View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Columbia University Academic Commons

Rescaling Capital:

The Potential of Small-Scale and Mass-Produced Physical Capital in the Energy and Materials Processing Industries

Eric Dahlgren

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2014

© 2013

Eric Dahlgren

All rights reserved

Abstract

Rescaling Capital:
The Potential of Small-Scale and Mass-Produced Physical Capital in the Energy and Materials Processing Industries

Eric Dahlgren

Observing the evolution of size of physical capital in fundamental infrastructure and processing industries such as energy, mining, and chemical processing, etc, over the last century suggests the prevalence of an unambiguous mantra – "bigger-is-better." This dissertation questions some of the underlying arguments supporting this apparent orthodoxy. Moreover, arguments are put forth highlighting the potential in substantially diverting from this monolithic approach to productive capital and instead focus on a route marked by mass production of small-scale units. Such a shift would most likely herald transformational technology solutions to industries that have long been considered mature.

One of the underlying drivers for scaling up in unit size rests on the empirical observation that fixed costs of productive capital generally increase only sub-linearly with size. Arguments suggesting that this trend, typically referred to as the "two-thirds-rule," inherently favors a large unit scale on the basis of material consumption are rejected on physical grounds in this dissertation. With the number of units produced a different form of cost reduction can be attained – through learning. Classifying technologies as either small or large based on the number of end consumers, a meta-study concludes that small-scale technologies learn substantially faster. In fact, comparing the two empirical formulations of cost reductions that typically accompany scaling up in size and scaling up in numbers reveals

almost identical levels of cost scaling with aggregate capacity.

To investigate the possible existence of operational returns to unit scale a case study in four different electricity generating technologies in the U.S. (coal, combined cycle, gas turbine and nuclear) is performed. With only one exception, these technologies exhibit a weak but significant trend of decreasing operational costs with unit (generator) size. However, this trend disappears, or is even reversed, once labor costs are subtracted from total cost. Thus, the relatively recent advent of low-cost automation technologies removes the main impetus to keep increasing unit scale from the perspective of operational cost. This conclusion from a statistical analysis of internally very different technologies suggests wider applicability. At least, it cannot be dismissed outright in other sectors.

Abandoning large-scale and custom-made capital in favor of a small-scale and mass-produced variety will likely be accompanied by several heretofore new features. Two foreseen such features are shorter lifetime and lead time of investments. These two features will bring increased flexibilities of engagement and disengagement in a given market. The introduction herein of a real options model aims to quantify this flexibility. Among other applications, the introduced framework can be deployed to estimate the critical investment cost to render a small-scale solution competitive with a large-scale counterpart of known cost.

A more detailed analysis of reverse osmosis desalination technology is performed from the perspective of unit scale. Studying transfer phenomena in a thin rectangular channel with semipermeable walls, simulating the conditions in commercial operation, reveals non-intuitive conclusions regarding optimal operating conditions in this technology. Not only would a shorter feed channel (small scale) result in reduced specific energy consumption in the separation stage, it would also suggest operating at lower recovery rates. The findings here suggest that operating at a smaller unit scale entails more than simply scaling down existing process units, rather, all steps need to be reevaluated.

Contents

List of Figures

List of Tables

Acknowledgements

I like to acknowledge several people without whose support, advice and friendship the completion of this dissertation would have been a much more difficult journey.

My co-authors on the papers (Dahlgren et al., 2013; Dahlgren and Lackner, 2012; Dahlgren and Leung, 2013; Dahlgren and Lackner, 2013), Caner Göçmen, Professor Tim Leung and Professor Garrett van Ryzin have with great stride and patience introduced me to their respective disciplines. Their unique insights and assistance cannot be overstated. I would also like to thank Yoachim Haynes for the help in collecting and analyzing volumes of data.

Fellow students in the Lackner group, past and present, have not only provided intellectual support but also made the everyday experiences at the Lenfest Center memories for life. A special thanks to Carey Russell for her help in proof reading this dissertation, but more importantly for being a great friend over the years. Thank you, Allen Wright, Christoph Meinrenken, and Sarah Brennan for letting me experience your genuinely positive personalities.

Finally, with unmatched fervor and joy for everything scientific, my adviser Professor Klaus Lackner has been a unique source of inspiration and motivation. His unique approach to science will forever serve as a model for the exceptional. I am proud to have worked in his great shadow and I am honored to call Klaus my mentor and friend.

Chapter 1

Introduction

Are the basic energy and materials processing industries bound to follow their historic general path of scaling up the size of individual technologies? Based on physical arguments, general economic arguments, financial arguments and with the help of individual technology case studies, as the title implies, I posit that there are today no sound arguments why this question should be answered in the affirmative. In fact, several benefits can be attained through a paradigm shift towards radically smaller-scale units.

To reduce this question in tractable components it is necessary to first define the key concept of *unit scale*. By unit scale, I refer to the capacity of an isolated unit of technology, e.g. the nameplate rating of a single generator in a power plant, the payload capacity of an individual mining haul truck etc. Importantly, the unit scale is the scale of the irreducible functioning component in an industrial implementation. Generally, unit scale thusly defined is positively correlated with the physical size of the equipment, which allows me to use unit scale and unit size synonymously unless further specification is required.

Some technologies are associated with a natural scale, or at least bounds on the unit scale. For instance, the transportation sector finds a natural minimum scale tied to the size of a human being. The existence of such bounds introduces unavoidable *indivisibilities* in

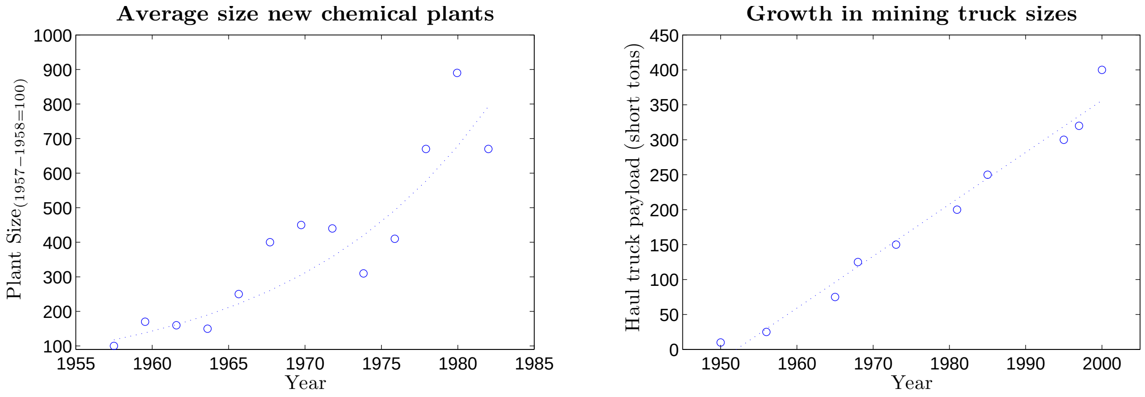

Figure 1.1: *Evolution of unit size in chemical processing industry (left), data adapted from (Lieberman, 1987), and the size of mining trucks (right), data adapted from (Koellner et al., 2004).*

the production of capital goods and a discussion on unit scale gets constrained. However, the energy and materials processing industries are almost exclusively devoid of such inherent indivisibilities. This observation makes the discussion of unit scale relevant. More precisely, while these industries have opted to provide additional capacity by scaling up in unit size, the underlying tenor in this thesis is the possibility of instead scaling up in unit numbers.

To exemplify the trend of “bigger-is-better”, consider the evolution of unit size in the two industries/technologies depicted in Figure 1.1. For all practical purposes, the chemical processing industry (which in the figure comprises the production of 22 different bulk chemicals (Lieberman, 1987)) is very different from mine-site ore transportation. Yet they both have trended towards very large sizes. Similar trends are noticeable in other sectors as well, suggesting the existence of general forces that favor large unit sizes.

The strategy of scaling up in unit size is often conflated with the strategy of gaining economies of scale to the point that they are sometimes used synonymously. However, many of the benefits that are attributed to economies of scale are, in fact, a function of firm-wide, or even industry-wide activities and not the scale of individual units of production. This realization calls for a reevaluation of the underlying causes. In this thesis, I will refer to

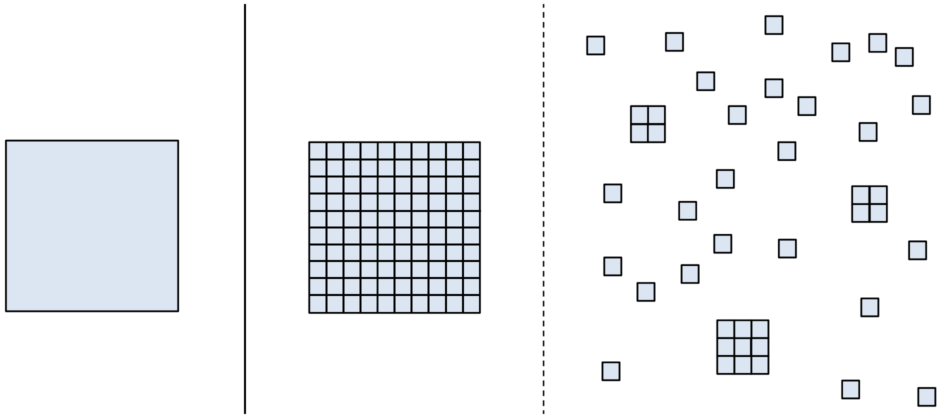

Figure 1.2: *Large-scale units are by necessity centralized. Small-scale modular units allows for distributed operation but retains the option of centralization.*

economies of unit scale, those benefits that are directly attributable to unit size. An example of economies of unit scale is the empirical observation that the investment cost in capacity tends to increase only sub-linearly with size. This observation has been elevated to the form of a rule, variably called the "two-thirds rule," "0.7 rule", names referring to the value of the exponent in the power law that is used to estimate how cost scale with size. Other benefits that in the existing paradigm could be attributable to increasing unit scale is increased labor productivity and increased conversion efficiency.

While the observed economies of unit scale and related cost reductions stemming from scaling up in unit size certainly are real, other means of cost reductions are available in industries that scale up in numbers instead. Similar in guise to the empirical rules stipulating how cost scale with size, *learning curves* are often used in mass production-oriented industries to estimate the cost as cumulative production grows. The most famous such example, albeit formulated slightly differently, is Moore's law, referring to the observation of a doubling of component density in integrated electronics every 18 months. More generally, the mass production process can exhibit exponentially decreasing cost as production increases.

Several benefits that are commonly lumped under the umbrella of economies of scale are not inherently dependent on unit scale. For instance, the operational benefits that accrue with increased centralization come naturally when scaling up, e.g. transportation cost of inputs and outputs, security, administration, etc. Nothing prevents those benefits to be garnered by aggregating a large number of small unit in the same location. However, potential benefits from decentralized operation are only possible if capacity comes in small-scale and modular units. The computer industry serves as an illustrative example of such benefits. Being on a trajectory towards increasing unit size, these super-machines suddenly found a superior competitor in the mass-produced PC and the industry diverted to smaller unit sizes. Improvements in usability meant that the user herself could tap the ever increasing utility of electronic devices, making the notion of labor cost moot. Except for consumer products, any strategy relying on scaling up in numbers is confronted by issues related to increased labor cost and complexity. Only recently have automation technologies reached sufficient levels of cost and ability to make this strategy viable.

Currently, custom-made capacity additions in the industries in question are usually preceded by several years in planning and construction. Once operational, these large-scale projects require decades of profit generation to reach financial viability. This inertia presents barriers to entry and consequently arguably also to innovation. In a paradigm marked by modularity and mass-production, capacity can be dispatched continuously and silo-based, cost reduction-driven innovation can more easily be replaced by collaborative innovation between different enterprises opening up new markets and uses. This concept, referred to as “pull innovation” by Weaver (2008), is much facilitated by a small unit scale.

Technologies that simultaneously exist commercially on both a large scale and on a small scale are rare. However, to illustrate the potential of mass-production and small scale, the observation made originally Klaus Lackner in the comparison of the internal combustion car engine and a single-cycle thermal power plant is thought provoking. Both technologies have existed since at least the early 20th century and both perform the same job of converting chemical energy into mechanical work. Moreover, under ideal conditions they do so at comparable efficiencies. These two technologies have two main distinguishing although related features. First, the car engine has a power rating on the order of 100 kW, whereas the individual generator in the power plant has evolved to sizes on the order of 100 MW. Additionally (and consequently), the small scale of the engine makes it amenable to mass-production and the large scale of the power plant implies more customized production. Interestingly, the car engine is about two orders of magnitude cheaper per unit capacity, $10/kW vs. $1000/kW, (Larminie and Dick, 2003; EIA, 2010).

1.1 Background

While the scope and implications of this thesis extends beyond the traditional purview of environmental engineering, the main motivation for investigating unit scale finds its roots in issues closely related to affairs of the Department of Earth & Environmental Engineering. Controlling mass and energy flows that are relatively dilute on a large scale forces the issue of unit scale into consideration. For instance, air capture of CO_2, likely a necessary technology to mitigate climate change in the long term, finds no inherent justification for a large unit scale. Such a technology ought instead be sized according the end use of the CO_2, whether it is sequestration or recycling. Resource extraction is another area which could benefit from small, modular and distributed operation. Such a trend has recently become visible with increased attention paid to hydrocarbons trapped in shale formations. However, the notion extends further. Ore bodies that today are currently too small to develop could see a resurgence with a paradigm shift to small-scale technologies.

Small-scale technologies have taken hold in the renewable energy sector. Indeed, the benefits of distributed electricity systems powered by renewables have been getting substantial attention in recent years, see e.g. Williams et al. (2012); Martinot et al. (2007); Lovins et al. (2002). By deviating from the old paradigm of large-scale centralized generating stations, several possibilities arise that are not immediately linked to any one technology, e.g. combined heat and power and smart grids. These kinds of synergistic technology solutions can be seen as examples of the “pull innovation” mentioned earlier.

Acknowledged as the linchpin to even contemplate small-scale technologies in the energy and materials processing industries is automation. Moreover, it also recognized that this term, as used in the context throughout this dissertation, encompasses many different technologies, e.g. sensor technology, robotics, data communication and responsive networks (Vrba, 2013; Luo and Chang, 2012; Luettel et al., 13; Petrina, 2011; Mohan and Ponnambalam, 2009). Studying these technologies more closely is outside the scope of this work. Nonetheless, it is here generally assumed that automation, as a strategy to reduce or even remove human labor, is technically feasible today or in the near future.

Part of this work relates to established neoclassical economic theory of economies of scale, see (Solow et al., 1966; Panzar and Willig, 1977; Edwards and Starr, 1987). More recent contributions, (Lipsey et al., 2005; Carlaw, 2004; Tone and Sahoo, 2003), suggesting complementing the notion of economies of scale beyond the concept of the production function are also reviewed. Engineering aspects of economies of scale, more appropriately referred to as economies of unit scale, have been examined in part by Srikanth and Funk (2011); Funk (2010); Jack (2009); Hisnanick and Kymn (1999); Humphreys and Katell (1981). The concept of learning curves has been studied extensively, see e.g. (Ferioli and van der Zwaan, 2009; Argote and Epple, 1990; McDonald and Schrattenholzer, 2001). Ample data on learning available in the existing literature made possible a meta study of how learning relates to unit scale.

The notion of adopting a financial asset pricing framework to value real investments have been established for quite some time, see e.g. (Dixit and Pindyck, 1994; Pindyck, 1986; McDonald and Siegel, 1986). However, actual implementation of real options analysis as a valuation tool in infrastructure industries, such as energy, is only recently gaining traction. The framework herein is focused on features of small scale and is therefore a natural extension to recent studies (Westner and Madlener, 2012; Herder et al., 2011; Frayer and Uludere, 2001; Kaslow and Pindyck, 1994). Lastly, the contribution made in analyzing the mass transport and energy consumption in reverse osmosis desalination firmly builds on recently published work, e.g. (Song, 2010; Greenlee et al., 2009; Guillen and Hoek, 2009).

1.2 Overview

It is the general purpose of this thesis to highlight the possibilities that accompany a small unit scale. These possibilities are often overlooked due an institutional bias toward large scale. 'Scale-up,' referring to size, is an ingrained concept among engineers, which suggests that technologies that do not exhibit a propensity for ballooning in size are overlooked. Instead, any given technology should be investigated and researched without blinders to the small side of the size spectrum.

This dissertation can considered to consist of two parts. In Chapters 2 through 4, general economic and technology-agnostic engineering aspects are addressed as they relate to unit scale. More precisely, the contribution made here is on account of the following questions:

1. Does the 'two-thirds law' find support in fundamental physics?

2. Is there a significant difference in learning rates between large and small scale technologies?

3. To what extent do increased labor productivity and conversion efficiency account for operational economies of unit scale?

The first question addresses claims made in the literature that would render a small scale approach inferior due to increased resource consumption in the production of capital goods. In Chapter 2, such claims are generally dismissed through a detailed analysis on the basis of continuum mechanics. Later in the same chapter, the second question is resolved through a meta analysis on published learning rates for a wide variety of different technologies. By classifying technologies as either large or small based on the number of end users connected to individual units, it is concluded that small technologies, on expectation, have a 10 percentage points higher learning rate. Operational or variable cost are less tractable to address from a general perspective due to the distinct nature of different technologies. However, a sweeping discussion on operational cost as they relate to scale, including a simple model illustrating the potential for distributed operation, is brought forward in Chapter 3. Connected to the issues of scale-related labor intensity and conversion efficiency, a statistical analysis on the operational cost in the four main electricity generating technologies is performed in Chapter 4. The main conclusion here is that there are indeed positive operational returns to unit scale in these technologies. However, these are mainly explained by increasing labor productivity alone. Once the labor cost is removed these scale economies vanish. The technologies in this case study are relatively diverse, suggesting that this conclusion can not be immediately dismissed in other industries.

The second part of this dissertation, Chapters 5 and 6, focuses on the following two queries:

4. What financial flexibilities arise due to the perceived shorter lifetime and shorter lead time of mass-produced physical capital?

5. What benefits arise when scaling down reverse osmosis desalination technologies?

A paradigm shift towards mass production of small-scale technologies is likely to entail capacity with shorter lifetimes and more rapid deployment. This presents the firm with increased flexibility to engage a given market due to shorter lead times. It also allows the firm to more easily disengage without stranding a significant investment meant to last for decades. To capture these flexibilities a real options model is introduced. This model can be employed to estimate the critical cost where capital marked by short lifetimes and lead times (i.e. small-scale technologies) is competitive with its current paradigm counterpart.

To complement the technology-independent tenor in this dissertation, a detailed study of a specific technology, reverse osmosis desalination, is performed in Chapter 6. The focus is on the actual separation stage in the process, which is the main energy-consuming step. By carefully investigating transport phenomena in the cross-flow filtration, it is concluded that reducing size (shorter feed channels) presents the opportunity of reducing specific energy consumption in this technology that is foreseen to increase in demand.

The greatest benefits of properly incorporating small-scale thinking will likely be captured in novel technologies, rather than by scaling down replicas of existing implementations. However, with no intention of being exhaustive, three specific technologies or technology classes are briefly reviewed in Chapter 7 (ammonia synthesis, fuel synthesis and mining), which would benefit from scaling down. The common theme among these technologies is that they have today evolved into very large unit sizes even though the physics involved would favor a smaller scale. Moreover, the geographically distributed nature of the inputs and/or the demand for the outputs in these technologies suggests a propensity for distributed operation.

Chapter 2

Fixed Costs and Unit Scale

The energy and materials processing industries share one crucial feature; trafficking in commodities, they all enjoy complete divisibility of material inputs and outputs. No physical law outright prohibits natural gas to be extracted, processed and oxidized for electricity generation one mol at a time. Furthermore, there is nothing that prevents aggregation and disaggregation of the products at any step. Yet, notwithstanding this extreme example, the individual process units in these industries have, with only a few and relatively recent exceptions, evolved into larger and larger sizes. This suggest the existence of general size economies at work with the consequence that scaling up garners competitive advantages and, therefore, has dictated technological evolution.

Technical progress, as it relates to the scale of a unit technology, can be viewed as the confluence of three different forms of innovations according to Sahal (1985). Scaling up (or down) a process, or a single piece of equipment, consisting of different components requires *structural innovation* to accommodate the different internal scaling behaviors. Additionally, *material innovation* is typically required to facilitate the physics of the process over different size ranges. For instance, as will be discussed further below, building larger wind turbines requires lighter and stronger material in order for the larger structure not to succumb to the relentless pull of gravity and other inertial forces. The third form of innovation mentioned is *systems innovation*, which is related to the integration of symbiotic technologies.

