cross}. We have to consider the following three cases:

- For $w_2 \leq w_2^L$: In this case, we obtain $\Gamma_2^{\text{lim}}(A_2) < 0$ and $\Gamma_2^{\text{total}}(A_2) < 0$ for all $A_2 \geq 0$. Thus, A_2^{cross} provides the only critical level of initial allocation (remember we defined $A_2^{\text{lim}} = \infty$ and $A_2^{\text{total}} = \infty$), and we thus obtain $A_2^{\text{cross}} < A_2^{\text{lim}}$.
- For $w_2^L < w_2 \leq w_2^E$: In this case, we obtain $\Gamma_2^{\text{total}}(A_2) < 0$ for all $A_2 \geq 0$ (remember we defined $A_2^{\text{total}} = \infty$). Observe, furthermore, that $\Gamma_2^{\text{lim}}(A_2) = \Gamma_2^{\text{total}}(A_2) + \Gamma_2^{\text{cross}}(A_2)$ for all A_2. This directly implies, however, that $\Gamma_2^{\text{cross}}(A_2^{\text{lim}}) > 0$ and thus $A_2^{\text{cross}} < A_2^{\text{lim}}$.

- For $w_2 \geq w_2^E$, we evaluate $\Gamma_2^{\text{total}}(A_2^{\text{cross}})$ which yields [compare expressions (29) and (34)]:

$$\Gamma_2^{\text{total}}(A_2^{\text{cross}}) = \Psi_{I I 2} - \Psi_{I 1} \frac{A_2^E}{A_1^E} = \overline{P_q^2} \left(F(\theta_{\underline{P}}) - F(\theta_{\overline{B}}) \right) - \overline{P_q^1} \frac{(1 - F(\theta_{\overline{B}}))w_2 - (1 - F(\theta_{\underline{P}}))w_1}{w_1} \tag{35}$$

Observe that we denote by $\overline{P_q^1}$ the average slope of demand for those demand levels where total investment is binding and by $\overline{P_q^2}$ the average slope of demand for those demand levels where base load investment is binding, i.e.:

$$(i) \quad \overline{P_q^1} := \frac{\int_{\theta_{\overline{P}}}^{\overline{\theta}} P_q(X_1^*, \theta) \mathrm{d}F(\theta)}{1 - F(\theta_{\overline{P}})} \qquad (ii) \quad \overline{P_q^2} := \frac{\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} P_q(X_2^*, \theta) \mathrm{d}F(\theta)}{F(\theta_{\underline{P}}) - F(\theta_{\overline{B}})} \tag{36}$$

Rearranging this yields:

$$\Gamma_2^{\text{total}}(A_2^{\text{cross}}) = \frac{-\overline{P_q^1}(1 - F(\theta_{\overline{B}}))}{w_1} \left(w_2 - \left(\frac{(1 - F(\theta_{\underline{P}})) + (F(\theta_{\underline{P}}) - F(\theta_{\overline{B}}))\frac{\overline{P_q^2}}{\overline{P_q^1}}}{1 - F(\theta_{\overline{B}})} \right) w_1 \right) \tag{37}$$

Now, define

$$w_2^S := \frac{(1 - F(\theta_{\underline{P}})) + (F(\theta_{\underline{P}}) - F(\theta_{\overline{B}}))\frac{\overline{P_q^2}}{\overline{P_q^1}}}{1 - F(\theta_{\overline{B}})} w_1 \tag{38}$$

Observe that $w_2^E < w_2^S \leq w_1$ since $F(\theta_{\underline{P}}) - F(\theta_{\overline{B}}) > 0$ and $0 < \frac{\overline{P_q^2}}{\overline{P_q^1}} \leq 1$; notice that for $\overline{P_q^2} = \overline{P_q^1}$ (e.g., for $P_{qq} = P_{q\theta} = 0$), we obtain $w_2^S = w_1$. Furthermore, for $w_2 > w_2^S$ we obtain $\Gamma_2^{\text{total}}(A_2^{\text{cross}}) > 0$ and for $w_2 < w_2^S$, we obtain $\Gamma_2^{\text{total}}(A_2^{\text{cross}}) < 0$. Since $\Gamma_2^{\text{lim}}(A_2) = \Gamma_2^{\text{total}}(A_2) + \Gamma_2^{\text{cross}}(A_2)$, for all A_2 we obtain:

$$\begin{array}{ll} 0 < A_2^{\text{total}} < A_2^{\text{lim}} < A_2^{\text{cross}} & \text{if } w_2 > w_2^S \\ 0 < A_2^{\text{total}} = A_2^{\text{lim}} = A_2^{\text{cross}} & \text{if } w_2 = w_2^S \\ 0 < A_2^{\text{cross}} < A_2^{\text{lim}} < A_2^{\text{total}} & \text{if } w_2^E < w_2 < w_2^S \end{array} \tag{39}$$

C.4 Proof of Lemma 3(iii)

Observe that we have derived $\frac{\mathrm{d}X_1^*}{\mathrm{d}T}$ and $\frac{\mathrm{d}X_2^*}{\mathrm{d}T}$ in expression (24). The statements of Lemma 3(iii) follow directly since $\Psi_{I1} < 0$, $\Psi_{II2} < 0$ and $\Psi_{\mathrm{Ee}} < 0$ (see Definition 1).