To this observer, it is very likely, if not even certain, that the relatively recent advent of low-cost automation technologies has made possible a quantum leap in the last category. Thirty years ago, the notion of simultaneous control and operation of thousands of parallel processes was generally not practically conceivable. Today's ability of distributed computing, improved sensory technologies and wireless data transmission, to name a few of the many features involved in automation and robotics, have dramatically altered that view. It is therefore highly plausible that scaling up in numbers rather than size today offers economies of scale different from those enjoyed with the scale-up strategy of the last century.

Neoclassical economic theory introduces the concept of economies of scale in terms of the degree of homogeneity of the production function and the related cost function, see e.g. (Panzar and Willig, 1977; Solow et al., 1966). This homogeneity measures the response in output when all production inputs (labor, capital, land, fuel, etc.) are increased with the same factor λ. Economies of scale, or increasing returns to scale, are said to occur if the output increases more than the factor λ, and dis-economies of scale are exhibited if the output scales less than λ. Such an approach overlooks the importance of potential indivisibilities of the inputs, particularly labor. The economies of scale resulting from indivisibility of labor was noted already by Adam Smith (Edwards and Starr, 1987) and have arguably been a key driver of the trend of scaling up in size. At least, this indivisibility has conspired against scaling up in numbers. This discrimination against large numbers is voided by automation to a great extent.

Another shortcoming of classical theory is that it tends to conflate the notions of the firm and the production unit. Indeed, more recent literature suggest expanding the view on factors effected by scale in production (Lipsey et al., 2005; Carlaw, 2004; Tone and

Sahoo, 2003). Crucially, the positive economies that can be attributed to the operation and acquisition of the physical capital, here referred to as *economies of unit scale*, ought to be distinguished from those stemming from firm-wide undertakings. Structuring such a taxonomy on the basis of the physical scale of individual pieces of equipment is complicated by non-exclusiveness. For instance, the reduced overhead and indivisibilities that accompany centralization is attained automatically when scaling up in size. However, the same benefits can be achieved by agglomerating many small units in one location. On the other hand, the possibility of decentralized operation is only achievable by scaling up in numbers. It is the goal here to investigate and catalog those attributes that are inherently scale dependent.

2.1 Cost Reductions by Scaling Up in Unit Size

Focusing first on the production cost of capital goods, it is a well-known observation that many pieces of process equipment increase only sub-linearly in cost when scaling up in size. A traditional method of estimating the cost $k(c)$ of a piece of equipment with capacity c uses a power law:

$$k(c) = k_{\text{ref}} \left(\frac{c}{c_{\text{ref}}} \right)^{\alpha}, \tag{2.1}$$

where k_{ref} is the cost of a reference unit with capacity c_{ref}. Positive returns to scale in construction are signified with values of α less than 1. The fact that the scale parameter α for many different pieces of equipment, as well as aggregated pieces of machinery, has been estimated in the range $0.6 - 0.8$ has deputized the empirical relationship in (2.1) to the form of a rule with names like "0.6 rule," "0.7 rule" or sometimes "two-thirds rule," frequently occurring in the literature on engineering cost estimates (Humphreys and Katell, 1981; Jenkins, 1997; Euzen et al., 1993).

Several factors contribute to this trend of cost decreasing with size. These are primarily the overheads and indivisibilities that go into the production of custom-made capital goods. Within a given size regime, the fixed cost of planning and design, as well as ancillary components like control and monitoring equipment are insensitive to unit size. Spreading out these costs over a large capacity naturally reduces their contributions on a per unit output basis. One commonly encountered explanation to the observed scaling law in (2.1) rests on the geometric relationship between surface area and enclosed volume at different sizes (Srikanth and Funk, 2011; Funk, 2010; van Mieghem, 2008; Lipsey et al., 2005; Haldi and Whitcomb, 1967; Husan, 1997; Tribe and Alpine, 1986). With capacity typically being proportional to the enclosed volume, this argument posits that the amount of material, and therefore cost, necessary to construct the equipment scales with the surface area. Increasing the capacity with a factor λ^3 by uniformly scaling all linear dimensions with a factor λ would consequently increase cost by λ^2, offering a tidy explanation to the often observed values of $\alpha = 2/3$ in (2.1).

Were such an argument indeed valid, any strategy that seeks to scale up in numbers would run afoul in terms of material consumption, a potentially insurmountable constraint. However, scaling observed in nature indicates critical flaws with this logic, where the bones of an elephant are disproportionately thicker than those in a mouse. A careful investigation of the mechanics involved actually reveals that uniform scaling is generally not possible and that scaling up in size, just like with the elephant, necessitates a dis-proportionate increase in materials.

2.1.1 Scaling of Linear Elastic Structures

The implications of trying to uniformly scale a solid structure are here investigated from first principles. The underlying theory can be found in any standard mechanics text book, see e.g. (Lai et al., 1993). Without making any greater sacrifice to the generality of the conclusion, the discussion is here limited to statics. A more detailed presentation, including the dynamics of the system, can be found in (Dahlgren and Lackner, 2012).

Problem definition

A continuum can be characterized by its material distribution, or density field, in three-dimensional space. The equations of motion, which in the case of statics merely take the form of a force balance, can be stated as

$$\partial_i \sigma_{ij} + F_j \;\; = \;\; 0, \tag{2.2}$$

where σ is the stress tensor and where F denotes the body forces, e.g. gravitational. Considering the continuum as a solid structure, the stresses throughout the solid depends on material properties. More generally, when the structure is subjected to forces, internal and/or external, strains will appear in the solid. The structures of interest here are engineered pieces of equipment designed to operate within the elastic limit. That is, a structure operated in the elastic regime will revert back to its original shape once the applied forces are removed. If a particle of this distribution originally is located at x_0, the position $x(x_0)$ of the same particle after the structure is distorted can be described with the help of the displacement field $u = x(x_0) - x_0$. Most materials encountered in practice exhibit a linear elastic behavior, and the displacements under normal operation can be considered very small. Under those conditions, the stresses and strains in the structure are given by

$$\begin{cases} \sigma_{ij} & = & C_{ijkl}\epsilon_{kl}, \\ \epsilon_{ij} & = & \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right), \end{cases} \tag{2.3}$$

where the strain, expressed through the tensor ϵ, is related to the stresses by the elasticity tensor C. The partial differential equation in u given by (2.2)-(2.3) is defined on the domain $\mathcal{U}$, with boundary $\partial\mathcal{U}$. To completely describe this problem, boundary conditions are required. These conditions are typically of both Neumann and Dirichlet type

$$\begin{cases} \sigma(x)\vec{n}(x) & = & T(x), \quad x \in \partial\mathcal{U}^T, \\ u(x) & = & \overline{u}(x), \quad x \in \partial\mathcal{U}^u, \end{cases} \tag{2.4}$$

where a traction, or normal contact force, T is specified on one part of the boundary $\mathcal{U}^T$ and where the displacement is given on the complementary part $\mathcal{U}^u$. Equations (2.2)-(2.4) completely determines the problem of static deformation of a body in the linear elastic regime. Note that no conditions on isotropy of the material have been specified, i.e. both ρ and C are both allowed to vary over $\mathcal{U}$.

The scaled problem

Given a structure contained in $\mathcal{U}$, subject to equations (2.2) - (2.4) and permitting a solution u, the goal is now to investigate the behavior of a similar structure where all spatial dimensions are scaled by a factor λ. The scaled structure occupies the domain $\mathcal{U}_\lambda$. The definition of a scaled domain can be made precise by imposing that any function $\tilde{f}$ defined on $\mathcal{U}_\lambda$ satisfies

$$\tilde{f}(\xi) = \tilde{f}(\lambda x), \tag{2.5}$$

where $\xi \in \mathcal{U}_\lambda$ and $x \in \mathcal{U}$. The purpose is to study geometrically similar structures subjected to identical operating conditions. These operating conditions are expressed primarily through the traction on the boundary, but conceivably also through the body forces if the process carries a significant electromagnetic signature. A poignant example is the chemical reactor in which the pressure (and temperature) is a parameter that affects the reaction conditions and therefore has to be kept constant. Since the discussion is predicated on uniform scaling using the same material, the material properties embodied in the density field and the elasticity tensor should also be identical. Consequently, the scaling performed in (2.5) gives rise to a physically different system.

As a matter of definition, a function $\tilde{f}$ on $\mathcal{U}_\lambda$ is said to be *generated by a function* f *on* $\mathcal{U}$ if

$$\tilde{f}(\xi) = kf(x), \tag{2.6}$$

where k is some constant. If $k = 1$, then $\tilde{f}$ is said to be *unitarily generated* by f. From (2.5) and (2.6) it follows that the spatial derivatives of a generated function evaluate to

$$\partial_i \tilde{f}|_{(\xi)} = \frac{k}{\lambda}\partial_i f|_{(x)}, \quad \text{or simpler,} \quad \partial_i \tilde{f} = \frac{k}{\lambda}\partial_i f. \tag{2.7}$$

With these definitions, a uniformly scaled structure, using the same material, is one where both density $\tilde{\rho}$ and the elasticity $\tilde{C}$ on $\mathcal{U}_\lambda$ are unitarily generated by their respective counterparts on $\mathcal{U}$. Moreover, keeping the boundary conditions constant implies that also $\tilde{T}(\xi) = T(x)$. With these restriction the scaled problem on $\mathcal{U}_\lambda$ can still be given in its general form:

$$\begin{cases} \partial_i \tilde{\sigma}_{ij} + \tilde{F}_j &= 0, \\ \tilde{\sigma}_{ij} &= \tilde{C}_{ijkl}\tilde{\epsilon}_{kl}, \\ \tilde{\epsilon}_{ij} &= \frac{1}{2}\left(\partial_i \tilde{u}_j + \partial_j \tilde{u}_i\right), \\ \tilde{\sigma}\vec{n} &= \tilde{T}, \end{cases} \tag{2.8}$$

where the Dirichlet boundary condition has been omitted.

Choosing a solution candidate $\tilde{u}(\xi) = \lambda u(x)$, where λ is the spatial scaling coefficient, leads to the same strain as in the original problem. This can be seen using (2.7),

$$\tilde{\epsilon}_{ij}(\xi) = \frac{1}{2}\left(\partial_i \tilde{u}_j(\xi) + \partial_j \tilde{u}_i(\xi)\right) = \frac{1}{2}\left(\frac{\lambda}{\lambda}\partial_i u_j(x) + \frac{\lambda}{\lambda}\partial_j u_i(x)\right) = \epsilon_{ij}(x). \tag{2.9}$$

That is, the strain $\tilde{\epsilon}$ is unitarily generated by ϵ and if the original structure was operated within specified safeguards then the same would apply to the scaled structure. Consequently, from (2.9) it follows that the stresses are unitarily generated as well:

$$\tilde{\sigma}_{ij}(\xi) = \tilde{C}_{ijkl}(\xi)\tilde{\epsilon}_{kl}(\xi) = C_{ijkl}(x)\epsilon_{kl}(x) = \sigma_{ij}(x). \tag{2.10}$$

This means that the pressure boundary condition in (2.8) is also satisfied, leaving only the force balance of the scale system to be verified. Assuming that the body forces are generated according to $\tilde{F}(\xi) = aF(x)$, the force balance can be stated as

$$\partial_i \tilde{\sigma}_{ij} + \tilde{F}_j = \frac{1}{\lambda}\left(\partial_i \sigma_{ij} + \lambda a F_j\right). \tag{2.11}$$

Thus, the candidate $\tilde{u}(\xi) = \lambda u(x)$ is indeed a solution to the scaled problem provided that $a = 1/\lambda$.

A physical interpretation of the result above is that as long as body forces decrease with the same factor as the spatial dimensions increase, a body can be scaled up and subjected to the same boundary conditions while exhibiting identical strains. Another case where the solution $\tilde{u}(\xi) = \lambda u(x)$ still holds is in the limit where boundary forces are of such a magnitude that body forces can be neglected. This would be the situation in some practical situations as e.g. high-pressure vessels. However, in most engineering applications, body forces are non-negligible. Moreover, since these are typically gravitational, scaling them down is not possible when using the same material. Note that if either body forces can be scaled down or ignored, the amount of material used is in the scaled structure is still

$$\int \tilde{\rho} d\mathcal{U}_\lambda = \lambda^3 \int \rho d\mathcal{U}.$$

That is, the amount of material scales with the volume and not, as suggested in existing orthodoxy, with surface area λ^2. In fact, if body forces are gravitational it seems likely that a scaled-up structure need to be augmented in the structural elements to be able hold the increased weight. Thus, whenever body forces like gravitational forces matter, scaling up comes at a price and symmetric scaling is not possible. Wall thickness increases faster than volume, and typically, there comes a point where the system cannot be made any larger. Only by employing lighter and stronger materials has this "scale frontier" has been pushed to larger sizes (Lieberman, 1987; Levin, 1977).

Expounding on the idea of structural innovation made by Sahal (1985) further questions the validity of uniform scaling, and consequently also the veracity of the explanations to the observed scaling behavior of cost as a function of surface area to volume ratios. Mentioned in his work is the observation that larger organisms need increased differentiation of internal functionality. For instance, with the amount of respiratory tissue scaling with the cube of the linear dimension, the capacity for gas transport is a surface phenomenon and therefore scales only with the square of the linear dimension. This warranted the evolution of respiratory organs in larger animals, a functionality that is not required in smaller organisms. The same analogy can be extended to industrial equipment getting increasingly more complicated when increasing the unit scale. Housing a strongly exothermic reaction, large reactors generally require internal heat exchangers to maintain process conditions. Smaller reactors can potentially shed the generated heat through the reactor walls.

In summary, while uniform scaling is shown to be infeasible, increasing equipment size in many cases is accompanied not only by increased material consumption for the mechanical members but also for increased complexity. It may be empirically true that more effort has gone into improving large units and that they therefore operate closer to the mechanical optimum compared to smaller pilot sized units. Also, indivisibilities in certain components and labor during production could partly explain the observed relative reductions in material consumption and labor required when scaling up. However, this is not an inherent physical phenomenon and could therefore be addressed if the focus is on scaling up in numbers rather than unit size.

2.2 Cost Reduction Through Learning

The notion that production costs decline over time dates back almost a century to the analysis by the aeronautical engineer Wright (1936). More specifically, in the production of airplane frames, Wright predicted that the labor required to produce the Nth frame is proportional to $N^{-1/3}$, see (Arrow, 1962). Later studies isolated the effect of productivity increases by studying processes were no new capital investments were made. The so called 'Horndal effect', coined by Lundberg (1961), describes a 2% annual productivity increase over a period of 15 years at a Swedish steel mill despite being "neglected from a capital investment perspective" during the same period. The offered explanation for such an increase in productivity included training of work force, improved working conditions and improved organization. Indeed, subsequent studies argue that this form of learning should be viewed as a managerial possibility to increase productivity (Lazonick and Brush, 1985; Dutton and Thomas, 1984). Conversely, as mentioned by McDonald and Schrattenholzer (2001), learning might reverse itself if a process is stalled over longer periods of time. Rather than experiencing a decrease in cost, such processes may exhibit increases in cost over time.

In addition to labor productivity gains and accumulation of experience, the incorporation of exogenous technological improvements in new capital equipment also tend to result in cost reductions over time. Such effects were investigated by Arrow (1962), who acknowledged the Horndal effect but posited that capital goods had a fixed productivity once installed. In his study, the cumulative gross investment, rather than time or output, served as an index of experience and learning. Importantly, such an approach would include benefits of economies of unit scale mentioned earlier. That is, cost reductions that ensue simply by increasing unit size would be attributed to learning as well.

Following more recently published work , see e.g. (Ferioli and van der Zwaan, 2009; McDonald and Schrattenholzer, 2001; Argote and Epple, 1990), the effect of learning will here be expressed as a cost decrease for every doubling of cumulative production. That is, denoting the cost of the nth unit produced k_n, the cost decrease through learning ε is expressed as

$$\frac{k_{2n}}{k_n} = \varepsilon. \tag{2.12}$$

The learning parameter ε is sometimes called the progress ratio in the literature. Also, references will be made to the learning rate (LR), defined by $1-\varepsilon$. For example, a technology with a learning rate of 15% experiences a cost reduction of the same fraction every time cumulative production doubles. Given the cost of the first unit, k_{ref}, and the learning rate $1-\varepsilon$, future costs can be forecast by

$$k_n = k_{\mathrm{ref}}\varepsilon^{\log_2 n} = k_{\mathrm{ref}} n^{\log_2 \varepsilon}. \tag{2.13}$$

Such a semi-continuous approximation, akin to the one presented by Wright discussed above, give rise to the so-called learning curve.

This use of learning curves and imputed learning rates is frequently employed when trying to forecast future cost of various energy technologies, see e.g. (Kim and Chang, 2012; Lindman and Söderholm, 2012; Neij, 2008). Based on past cost data, learning rates can be found, and with them, the amount of time and investment necessary to reach a certain cost target can be estimated. One of the most famous example of such an analysis regards the production of PV modules, see Figure 2.1. With remarkable accuracy the production of PV modules is seen to follow a 20% learning curve.

Using a relatively simple one-factor model of learning, like the one in (2.12), combines

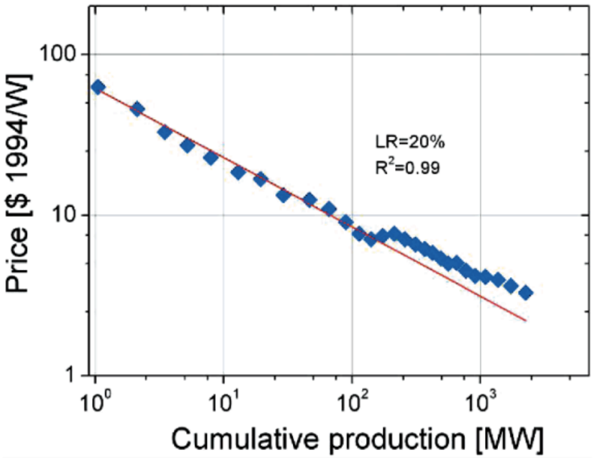

Figure 2.1: *Observed price decline of PV modules during the period 1968-1998 as function of global cumulative production Harmon (2000). Figure from Ferioli and van der Zwaan (2009), courtesy of Elsevier.*

the effect of a large number of factors. In addition to possible benefits of scaling up, these factors include process innovations and the various types of learning effects mentioned. Some models attempt to endogenize technological change by also including an index of R&D expenditures by the firm or the industry as a whole, see e.g. (Castelnuovo et al., 2005). However, the vast majority of published studies have converged on the use of the model in (2.12). Lastly, it should be noted that such a model treats intergenerational products equally. For instance, when studying the production of color television sets, the added utility of remote control, larger screens etc, are not captured but are still included in the cost. Similarly, larger scale energy technologies like coal-fired generation have over the past decades been forced to include various environmental mitigation technologies. Compared to previous implementations, these added features are typically not factored out when studying cost over a longer timespan.

2.2.1 Scale Dependence on Learning Rates

It has been noted in the literature that large-scale technologies generally exhibit lower learning rates than small-scale technologies (Neij, 1997; McDonald and Schrattenholzer, 2001). However, this observation is made in passing without any clear distinction of the difference between large and small technologies. Furthermore, as far as it is known to this observer, no claims have been brought forward in the literature regarding the statistical significance of such a difference, nor any quantification of the same.

In this section, a classification of large and small scale technologies is introduced. With such a classification and with ample data on learning curves for many different technologies available in the literature, a meta study on the influence of scale is possible. Specifically, with meaningful sample sizes, a statistical hypothesis test can be performed whether the samples of learning rates for small and large-scale technologies are drawn from the same distribution. Rejection of such an hypothesis would statistically solidify the difference in learning between large and small. Also, the difference between the respective sample mean can serve as an indicator of the difference in learning rates between the two broad classes of technologies.

A major part of the methodology of this study is the classification between large-scale and small-scale technologies. The classification made here is based on the number of end consumers that reasonable can be assumed to utilize the service or output of a single unit of the technology in question. Specifically, a unit designed to provide output or use for less than 100 consumers is labeled small. To make a clear distinction from small-scale technologies, those technologies that are produced on a scale where more than 10,000 end consumers can be tied to a single unit are labeled large. For instance, with a per capita electricity consumption at a rate of 1kW (this figure is slightly higher in the U.S. but lower in the rest of the world), those electricity generating technologies that are made with unit capacities less than 100 kW (e.g. PV modules) are considered small scale. Conversely, power plants with generators of capacities above 10 MW are considered large. In this specific technology class, it can be noted that wind-power technologies, which are extensively studied from a learning-curve perspective, see e.g. (Lindman and Söderholm, 2012; Junginger et al., 2005), and which typically have generator sizes on the scale of 1 MW are considered intermediary, and hence are not included in this study.

Data in the literature generally include, in addition to an estimated learning rate, a specification of the technology and a time period, as well as a geographical region, over which data has been collected. For the large-scale technologies, learning is sometimes stated both in terms of labor productivity (or variable cost decline) and investment cost. In those cases, only the learning rate referring to the investment cost is used. Since most of the small-scale technologies are consumer products, the notion of labor cost is moot. The learning rates for small-scale technologies is therefore presented either in terms of unit price or unit production cost. For a complete presentation of the small and large technologies, as well as their learning rates, see Tables 2.1 and 2.2.

Since some technologies are studied more frequently than other, e.g. PV modules, there is a tendency for over-representation of some technologies in the literature on learning. In order to consider the rates included in the samples as distinct and independent, the following methodology was implemented. Two technologies with the same specification studied in the same, or overlapping, regions but over different time periods are considered distinct and both rates are included. An example of such an instance is ‘Refrigerators’, which can be seen to enter twice in Table 2.1. Similarly, two or more instances of the same technology studied over similar time periods but separate geographic regions are also considered distinct, and all of them are included, as is the case for ‘Ethanol’ in USA and Brazil, see Table 2.2. These criteria are justifiable since production techniques generally vary over time and, albeit to a lesser extent perhaps, between continents. When the time periods overlap partially, with similar specification and geographical origin, their average learning rate is used, see e.g. 'Lignite conventional' in Table 2.2. Lastly, if one time period includes the other, only the reported learning rate studied over the longer time period is used. This choice is motivated by the assumption of greater accuracy the more data were available for the study in question.

The sample population of small-scale technologies included 41 learning rates with a sample mean of 20.5% and a standard error of 9.5%. The population of large-scale technologies included 26 different learning rates with a sample mean of 10.8% and a standard error of 8.1%. By employing a Kolmogorov-Smirnov test[1], the hypothesis that the two samples are drawn from the same distribution can be rejected with a significance level greater than 99%. While this test is non-parametric, fitting the two samples to normal distributions with the respective mean and variance parameters provides a good fit in both cases and also illustrates the difference between the two samples, see Figure 2.2.

The above analysis suggests that the difference in learning of large and small technologies is indeed statistically significant. Moreover, smaller technologies on average learn 10 percentage points faster when considered as cost decrease versus cumulative production. Note that this difference includes any cost reductions that may be attributed to the scaling up of individual units, a feature arguably present only for the large-scale technologies.

There are several reasons why one would expect higher learning rates for smaller scale technologies. For instance, these technologies are generally mass produced in controlled environments that allow for continuous improvements of the process. Moreover, the incorporation of exogenous technological improvement, either in the materials used or in the production method, is arguably easier to facilitate gradually over time. Conversely, units of large-scale technologies generally come online at less frequent intervals. As mentioned, such a time lag

[1] Letting $F_n(x)$ and $G_m(x)$ denote the empirical cumulative distribution functions of the two samples, the test statistic $D_{mn} = \left(\frac{mn}{m+n}\right)^{1/2} \sup_x |F_n(x) - G_m(x)|$ can be checked against critical values of the Kolmogorov-Smirnov distribution. This test was implemented using the Matlab routine `kstest2`, giving a p-value of $5.6 \cdot 10^{-4}$.

Specification	LR(%)	Time	Region	Reference
AC	10	1972-1997	Japan	Weiss et al. (2010a)
Room AC	23	1958-1993	USA	Weiss et al. (2010a)
Central AC	24	1967-1988	USA	Weiss et al. (2010a)
Dishwashers	10	1947-1968	USA	Weiss et al. (2010a)
Freezers	$\overline{13.3}$	1970-2003	Global	Weiss et al. (2010a)
Laundry dryers (electric)	6	1950-1961	USA	Weiss et al. (2010a)
Laundry dryers	27	1969-2003	Global	Weiss et al. (2008)
Refrigerators	7	1922-1940	USA	Weiss et al. (2010a)
Refrigerators	9.1	1964-2008	Global	Weiss et al. (2010b)
TVs B&W	22	1948-1974	USA	Weiss et al. (2010a)
TVs Color	7	1961-1974	USA	Weiss et al. (2010a)
Washing machines	33	1965-2008	Global	Weiss et al. (2010a)
Ford (model T)	14	1910-1926	USA	Weiss et al. (2010a)
DRAM	$\overline{20.3}$	1974-1992	Global	Irwin and Klenow (1994)
4-Function calculators	30	1970s	USA	Weiss et al. (2010a)
Hand-held calculators	26	1975-1978	USA	Weiss et al. (2010a)
Digital watches	26	1975-1978	USA	Weiss et al. (2010a)
Sony laser diodes	23	1982-1994	Japan	Weiss et al. (2010a)
Integrated circuits	$\overline{26.3}$	1962-1972	USA	Weiss et al. (2010a)
MOS/LSI	20	1970-1976	USA	Cunningham (1980)
MOS dynamic RAM	32	1973-1978	USA	Cunningham (1980)
Disk memory drives	32	1973-1978	USA	Cunningham (1980)
Modular-electronic CFLs	20	1992-1998	Global	Weiss et al. (2010a)
Integral-electronic CLFs	16	1992-1998	Global	Weiss et al. (2010a)
Modular-magnetic CFLs	41	1992-1998	Global	Weiss et al. (2010a)
PAFC	25	1993-1998	n/a	Whitaker (1998)
PEMFC	30	2002-2005	n/a	Schoots et al. (2010)
SOFC	$\overline{33.7}$	n/a	n/a	Rivera-Tinoco et al. (2012)
Electrolysis (hydrogen)	18	1972-2004	n/a	Schoots et al. (2008)
Magnetic ballasts CFLs	16	1981-1988	USA	Weiss et al. (2010a)
Magnetic ballasts CFLs	41	1990-1993	USA	Weiss et al. (2010a)
Magnetic ballasts FLs	3	1977-1993	USA	Weiss et al. (2010a)
Electronic ballasts CFLs	13	1986-1998	USA	Weiss et al. (2010a)
Electronic ballasts FLs	11	1986-2001	USA	Weiss et al. (2010a)
PV panels	22	1959-1974	USA	McDonald and Schrattenholzer (2001)
PV modules	21	1976-1992	Japan	Neij (2008)
PV modules	$\overline{20}$	1976-1992	USA	Neij (2008)
PV BoS	$\overline{20.5}$	1992-2001	EU	Neij (2008)
Heat pumps	$\overline{32.5}$	1980-2004	EU	Weiss et al. (2010a)
Condensing gas boilers	4	1992-1999	Germany	Weiss et al. (2010a)
Condensing combi boilers	14	1988-2006	Netherlands	Weiss et al. (2008)

Table 2.1: *Learning rates of small-scale technologies. Population size – 41, sample mean – 20.5%, standard error – 9.5%. Over-lined entries denote average values and the years in these instances signify the total time span of the studies.*

Specification	LR(%)	Time	Region	Reference
Electricity from biomass	$\overline{28.25}$	1980-1998	Global	Kahouli-Brahmi (2008)
CHP biomass	11	1990-2002	Sweden	Junginger et al. (2006)
FB boilers	8.5	1975-2002	Global	Junginger et al. (2006)
CFC subst.	7	1988-1999	n/a	Laitner and Sanstad (2004)
Conventional coal	$\overline{8.8}$	1975-1998	Global	Kahouli-Brahmi (2008), McDonald and Schrattenholzer (2001)
Lignite conventional	$\overline{6.04}$	1975-2001	Global	Kahouli-Brahmi (2008), McDonald and Schrattenholzer (2001), Jamasb (2006)
Pulverized subcrit.	6	1942-1999	Global	Yeh and Rubin (2007)
Pulverized supercrit.	$\overline{4.83}$	1942-2010	Global	Yeh and Rubin (2007), McDonald and Schrattenholzer (2001), Jamasb (2006), Winkler et al. (2009)
CCGT	$\overline{11.02}$	1981-1998	Global	McDonald and Schrattenholzer (2001), Jamasb (2006)
CHP	0.23	1980-1998	Global	Kahouli-Brahmi (2008)
Ethanol	20	1975-2004	Brazil	Van Den Wall Bake et al. (2009),
Ethanol	13	1980-2005	USA	Hettinga et al. (2009)
Gas pipelines (on-shore)	3.7	1984-1997	USA	McDonald and Schrattenholzer (2001)
Gas pipelines (off-shore)	24	1984-1997	USA	McDonald and Schrattenholzer (2001)
Gas turbines	13	1958-1980	Global	McDonald and Schrattenholzer (2001)
Hydro power	$\overline{1.68}$	1975-2001	Global	Kahouli-Brahmi (2008), Winkler et al. (2009)
Coal gasification	-7	1942-2002	Global	Schoots et al. (2008)
SMR	11	1960-2003	Global	Schoots et al. (2008)
O2 production (cryogenic)	10	1980-2003	Global	Rubin et al. (2007)
LNG cryogenic	20	1972-2003	Global	Rubin et al. (2007)
Nuclear power	$\overline{21.05}$	1975-1998	Global	Kahouli-Brahmi (2008), McDonald and Schrattenholzer (2001)
Oil extraction (off-shore)	25	n/a	North Sea	McDonald and Schrattenholzer (2001)
Oil extraction (at well)	5	1869-1971	n/a	McDonald and Schrattenholzer (2001)
Scrubbers FGD	13	1976-1995	Global	Riahi et al. (2004)
Scrubbers SCR	14	1983-2000	Global	Yeh et al. (2005)
Solar Thermal	2.2	1985-2001	n/a	Kahouli-Brahmi (2008)

Table 2.2: *Learning rates of large-scale technologies. Population size – 26, sample mean – 10.8%, standard error – 8.1%. Over-lined entries denote average values and the years in these instances signify the total time span of the studies.*

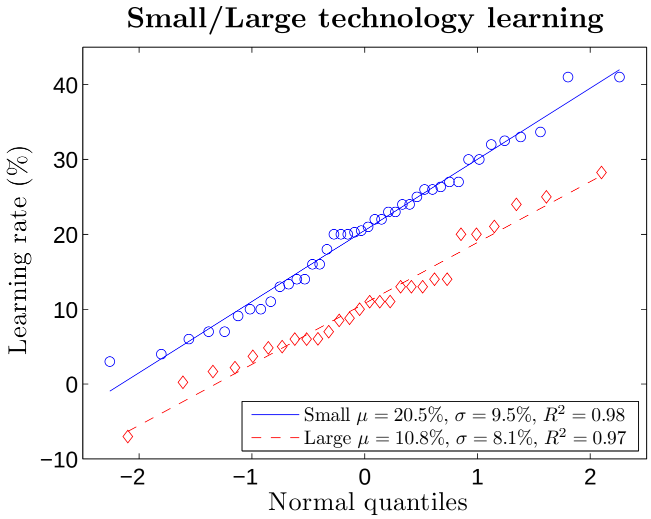

Figure 2.2: *Q-Q plot of normal distributions of the learning rates for small and large technologies. A Kolmogorov-Smirnov test rejects the hypothesis that the two samples come from the same distribution. Importantly, the average learning rate of small-scale technologies is 10 percentage points higher than for large-scale technologies.*

can erode learning accumulated within a firm, e.g. through labor turnover, further suggesting that larger scale technologies might exhibit lower learning rates.

It should be acknowledged that there possibly is an inherent bias towards higher learning rates overall when performing a meta study of this kind. Studies of technologies where little or no learning occurs are likely to attract less attention, with some exceptions, see e.g. (Schoots et al., 2008). However, since such a bias is likely to affect both large and small technologies, it should not substantially affect the conclusion of this study that smaller technologies learn at a higher rate.

2.3 Economies of Unit Scale vs. Numbers

In the previous sections, two distinctly different strategies of reducing fixed cost were investigated: scaling up in size and scaling up in numbers. Using the expressions in (2.1) and (2.13), the total cost of providing the same nominal capacity can be estimated in both scenarios.

That is, given a reference unit of capacity c_{ref} and cost k_{ref}, this unit can either be scaled to a capacity Nc_{ref} or reproduced N times with the same resulting aggregate capacity. The cost $k(Nc_{\text{ref}})$ of the scaled up version is straightforward to retrieve from (2.1). By integrating the learning curve in (2.13), the total cost $K(N)$ of the modular system can be found. These two estimates can be stated as

$$
\begin{aligned}
k(Nc_{\text{ref}}) &= k_{\text{ref}} N^{\alpha}, && (2.14)\\
K(N) &= \frac{k_{\text{ref}}}{1+\log_2 \epsilon} N^{\log_2 \varepsilon + 1}. && (2.15)
\end{aligned}
$$

Mentioned in Section 2.1 were the empirically observed values of the scale parameter α in the range of 0.6 to 0.8. In Section 2.2, it was concluded that small-scale technologies, which this reference unit is supposed to be, on average learn at a rate of $\varepsilon = 80\%$. At such a learning rate, the exponent in (2.15) evaluates to $\log_2 \varepsilon + 1 = 0.7$. The conclusion is that $\log_2 \varepsilon + 1 \approx \alpha$, and hence the reduction in fixed cost stemming from scaling up in size is on par with the cost reduction of scaling up in numbers.

The premise that scaling up in size inherently reduces material consumption over scaling up in numbers was rejected in the previous section. Observing similar scaling laws for the two strategies then invites the question whether the cost savings have the same underlying causes. Indeed, exploiting indivisibilities in production has been suggested to account for economies of scale both when scaling up in size and in numbers. However, the investigation of learning suggests another difference. Smaller scale technologies are produced in a more continuous fashion which more easily allows for gradual incorporation of exogenous technology improvement.

The capital allocation in the two strategies exhibit a crucial difference as well. In a large-scale scenario, the investment is typically considered sunk since alternative uses for this capital is scarce. For small-scale technologies, the main investment is upstream of the end product, generally in mass production facilities. These facilities can to a greater extent be retooled, thereby allowing for re-purposing of the capital. As mentioned in (Bernard et al., 2006), by an overwhelming majority, U.S. manufacturing firms that have switched products have done so using existing facilities rather than constructing new plants. The possibility of such flexible manufacturing allows for exploitation of possible indivisibility of upstream capital itself. The financial flexibility that arise for the section downstream of this mass-production step are further addressed in Chapter 5.

Another scale-dependent feature that enters into the fixed cost discussion is redundancy. That is, failure of a single component in a large-scale technology risks total loss of output. In a modular setting, a similar failure would only translate into a partial outage. As exemplified by Göçmen in (Dahlgren et al., 2013), if the possibility of critical failure are independent among similar process units, scaling up in size entails carrying more excess capacity to ensure the same level of overall availability. Conversely, the level of reliability of individual units in a modular setting can be reduced without sacrificing aggregated availability. Diversifying risk of critical failure among many parallel systems allows for further cost reductions, through less strict quality control, of small-scale technologies.

Chapter 3

Variable Costs and Unit Scale

Viewing process capacity as a black box, empirical observations discussed previously suggest that the strategies of scaling up in size and scaling up in numbers require comparable levels of up-front investments. Variable costs, on the other hand, are not as easily extrapolated into general trends across different technologies. Instead, this chapter strives to point to general features of labor productivity, conversion efficiency, and their scale dependence, which is further explored in the case study on U.S. electricity generating technologies in the following chapter. Moreover, based on the demonstrated equivalence in capital cost reductions when scaling up in size and scaling up in numbers, a simple model is introduced that demonstrates how the inclusion of transportation costs can impact the optimal size of a given technology.

3.1 Labor efficiency

Adam Smith notes in his discussion on the application of labor, "... everybody must be sensible how much labour is facilitated and abridged by the application of proper machinery. It is unnecessary to give any example" Smith (2000). Despite this comment, it should be mentioned that unit scale, as it influences labor productivity, has been studied in relevant industries (Garcia et al., 2001; Baily et al., 1985; Levin, 1977). The rather natural conclusion is that increasing the size (capacity) of a piece of machinery entails approximately zero marginal operating labor. One component of the overall economies of scale witnessed in commodity-based industries is therefore increased labor productivity. However, increased productivity offers diminishing returns with ever increasing sizes, as illustrated in the case of electricity production in the next chapter.

Scaling up in numbers rather than unit size would historically not be accompanied by similar increases in labor productivity. Rather the opposite, without automated control and operation, achieving stable output would have been a daunting task in a highly modular setting. Historically, labor can be posited as a discriminant against small-scale technologies, with one exception. Virtually all the small-scale technologies surveyed in the previous section on learning rates are consumer goods, or parts thereof. The notion of labor cost attached to the operation of these technologies, e.g. integrated circuits in personal computers, is subsumed in the utility of the user.

An alternative to reducing labor by either scaling up or shifting this factor input to the user is to employ automation. Undeniably, every technology faces different circumstances, and in some instances complete automation might not be feasible with today's technology. However, the progress made in sensor technology, robotics, data communication and network capabilities of automated agents, to name a few of the many facets that go into the umbrella concept of automation technologies, have made automation a viable strategy (Vrba, 2013; Luo and Chang, 2012; Luettel et al., 13; Petrina, 2011; Mohan and Ponnambalam, 2009).

As discussed by Göçmen and van Ryzin in (Dahlgren et al., 2013), automation has already today changed various industry dynamics and increased service that would have been too costly with manual labor, e.g. ATMs, electronic check-ins for flights and car-sharing services such as Zipcar. Moreover, the previously mentioned individual technologies have also made it possible to operate in inhospitable environments through remote control. Monitoring conditions in leaking oil wells and failing nuclear plants are two unfortunate yet notable examples. In summary, automation technologies can now, with little or no imagination, be deployed to partially or completely remove the need for human labor in continuous operation. Rather than an insurmountable technology barrier, automation can therefore be considered as an added fixed cost in the production of the technology component.

3.2 Conversion efficiency

The underlying physics of a process sometimes strongly favors a certain size in terms of conversion efficiency, defined as a ratio of inputs to desired output. The charge transfer in electronic components in computers is a process that clearly favors a small scale. Since the desired output here can be thought of as a binary change in electrostatic potentials across a circuitry, smaller components can achieve this with reduced ohmic heat and consequently reduced need for constant cooling. On the other hand, hot and cold storage favor large unit scales since heat losses scale with the surface area of the containment.

In general, when the capacity of a process can be related to the magnitude of convective transport or inertial motion, efficiency is typically benefited by a large unit scale. This can be related to a decreasing surface area to volume ratio. Consider, for instance, the convective transport of a fluid, perhaps with a valuable enthalpic content. Losses through friction and conductive heat transfer are surface area phenomena which decrease disproportionately to capacity when scaling up a unit in size. Likewise, the resistive losses in a conductor decreases with cross-sectional area. Similarly, when capacity is related to the inertial motion, e.g. a spinning shaft or the linear motion of a ship, the dissipative frictional losses scale with the surface area.

An entire process unit is however seldom composed of merely one process of the type mentioned above. While size may serve as a qualitative indicator of efficiency, the magnitude of efficiency gains (or losses) by scaling have to be evaluated case by case. The comparison of efficiencies in a large-scale single cycle power plant and in a small-scale internal combustion engine serves as a case in point. The car engine can under optimal conditiona attain efficiencies in the range of 30-35%, which is comparable to the efficiency of most steam plants (White et al., 2006).

This example also serves to highlight a crucial distinction between optimal conditions and practical conditions. Most pieces of industrial equipment exhibit variable efficiency with output. The time spent outside the designed operating parameters, such as transitioning from cold start to optimal conditions in power generation, acts to decrease the overall efficiency. The same inertial forces that favor large scale in terms of efficiency also tend to make the technology less able to respond to variations in demand. If the output of a process can be easily stored, then the process unit can be shielded from outside demand fluctuations. If not, as is currently the case with electric power, varying output of a single unit can decrease the average efficiency substantially from the efficiency in optimal conditions. A modular facility comprised of many small units that each have a lower efficiency under optimal conditions compared to a large unit may yet provide higher average efficiencies since demand variability can be met with individual units ramping up or down in sequence.

Lastly, the importance of physical efficiency depends strongly on the cost of inputs. If these costs are negligible, as is typically the case with renewable power generation, then the notion of efficiency is almost moot.

3.3 Locational flexibility

Small unit scale offers the possibility of distributing operation compared to the large-scale technologies. In the case of electricity generation, the potential benefits are well-documented, see e.g. (Pepermans et al., 2005; Lovins et al., 2002). In addition to increased reliability that comes from ability to detach from the grid, the main benefit comes in the form of reduced transportation cost, both of electric power and possibly also of heat. The reduced transportation cost is a feature that can be extended to a wider class of technologies. More specifically, the following rather simple cost model illustrates how decentralization interplays with unit scale.