D Proof of Lemma 4

To derive the optimal design of the cap and trade mechanism (A_1, A_2, T), we first differentiate welfare as given by expression (10) with respect to each of those parameters. We obtain for $\frac{\mathrm{d}W}{\mathrm{d}A_1}$:

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}A_1} &= \int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\left[P(Q^*,\theta)-c_2\right]\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\left[P(Q^*,\theta)-c_1\right]\mathrm{d}F(\theta)\\ &+\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{B}}}^{\theta_{\overline{P}}}\left[P(X_2^*,\theta)-c_2\right]\mathrm{d}F(\theta)+\int_{\theta_{\underline{P}}}^{\overline{\theta}}(c_1-c_2)\mathrm{d}F(\theta)-(k_2-k_1)\right]\\ &+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left[P(X_1^*,\theta)-c_1\right]\mathrm{d}F(\theta)-k_1\right]\end{aligned}$$

We can now plug in the equilibrium conditions for firms' investment choices given by expressions (2) and (3), and we can plug in the optimality conditions for the unconstrained spot markets whenever investment is not binding. [30] This yields:

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}A_1} &= \int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\left[-P_q\frac{Q^*}{n}+w_2e\right]\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\left[-P_q\frac{Q^*}{n}+w_1e\right]\mathrm{d}F(\theta)\\ &\quad\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}}\left[-P_q\frac{X_2^*}{n}+w_2e\right]\mathrm{d}F(\theta)+\int_{\theta_{\underline{P}}}^{\overline{\theta}}(w_2-w_1)e\mathrm{d}F(\theta)-(A_2-A_1)e\right]\\ &\quad+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left[-P_q+w_1e\right]\mathrm{d}F(\theta)-A_1e\right]\end{aligned}$$

This can be further simplified by making use of the derivative of the permit pricing given by expression 4 with respect to A_1 (i.e., $\frac{\Psi_E}{A_1}$). This allows to eliminate all terms containing the emission factors w_1 and w_2 from the above expression (shown explicitly in expression (41)). We thus obtain

$$\frac{\mathrm{d}W}{\mathrm{d}A_1} = \int_{\underline{\theta}}^{\theta_{\overline{B}}} \frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} \frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta)$$

[30] That is for spot markets $\theta \in [\underline{\theta}, \theta_{\overline{B}}] \cup [\theta_{\underline{P}}, \theta_{\overline{P}}]$; in those cases, the optimality conditions are simply given by $P(Q^*,\theta) + P_q\frac{Q^*}{n} - c_i - w_ie^* = 0$, for $i = 1, 2$.

$$\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}}\left[-P_q\frac{X_2^*}{n}\right]\mathrm{d}F(\theta)-(A_2-A_1)e\right]$$
$$+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left[-P_q\frac{X_1^*}{n}\right]\mathrm{d}F(\theta)-A_1e\right] \tag{40}$$

In order to show why indeed expression (40) is obtained, we now differentiate the permit pricing condition Ψ_E (compare Lemma 1) with respect to A_1, and this yields

$$-\int_{\underline{\theta}}^{\theta_{\overline{B}}} w_2\frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\mathrm{d}F(\theta)-\int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} w_1\frac{\mathrm{d}Q^*}{\mathrm{d}A_1}\mathrm{d}F(\theta)$$
$$=\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} w_2\mathrm{d}F(\theta)+\int_{\theta_{\underline{P}}}^{\overline{\theta}}(w_2-w_1)\mathrm{d}F(\theta)\right]+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}} w_1\mathrm{d}F(\theta)\right]. \tag{41}$$

Now, observe that $\frac{\mathrm{d}Q^*}{\mathrm{d}A_1}=\frac{\mathrm{d}Q^*}{de^*}\frac{\mathrm{d}e^*}{\mathrm{d}A_1^*}$, since unconstrained spot market output does not directly depend on the degree of free allocation A_1. By multiplying expression (41) with Δ (as defined in the lemma), we obtain:

$$\left(\int_{\underline{\theta}}^{\theta_{\overline{B}}}\frac{\mathrm{d}Q^*}{de^*}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta)+\int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}}\frac{\mathrm{d}Q^*}{de^*}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta)\right)\frac{\mathrm{d}e^*}{\mathrm{d}A_1}$$
$$=\frac{-\Delta}{n}\left(\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}A_2^E+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}A_1^E\right) \tag{42}$$

We can now plug expression (42) into expression (40), which yields

$$\frac{\mathrm{d}W}{\mathrm{d}A_1}=\frac{-\Delta}{n}\left(\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}A_2^E+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}A_1^E\right)$$
$$+\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}}\left(-P_q\frac{X_2^*}{n}\right)\mathrm{d}F(\theta)-(A_2-A_1)e\right]$$
$$+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left(-P_q\frac{X_1^*}{n}\right)\mathrm{d}F(\theta)-A_1e\right]$$