The agent in this model is a firm that produces a good and services a large area A_{tot} with uniformly distributed and constant demand. With a uniform demand density, the area A serviced by a single production facility is proportional to its capacity, suggesting that the capital cost is a function of A. The other area-dependent cost is the transportation cost of the output from the plant to the consumer. All other costs are assumed to be negligible or independent of the area[1]. Based only on capital and transportation costs, the firm is to determine how large an area, A, each individual facility should serve.

With capacity being proportional to the area A, the cost $k(A)$ of a facility can be estimated using the power law in (2.1):

$$k(A) = k_{\text{ref}} \left(\frac{A}{A_0} \right)^{\alpha},$$

where k_{ref} is the cost of a reference facility servicing an area A_0. Rather than scaling up a single facility in size to serve a large area, the firm can opt to mass produce capacity and distribute it over $N = A_{\text{tot}}/A$ locations. Assuming this can be done at a learning rate $1 - \varepsilon$ (see section 2.2) the total capital expenditure to provide capacity for all of A_{tot} can be stated as

$$k(A_{\text{tot}}) = k(A) N^{1+\log_2 \varepsilon} = k'_{\text{ref}} \left(\frac{A}{A_0} \right)^{-\gamma},$$

[1] This would be the case if the inputs can be sourced locally, e.g. from the ambient environment or if the transport of the inputs occurs infrequently enough to render their shipping cost much smaller than that of the output

where the notation $\gamma = -(\alpha - (1 + \log_2 \varepsilon))$ is introduced for brevity. The total capital expenditure can be distributed over the total output, giving a capital cost contribution per unit output of

$$K_C(A) = \eta k(A_{\text{tot}}) = K_C(A_0) \left(\frac{A}{A_0}\right)^{-\gamma}. \tag{3.1}$$

With this cost, the conversion from capital cost to the contribution per unit output is included in η. Such a conversion includes several parameters, including the lifetime of the plant. As mentioned in Section 2.3 typical values for the scaling parameter α ranges between 0.6 and 0.8, and the learning parameter ε around 0.8. Combined, this suggests that γ is close to zero.

Turning to the transportation cost per unit output, $K_T(A)$, it is reasonable to assume that this cost is increasing in A. This assumption stems from the expectation that transportation cost do not decrease with the shipping distance. On the other hand, when shipping over longer distances, the firm is likely able to take advantage of economies of scale by aggregation and using more efficient modes of transportation, suggesting that this cost rises no more than linearly in the distance. Assuming a power law as well for the transportation costs:

$$K_T(A) = K_T(A_0) \left(\frac{A}{A_0}\right)^{\beta}, \tag{3.2}$$

these arguments about scaling with distance imply $0 \leq \beta \leq 1/2$, where the multiplier $K_T(A_0)$ represents the unit shipping cost for a service area A_0. The model in (3.2) conforms with shipping cost models in the literature, see e.g. (Langevin et al., 1996).

With the expressions in (3.1) and (3.2) the relevant total area-dependent cost per unit output can be formulated as

$$K(A) = K_T(A) + K_C(A) = K_T(A_0) \left(\frac{A}{A_0}\right)^{\beta} + K_C(A_0) \left(\frac{A}{A_0}\right)^{-\gamma}. \tag{3.3}$$

If $\gamma < 0$, i.e. when the reductions from learning exceeds those from scaling up, then both capital and transportation costs are increasing functions of the area. In this case, the optimal area is $A = 0$, and the system is driven to the smallest possible size. On the other hand, if $\gamma > 0$, then the total cost $K(A)$ is convex with a finite optimum at A_{opt}. Furthermore, there is an area where the transportation cost K_T equals the capital cost contribution K_C per unit output. Choosing the reference area A_0 as this area, i.e. where $K_C(A_0) = K_T(A_0) = K_0$, delineates the total cost in a region where it is capital-cost dominated, $A < A_0$, and transportation-cost dominated, $A > A_0$. The optimal area A_{opt} is found from the first order condition on (3.3):

$$\frac{dK}{dA}(A_{\text{opt}}) = 0 \quad \Rightarrow \quad \frac{A_{\text{opt}}}{A_0} = \left(\frac{\gamma}{\beta}\right)^{\frac{1}{\beta+\gamma}}.$$

Introducing the parameter $\lambda = \gamma/\beta$, it can be seen that

$$\left(\frac{A_{\text{opt}}}{A_0}\right)^{\beta} = \left(\frac{\gamma}{\beta}\right)^{\frac{\beta}{\beta+\gamma}} = \lambda^{\frac{1}{1+\lambda}}. \tag{3.4}$$

Furthermore, the total cost in (3.3), with $K_C(A_0) = K_T(A_0) = K_0$, at the optimal area can also be expressed through λ:

$$K(A_{\text{opt}}) = K_0\left(\lambda^{\frac{1}{1+\lambda}} + \lambda^{-\frac{\lambda}{1+\lambda}}\right). \tag{3.5}$$

In summary, the simple model above takes only two cost items into account; capital cost and transportation cost of the output. As suggested in Section 2.3 and made explicit here, if the economies of learning overtake the economies of unit scale, manifested by $\alpha-(1+\log_2 \varepsilon) > 0$, then the system is driven to small unit scale. Furthermore, if demand is distributed over a large area and if indeed all costs other than the two mentioned can be neglected or considered independent of the service area, then the presence of a nondecreasing transportation cost of the output presents a clear impetus for distributed operation.

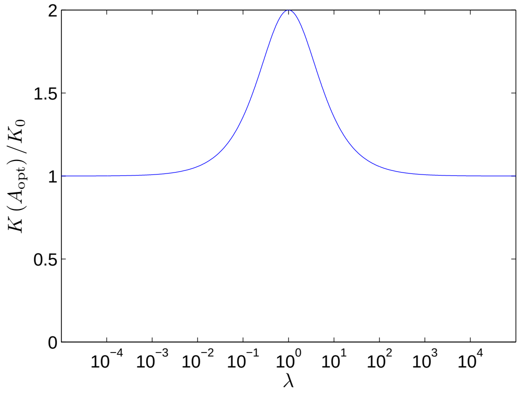

Figure 3.1: *Assuming the total cost $K(A)$ is convex, the optimal total cost is seen to be at best half the total cost when the facilities are sized using the heuristic that capital cost should equal transportation cost per unit output.*

On the other hand, if capital costs indeed favor larger sizes, i.e. if $\alpha - (1 + \log_2 \varepsilon) < 0$, this model brings a non-intuitive conclusion. With the exact values of the parameters β and γ (i.e. α and ε) being obscured, no definitive optimization can be performed. However, as seen in Figure 3.1, abiding by the heuristic to size the individual facility such that the capital cost contribution per unit output is equal to the output transportation cost will lead to costs that are at worst twice the optimal cost.

The assumption in the model that the transportation cost of inputs can be ignored is likely not true in some instances. Violating this assumption could lead to total transportation costs (inputs and outputs) being increasing with a higher degree of decentralization. If capital cost still favor smaller unit sizes, that is if $\alpha - (1 + \log_2 \varepsilon) > 0$, the optimal strategy would still be small-scale technologies, but now these should be aggregated at a few locations.

Chapter 4

Case Study: U.S. Electricity Sector

The U.S. electric power generation sector presents an interesting area for a case study in unit scales for several reasons. With access to the national grid, the total market far exceeds the largest possible scale in any technology. Transmission constraints certainly do exist but they result in increased cost, rather than insurmountable barriers for delivery in most cases. Hence, the scales of existing installations arguably represent an optimal size choice with no substantial demand constraints. Moreover, acting in a partially regulated industry, owners of generating capacity have over the last century been required to provide state and federal institutions with information regarding both operational details and the status on all installed capacity. Thus, data is readily available. Lastly, even though the purpose of providing electric power to the grid is the same for all technologies, these are internally very different. For instance, handling a gaseous fuel in natural gas-fired power plants differs significantly from the solids handling in coal-fired plants, which in turn is distinctly different from processing the radioactive waste in a nuclear power plant. Such significant differences within one industry would suggest that any conclusions regarding scale have a broader applicability. At the very least, these conclusions should not immediately be dismissed in other process industries.

This chapter begins with a review of the observed size choices in the main generating technologies in the U.S, thus confirming the trend of “bigger–is–better”. In order to determine if, or to what extent, operating cost motivates this trend, operating data from power plants around the country is studied. By regressing operational costs on unit scale, while controlling for other pertinent variables as well, it is determined that decreased operational cost only weakly supports the observed trend. Furthermore, it is also determined that the decrease in operational cost with scale is chiefly explained by decreased labor cost. Operating costs net of labor costs show no significant correlation with unit size. Thus, in a paradigm of low-cost automation, the only impetus for increasing unit scale is likely to be found in traditional economies of unit scale.

4.1 Historical Trends of Generator Sizes

In order to observe the evolution of unit size in the electricity generating sector six distinct technologies are investigated. For each technology, a time series of the average generator size (nameplate capacity) installed per year is constructed from generator-level data compiled by the Energy Information Administration, c.f. (EIA, 2004, 2011). (This data includes retired capacity, meaning that the presentation is not merely a summation of capacity that has survived in operation till this day.) The samples of the technologies labeled ‘Combined cycle’ and ‘Gas turbine’ are limited to those powered by natural gas, which constitute significant majorities of the generators in use in both cases. Moreover, the selection of generators in all technology classes is limited to those that deliver power trans-grid, which suggest that they are sized without any immediate demand constraints. This selection excludes certain dedicated industry power sources. Still, combined these technologies account for more than 90% of the currently installed total capacity of almost 1.1 TW. The time series of average generator size installed are presented in Figure 4.1 alongside the total installed capacity of the same technology, including those facilities that are on stand-by.

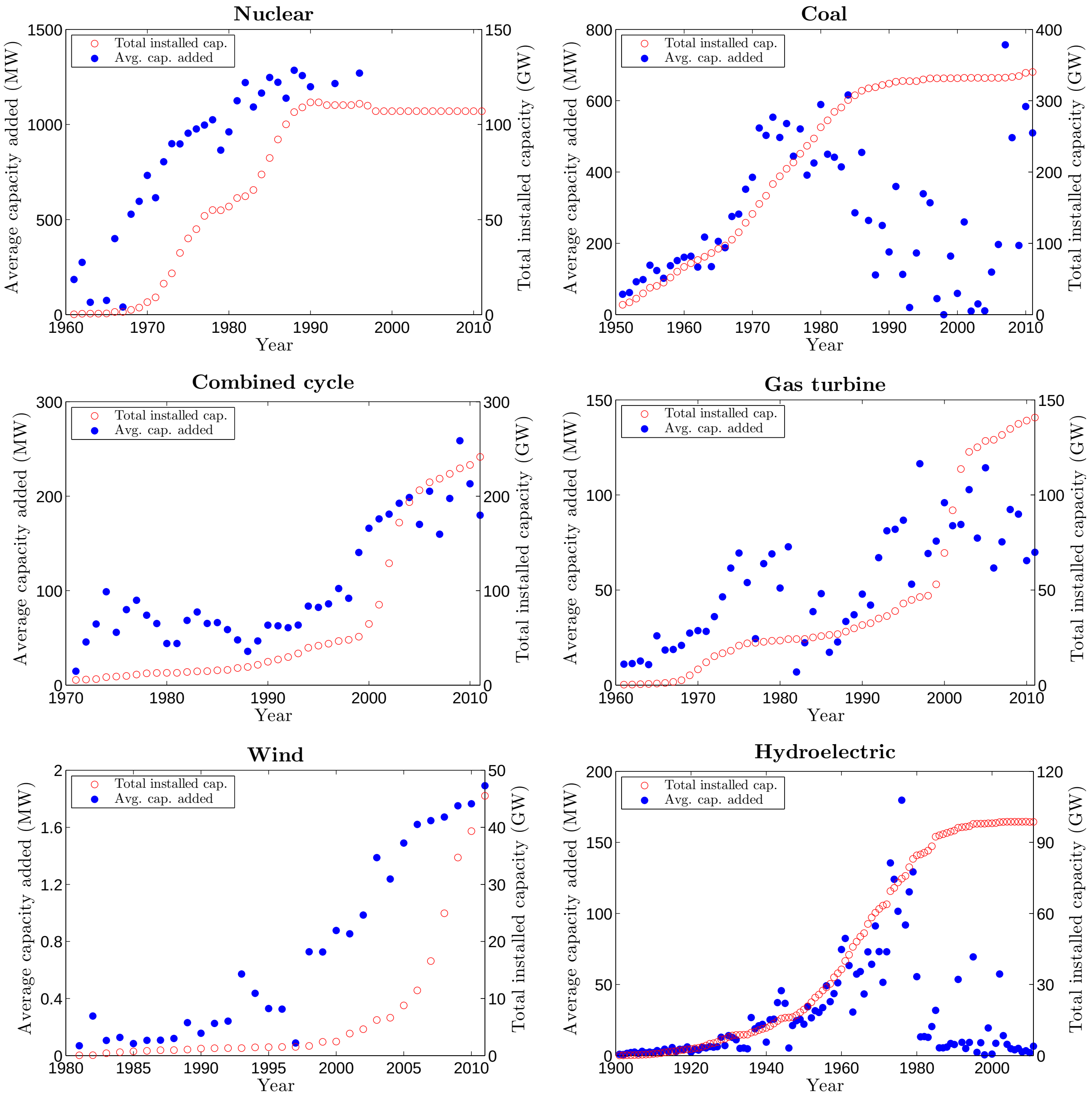

Figure 4.1: *Historically, the average generator sizes have increased with the cumulative growth of the specific technology. The technology labeled 'Gas turbine' comprises only those turbines powered by natural gas. Also, the average generator size in 'Combined cycle' is the average size of the both the gas turbines and the steam turbines in the same cluster. The year assigned to this class is the year the latest generator in a cluster was added.*

Applied to the electricity generating sector, Figure 4.1 confirms the general assumption in this dissertation of a "bigger–is–better"–trend. The impact of economies of unit scale were first realized for hydroelectric generation in the early part of the twentieth century, which was made possible by the emergence of high-voltage transmission lines capable of delivering the output over long distances. The same strategy of scaling up was then carried over to the thermal technologies, initially coal but later also nuclear and natural gas-fired generation. Until the relatively resent restructuring of certain power markets in the U.S., the ownership of these large-scale, and by necessity, centralized power stations were limited to few large investor-owned utilities with access to the capital necessary for these huge investments (Carley and Andrews, 2012; Carley, 2011).

From Figure 4.1, it can be deduced that during growth phases of a certain technology, the average size of the generators installed generally increases as well. For instance, between 1970 and 1990, the average nameplate capacity of nuclear reactors increased three-fold and the average coal-fired generator increased almost ten-fold during the 30 years following 1950. The apparent decrease in average sizes in some of the technologies, e.g coal after mid 1980s, is explained by a stagnation in the demand of that technology in general. The impact of smaller, niche applications (still with the capability of delivering power trans-grid) is marginalized by large installations during periods of growth but dominate the mean in times of demand stagnation. That is, the validity of the average as an appropriate indicator of optimal size at the specific time increases with the slope of the cumulative capacity at the same year.

The history of natural gas-fired generation is interrupted by the “The Power Plant and Industrial Fuel Use Act” enacted by the U.S. Congress in 1978 and later repealed in 1987. Driven by concerns over what was believed to be very limited supplies, the “Fuel Use Act” limited industrial use of natural gas and banned completely the construction of new natural gas-fired power plants. This ban severely hampered R&D effort on gas turbine technologies in the U.S.(Norberg-Bohm, 2000), explaining the slump in average sizes in natural-gas fired technologies, ‘Combined cycle’ and ‘Gas turbine’ in Figure 4.1, over the same period. The

Technology	α
Gas turbine + HRSG[1]	0.7 (Hamelinck and Faaij, 2002)
Steam turbine + steam system	0.7 (Hamelinck and Faaij, 2002)
Nuclear power plant	0.619 (Locatelli and Mancini, 2010)
Hydroelectric plant	0.82 (Hreinsson, 1987)

Table 4.1: *Scale parameters for various electricity generating technologies.*

one possible exception to the apparent trend of a positive correlation between average nameplate capacity and demand growth within the same technology can be found in gas turbine technology since the year 2000. This technology is generally intended for peaking plants, where the ability to respond to rapid demand fluctuations is crucial. The increased inertia of larger turbines impairs this ability, offering a possible explanation to a stagnation in average sizes despite continued growth during the last decade.

Wind energy, alongside most other renewable technologies, is generally considered small-scale. However, while the average size of wind-powered generators is orders of magnitude smaller than those found in thermal generation the same trend emerges. In the past two decades, wind turbines have increased five-fold in nameplate capacity. Building higher towers not only make larger turbines possible but also gives access to higher wind speeds, thereby increasing the power output per swept area. As to hydroelectric generation, most of the waterways suitable for development have already been exploited today, effectively capping growth (EIA, 1998). Still, before 1980 generator sizes followed the same trend as the other technologies and grew alongside the total installed capacity.

The trend of increasing sizes is at least partially explained by well-documented economies of unit scale in the electricity generating industry in general (Locatelli and Mancini, 2010; Hamelinck and Faaij, 2002; Hreinsson, 1987; Christensen and Greene, 1976). Numerical values of the power, or scale parameter α, introduced in equation (2.1), used for capital cost estimates in various technologies are presented in Table 4.1. Clearly, reducing cost by scaling up in unit size has been a strategy available to the industry as a whole.

The trend of building ever larger units has had some interesting consequences on the U.S. electricity market. With electricity generally being a non-storable commodity, over-building capacity translates into under-utilized capital, which highlights the need for accurate forecasting of future market conditions. The cost reductions stemming from economies of unit scale by increasing the unit size risk being off-set by unforeseen revenue shortfall. For instance, in the early 1970s it was generally believed that electricity demand would keep growing at an exponential pace well into the 21st century (Anderson, 1973). With hindsight, it can be concluded that such growth patterns failed to materialize, see Figure 4.2. Elusive demand growth and concurrent energy crises caused financial setbacks for the electric utilities with capital intensive large-scale assets on their balance sheets (Carley, 2011; EIA, 2000). Furthermore, the generating technologies of choice at the time, coal and nuclear (see Figure 4.1), were, and still are today, endowed with long lead times in construction and planning of 7-10 years, or even more (Keeney and Sicherman, 1983; Joskow and Baughman, 1976). Thus, the over-building was compounded by projects that were under construction and too costly to abandon, which came online once the decline in demand had already been experienced, see Figure 4.3. The second and more pronounced peak in capacity added occurred in the early 2000s. This spurt of capacity expansion was preceded by a period of tightening power supply and substantial statewide and federal deregulation of power markets (Joskow, 2005; EIA, 2003).

The lumpy investment behavior observed in electric power generation is amplified by the large scales involved. The time for demand to 'grow into capacity' naturally increases with scale, and any idle capacity translates into higher capital cost per unit output produced (Lovins et al., 2002; Manne, 1961). Moreover, the long lead times and long amortization times involved raise the difficulty to plan future investments. As observed in the electricity generating industry, erroneous demand forecasts have substantial economic impact.

[1] Heat recovery steam generator

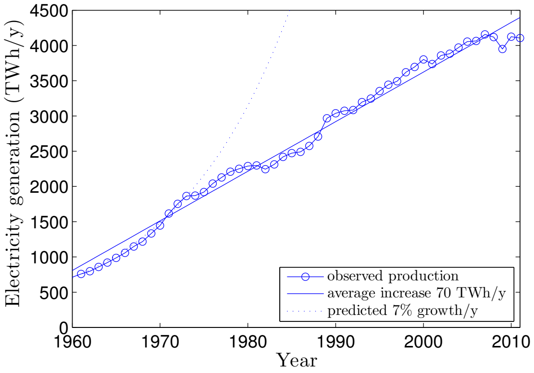

Figure 4.2: *The observed exponential growth in demand at 7% per year lasted until the early 1970s. With recent exceptions, the growth has been linear with about 70TWh/y.*

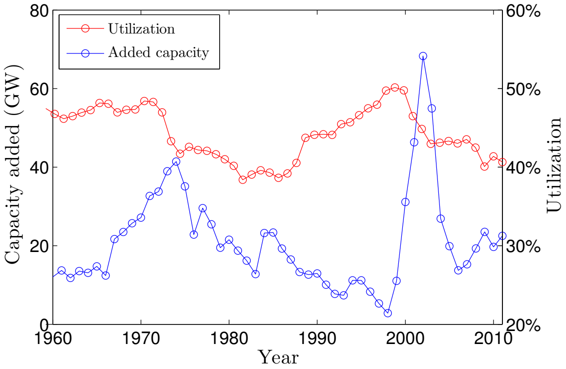

Figure 4.3: *The two periods of major capacity expansion in the U.S. electric power sector (1970s and the early 2000s) resulted in significant decrease in utilization.*

4.2 Operational Costs

As exhibited in the previous section, the U.S. electric power industry has followed a general trend of increasing unit scale. To what extent this trend is motivated by decreasing operational costs, in addition to the economies of unit scale realized in construction (Carley and Andrews, 2012; Carley, 2011), is the focus of this section. To address this question, operational data on a generator level is studied from power plants distributed over the United States. This data allows for a regression of operational cost on unit size while controlling for other pertinent variables influencing cost. The analysis is limited to the non-renewable technology classes described in the previous section, in part due to lack of sufficient data, but also because the renewable technologies do not incur fuel cost, and thereby require different cost models.