Rearranging finally yields

$$\frac{\mathrm{d}W}{\mathrm{d}A_1}=\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}}\left(-P_q\frac{X_2^*}{n}\right)\mathrm{d}F(\theta)-(A_2-A_1)e-\frac{\Delta}{n}A_2^E\right]$$
$$+\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left(-P_q\frac{X_1^*}{n}\right)\mathrm{d}F(\theta)-A_1e-\frac{\Delta}{n}A_1^E\right]$$

which corresponds exactly to the expression for $\frac{\mathrm{d}W}{\mathrm{d}A_1}$ stated in the lemma. The very same steps yield $\frac{\mathrm{d}W}{\mathrm{d}A_2} = \frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}\Omega_I + \frac{\mathrm{d}X_2^*}{\mathrm{d}A_2}\Omega_{II}$, as stated in the lemma.

We finally determine $\frac{\mathrm{d}W}{\mathrm{d}T}$. Analogous to expression (40), we obtain:

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}T} &= \frac{\mathrm{d}X_2^*}{\mathrm{d}T}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}}\left[-P_q\frac{X_2^*}{n}\right]\mathrm{d}F(\theta) - (A_2 - A_1)e\right] \\ &\quad + \frac{\mathrm{d}X_1^*}{\mathrm{d}T}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left[-P_q\frac{X_1^*}{n}\right]\mathrm{d}F(\theta) - A_1 e\right] \\ &\quad \int_{\underline{\theta}}^{\theta_{\overline{B}}}\frac{\mathrm{d}Q^*}{\mathrm{d}T}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}}\frac{\mathrm{d}Q^*}{\mathrm{d}T}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta) + e^* - D_T(T)\end{aligned} \tag{43}$$

which is obtained since all terms containing the emission factors w_1 and w_2 integrate to 1. Why this is the case becomes clear when differentiating the permit pricing condition Ψ_E with respect to T:

$$\begin{aligned}&-\int_{\underline{\theta}}^{\theta_{\overline{B}}} w_2\frac{\mathrm{d}Q^*}{\mathrm{d}T}\mathrm{d}F(\theta) - \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}} w_1\frac{\mathrm{d}Q^*}{\mathrm{d}T}\mathrm{d}F(\theta) \\ &= \frac{\mathrm{d}X_2^*}{\mathrm{d}T}A_2^E + \frac{\mathrm{d}X_1^*}{\mathrm{d}T}A_1^E - 1.\end{aligned}$$

Now, observe that $\frac{\mathrm{d}Q^*}{\mathrm{d}T} = \frac{\mathrm{d}Q^*}{de^*}\frac{\mathrm{d}e^*}{\mathrm{d}T}$ since unconstrained spot market output does not directly depend on the total emission cap T. By multiplying expression (44) with $\frac{\Delta}{n}$, we obtain:

$$\begin{aligned}&\left(\int_{\underline{\theta}}^{\theta_{\overline{B}}}\frac{\mathrm{d}Q^*}{de^*}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta) + \int_{\theta_{\underline{P}}}^{\theta_{\overline{P}}}\frac{\mathrm{d}Q^*}{de^*}\left[-P_q\frac{Q^*}{n}\right]\mathrm{d}F(\theta)\right)\frac{\mathrm{d}e^*}{\mathrm{d}T} \\ &= \frac{-\Delta}{n}\left(\frac{\mathrm{d}X_2^*}{\mathrm{d}T}A_2^E + \frac{\mathrm{d}X_1^*}{\mathrm{d}T}A_1^E - 1\right)\end{aligned} \tag{44}$$

We can now plug expression (44) into expression (43) which yields after rearranging:

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}T} &= \frac{\mathrm{d}X_2^*}{\mathrm{d}T}\left[\int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}}\left(-P_q\frac{X_2^*}{n}\right)\mathrm{d}F(\theta) - (A_2 - A_1)e - \frac{\Delta}{n}A_2^E\right] \\ &\quad + \frac{\mathrm{d}X_1^*}{\mathrm{d}T}\left[\int_{\theta_{\overline{P}}}^{\overline{\theta}}\left(-P_q\frac{X_1^*}{n}\right)\mathrm{d}F(\theta) - A_1 e - \frac{\Delta}{n}A_1^E\right] + \frac{\Delta}{n} + e^* - D_T(T)\end{aligned}$$

which corresponds exactly to the expression for $\frac{\mathrm{d}W}{\mathrm{d}T}$ stated in the lemma.

E Proof of Theorems 1 and 2

The optimality conditions established in Lemma 4 are satisfied if the following conditions hold:

$$\Omega_I = \int_{\theta_{\overline{P}}}^{\overline{\theta}} \frac{-P_q X_1^*}{n} \mathrm{d}F(\theta) - A_1 e^* - \frac{\Delta}{n} A_1^E = 0$$
$$\Omega_{II} = \int_{\theta_{\overline{B}}}^{\theta_{\underline{P}}} \frac{-P_q X_2^*}{n} \mathrm{d}F(\theta) - (A_2 - A_1) e^* - \frac{\Delta}{n} A_2^E = 0$$
$$e^* = D_T(T) - \frac{\Delta}{n}.$$

The case of perfect competition as analyzed in Theorem 1 is obtained for $n- > infty$. Observe that elimination of all terms involving the number of firms n in the denominator in the above conditions yields the characterization of the first best solution stated in Theorem 1. In order to obtain the solution obtained for the case of imperfect competition as established in Theorem 2, we solve the first two conditions for the levels of free allocation A_1 and A_2.

F Proof of Theorem 4

The optimality condition for $\mathcal{T}$ has been derived in Lemma 4 (iii). After plugging in the results of comparative statics for $\frac{\mathrm{d}X_1}{\mathrm{d}\mathcal{T}}$ and $\frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}}$ derived in expression (24), we obtain:

$$e^* - \mathcal{D}_{\mathcal{T}} = \left((A_2 - A_1) e \Psi_{\mathrm{IIe}} \frac{\Psi_{I1}}{-\Psi_{\mathrm{Ee}}\ C} + A_1 e \Psi_{\mathrm{Ie}} \frac{\Psi_{II2}}{-\Psi_{\mathrm{Ee}}\ C} \right) \tag{45}$$

We now make use of the notation introduced in Definition 1 which allows us to rewrite expression (45) as follows:

$$e^* - \mathcal{D}_{\mathcal{T}} = \left((A_2 - A_1 - A_2^E)(A_2 - A_1)\Psi_{II2} + (A_1 - A_1^E) A_1 \Psi_{I1} \right) \frac{e^*}{-\Psi_{\mathrm{Ee}}\ C} \tag{46}$$

Notice that $\frac{e^*}{-\Psi_{\mathrm{Ee}}\ C} > 0$ as established in Appendix A. The remainder of the right-hand side of expression (46) states $\Gamma_0(A_1, A_2)$, as defined in expression (14). The expression $e^* - D_T(T)$ and Γ_0 do thus exhibit the same sign which proofs the theorem.

G Proof of Theorem 5

As a **first step**, we determine the properties of the optimal allocation A_2^*. Observe that the optimality condition $A_2^* = \frac{\mathrm{d}X_2^*/\mathrm{d}A_2 - \mathrm{d}X_1^*/\mathrm{d}A_2}{\mathrm{d}X_2^*/\mathrm{d}A_2} A_1$ stated in the theorem directly is obtained by rearranging expression (12). In Lemma 3 (i), we have established $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} > 0$ for all $A_1 < A_1^{\lim}$. We thus obtain $A_2^* > A_1$ if and only if $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_2} < 0$. Furthermore, we obtain $A_2^* = 0$ if $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_2} < \frac{\mathrm{d}X_1^*}{\mathrm{d}A_2}$ since we only consider nonnegative levels of free allocation. By making use of the properties of comparative statics established in Lemma 3 (i), we directly obtain the properties of A_2^* as stated in the theorem.

As a **second step**, we determine the optimal emission cap T^*. The optimality condition for T^* has been derived in expression (13) and yields after substituting for A_2^*:

$$e^* - \mathcal{D}_{\mathcal{T}} = \frac{\mathrm{d}X_1^*}{\mathrm{d}\mathcal{T}} A_1 e - \frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}} A_1 e \left[\frac{\mathrm{d}X_1/\mathrm{d}A_2}{\mathrm{d}X_2/\mathrm{d}A_2}\right].$$

After substituting for $\frac{\mathrm{d}X_1}{\mathrm{d}\mathcal{T}}$, $\frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}}$, $\frac{\mathrm{d}X_1}{\mathrm{d}A_2}$, and $\frac{\mathrm{d}X_2}{\mathrm{d}A_2}$ [expression (24)], this reads as follows:

$$e^* - \mathcal{D}_{\mathcal{T}} = \left((\Psi_{I12}\Psi_{\mathrm{Ie}}) - (\Psi_{I1}\Psi_{\mathrm{IIe}}) \left[\frac{\left(-\Psi_{\mathrm{Ie}} \frac{\Psi_{E2}}{\Psi_{\mathrm{Ee}}}\right)}{-\left(\Psi_{I1} - \Psi_{\mathrm{Ie}} \frac{\Psi_{E1}}{\Psi_{\mathrm{Ee}}}\right)} \right] \right) \frac{A_1 e}{-\Psi_{\mathrm{Ee}}\, C}$$

Rearranging and plugging in for Ψ_{Ie} (compare Appendix A), we obtain:

$$e^* - \mathcal{D}_{\mathcal{T}} = \left(A_1^E - A_1\right) \frac{\left(\Psi_{I1}\left(\Psi_{\mathrm{IIe}} \frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\right) + \Psi_{I12}\left(\Psi_{I1} + \Psi_{\mathrm{Ie}} \frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right)\right) A_1 e}{\left(\Psi_{I1} + \Psi_{\mathrm{Ie}} \frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right) \Psi_{\mathrm{Ee}}\, C} \tag{47}$$

Observe that for $A_1 < A_1^{\lim}$, the sign of the right-hand side of expression (47) is entirely determined by the expression $(A_1^E - A_1)$, the remainder of expression (47) is strictly positive since $\left(\Psi_{I1} + \Psi_{\mathrm{Ie}} \frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right) < 0$ and $\left(\Psi_{\mathrm{IIe}} \frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\right) < 0$ (as shown further below in step three).