The data is collected from those utilities and power producers that were required to file operational information to the Federal Energy Regulatory Commission (FERC) in Form No. 1 (FERC, 2010) in 2010. Upon retrieval, a single data point, in addition to generator nameplate capacity, contains information on total production cost (capital charges excluded), fuel cost, annual generation volume, average heat rate (efficiency), age and employee head count. Since the labor head count occasionally is presented on a plant level in the source material, averaging over the number of generators in the plant is required. This averaging is complicated by the fact that a single plant may comprise generators of different technologies. Additionally, even if the same technology is used in a multi-generator plant, these generators can differ substantially in both size and age. To obtain reasonably accurate generator-level data points (without sacrificing sample size), the following criteria were imposed on plant-level data before averaging and inclusion in the sample:

1. The plant comprises generators of only one technology.
2. The difference in age between the oldest and the newest generator is at most 10 years.
3. The difference in nameplate capacity of the largest and smallest generator is less than a factor 2.

With these restrictions the total sample encompasses 25% of total installed capacity in the U.S in the year 2010. The operational data contains employee head count but not salary levels. Instead, industry average salary levels are used as reported in (U.S. Census Bureau, 2007). The variables included in the regression analysis are detailed in Table 4.2.

A multiplicative model[2] is stipulated for the dependent variable $Y_{\text{tot cost}}$ (total operational cost) and the independent variables denoted X_i. This allows for a log-linear regression:

$$\log Y_{\text{tot cost}} = \log \beta_0 + \sum_i \beta_i \log X_i. \tag{4.1}$$

The null hypothesis in this multiple regression is that $\beta_i = 0$, i.e. the independent variable X_i does not influence total operational cost. Rejection of this hypothesis permits a qualitative

[2] Anticipated non-linear behavior in some or all of the dependent variables make a multiplicative model better suited than a linear model. Moreover, the scope of this study is limited to investigate causality of observed factors. No claims are made as to the predictive power of this model outside given statistics.

	Variable	Comment
$Y_{\text{tot cost}}$	Operational cost (\$/kWh)	Total operational cost divided by total generation (kWh)
Y_{labor}	Labor cost (\$/kWh)	Reported employee head count multiplied by industry salary levels (\$/y) and divided by total generation (kWh)
X_i	Nameplate capacity (MW)	Reported value
	Capacity factor (%)	Total generation divided by nameplate capacity (kW) and by 8760 (h/y)
	Efficiency (%)	3412 BTU/kWh divided by the reported heat rate (BTU/kWh)
	Fuel cost (\$/MMBTU)	Total fuel cost divided by total generation adjusted for efficiency
	Age (y)	Defined by the year the generator became operational

Table 4.2: *Variables included in the regression analysis of operational cost for non-renewable technologies in the U.S. electricity generating sector. The two dependent variables, $Y_{tot\ cost}$ and Y_{labor}, are regressed separately on the set of independent variables X_i. Moreover, to cancel the effect of labor, e.g. by assuming complete automation, the difference $Y_{tot\ cost} - Y_{labor}$ is also regressed on X_i.*

influence (increasing/decreasing) of the independent variable at given level of significance. A graphical display of the operational cost versus generator size can be found in Figure 4.4 and the results from the multiple regression analysis is found in Table 4.3.

The analysis shows that in three of the four technologies surveyed, increasing unit size significantly decreases operational cost. However, this trend is generally weak. For instance, all else being constant, a doubling of the generator size in a coal-fired power plant is expected to yield a decrease in cost of only 4%, see Table 4.3.

Thermal generation technologies, like the ones studied, require time to ramp up from cold start to optimal operating conditions, during which time efficiencies are lower. Moreover, since labor cost typically is not considered a short-run variable cost, operating at a lower capacity factor would still incur the same labor expenditures. Combined, this explains why operating cost declines with the capacity factor[3].

With the cost of fuel being the main variable cost item for all technologies, it is not surprising that efficiency and fuel cost both significantly affect total operational cost. Lastly, technological improvement would suggest that costs are lower for newer plants. However, the analysis shows a weak decrease in cost with age for only two of the four technologies.

[3] It should be noted that the capacity factor used in this analysis is the annual average, which may obscure seasonal or even diurnal trends.

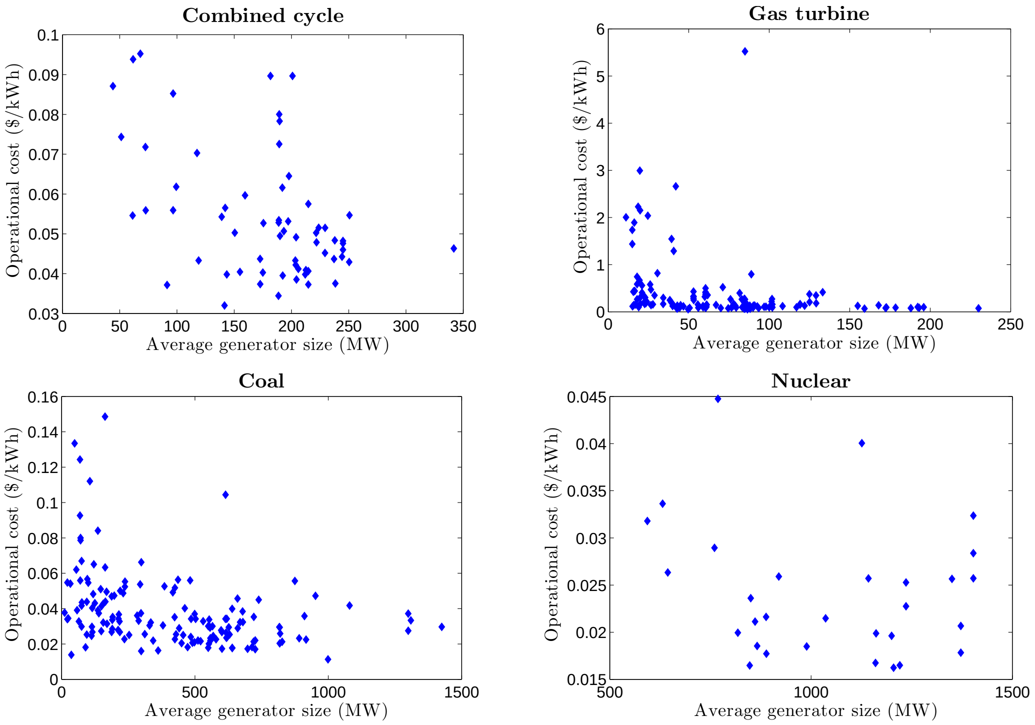

Figure 4.4: *Total operation cost per kWh produced for the for main, non-renewable, electric generating technologies. With the exception of nuclear, all technologies show a weak but significant trend of decreasing operational cost with increasing generator size.*

	Combined cycle	Coal	Gas Turbine	Nuclear
log(Nameplate capacity)	-0.09^{*}	-0.06^{***}	-0.20^{***}	-0.28
	(0.04)	(0.02)	(0.07)	(0.2)
log(Capacity factor)	-0.15^{***}	-0.19^{***}	-0.30^{***}	-1.50^{***}
	(0.02)	(0.03)	(0.03)	(0.31)
log(Efficiency)	-0.42^{*}	-1.01^{***}	-0.73^{***}	0.07
	(0.21)	(0.21)	(0.10)	(0.98)
log(Fuel cost)	0.82^{***}	0.78^{***}	0.54^{***}	0.61^{***}
	(0.06)	(0.03)	(0.09)	(0.18)
log(Age)	-0.00	-0.06^{***}	-0.10^{*}	-0.15
	(0.02)	(0.02)	(0.06)	(0.27)
Adj. R^2	0.85	0.90	0.78	0.55
Num. obs.	63	149	140	30

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$, (standard error)

Table 4.3: *Regression of total operational costs on select independent parameters. With the exception of nuclear power, all technologies show a significant, albeit weak, decreasing trend in operational cost with increasing generator nameplate capacity. Moreover, again with the exception of nuclear, efficiency, fuel cost and capacity factor are all significant predictors of operational cost.*

This trend could be explained by more stringent environmental controls for newer plants, e.g. sulfur scrubbing of the flue gas in coal plants, which increase operational cost and from which some of the older plants are exempted through grandfathering (Ackerman et al., 1999).

4.3 Labor and Efficiency as Functions of Unit Scale

It was suggested in Chapter 3 that decreasing labor cost has been one of the main historical drivers for increasing unit scale. In a paradigm without automation, at first approximation, the amount of labor scales primarily with the number of units in operation. On the other hand, increasing unit size requires approximately zero marginal labor. To test this theory, labor cost per unit output is regressed on the same independent variables as in the previous analysis,

$$\log Y_{\text{labor}} = \log \beta_0 + \sum_i \beta_i \log X_i.$$

As can be seen in Table 4.4, the assumption of decreasing labor cost with size is verified for coal, combined cycle and nuclear technologies. Moreover, since labor typically is not considered a short-run production factor within the time scale of one year, operating at a higher capacity factor naturally decreases labor cost per unit output, which is evidenced in the result for these two technologies. Interestingly, combined cycle also exhibits significantly decreasing labor cost with efficiency. One explanation to this trend could be the decreased mass flows (fuel but also possibly coolant) that accompany a higher efficiency, which lessens the need for maintenance and consequently also labor.

The overall strength of the test is much lower for nuclear compared to the other technologies. This suggests that the independent variables only account for a small portion of the total variance of the labor cost per unit output. Compared to fossil fuel-powered generation, security and the handling of radioactive material at nuclear power facilities requires

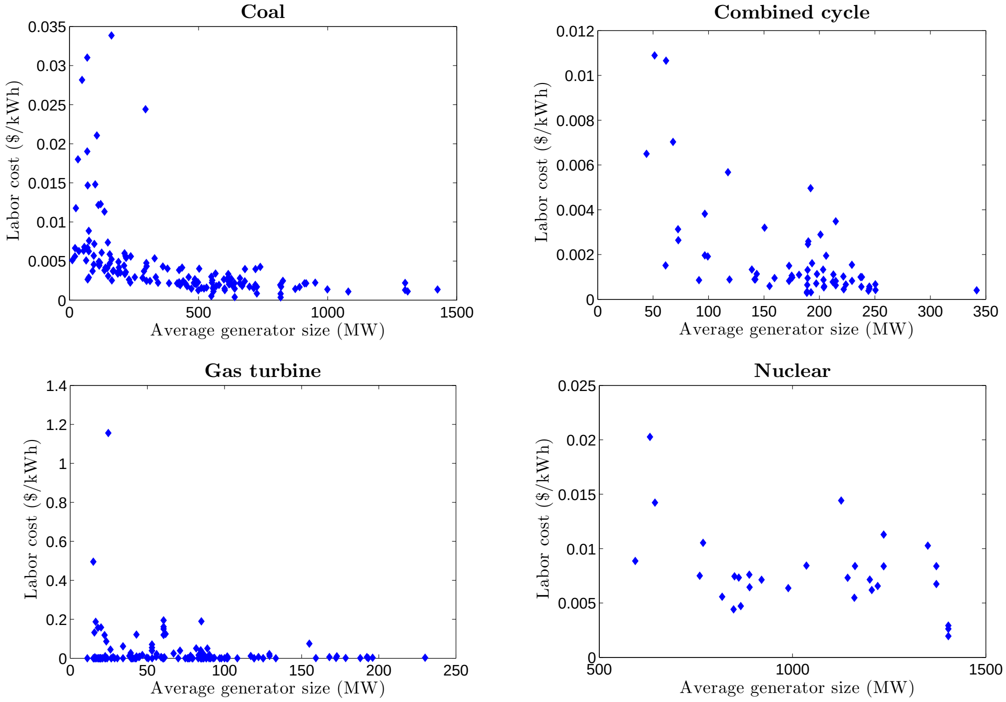

Figure 4.5: *With the exception of gas turbine technology, all technologies exhibit a significant strong decreasing trend of labor cost with unit size.*

	Combined Cycle	Coal	Gas Turbine	Nuclear
log(Nameplate capacity)	-0.79^{***}	-0.44^{***}	0.24	-1.03^{*}
	(0.12)	(0.05)	(0.59)	(0.51)
log(Capacity factor)	-0.93^{***}	-0.71^{***}	0.40	-1.14
	(0.07)	(0.08)	(0.25)	(0.78)
log(Efficiency)	-0.45^{**}	0.27	-0.78	0.26
	(0.60)	(0.57)	(0.86)	(2.47)
log(Fuel cost)	0.05	0.13	0.72	-0.30^{***}
	(0.17)	(0.10)	(0.79)	(0.45)
log(Age)	-0.01	0.01	0.04	-0.30
	(0.07)	(0.06)	(0.52)	(0.68)
Adj. R^2	0.89	0.74	-0.01	0.14
Num. obs.	63	149	140	30

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$, (standard error)

Table 4.4: *Results of a log-linear regression of labor cost per unit output on select independent parameters. With the exception of gas turbines, all other technologies exhibit both significant and strongly decreasing labor cost with unit size. The marked weakness of the regression for gas turbines is likely explained by the peaking nature of this technology, which causes reported employee count to be zero in many instances.*

additional labor, the cost of which is not adequately captured by the available independent variables. The significant trend of decreasing labor cost with fuel cost for nuclear power may find its explanation in the different assemblies of the fuel that command different prices and also different levels of handling in the plant.

Considering the R^2 values for the regression for gas turbine technology, these are very low. The most likely explanation to this bad fit is the peaking nature of this technology and consequently, the low average capacity factors encountered compared to the other technologies. At low capacity factors the reported whole number of employees attributable to each generator, were generally either zero or one, which arguably overshadows significant arbitrariness, causing a low accuracy of the regression.

The labor cost per unit output of the four technologies are displayed in Figure 4.5 as function of generator size. As part of overall operational expenditures presented in Figure 4.4, labor accounts for roughly 10% and even less for the very largest installations. Continuing the trend of increasing unit scale will therefore yield diminishing returns to scale. On the other hand, operating very small generators (on the order of kW) would, by extrapolating the data, suggest incurring prohibitively high labor cost.

To test if factors other than labor would strongly discriminate against smaller sizes an additional regression analysis is performed. In this analysis, the total operating cost net of the labor component is regressed on the same independent variables as before:

$$\log\left(Y_{\text{tot cost}} - Y_{\text{labor}}\right) = \log\beta_0 + \sum_i \beta_i \log X_i. \tag{4.2}$$

The regression result of this model is displayed in Table 4.5. With the labor component taken out, the unit size either fails to be a significant predictor of cost, or it even suggests a increasing trend, as in the case of coal. Based on this analysis, the possibility of fully automating processes and thereby eliminating, or at least drastically reducing the need for

	Combined Cycle	Coal	Gas Turbine	Nuclear
log(Nameplate capacity)	-0.06	0.00^{*}	-0.11	0.03
	(0.05)	(0.02)	(0.11)	(0.30)
log(Efficiency)	-0.34	-1.28^{***}	-0.78^{***}	-0.25
	(0.22)	(0.25)	(0.16)	(1.45)
log(Capacity factor)	-0.12^{***}	-0.09^{**}	-0.22^{***}	-1.71^{***}
	(0.03)	(0.04)	(0.05)	(0.46)
log(Fuel cost)	0.84^{***}	0.89^{***}	0.65^{***}	0.77^{***}
	(0.06)	(0.04)	(0.14)	(0.26)
log(Age)	0.01	-0.04	-0.01	0.02
	(0.03)	(0.02)	(0.1)	(0.40)
Adj. R^2	0.883	0.86	0.56	0.40
Num. obs.	63	149	140	30

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$, (standard error)

Table 4.5: *Results of a log-linear regression of operating cost net of labor cost per unit output on select dependent parameters. Without the labor component, unit scale is not a significant predictor of variable cost among the technologies studied.*

labor, would open up the possibility of operating smaller units without significant operational cost increases.

In addition to labor, it was suggested in 3.2 that efficiency gains is another driver for increasing unit size. For these thermal technologies, larger units likely suffer reduced heat losses as well a lower relative frictional losses in the spinning components. Observing the visual display of efficiency versus size, see Figure 4.6, seems to support that argument. However, regressing efficiency on the other independent variables and total operating cost,

$$\log X_{\text{eff}} = \log \beta_0 + \sum_i \beta_i \log X_i + \beta' \log Y_{\text{tot cost}}, \tag{4.3}$$

reveals that this trend is weaker than appearances suggest. Importantly, all else being constant, efficiency is strongly correlated with fuel cost. While there is a quality difference among different types of coal, which could explain some of the variability in efficiency, this suggests that operational factors to a high degree can influence the observed efficiency. It should be reminded that the efficiency is computed from annual averages which may obscure the impact of short-term output fluctuation, which presumably lowers efficiency.

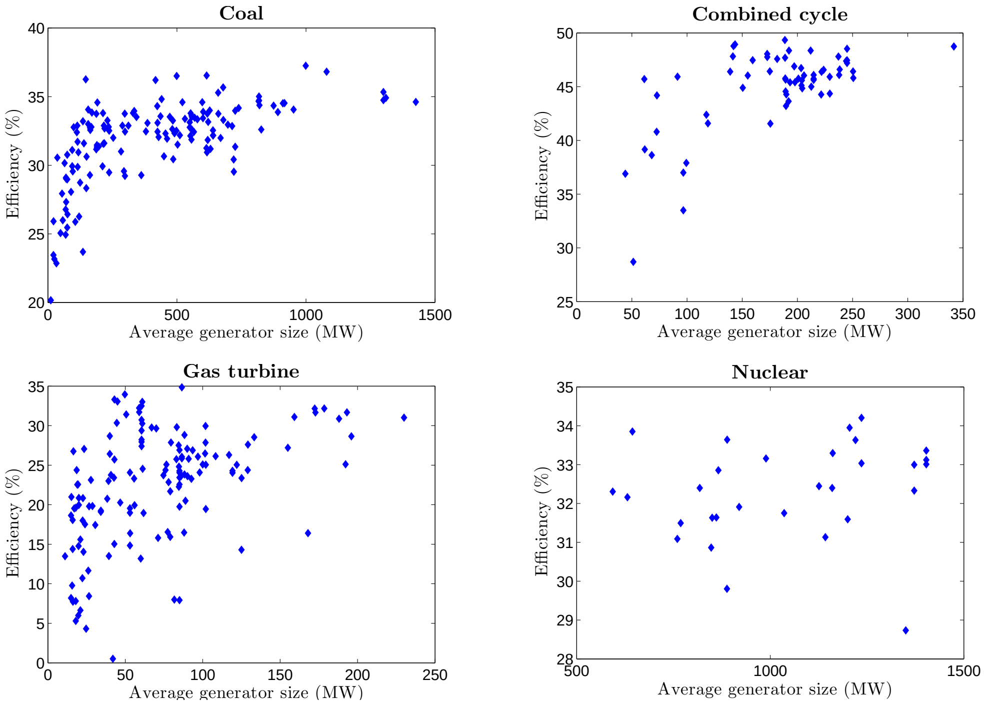

Figure 4.6: *The seemingly strong trend of increasing efficiency with size obscures the impact of other factors, primarily fuel cost.*

	Combined Cycle	Coal	Gas Turbine	Nuclear
log(Nameplate capacity)	0.09***	0.05***	0.00	0.02
	(0.02)	(0.01)	(0.05)	(0.04)
log(Capacity factor)	0.02	−0.09**	0.03**	0.13
	(0.02)	(0.01)	(0.03)	(0.09)
log(Fuel cost)	0.10***	0.89***	0.17***	0.05
	(0.07)	(0.02)	(0.07)	(0.04)
log(Age)	−0.02	−0.03***	−0.13***	−0.01
	(0.01)	(0.01)	(0.04)	(0.05)
log(Operational cost)	−0.15*	−0.14***	−0.40***	0.00
	(0.08)	(0.03)	(0.05)	(0.05)
Adj. R^2	0.58	0.72	0.59	0.09
Num. obs.	63	149	140	30

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$, (standard error)

Table 4.6: *Regressing efficiency on the other independent variables and total operational cost reveals a positive correlation on size, all else being constant. Not surprisingly, increasing fuel cost puts an impetus on more efficient operation. Also, the negative correlation between efficiency and cost is preserved from Table 4.3*

4.4 Summary

In this chapter, the suggested trend of “bigger is better” is established for six of the largest generating technologies (by installed capacity). Almost without exception, the strategy across all technologies in the U.S. electricity generating industry has been to scale up the sizes of the individual generating units. This strategy has led to substantial increases in nameplate capacity, up to orders of magnitude in some instances. While a larger unit scale reduces fixed cost, the large scales involved together with the long lead times in construction of new capacity, have arguably made it more difficult to plan capacity additions in accordance with demand growth.