Finally, notice that expression (47) has been derived without nonnegativity constraint on A_2^*. As shown above, however, for $w_2 > w_2^E$ and $A_1 \in [A_1^E, A_1^{\lim}]$, we obtain $A_2^* = 0$ [instead of a negative value as resulting in the computations leading to expression (47)]. In this case, the optimality condition given by expression (13) is simplified as follows:

$$e^* - \mathcal{D}_{\mathcal{T}} = \left(\frac{\mathrm{d}X_1^*}{\mathrm{d}\mathcal{T}} - \frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}}\right) A_1 e = (\Psi_{I12}\Psi_{\mathrm{Ie}} - \Psi_{I1}\Psi_{\mathrm{IIe}}) \frac{A_1 e}{-\Psi_{\mathrm{Ee}}\, C} < 0 \tag{48}$$

The inequality is obtained since $\Psi_{\mathrm{Ie}} > 0$ and $\Psi_{\mathrm{IIe}} < 0$ for $w_2 > w_2^E$ and $A_1 \in [A_1^E, A_1^{\lim}]$ (compare Appendix A). We thus summarize the results obtained in expressions (47) and (54) as follows:

$$\begin{cases} e^* > D_T & \text{if } A_1 < A_1^E \\ e^* = D_T & \text{if } A_1 = A_1^E \\ e^* < D_T & \text{if } A_1 > A_1^E \end{cases} \tag{49}$$

As a **third step**, we finally show that the second-order conditions established in Lemma 2 (i) are indeed satisfied for all $A_1 \in [0, A_1^{\lim}]$, A_2^*. Since $\left(\Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right) < 0$ for $A_1 < A_1^{\lim}$, we just need to show that $\left(\Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right) < 0$ for $A_1 < A_1^{\lim}$. Notice first that we can rewrite Ψ_{IIe} (compare Appendix A) and thus obtain:

$$\Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} = \left(A_2^* - A_1 - \Psi_{E2}\right) \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} \tag{50}$$

Furthermore, we obtain for $(A_2^* - A_1)$ [compare expressions (12) and (24)]:

$$(A_2^* - A_1) = -\frac{\frac{\mathrm{d}X_1}{\mathrm{d}A_2}}{\frac{\mathrm{d}X_2}{\mathrm{d}A_2}} A_1 = \frac{\left(\Psi_{\text{Ie}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}}\right) A_1}{\left(\Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right)}$$

Plugging in allows us to rewrite expression (50) as follows:

$$\Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} = \frac{(\Psi_{E2})^2}{\left(\Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right)(-\Psi_{\text{Ee}})} \left(\Psi_{\text{Ie}} \frac{A_1}{-\Psi_{\text{Ee}}} - \left(\Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right)\right)$$

Substituting for $\Psi_{\text{Ie}} = A_1 - A_1^E$ and $\Psi_{E1} = A_1^E$ (see Appendix A) then yields:

$$\Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} = \frac{(\Psi_{E2})^2}{\left(\Psi_{I1} + \Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}}\right)(\Psi_{\text{Ee}})} \left(\Psi_{I1} - \frac{(A_1 - A_1^E)^2}{-\Psi_{\text{Ee}}}\right) < 0$$

We can thus conclude that the second-order conditions established in Lemma 2 (i) are satisfied if and only if $A_1 < A_1^{\lim}$ [Notice that the "only if" part follows directly from Lemma 2 (iii)].

H Proof of Theorem 6

As a **first step**, we determine the properties of the optimal allocation A_1^*. Observe that the optimality condition $A_1^* = \frac{\mathrm{d}X_2^*/\mathrm{d}A_1}{\mathrm{d}X_2^*/\mathrm{d}A_1 - \mathrm{d}X_1^*/\mathrm{d}A_1} A_2$ stated in the theorem directly is obtained by rearranging expression (11). In Lemma 3 (ii), we have established $\frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} > \frac{\mathrm{d}X_2^*}{\mathrm{d}A_1}$ for all $A_2 < A_2^{\lim}$. We thus obtain $A_1^* > 0$ if and only if $\frac{\mathrm{d}X_2^*/\mathrm{d}A_1}{<} 0$. Furthermore, we obtain $A_1^* > A_2$ if and only if $\frac{\mathrm{d}X_2^*}{\mathrm{d}A_1} < \frac{\mathrm{d}X_1^*}{\mathrm{d}A_1} < 0$. By making use of the properties of comparative statics established in Lemma 3 (ii), we directly obtain the properties of A_1^* as stated in the theorem.