With well-documented economies of unit scale for all technologies surveyed, the strategy of scaling up in size has reduced the investment costs of generating capacity. Moreover, for all technologies, except nuclear, a statistically significant albeit weak trend of decreasing operational cost with increasing size is observed. Thus, general operational economies of unit scale have been present in the industry as well, further motivating the drift to larger sizes.

Analyzing labor cost as function of unit scale confirms the rather natural assumption of lower labor cost with increasing size. The extent to which this increase in productivity drives the operational economies of unit scale is determined through another analysis. Specifically, taking as dependent variable the difference between total cost and labor cost allows for a regression where labor has been removed. Interestingly, the trend of decreasing operational cost with size that was visible previously is now absent. This suggests that labor is indeed the main driver of operational economies of unit scale. The emerging availability of low-cost automation technologies can therefore actively remove one of the main motives to build larger units.

The general assumption of increasing conversion efficiency with size finds some support in the analysis. However, even though fuel cost is the largest single variable cost item in all technologies, this trends does not drive total cost as evidenced in the previous analysis. The price of fuel is found to correlate with the observed efficiency. This trend is strongest for coal, which may be explained by the different quality of fuels available. That the trend is visible for natural gas-fired capacity, where the fuel is more standardized, reveals that operating conditions affect efficiency more than inherent size effects.

Chapter 5

Real Options of Small-Scale Investments

The application of theory, as well as tools and nomenclature, developed in the financial industry to problems regarding both operation and investment in real assets has grown more ubiquitous over the past decades. A correct implementation of these real options has the potential to increase the value of a firm (Sick and Gamba, 2010). Furthermore, making investment decisions based on a real options analysis can increase value above and beyond those based merely on classical net present value considerations.

In the simplest case, the value of a project is influenced by only one stochastic variable that changes over time. Equipped with an estimate of the first moment (the mean) of this stochastic process, a manager can determine if the sum of the discounted cash flow from the project supersedes the cost of making the investments, i.e. if the net present value is positive. In this case, the manager should recommend proceeding with the investment according to classical net present value analysis. As explained succinctly in (McDonald and Siegel, 1986), in contrast to this binary outcome of 'invest now or don't invest now', a real options analysis evaluates the expected value of an investment over all future times. This results in an optimal timing rule when to invest. In order to undertake such an analysis, information about both the first and the second moment (variance) of the process is required. A further introduction to real options can be found in (Dixit and Pindyck, 1994).

The notion of investment timing is particularly prescient in the commodity-based industries due to their sheer size in the current paradigm of "bigger-is-better". If market or regulatory conditions change adversely, then there is generally little or no secondary use for the physical capital. Terminating operation, or perhaps even construction, therefore entails stranding most or all of the invested capital. An example that vividly illustrates the irreversibility of investments in these industries, magnified by the large scales involved, is the Shoreham Nuclear Power station on Long Island. After 12 years of construction the plant's operating license was revoked in the mid 1980s and a $5.6 billion dollar investment wiped out before a single kWh of power was sold (Levendis et al., 2006).

By employing a real option analysis Pindyck (1986) showed that the existence of uncertainty regarding future profits makes the optimal capacity much smaller than if the investment were reversible. More recent real option analyses regarding infrastructure investments can be found in (Westner and Madlener, 2012; Frayer and Uludere, 2001; Kaslow and Pindyck, 1994). For large scale projects, the possibility of sequencing investment decisions in a single project to get additional information further enhances to option value of these kinds of investments (Carelli et al., 2010; Lumley and Zervos, 2001).

In addition to the high degree of irreversibility that the large individual investments in productive capital have in the industries in question, they also have very long pay-back times. Of the 'large projects', defined by a capital cost of over $1 billion, that were funded through project finance[1] during the years 1997-2001, more than half had a concession agreement

[1] This structure lets the sponsoring companies raise the necessary funds 'off-balance sheet' by creating a separate entity associated with the project alone. Typically, these projects have a high degree of leverage (up to 80%), made possible by long-term offtake agreements of necessary inputs, as well as the produced output (Fight, 2005; Brealey et al., 2008).

lasting more than 25 years (Finnerty, 2007). Thermal power plants are typically planned to have even longer economic life, of up to 40 years (Deutch et al., 2003). Moreover, in addition to the long lifetimes that are required to yield financial viability, the lead times in planning and construction of these large projects are also long, up to 10 years.

Whether it be by design in construction or in operation with a limited maintenance schedule, small and mass-produced equipment might be endowed with a much shorter physical lifespan than these large-scale investment. Moreover, with mass-production of small-scale units comes the ability to build to stock and drastically shorten lead times between the investment and start of operation. Deploying small-scale capital therefore provides additional flexibility to engage and disengage a given activity. This increased flexibility with shorter lifetimes and lead times is the focus of this chapter.

To capture this increased flexibility, a framework that incorporates operational flexibility, lead time, capital lifespan, and multiple (finite/infinite) future investment opportunities is introduced below. The investment problem is formulated as an optimal multiple stopping problem, where the firm maximizes the expected discounted reward from sequential investments. In particular, embedded in the formulation is the operational flexibility of temporarily suspending production to avoid negative cash flows. Importantly, the problem depends explicitly on lifetime and lead time. Under a capacity constraint, the firm's consecutive investments are separated (or refracted) by the lifetime of the capital. Hence, the firm's investment decision bears similarity to the valuation of a forward-starting swing call option written on the expected reward.

Swing options have frequently been used to price energy delivery contracts where the holder has the right to alter, or 'swing,' volumes up or down at the start of each time period (Jaillet et al., 2004; Deng and Oren, 2006). Instead of fixed-period contracts, the only constraint in the model below is a minimum separation time between consecutive investments.

This allows for the valuation of the swing contract as a refracted optimal multiple stopping problem (see e.g. Carmona and Touzi (2008); Carmona and Dayanik (2008)). Other financial applications involving multiple exercises include employee stock option valuation (Leung and Sircar, 2009; Grasselli and Henderson, 2009), and the operation of a physical asset (Ludkovski, 2008). In contrast to the simulation approaches commonly found in existing literature for swing-type options (Meinshausen and Hambly, 2004; Chiara et al., 2007; Bender, 2011) the problem formulation here lends itself to an iterative algorithm that can be solved numerically.

Resulting from the model below is both the firm's value function and its optimal stopping rule, which is described by a sequence of critical price thresholds. This sequence is shown to be decreasing and converging to the threshold corresponding to the case with infinite investment opportunities. Moreover, this framework is also useful for analyzing the critical investment cost that makes small-scale (short lead time, short lifetime) alternatives competitive with traditional large-scale investments of the same or similar technology. Returning to the comparison between the car engine and the power plant, the former is mass-produced in days and likely to last on the order of years, whereas the power plant follows conventional time pattern described above. With the investment cost known for the large-scale power plant this model can be used to estimate the maximal investment cost for a car engine, retrofitted to run on the same fuel and equipped with a generator to produce electricity, to be competitive with the conventional power plant.

Proofs of propositions, lemmas, and theorems below can be found in Appendix A.

5.1 Problem Formulation

The agent in this formulation is a firm that has the ability to invest in capital equipment that produces a single good in a given market. Furthermore, this firm is assumed to be acting as a price taker in this market with a finite capacity constraint. These assumptions void the need for incorporating demand elasticities and instead implies that the investment decision will be of 'bang-bang' type. That is, if an investment is made, it will be up to the given capacity constraint. Moreover, the operational cost is assumed to be independent of individual unit size. Under such circumstances, the investment can be analyzed on a per-unit-capacity basis.

The investment cost, $I(T,\nu)$, is considered exogeneous and deterministic and depends on lifetime, T, and lead time, ν, among other factors. These two parameters, both considered deterministic, also affect directly the net present value, ψ, of a single investment. Moreover, the discount rate, r, used by the firm is also considered to be exogenous and constant. In the background, a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, equipped with a filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ is fixed. Let $(X_t)_{t\geq 0}$ be an $\mathcal{F}_t$-adapted output price process. An investment generates a random cash flow-process of the form $(f(X_t))_{t\geq 0}$, where f is a function known to the firm.

The expected discounted stream of future cash flows, minus the initial investment cost, yields the net present value

$$\psi(x;T,\nu) = -I(T,\nu) + \int_{\nu}^{\nu+T} e^{-rt}\,\mathbb{E}\left\{f\left(X_t^{0,x}\right)\right\} dt, \qquad 0 < x < \infty, \tag{5.1}$$

where the conditional notation $X_t^{0,x} \equiv \{X_t | X_0 = x\}$ is introduced. The expression in (5.1) helps clarify the role of the lead time ν as the time between the expenditure I and start of operation and access to the cash flow $f(X_t)$.

The limitation on the amount of capacity that can be active at any given point in time, as implied by the assumption of the role as a price taker, naturally introduces the notion of a refraction time. That is, two consecutive investments have to be separated with at least a time T, the lifetime of the capital. With this being the only restriction on the investment strategy of the firm, the value $v^{(k)}(x;T,\nu)$ of k consecutive investments can be formulated as an optimal multiple stopping problem

$$v^{(k)}(x;T,\nu) = \sup_{\vec{\tau}\in\mathcal{S}^k} \mathbb{E}\left\{\sum_{i=1}^{k} e^{-r\tau_i}\psi(X_{\tau_i}^{0,x},T,\nu)\right\}, \qquad 0 < x < \infty. \tag{5.2}$$

The set $\mathcal{S}^k$ is the set of all refracted stopping times $\vec{\tau} = (\tau_1,\tau_2,\ldots,\tau_k)$, i.e. $\tau_i - \tau_{i-1} \geq T$ for $i = 2,3\ldots,k$. Under this requirement, nothing prevents the firm from having two consecutive investments operating with a seamless transition. This is because the decision to invest in future capital can be made before the current investment expires. Provided that a least upper bound in (5.2) actually exists, the stopping vector $\tau_p = \tau_{p+1} = \cdots = \tau_k = \infty$, $p \leq k$ can be included in $\mathcal{S}^k$. For such a stopping rule, $v^{(k)}$ can be interpreted as the value of the contract when not every exercise is being called.

The tacit assumption that the supremum in (5.2) exists implies that $e^{-rt}\psi(X_t;T,\nu)$ is integrable and $\lim_{t\to\infty}\mathbb{E}\left\{e^{-rt}\psi(X_t;T,\nu)\right\} = 0$ for every choice of T and ν. For brevity, the dependence on the parameters T and ν in ψ, $v^{(k)}$ and I is suppressed unless such a dependence is specifically required. Since the firm is free to choose the timing of every investment, up to the condition of a refraction time T, the following proposition highlights the optionality embedded in the investment decision.

Proposition 1. *The value function $v^{(k)}(x)$, $k \geq 1$, satisfies,*

$$v^{(k)}(x) = \sup_{\vec{\tau}\in\mathcal{S}^k} \mathbb{E}\left\{\sum_{i=1}^{k} e^{-r\tau_i}\psi^{+}(X_{\tau_i}^{0,x})\right\}, \tag{5.3}$$

with $\psi^{+}(x) = \max[0,\psi(x)]$.

This proposition reveals that it is never optimal to make the investment in the negative region of $\psi(x)$. The firm has the ability to delay investment, and therefore, can always avoid value destruction.

The above problem formulation, as well as Proposition 1, holds for any underlying stochastic process. Proceeding, the formulation in (5.1) and (5.2) are considered under the geometric Brownian motion (GBM) model. Specifically, the output price process is modeled by the stochastic differential equation (SDE):

$$dX_t = \alpha X_t \, dt + \sigma X_t \, dB_t, \qquad X_0 \in (0, \infty), \tag{5.4}$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion. The drift rate α and the variance parameter σ are both assumed constant. The filtration generated by B is denoted $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$.

From the firm's perspective, the investment cash flow is assumed to be the difference between an uncertain output price X_t, modeled by (5.4), and a constant operational cost c. Moreover, the firm has the *operational flexibility* to temporarily suspend production if cost exceeds output price. Hence, the effective investment cash flow $f(X_t^{0,x})$ at any future time t is given by

$$f(X_t^{0,x}) = \left(X_t^{0,x} - c\right)^+, \tag{5.5}$$

where $(X_t - c)^+ = \max[0, X_t - c]$. That is, at time t the firm has the right, but not the obligation, to claim a revenue X_t at a cost c. This type of option is called a European call option on X_t with a strike price c and maturity t. The expectation of $f(X_t^{0,x})$ in (5.5) can be seen as the undiscounted price of such a European call option (McDonald and Siegel, 1985), namely,

$$\mathbb{E}\left\{\left(X_t^{0,x} - c\right)^+\right\} = x\Phi(d_+(t))e^{\alpha t} - c\Phi(d_-(t)), \tag{5.6}$$

where Φ is the standard normal cumulative distribution function, and where

$$d_{\pm}(t) = \left[\ln\left(\frac{x}{c}\right) + \left(\alpha \pm \frac{1}{2}\sigma^2\right) t\right] / \sigma\sqrt{t}\,.$$

With the above expressions the reward function $\psi(x)$ becomes

$$\psi(x) = -I + \int_{\nu}^{\nu+T} \left(x\Phi(d_+(t))e^{\alpha t} - c\Phi(d_-(t))\right) e^{-rt} dt \,.$$

Under this setting, the optimal multiple stopping problem in (5.2) is recasted as a sequence of optimal single stopping problems. More precisely, the value $v^{(k)}(x)$ is expressed as

$$v^{(k)}(x) = \sup_{\tau_k \in \mathcal{S}} \mathbb{E}\left\{e^{-r\tau_k}\psi^{(k)}(X^{0,x}_{\tau_k})\right\} . \tag{5.7}$$

The set $\mathcal{S}$ is the set of all $\mathbb{F}$-stopping times and

$$\psi^{(k)}(x) = \psi(x) + e^{-rT}\mathbb{E}\left\{v^{(k-1)}(X^{0,x}_T)\right\} , \quad v^{(0)} \equiv 0 \,.$$

The entity $\psi^{(k)}(x)$ can be interpreted as the value of a single investment plus the value of $k-1$ future investment opportunities. By construction, the admissible stopping times $(\tau_1, \ldots, \tau_k)$ are again refracted by at least T (years). By standard optimal stopping theory, the process $\left(e^{-rt}v^{(k)}(X_t)\right)_{t\geq 0}$, called the Snell envelope of the process $\left(e^{-rt}\psi^{(k)}(X_t)\right)_{t\geq 0}$, is constructed to be the smallest supermartingale dominating the $\left(e^{-rt}\psi^{(k)}(X_t)\right)_{t\geq 0}$, for every k. For further details regarding this approach under a more general framework, see Carmona and Touzi (2008).

5.2 Analytical Results

This section presents an analytical study of the optimal stopping problem in (5.2.1). The main result, given in Theorem 1, is the characterization of the value function and the optimal stopping rule, for every $k \geq 1$ opportunities to invest. With these results an iterative approach to finding the value functions $v^{(k)}(x)$ and the optimal exercise boundaries x^*_k is developed in Corollary 1.

In the context of real investments, some general conditions are imposed on the reward function ψ. First, ψ is assumed to be continuous, increasing and sufficiently smooth. Additionally, it is assumed that there is a unique break-even point, $x_0 > 0$, where $\psi(x_0) = 0$, and $\psi(x) < 0$ for $x < x_0$ and $\psi(x) > 0$ for $x > x_0$. Finally, the function $\psi(x)$ is assumed to be bounded by some affine function of x. Such a bound, together with an assumption that the discount rate r exceeds the drift rate α of the underlying process, ensures that perpetual waiting will lead to zero expected reward, i.e. $\lim_{t\to\infty} \mathbb{E}\left\{e^{-rt}\psi(X_t)\right\} = 0$. Under these conditions, the real option problem with a single investment opportunity is analyzed, which subsequently is the building block for solving the optimal multiple stopping problem.

5.2.1 Optimal Single Stopping Problem

With only one investment opportunity, i.e. $k = 1$, the stopping problem in () is given by

$$v^{(1)}(x) = \sup_{\tau_1 \in \mathcal{S}} \mathbb{E}\left\{e^{-r\tau_1}\psi(X^{0,x}_{\tau_1})\right\}. \tag{5.8}$$

As opposed to the previously introduced European option (possible exercise at maturity only), the formulation in (5.8) considers the optimal present value of an investment over all future times. An option with the right to exercise at any future time to recieve the payoff ψ is called a perpetual American option on ψ, the price of which is given by $v^{(1)}(x)$.

Let the first passage time $\tau_{x_1^*}$ be a candidate stopping time for the problem in (5.8):

$$\tau_{x_1^*} = \inf\{t \geq 0 \,|\, X^{0,x}_t \geq x_1^*\,,\, x > 0\,\},$$

with threshold $x_1^* > 0$. Taking the Laplace transform of the first passage time of X (see e.g.

(Shreve, 2004, p.346)), for $x \leq x_1^*$, gives

$$\mathbb{E}\left\{e^{-r\tau_{x_1^*}}\psi(X^{0,x}_{\tau_{x_1^*}})\right\} = \psi(x_1^*)\left(\frac{x}{x_1^*}\right)^{\gamma}. \tag{5.9}$$

The power γ in (5.9) is the positive solution to the second order equation

$$\frac{1}{2}\sigma^2\gamma(\gamma-1)+\alpha\gamma-r=0, \tag{5.10}$$

which arises from the time-independent Black-Scholes differential equation (see Appendix A). It can be seen from (5.10) that the condition $r > \alpha$ implies that $\gamma > 1$. If $x \geq x_1^*$, then $\tau_{x_1^*} = 0$ and $v^{(1)}(x) = \psi(x)$. From (5.9) it can be seen that a necessary condition for x_1^* to be an optimal stopping boundary is that x_1^* maximizes $\psi(x)/x^{\gamma}$. The assumption of a linear bound on $\psi(x)$ ensures the existence of such a maximum.

The first order condition for a maximum at $x = x_1^*$ is given by

$$\left.\frac{d}{dx}\frac{\psi(x)}{x^{\gamma}}\right|_{x=x_1^*} = \frac{x_1^*\psi'(x_1^*)-\gamma\psi(x_1^*)}{(x_1^*)^{\gamma+1}} \equiv -\frac{\Lambda\psi(x_1^*)}{(x_1^*)^{\gamma+1}} = 0 \quad \Leftrightarrow \quad \Lambda\psi(x_1^*) = 0, \tag{5.11}$$

where the operator notation $\Lambda = \left(\gamma - x\frac{d}{dx}\right)$ is introduced. The second order condition, sufficient to prove a maximum together with the first order condition above, can then be stated as

$$\left.\frac{d^2}{dx^2}\frac{\psi(x)}{x^{\gamma}}\right|_{x=x_1^*} = -\frac{\frac{d}{dx}\Lambda\psi(x_1^*)}{(x_1^*)^{\gamma+1}} < 0 \quad \Leftrightarrow \quad \frac{d}{dx}\Lambda\psi(x_1^*) > 0.$$

Note that a maximum of $\psi(x)/x^{\gamma}$ at x_1^* is not a sufficient condition for an optimal stopping rule $\tau_{x_1^*}$ for a general reward function ψ. One would also have to prove that $\left(e^{-rt}v^{(1)}(X_t)\right)_{t\geq 0}$ satisfies the supermartingale property for this choice of $\tau_{x_1^*}$. The following lemma gives the sufficient conditions for $\tau_{x_1^*}$ to be an optimal stopping rule for problem (5.8).

Lemma 1. *Let $\psi : \mathbb{R}^+ \to \mathbb{R}$ be a reward function in the single stopping problem (5.8). If x_1^**

is a global maximum for $\psi(x)/x^\gamma$ on $\mathbb{R}^+$ and if

$$\frac{d}{dx}\Lambda\psi(x) \equiv (\gamma-1)\psi'(x) - x\psi''(x) \geq 0, \quad x \geq x_1^*, \tag{5.12}$$

then

$$v^{(1)}(x) = \psi(x \vee x_1^*)\left[1 \wedge \left(\frac{x}{x_1^*}\right)^\gamma\right], \quad x \in \mathbb{R}^+,$$

where $v^{(1)}(x)$ is continuous on $\mathbb{R}^+$.