As a **second step**, we determine the optimal emission cap T^*. The optimality condition for T^* has been derived in expression (13) and yields after substituting for A_1^* (compare the first step above):

$$e^* - \mathcal{D}_{\mathcal{T}} = \frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}}\left(A_2 - A_1^*\right)e^* + \frac{\mathrm{d}X_1^*}{\mathrm{d}\mathcal{T}}A_1^* e^* = \frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}}\frac{\frac{\mathrm{d}X_1}{\mathrm{d}A_1}A_2 e^*}{\frac{\mathrm{d}X_1}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2}{\mathrm{d}A_1}} + \frac{\mathrm{d}X_1^*}{\mathrm{d}\mathcal{T}}\frac{-\frac{\mathrm{d}X_2}{\mathrm{d}A_1}A_2 e^*}{\frac{\mathrm{d}X_1}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2}{\mathrm{d}A_1}}$$

We can now plug in for $\frac{\mathrm{d}X_1}{\mathrm{d}A_1}$ and $\frac{\mathrm{d}X_2}{\mathrm{d}A_1}$ as derived in expressions (22) and (23) and for $\frac{\mathrm{d}X_1}{\mathrm{d}\mathcal{T}}$ and $\frac{\mathrm{d}X_2}{\mathrm{d}\mathcal{T}}$ as derived in expressions (24) which yields:

$$\begin{aligned} e^* - \mathcal{D}_{\mathcal{T}} = {} & \frac{(-1)\ A_2\ (e^*)^2}{\left(\frac{\mathrm{d}X_1}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2}{\mathrm{d}A_1}\right)(-\Psi_{\mathrm{Ee}})\,C^2}\Bigg((\Psi_{\mathrm{Ie}}\Psi_{I I 2}) \\ & \left[\left(\Psi_{I1} + \Psi_{\mathrm{Ie}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right) + \left(\Psi_{\mathrm{IIe}}\frac{\Psi_{E1}}{-\Psi_{\mathrm{Ee}}}\right)\right] \\ & + (\Psi_{\mathrm{IIe}}\Psi_{I1})\left[\left(\Psi_{I I 2} + \Psi_{\mathrm{IIe}}\frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\right) + \left(\Psi_{\mathrm{Ie}}\frac{\Psi_{E2}}{-\Psi_{\mathrm{Ee}}}\right)\right]\Bigg) \\ = {} & (\Psi_{\mathrm{Ie}} + \Psi_{\mathrm{IIe}})\frac{(-1)\ A_2\ (e^*)^2}{\left(\frac{\mathrm{d}X_1}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2}{\mathrm{d}A_1}\right)(-\Psi_{\mathrm{Ee}})\,C} \end{aligned} \tag{51}$$

Now, define:

$$\Gamma_2^{\mathrm{em}}(A_2) := \Psi_{\mathrm{Ie}} + \Psi_{\mathrm{IIe}} = A_2 - (A_1^E + A_2^E). \tag{52}$$

Observe that $\Gamma_2^{\mathrm{em}}(A_2^{\mathrm{em}}) = 0$. Furthermore, notice that $A_2^{\mathrm{em}} < 0$ for $w_2 < w_2^L$, that is, $\Gamma_2^{\mathrm{em}}(A_2) > 0$ for all $A_2 \geq 0$ whenever $w_2 < w_2^L$. In order to compare A_2^{em} to the previously established critical levels of initial allocation, we make the following two observations:

$$\Gamma_2^{\mathrm{total}}(A_2^{\mathrm{em}}) = \Psi_{I I 2} < 0 \qquad \text{and} \qquad \Gamma_2^{\mathrm{cross}}(A_2^{\mathrm{em}}) = \Psi_{I1} < 0$$

This allows to directly conclude that $A_2^{\mathrm{em}} < A_2^{\mathrm{total}}$ and $A_2^{\mathrm{em}} < A_2^{\mathrm{cross}}$.

By making use of the newly introduced Γ_2^{em}, we can rewrite expression (51) as follows:

$$e^* - \mathcal{D}_{\mathcal{T}} = -\Gamma_2^{\mathrm{em}}(A_2)\frac{A_2 e^*}{\left(\frac{\mathrm{d}X_1}{\mathrm{d}A_1} - \frac{\mathrm{d}X_2}{\mathrm{d}A_1}\right)(-\Psi_{\mathrm{Ee}})\,C} \tag{53}$$

Finally, notice that expression (53) has been derived without nonnegativity constraint on A_1^*. As shown in step one of the present proof, however, for $w_2 < w_2^S$ and $A_2 \in [A_2^{\text{cross}}, A_2^{\text{lim}}]$ we obtain $A_1^* = 0$ [instead of a negative value as resulting in the computations leading to expression (53)]. In this case, the optimality condition given by expression (13) is simplified as follows:

$$e^* - \mathcal{D}_{\mathcal{T}} = \frac{\mathrm{d}X_2^*}{\mathrm{d}\mathcal{T}} A_2 e^* = \frac{\Psi_{I1}\Psi_{\text{IIe}}}{-\Psi_{\text{Ee}}\, C} A_2 e^* = \frac{\mathrm{d}X_2^*}{\mathrm{d}\mathcal{T}} A_2 e^* = \left(A_2 - A_1 - A_2^E\right) \frac{\Psi_{I1}\, A_2 e^*}{-\Psi_{\text{Ee}}\, C} < 0 \tag{54}$$