Remark 1. The first order condition $\Lambda\psi(x_1^*) = 0$ in (5.11), together with the condition that $(d/dx)\Lambda\psi(x) \geq 0$, $x \geq x_1^*$, bounds the second derivative (convexity) of the reward function $\psi(x)$ for large x. Up to the condition of a maximum of $\psi(x)/x^\gamma$ at $x = x_1^*$, the behavior of the function $\psi(x)$ to the left of $x = x_1^*$ is irrelevant. The conditions in Lemma 1 are therefore less restrictive than requiring that the drift term of $e^{-rt}\psi(X_t)$ be monotonic for all X_t, as presented in (Dixit and Pindyck, 1994, pp.128-130) as a part of the sufficient conditions for a connected stopping boundary at $x = x_1^*$ for a perpetual American call on $\psi(x)$.

5.2.2 Optimal Multiple Stopping Problem

With Lemma 1 and Proposition 1, the main result can be stated and proved.

Theorem 1. *Let $\psi : \mathbb{R}^+ \to \mathbb{R}$ be a reward function with a break-even point x_0. If $\Lambda\psi(x)$ is convex for $x \in (x_0, \infty)$, with $\Lambda\psi(x)$ increasing for large x, then, for every $k \geq 1$, there exists an $x_k^* > x_0$ such that*

$$v^{(k)}(x) = \psi^{(k)}(x \vee x_k^*)\left[1 \wedge \left(\frac{x}{x_k^*}\right)^\gamma\right], \quad k \geq 1, \tag{5.13}$$

where

$$\psi^{(k)}(x) = \psi(x) + e^{-rT}\mathbb{E}\left\{v^{(k-1)}(X_T^{0,x})\right\}. \tag{5.14}$$

Moreover, the sequence $(x_k^*)_{k\geq 1}$ is strictly decreasing, and $\left(v^{(k)}\right)_{k\geq 1}$ is a strictly increasing sequence of continuous functions on $\mathbb{R}^+$. Also, for any bounded subset $D \in \mathbb{R}^+$ there exists a constant K_D, such that $v^{(k)}(x) \leq K_D$, for $x \in D$ and $k \geq 1$.

From a computational perspective it is convenient to introduce the auxiliary function $u^{(k)}$ through

$$u^{(k)}(x) = \Lambda v^{(k)}(x) = \Lambda \psi^{(k)}(x) 1_{\{x \geq x_k^*\}}. \tag{5.15}$$

With the conditions imposed on $\psi(x)$ (bounded by a linear function, convexity of $\Lambda\psi(x)$ on (x_0, ∞) and $\Lambda\psi(x)$ being increasing for large x), one can show that $\left(u^{(k)}\right)_{k\geq 1}$ is an increasing sequence of continuous functions bounded on every compact subset $D \subset \mathbb{R}^+$. Given the function $u^{(k)}(x)$ the value function $v^{(k)}(x)$ can be reconstructed through

$$v^{(k)}(x) = x^\gamma \left(\frac{\psi^{(k)}(x_k^*)}{(x_k^*)^\gamma} - \int_0^x y^{-\gamma-1} u^{(k)}(y) dy \right). \tag{5.16}$$

Since $u^{(k)}(x) = 0$ for $x < x_k^*$ there are no convergence issues in (5.16). The following corollary follows from Theorem 1.

Corollary 1. *The functions $u^{(k)}(x)$, for $k = 1, 2, \ldots,$ satisfy*

$$u^{(k)}(x) = \left(\Lambda\psi(x) + e^{-rT}\mathbb{E}\left\{u^{(k-1)}(X_T^{0,x})\right\}\right) 1_{\{x \geq x_k^*\}}, \quad u^{(0)}(x) \equiv 0, \tag{5.17}$$

The boundary point x_k^* is the unique solution to

$$\begin{cases} \Lambda\psi(x_k^*) + e^{-rT}\mathbb{E}\left\{u^{(k-1)}(X_T^{0,x_k^*})\right\} & = 0, \\ \frac{d}{dx}\left(\Lambda\psi(x) + e^{-rT}\mathbb{E}\left\{u^{(k-1)}(X_T^{0,x})\right\}\right)\Big|_{x=x_1^*} & > 0. \end{cases} \quad \square \tag{5.18}$$

Corollary 1, together with equation (5.16), outlines the inductive algorithm used to find the solution to (5.2.1). Proof of uniform convergence $v^{(k)} \to v^{(\infty)}$ and details of the numerical implementation can be found in Appendix A.

5.3 Application to Infrastructure Investments

The above framework is here applied to a general infrastructure investment. Importantly, the result allows for a sensitivity study of both the value function $v^{(\infty)}$ and the stopping boundary x_∞ with respect to the key parameters of lifetime T and lead time ν, as well as the process parameters. Moreover, the framework is employed to compare investment scenarios with relatively short lifetimes and lead times to a scenario with both a long lifetime and lead time. Such a comparison will reveal a critical investment cost of the short-lived scenario below which it will be competitive with the long-lived counterpart.

A common feature of many projects is that the cash flow is determined, to a great extent, by the price of one or more commodities. For instance, the the 'spark spread', i.e. the difference between the price of electricity and the price of natural gas, dominates the cash flow for natural gas-fired power plants. For desalination, the difference in the price of fresh water and electricity gives rise to a similar spread, etc. It is therefore reasonable to assume that the observable cash flow, Z_t, can be written as the difference between two price processes, denoted U_t and Y_t, i.e. $Z_t = U_t - Y_t$. It will here be demonstrated that if by U and Y are correlated GBMs, the process Z can still be represented by a single GBM.

Suppose that U and Y evolve according to

$$\begin{aligned} dU_t &= \alpha_U U_t \, dt + \sigma_U U_t \, dB_t^{(1)}, \\ dY_t &= \alpha_Y Y_t \, dt + \sigma_Y Y_t \, dB_t^{(2)}, \end{aligned}$$

where $B^{(1)}$ and $B^{(2)}$ are two standard Brownian motions with $\mathbb{E}\{dB_t^{(1)} dB_t^{(2)}\} = \rho dt$, with correlation coefficient $\rho \in [-1, 1]$. Let X_t be the quotient between output and input price,

$X_t = U_t/Y_t$. With the Itô-Doeblin formula one obtains

$$\begin{aligned} dX_t &= \frac{1}{Y_t}dU_t - \frac{U_t}{Y_t^2}dY_t - \frac{1}{Y_t^2}dU_t dY_t + \frac{U_t}{Y_t^3}dY_t^2 \\ &= \left(\alpha_U - \alpha_Y + \sigma_Y^2 - \rho\sigma_U\sigma_Y\right) X_t dt + X_t \left(\sigma_U dB_t^{(1)} - \sigma_Y dB_t^{(2)}\right) \\ &= \alpha X_t dt + \sigma X_t dB_t. \end{aligned} \tag{5.19}$$

As a result, the ratio $X_t = U_t/Y_t$ is also a GBM with

$$\alpha = \alpha_U - \alpha_Y + \sigma_Y^2 - \rho\sigma_U\sigma_Y, \quad \sigma = \sqrt{\sigma_U^2 + \sigma_Y^2 - 2\rho\sigma_U\sigma_Y}\,, \tag{5.20}$$

and with the standard Brownian motion

$$B_t = \frac{\sigma_U B_t^{(1)} - \sigma_Y B_t^{(2)}}{\sigma}\,, \quad t \geq 0.$$

From (5.20) it is noted that a negative (resp. positive) correlation of $B_t^{(1)}$ and $B_t^{(2)}$ can increase (resp. decrease) the volatility and drift of X. The cash flow is given by

$$Z_t = U_t - Y_t = Y_t\left(X_t - 1\right), \quad t \geq 0,$$

where

$$Y_t = Y_0 e^{\alpha_Y t} e^{-\frac{1}{2}\sigma_Y^2 t + \sigma_Y B_t^{(2)}}.$$

This motivates the definition of the new probability measure $\tilde{\mathbb{P}} \sim \mathbb{P}$ according to

$$\left.\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}\right|_{\mathcal{F}_t} = e^{-\frac{1}{2}\sigma_Y^2 t + \sigma_Y B_t^{(2)}}. \tag{5.21}$$

Applying a change of measure, the expected future discounted cash flows can be written as

$$\mathbb{E}\left\{e^{-rt}\left(U_t - Y_t\right)\right\} = \tilde{\mathbb{E}}\left\{e^{-(r-\alpha_Y)t}\left(Y_0 X_t - Y_0\right)\right\},$$

where $\tilde{\mathbb{E}}$ indicates the expectation under the measure $\tilde{\mathbb{P}}$.

To summarize, the change of measure turns the observable cash flow of two uncertain parameters back to the original form for X, whose drift α and σ are given in (5.20) under the measure $\tilde{\mathbb{P}}$, along with $c = Y_0$. The requirement that the effective discount rate $r - \alpha_Y$ must exceed the drift α of X amounts to

$$r > \alpha_U + \sigma_Y^2 - \rho\sigma_U\sigma_Y. \tag{5.22}$$

Interestingly, the drift of Y is irrelevant in regards to condition (5.22) and the integrability of $e^{-rt}Z_t$ under $\tilde{\mathbb{P}}$.

Recapitulating the choice of cash flow $f(X_t)$, introduced in Section 5.1, where the flexibility to temporarily suspend production to avoid a negative cash flow was incorporated, namely,

$$f(X_t) = (X_t - c)^+ .$$

The reward function associated with this cash flow is

$$\psi(x) = -I + \int_{\nu}^{\nu+T}\left(x\Phi(d_+(t))e^{\alpha t} - c\Phi(d_-(t))\right)e^{-rt}dt, \tag{5.23}$$

where

$$d_\pm(t) = \left[\ln\left(\frac{x}{c}\right) + \left(\alpha \pm \frac{1}{2}\sigma^2\right)t\right]/\sigma\sqrt{t}. \tag{5.24}$$

Since $f(X_t)$ is bounded by X_t, one realizes that there is a linear function bounding $\psi(x)$.

Investigating the derivatives of $\psi(x)$ yields[2]

$$\psi'(x) = \int_{\nu}^{\nu+T} \frac{d}{dx}\left(x\Phi(d_+(t))e^{\alpha t} - c\Phi(d_-(t))\right)e^{-rt}dt = \int_{\nu}^{\nu+T} \Phi(d_+(t))e^{-(r-\alpha)t}dt > 0. \tag{5.25}$$

From (5.23)-(5.25) it is seen that $\psi(x)$ is continuous and increasing in x. Furthermore, since $\lim_{x\to 0}\psi(x) = -I < 0$ there exists a unique break-even point x_0. Thus, $\psi(x)$ satisfies all the conditions in the definition of a reward function presented in Section 5.2.1. By differentiation, one obtains

$$\frac{d^2}{dx^2}[\Lambda\psi(x)] = \int_{\nu}^{\nu+T} \frac{\phi(d_+(t))}{x\sigma^2 t}\left((\gamma-1)\sigma\sqrt{t} + d_+(t)\right)e^{-(r-\alpha)t}dt,$$

where $\phi(x)$ is the density function of the normal distribution. It can be seen that $\Lambda\psi(x)$ is convex for all $x \geq x'$, where $x' = c\exp\left[-\left(\alpha + \gamma\sigma^2 - \frac{1}{2}\sigma^2\right)\nu\right]$. In turn, if $x_0 \geq x'$, which is ensured by imposing a lower bound on I, then $\Lambda\psi(x)$ is convex on (x_0, ∞), and consequently, all the conditions for Theorem 1 are satisfied and the algorithm outlined in Corollary 1 can be applied.

5.3.1 Sensitivity Analysis

This analysis starts with an investigation of the sensitivity of x^*_∞ with respect to the process parameters α and σ. Subsequently, the sensitivities of both x^*_∞ and the value function $v^{(\infty)}(x)$ with respect to the main model parameters of lifetime T and lead time ν are analyzed. Default parameter values can be found in Table 5.1.

Low values of the drift rate α corresponds to a higher 'effective' discount rate $r-\alpha$, sometimes called the convenience yield. This decreases the present value of future invest-

[2] It was previously noted that the integrand in the expression of $\psi(x)$ is the price of a European call option. The derivative of such an option with respect to current price x is a well known entity in stochastic calculus called the 'Delta' and is equal to $\Phi(d_+(t))e^{-(r-\alpha)t}$.

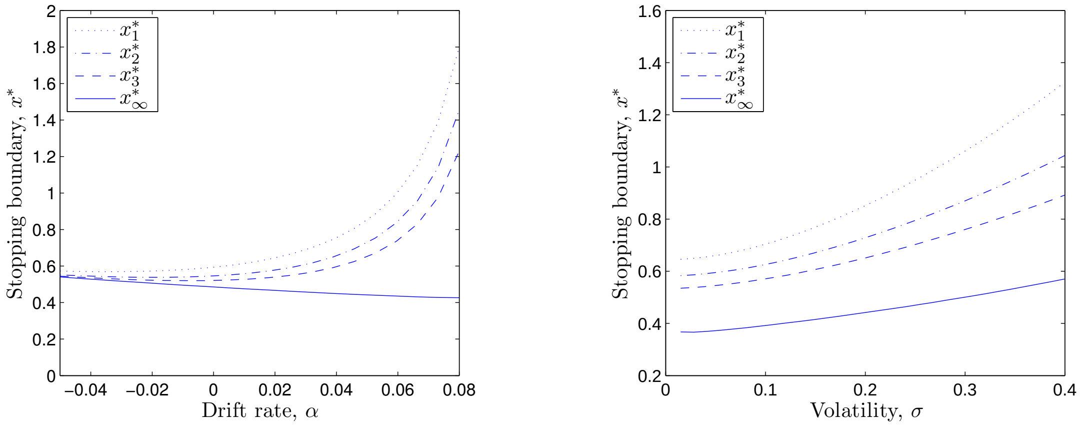

Figure 5.1: *Sensitivity of the optimal stopping boundaries x_k^* w.r.t the drift rate α (Left), and w.r.t the volatility σ (Right).*

ment, which explains the minor difference between x_1^* and x_∞^* for small α, see Figure 5.1 (left). Conversely, a higher drift rate of the underlying process enhances the value of future investments. This drives a greater wedge between the stopping boundary x_1^* of a single investment compared to the stopping boundary x_∞^* of multiple consecutive investments for large α. Consequently, this emphasizes the importance of including future investments in current decisions in environments with a high drift rate. Note that the same result is to be expected if decreasing the discount rate r, still with the condition that $r - \alpha > 0$. In Figure 5.1 (left), it can be observed that as α approaches r (10%) the exercise boundaries x_k^*, $k = 1, 2, 3$, increases rapidly. This is intuitive because theoretically the optimal exercise boundary would be infinite for $\alpha \geq r$. The same phenomenon also occurs for x_∞^* when α is very close to r.

Description	Parameter	Value
Lifetime	T	5
lead time	ν	1
Investment cost	I	1
Operational cost	c	0.1
Discount rate	r	10%
Drift rate	α	5%
Volatility	σ	20%

Table 5.1: *Default parameter values used in the calculations.*

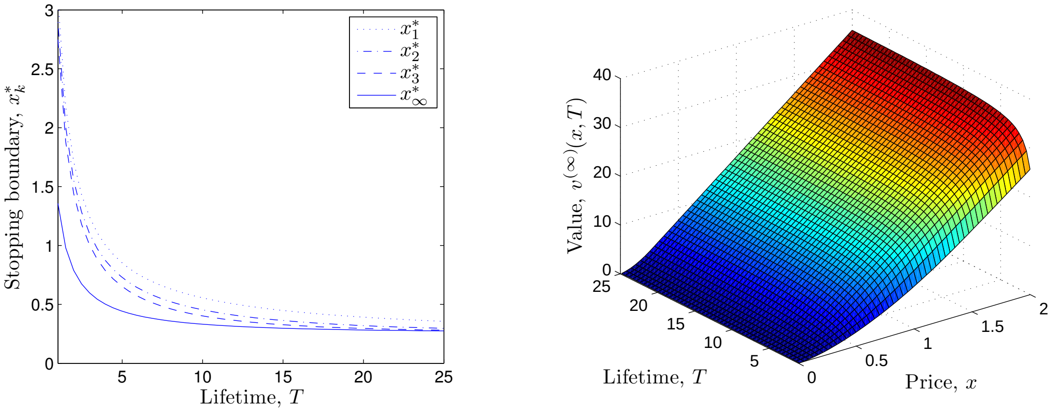

Figure 5.2: *(Left) Optimal stopping boundaries x_k^* decrease as lifetime T increases. (Right) The value $v^{(\infty)}(x,T)$ is increasing with respect to the underlying price levels x and lifetimes T.*

In Figure 5.1 (right), it can be seen that the optimal exercise boundary x_∞^* increases with volatility σ. This suggests that in a more volatile environment the firm will demand a higher output price level in order to enter the market, hence delaying the investment decision. Observe that the increasing pattern holds for finite k, $k = 1, 2, 3$, as well as for the infinite case.

Turning now to the the impact of lifetime T, the parameter of main interest, on the stopping boundaries x_k^* and value $v^{(k)}(x)$. First, in Figure 5.2 (left), a minor difference between x_1^* and x_∞^* is observed for long lifetimes (for $T \geq 20$). Intuitively, the incremental value of an additional investment 20 years or more from the present is minimal due to the discount factor e^{-rT} in (5.17) and (5.18). Therefore, for very long lifetimes, future investments decisions will not significantly influence the current decision to invest. However, for very short lifetimes, e.g. $T < 2$, there is a substantial difference between the optimal exercise levels with one and infinite investment opportunities (i.e. x_1^* vs. x_∞^* for small T). In an intermediate regime, with lifetimes of 5 to 15 years, it can be observed that including only one or two future investment decisions will significantly affect the decision regarding the first investment. In both finite and infinite cases, the optimal exercise boundary decays

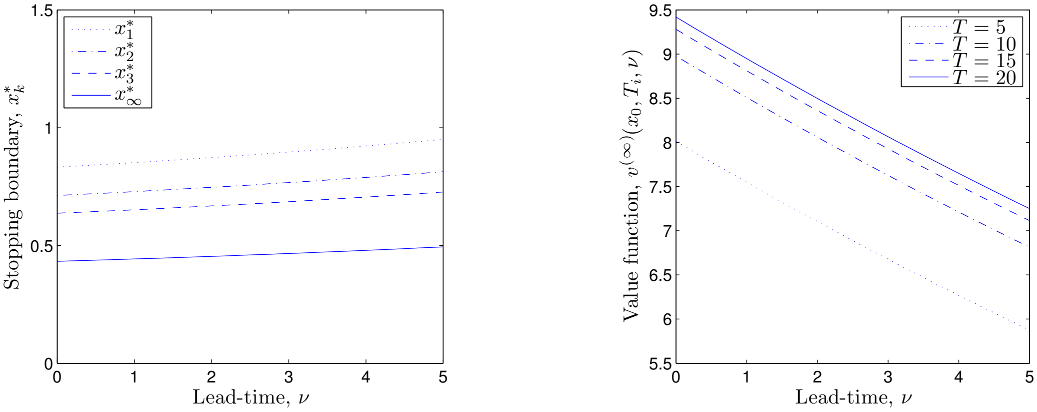

Figure 5.3: *(Left) Optimal stopping boundaries x_k^* for different lead times ν. (Right) The value $v^{(\infty)}(x_0, T_i, \nu)$ evaluated for different lead times ν at $x_0 = 0.5$, and for $T_i = 5, 10, 15, 20$.*

rapidly with respect to lifetime. For instance, x_∞^* is almost flat for lifetimes of 10 years or longer.

Considering the value $v^{(\infty)}(x)$ of the multiple investment scenario for different lifetimes a trend similar to the one regarding the stopping boundary is observed. The marginal impact of adding one year of life to short-lived capital is significant. However, this impact drastically decreases for longer-lived capital. For instance, as seen in Figure 5.2 (right), the value of multiple investments in capital with a 25 year lifespan is virtually the same as capital with a life of 5 years, at the same investment cost I and the same lead time ν. Even though the investment cost is the same, the increased flexibility in timing future investments in the scenario with the shorter lifetime makes them almost equally attractive.

As seen in Figure 5.3 (left), the lead time does not substantially influence x_∞^* nor the difference between x_1^* and x_∞^* at fixed ν. In other words, varying lead times (within the given range) will barely affect the firm's decision to invest. This is expected since seamless consecutive investments are possible regardless of the length of lead time. On the other hand, a longer lead time delays revenue generation relative to the outlay of the investment cost I. Increasing lead time therefore decreases the net present value of every future investment.

This explains the decreasing trend of the value function $v^{(k)}$ with respect to lead time in Figure 5.3 (right). In summary, the firm will realize a lower value with longer lead times of the investments, even though the corresponding exercise boundary changes only marginally. Displaying the value $v^{(\infty)}$ at different lead times ν and at a fixed price level but for several values of the lifetime T again shows the diminishing returns of adding lifetime to capital.