Observe that the above inequality is satisfied since $A_2 \geq A_1 + A_2^E$ whenever $A_2^{\text{cross}} \leq A_2 \leq A_2^{\text{lim}}$ [compare expression (34); remember that $A_1 = A_1^* = 0$ in the case considered].
We can thus establish the following results for the optimal cap on total emissions:

$$\begin{cases} e^* > D_T & \text{if } A_2 < A_2^{\text{em}} \\ e^* = D_T & \text{if } A_2 = A_2^{\text{em}} \\ e^* < D_T & \text{if } A_2 > A_2^{\text{em}} \end{cases} \tag{55}$$

As a **third step**, we finally show that the second-order conditions established in Lemma 2(i) are indeed satisfied for all $A_2 \in [0, A_2^{\text{lim}}]$, A_1^*. Remember we obtained for A_1^*:

$$A_1^* = \frac{\Gamma_2^{\text{cross}}}{\Gamma_2^{\text{lim}}} A_2 = \frac{\Gamma_2^{\text{cross}}}{\Gamma_2^{\text{cross}} + \Gamma_2^{\text{total}}} A_2$$

In order to verify the second-order conditions established in Lemma 2(i) (a), (b), and (c), we now separately analyze the following cases:

- First, observe that

$$\Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} = (A_2 - A_1 - \Psi_{E2}) \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} = \left(\frac{\Gamma_2^{\text{total}}}{\Gamma_2^{\text{lim}}} A_2 - \Psi_{E2}\right) \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} \tag{56}$$

  – For $w_2 < w_2^E$ (i.e., $\Psi_{E2} < 0$), expression (56) is negative since $\Gamma_2^{\text{total}} < 0$ and $\Gamma_2^{\text{lim}} < 0$ if $A_2 < A_2^{\text{lim}}$ and $w_2 < w_2^E$ (compare Appendix C).
  – For $w_2 > w_2^E$ (i.e., $\Psi_{E2} > 0$) and $A_2^{\text{total}} \leq A_2 \leq A_2^{\text{lim}}$, expression (56) is negative since $\Gamma_2^{\text{total}} \geq 0$ and $\Gamma_2^{\text{lim}} < 0$.

  Whenever expression (56) is negative, this directly implies that condition (b) is satisfied. Since, furthermore, $A_2 < A_2^{\text{lim}}$, also conditions (a) and (c) are satisfied.
- Second, for $A_2^{\text{cross}} \leq A_2 \leq A_2^{\text{lim}}$ we obtain $A_1^* = 0$ (compare step one of the present proof). We thus directly obtain:

$$\Psi_{\text{Ie}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}} = -A_1^E \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}} < 0 \tag{57}$$

This directly implies that condition (a) is satisfied. Since, furthermore, $A_2 < A_2^{\text{lim}}$, also conditions (b) and (c) are satisfied.

- Third, for $w_2 > w_2^E$ (i.e., $\Psi_{E2} > 0$) and $0 \leq A_2 \leq min(A_2^{\text{total}}, A_2^{\text{cross}})$
  - Whenever $A_2 < A_1 + A_2^E$ (i.e., $\Psi_{\text{IIe}} < 0$), we directly obtain $\Psi_{\text{IIe}} \frac{\Psi_{E2}}{-\Psi_{\text{Ee}}} < 0$. This directly implies that condition (b) is satisfied. Since, furthermore, $A_2 < A_2^{\text{lim}}$, also conditions (a) and (c) are satisfied.
  - Whenever $A_2 \geq A_1 + A_2^E$ (i.e., $\Psi_{\text{IIe}} \geq 0$), then $\Psi_{\text{IIe}} \frac{\Psi_{E1}}{-\Psi_{\text{Ee}}} > 0$. Since $\Gamma_2^{\text{cross}} < 0$, in the region considered this directly implies that condition (a) is satisfied. Since, furthermore, $A_2 < A_2^{\text{lim}}$, also condition (b) is satisfied. Finally, since conditions (a) and (b) are satisfied and $\Gamma_2^{\text{cross}} < 0$ and $\Gamma_2^{\text{total}} < 0$, also condition (c) is satisfied.

We can thus conclude that the second-order conditions as established in Lemma 2(i) are satisfied for all $0 \leq A_2 \leq A_2^{\text{lim}}$.

References

Anouliès L (2017) Heterogeneous firms and the environment: a cap-and-trade program. J Environ Econ Manag 84:84–101

Bovenberg AL, Goulder LH, Jacobsen MR (2008) Costs of alternative environmental policy instruments in the presence of industry compensation requirements. J Public Econ 92(5):1236–1253

Bovenberg L, Goulder LH, Gurney DJ (2005) Efficiency costs of meeting industry-distributional constraints under environmental permits and taxes. RAND J Econ 36(4):950–970

Böhringer C, Lange A (2005) On the design of optimal grandfathering schemes for emission allowances. Eur Econ Rev 49(8):2041–2055

Böhringer C, Rosendahl KE, Storrøsten HB (2017) Robust policies to mitigate carbon leakage. J Public Econ 149:35–46

Coase RH (1960) The problem of social cost. J Law Econ 3(1):1–44

Convery F (2009) Origins and development of the eu ets. Environ Resour Econ 43:391–412

Cramton P, Ockenfels A (2012) Economics and design of capacity markets for the power sector. Zeitschrift für Energiewirtschaft 36

Cramton P, Stoft S (2008) Forward reliability markets: Less risk, less market power, more efficiency. Utilities Policy 16(3):194–201

Crew MA, Fernando CS, Kleindorfer PR (1995) The theory of peak-load pricing: A survey. J Regul Econ 8(3):215–48

Damon M, Cole DH, Ostrom E, Sterner T (2019) Grandfathering: environmental uses and impacts. Rev Environ Econ Policy 13(1):23–42

Dardati E, Saygili M (2020) Aggregate impacts of cap-and-trade programs with heterogeneous firms. Energy Econ 92:104924

Fabra N (2018) A primer on capacity mechanisms. Energy Econ 75:323–335

German-Parliament (2007) Gesetz zur Aenderung der Rechtsgrundlagen zum Emissionshandel im Hinblick auf die Zuteilungsperiode 2008 bis 2012, see https://www.buzer.de/gesetz/7848/index.htm

Grimm V, Schewe L, Schmidt M, Zöttl G (2017) Uniqueness of market equilibrium on a network: A peak-load pricing approach. Eur J Oper Res 261(3):971–983

Grubb M, Neuhoff K (2006) Allocation and competitiveness in the eu emissions trading scheme: Policy overview. Climate Policy 6:7–30

Harstad B, Eskeland GS (2010) Trading for the future: Signaling in permit markets. J Public Econ 94(9):749–760

ICAP (2020) Emissions trading worldwide, international carbon action partnership (icap) status report 2020, see https://icapcarbonaction.com/en/icap-status-report-2020

IEA (2010) Reviewing existing and proposed emissions trading systems, see https://www.oecd-ilibrary.org/content/paper/5km4hv3mlg5c-en

IEA (2020) Implementing effective emissions trading systems, see https://www.oecd-ilibrary.org/content/publication/b7d0842b-en
Mackenzie I, Hanley N, Kornienko T (2008) The optimal initial allocation of pollution permits: a relative performance approach. Environ Resour Econ 39(3):265–282
Meunier G, Montero J-P, Ponssard J-P (2017) Using output-based allocations to manage volatility and leakage in pollution markets. Energy Econ 68:57–65
Meunier G, Montero J-P, Ponssard J-P (2018) Output-based allocations in pollution markets with uncertainty and self-selection. J Environ Econ Manag 92:832–851
Moledina AA, Coggins JS, Polasky S, Costello C (2003) Dynamic environmental policy with strategic firms: prices versus quantities. J Environ Econ Manag 45:356–376. https://doi.org/10.1016/S0095-0696(02)00055-4
Montero J-P (2002) Permits, standards, and technology innovation. J Environ Econ Manag 44(1):23–44
Narassimhan E, Gallagher KS, Koester S, Alejo JR (2018) Carbon pricing in practice: a review of existing emissions trading systems. Climate Policy 18(8):967–991
Neuhoff K, Martinez KK, Sato M (2006) Allocation, incentives and distortions: the impact of eu ets emissions allowance allocations to the electricity sector. Climate Policy 6(1):73–91
Pahle M, Fan L, Schill W-P (2011) How emission certificate allocations distort fossil investments: The german example. Energy Policy 39(4):1975–1987
Qiu R, Xu J, Zeng Z (2017) Carbon emission allowance allocation with a mixed mechanism in air passenger transport. J Environ Manage 200:204–216
Requate T (2005) Dynamic incentives by environmental policy instruments–a survey. Ecol Econ 54(2):175–195
Requate T, Unold W (2001) On the incentives created by policy instruments to adopt advanced abatement technology if firms are asymmetric. J Inst Theor Econ 157:536
Requate T, Unold W (2003) Environmental policy incentives to adopt advanced abatement technology: Will the true ranking please stand up? Eur Econ Rev 47(1):125–146
Rosendahl KE (2008) Incentives and prices in an emissions trading scheme with updating. J Environ Econ Manag 56(1):69–82
Tietenberg T (2006) Emissions trading: principles and practice, 2nd edn. Routledge, UK
UNO (2015) Paris agreement, see https://unfccc.int/process/conferences/pastconferences/paris-climate-change-conference-november-2015/paris-agreement
Zöttl G (2010) A framework of peak load pricing with strategic firms. Oper Res 58(6):1637–1649