From Section 5.1, the reward function ψ can be interpreted as the sum (integral) of European call options on the uncertain output price X with strike c and maturities ranging over the lifetime. Since increasing the strike price of a European call decreases its value, the cost parameter c has the same effect on the value $v^{(k)}$, which in turn raises the stopping boundary x_k^*. By similar reasoning, the same holds for I.

5.3.2 Critical Investment Cost

Experience from a given industry gives the investment cost (per unit of capacity) I_{large}, as well as the lifetime T_{large} and lead time ν_{large} of large-scale capital in the current paradigm. Transitioning to a paradigm of small-scale, mass-produced, and modular equipment will give rise to a new parameter ensemble, $\{I_{small}, T_{small}, \nu_{small}\}$, where one, a priori, only can infer shorter lifetimes and lead times. By considering an infinite time horizon, the framework above can be used to find the critical investment cost I_{crit} that would render a small-scale approach competitive. That is, for a given reference ensemble $\{I_{large}, T_{large}, \nu_{large}\}$ an I_{crit} can be found, such that $v_{small}^{(\infty)} \geq v_{large}^{(\infty)}$, provided that $I_{small} \leq I_{crit}$, for every choice of T_{small} and ν_{small}.

The observed historical trend of increasing unit sizes seemingly suggests that *total* cost decreases with size. Nevertheless, as shown in Chapter 4 this trend need not apply for operational costs, especially not if sufficient levels of automation can be employed. This justifies the assumption of letting the *operational* cost c be independent of the parameters

I, T and ν. As an illustration, compare a single-cycle thermal power plant to an internal combustion engine, both performing fundamentally the same task of converting chemical energy into mechanical work. Under reasonable circumstances, they can do so at comparable efficiencies. The car engine is mass-produced on the order of days. The power plant, on the other hand, is typically not ready for operation for several years. Moreover, the power plant is designed to last for decades while the car engine presumably will have a lifetime on the order of years under constant operation. How much would one be willing to pay for an engine that is fully automated, retrofitted to run on the same fuel and equipped with a generator to produce electricity? Such a mini-power plant can reasonably be assumed to incur similar levels of operational costs per kWh produced as its large-scale counterpart. This leaves investment cost, lifetime and lead time as the main distinguishing features from the large-scale power plant.

Analyzing the reward function in (5.23) for large x, it can be seen that $\Phi(d_\pm) \approx 1$ for $x \gg c$, and therefore, $\psi(x)$ is asymptotically affine with

$$\begin{aligned}\psi(x) &\approx -I + \int_{\nu}^{\nu+T} \left(xe^{\alpha t} - c\right) e^{-rt} dt \\ &= \frac{e^{-(r-\alpha)\nu}}{r-\alpha}(1 - e^{-(r-\alpha)T})x - \left(I + \frac{e^{-r\nu}}{r}(1 - e^{-rT})c\right) = ax - b. \qquad (5.26)\end{aligned}$$

For large enough values of x the function $u^{(k)}$ is affine as well and

$$\begin{aligned}u^{(k)}(x) &= \Lambda\psi(x) + e^{-rT}\mathbb{E}\left\{u^{(k-1)}(X_T^{0,x})\right\} \approx \Lambda\psi(x) + e^{-rT}u^{(k-1)}\left(\mathbb{E}\left\{X_T^{0,x}\right\}\right) \\ &= \sum_{i=0}^{k}\left(e^{-(r-\alpha)iT}(\gamma-1)ax - e^{-riT}\gamma b\right),\end{aligned}$$

where $\mathbb{E}\left\{X_T^{0,x}\right\} = xe^{\alpha T}$. Using (5.26) to substitute for a, one obtains for large x:

$$u^{(\infty)}(x) \quad = \quad \frac{e^{-(r-\alpha)\nu}}{r-\alpha}x - \frac{\gamma b}{1-e^{-rT}} \, ,$$

where the slope is independent of the lifetime T and decreasing in the lead time ν. From the definition of $u^{(k)}$ in (5.15), it follows that the same properties can be attributed to the value function $v^{(k)}(x;T,\nu)$ for large $x \gg c$. Consequently, comparing different scenarios, i.e. different T, ν and I, the scenario with the shortest lead time will always have the higher value for large enough values of x.

Next, the value of small-scale capital, here characterized by $T_{small} \leq 5$ years, $\nu_{small} \leq 3$ years, is compared against a benchmark of $T_{large} = 25$ years, $\nu_{large} = 5$ years, representing traditional large-scale capital. Furthermore, as a point of reference, the investment cost for large-scale capital is $I_{large} = 1$, which together with the constant operational cost $c = 0.1$ (assumed the same for every choice of I, T and ν) provides a relative monetary scale.

In Figure 5.4, the value $v_{large}^{(\infty)}(x)$ of the large-scale parameter ensemble is displayed alongside $v_{small,j}^{(\infty)}(x)$ with $T_{small} = 3$, $\nu_{small} = 0.25$ for three different investment costs, $I_{small,j} = 0.5, 1, 1.5$. With their shorter lead time, the values $v_{small,j}^{(\infty)}(x)$ is seen to exceed $v_{large}^{(\infty)}(x)$ for large values of x, verifying the analysis above.

From Theorem 1, it is known that $v^{(\infty)}(x)$ is proportional to x^γ for small enough values of x. This explains the constant appearance of the ratio $v_{small,j}^{(\infty)}(x)/v_{large}^{(\infty)}(x)$ in Figure 5.4 for small x. The value of the scenario with the lowest investment cost $v_{small,1}^{(\infty)}$ clearly exceeds the value of the 'large' scenario for all prices x. Indeed, given T_{large}, ν_{large} and I_{large} one can find a critical value $I_{crit}(T_{small}, \nu_{small})$, such that the value of the 'small' scenario is greater at any price level, as long as $I_{small} < I_{crit}$. This leads to the contour plot in Figure 5.5, which displays the critical values for different lifetimes, T_{small} and lead times, ν_{small}. In Section 5.3.1, it has been observed that the value $v^{(\infty)}$ increases with lifetime (see Figure 5.2) and

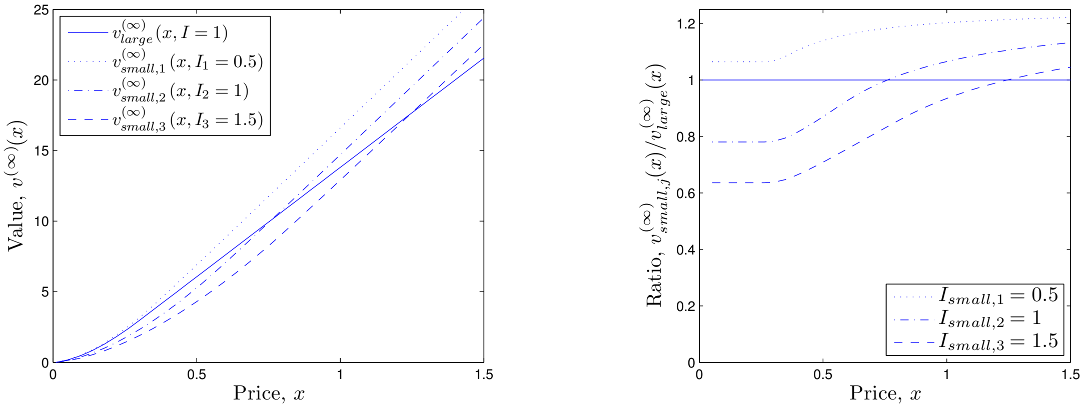

Figure 5.4: *(Left) The value of a large-scale investment scenario ($T_{large} = 25$ and $\nu_{small} = 5$) together with small-scale analogues ($T_{small} = 3$ and $\nu_{small} = 0.25$) at different investment costs. The figure verifies the result that the scenario with the shorter lead time has the higher value for large prices, x. (Right) Displaying the ratio of the previous value functions $v^{(\infty)}_{small,j}(x)/v^{(\infty)}_{large}(x)$ more clearly reveals that $v^{(\infty)}_{small,1}(x) > v^{(\infty)}_{large}(x)$ for every x. This suggests the existence of a critical investment cost I_{crit} such that the value of a small-scale investment scenario exceeds the traditional large-scale counterpart for any price x, as long as $I_{small} \leq I_{crit}$.*

decreases with lead time (see Figure 5.3). This helps explain the trend of I_{crit} with respect to T and ν in Figure 5.5. Precisely, in this domain of short lifetimes, a reduction in lead time yields a higher critical investment cost, which in turn increases the competitiveness of the small-scale approach.

As an example, with $T_{small} = 2.5$ years and $\nu_{small} = 0.3$ years, one can observe from Figure 5.5 that $I_{crit} = 0.5$. That is, despite a difference of a factor 10 in lifetime ($T_{large} = 25$ years), the investment cost is required to differ only by a factor of 2 between the short and long lifetime scenarios in order for the short-lived one to be preferable.

The closest resemblance of an estimation of I_{crit} using traditional NPV arguments without any optionality would be to compare the discounted investment costs over an infinite horizon. That is, assuming that capital is replaced every T_{large} (resp. T_{small}) years under the long (resp. short) lifetime scenarios. For simplicity, disregarding lead times, the discounted costs

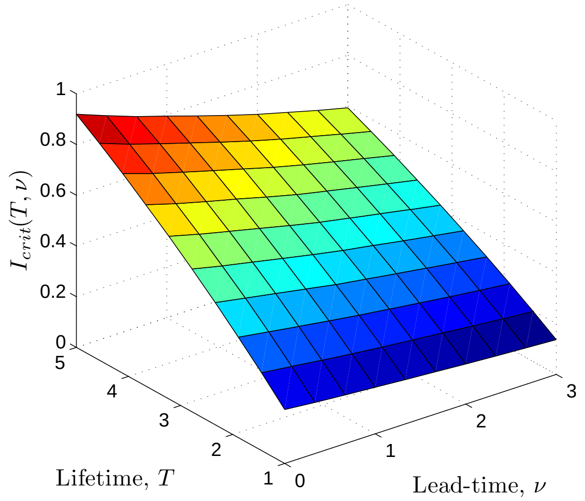

Figure 5.5: *Critical investment cost $I_{crit}(T,\nu)$ of a single small-scale investment, with the given T and ν, in order for the value of multiple consecutive such investments to exceed the value of multiple consecutive large-scale investments at any price level. Since the value function, for any choice of T and ν, is decreasing in the investment cost, we have $v^{(\infty)}(I,T,\nu) \geq v^{(\infty)}(I_{large},T_{large},\nu_{large})$, provided that $I < I_{crit}(T,\nu)$. The large-scale investment was characterized by the parameter values: $I_{large} = 1$, $T_{large} = 25$ years and $\nu_{large} = 5$ years.*

over an infinite horizon can be equated, yielding

$$\sum_{k=0}^{\infty} e^{-rT_{large}} I_{large} = \sum_{k=0}^{\infty} e^{-rT_{small}} I'_{crit} \quad \Leftrightarrow \quad \frac{I'_{crit}}{I_{large}} = \frac{1-e^{-rT_{small}}}{1-e^{-rT_{large}}},$$

where I'_{crit} denotes the critical investment cost of the short-lived scenario under the NPV argument. Using the same example as above, with $T_{small} = 2.5$ years, $T_{large} = 25$ years, and $I_{large} = 1$, we find that $I'_{crit} = 0.24$. Comparing to the critical investment cost $I_{crit} = 0.50$ above, the value of the optionality embedded in the framework permits twice the investment cost that would be suggested by standard net present cost arguments.

Lastly, note that, in the example with the car engine and the power plant, the cost per kW of capacity of the car engine is almost two orders of magnitude less than that of the power plant (Larminie and Dick, 2003). Clearly, the critical costs suggested in Figure 5.5

are not nearly as dramatic. This suggests great potential in abandoning the customized, large-scale investments in favor of mass-produced and modular capital.

5.4 Summary

In order to capture the distinct features of lifetime and lead time of an investment, a framework was introduced that valued consecutive investments over a possibly infinite time horizon. Naturally, including future investments significantly affects the exercise boundary, especially when the individual investment is short-lived. Moreover, the marginal benefit of increasing lifetime drastically diminishes when considering multiple investment opportunities. With such a valuation framework in place, critical investment costs of capital with lifetimes and lead times that deviate from industry standards can be estimated. For instance, such an estimation reveals that reducing lifetime of capital from a typical 25 years to 2.5 years need only be accompanied by a decrease of a factor 2 in investment cost in order to be superior, in overall value terms.

The analysis has assumed a log-normal stochastic cash flow. More realistic representations of commodity price dynamics would include other processes, e.g. with mean reversion and jumps. While the model formulation can handle any such processes, the analysis would likely have to resort to simulation rather than iterative solution procedures. However, some qualitative conclusions will likely still hold. For instance, under mean reversion, the value of an investment with considerable lead time and lifetime is expected to generate a cash flow close to the finite long-run mean. Shorter lived, and more quickly deployed capital would, on the other hand, be more suited to both exploit positive deviations from the mean, and also avoid periods of low prices. In a framework where the process returns to a mean in a timespan comparable to the lifetime of small-scale capital, the critical investment cost to achieve parity with large-scale scenarios would therefore be expected to be even higher than those found in Figure 5.5.

There are a number of ways this framework can be extended. First, one can incorporate the firm's aversion to risk and ambiguity associated with the stochastic investment returns, as discussed in (Henderson, 2007) and (Jaimungal, 2011) in the case of single investment. Moreover, additional information on the technological change, industry outlook, and future investment cost will enhance the decision analysis. For instance, intuition suggests that frequent technological advancement is more easily harnessed with a higher turnover rate since outdated technology can be abandoned without sacrificing investments with a long remaining horizon. Similarly, regulatory risk is less of an issue with shorter investment windows.

5.5 A comment on Discount Rates

How to properly place a current value on future events through discounting is a contested issue. The proposition of valuating the effects of climate change by using very low or even negative discount rates, as suggested in the 'Stern Report', met firm opposition among economists (Stern et al., 2007; Nordhaus, 2007). Even within the financial community, there are different opinions on how to properly account for the opportunity cost of capital in the discount rate in various situations. For instance, debt to equity ratios may change during the lifetime of a project (Esty, 1999) or the risk profile changes after the completion of construction and commencement of operation (Garvin and Cheah, 2004; Brealey et al., 2008).

Regardless of underlying factors, the discount rate used when evaluating the economic viability of a project ex ante is intrinsically linked to the expected economic life of the investment. As illustrated in Figure 5.6, assuming a constant cash flow (in real terms) over a 40 year investment, a discount rate of 15% means that half the present value of the project is

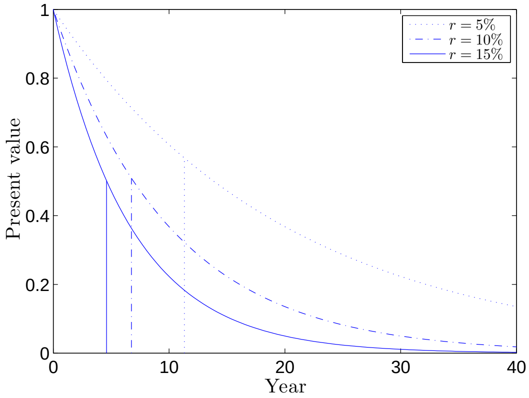

Figure 5.6: *This simple graph of the discount factor with a continuously compounded rate, e^{-rt}, illustrates the impact of the rate r used in valuations of long-lived investments. Discounting at 15% p.a. requires a 40-year investment to generate half its value in the first 4 years.*

generated during the first 4 years. Using a lower discount rate in the valuation of long-lived projects puts a higher value on revenues generated further out in the future. Consequently, a perceived low risk of future revenue shortfall is manifested by using a low discount rate. In the context of energy investments, which almost exclusively rely on long pay-back periods (or equivalently, uses a low discount rate), an uncertain regulatory environment is suggested to impede investments. These uncertainties could take the form of an anticipated price on carbon (Fan et al., 2010), or uncertain continuance of policy support for renewables (Lüthi and Prässler, 2011), or of course, a combination of the two. Regardless, the resolution of such uncertainties would not result in rampant investments if these cannot be assumed viable over a long period.

The use of a low discount rate, e.g. 5-6% in the valuation of nuclear and other thermal power generation investments (OECD/IEA, 2010; Deutch et al., 2003) reveals the perception of another risk. Capital intensive projects that rely on a long economic life for financial viability would not be initiated if the emergence of competing technologies were likely. Albeit without further substantiation, this observation does suggest that industries accustomed to large-unit scale, and consequently long lifetimes, are less amenable or prone to innovation.

Chapter 6

Small-Scale Reverse Osmosis Desalination

Providing large, stable and affordable supplies of fresh water from the oceans has been suggested as one of the main engineering challenges for the coming century (National Academy of Engineering, 2012). Among the various desalination technologies available for this purpose, RO has seen the largest growth in recent years and represents almost half of world desalination capacity today (Greenlee et al., 2009). Improvements in membrane materials and process designs over the past decades have substantially lowered the energy consumption in seawater reverse osmosis (SWRO) desalination. Depending on salinity and overall feed water quality, current state-of-the art SWRO facilities require between 3-4 kWh per cubic meter of produced water, down from around 20 kWh per cubic meter 30 years ago (Elimelech and Phillip, 2011; Fritzmann et al., 2007). However, energy still accounts for the largest fraction (typically almost half) of the total cost of desalinated seawater through RO today (Wittholz et al., 2008). With the thermodynamic limit around one fifth of the energy requirement in current processes, there is room for further improvement.

The expected increase in demand for desalinated water together with the nature of reverse osmosis, being a membrane-based process, makes this technology an appropriate candidate to study more closely. The focus of the investigation here is on the actual reverse osmosis (RO) stage, the main part of this desalination technology. More specifically, investigating the feed flow reveals two main hydrodynamic effects that impacts energy consumption in the separation stage: pressure drop of the feed and concentration polarization above the membrane surface.

The RO stage typically comprises three components: energy delivery (pumps), active membrane surface area and some form of an energy recovery device. The membrane area is manufactured in modules, which are then assembled in series in a single pressure vessel. Deviating from the industry norm of large-scale implementations would entail smaller pressure vessels, and consequently shorter feed channels. It will be shown below that scaling down accordingly mitigates the two hydrodynamic effects mentioned above and results in slight energy savings while keeping the same level of productivity, i.e. flux rate through the membranes.

Moving into a paradigm of mass-production of small-scale and standalone RO units will presumably lower the fixed cost of desalination hardware. With more expensive energy supplies (e.g. renewable power) this will emphasize the role of specific energy consumption as an overall cost driver. Therefore, further reductions of specific energy consumption by decreasing permeate flux through the membrane is also demonstrated.

The transport mechanisms in this study are modeled in a thin rectangular feed channel with two permeable walls under both laminar and turbulent conditions. These models allow for a calculation of the specific energy consumption in RO. Moreover, they also permit investigation of different geometries and flow conditions, including those observed in commercial operation. Compared to models in the literature, the laminar model developed here is applicable to a wider range of conditions, which allows for study of different feed

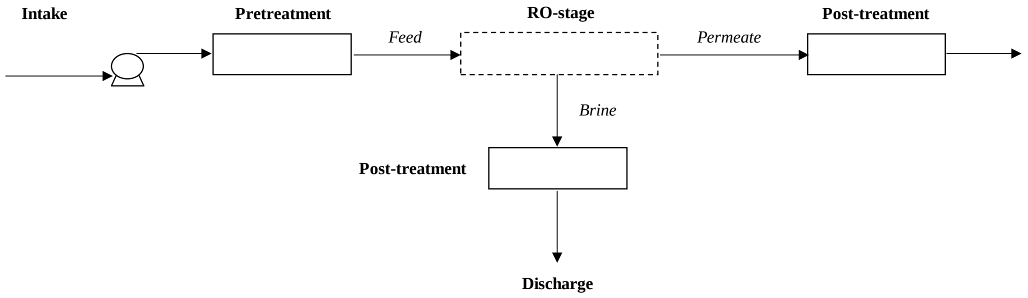

Figure 6.1: *Schematic overview of a typical SWRO plant (Fritzmann et al., 2007; Cerci, 2002)*

flow rates, permeation flux rates and different channel geometries. In order to reduce the dimensionality of the optimization problem, this study limits the channel length to 1 and 8 meters. The former represents the length of a single membrane module as they are currently manufactured and the latter number corresponds to the length of the feed channel typically observed in commercial RO implementations.

6.1 SWRO – An overview

The different processes in a typical SWRO plant can be categorized in six stages – intake, pretreatment, RO-separation, brine post-treatment, brine discharge and permeate post-treatment, see Figure 6.1. While the RO-stage is the main focus here, the other steps are briefly reviewed.

Water can be extracted from the ocean in one of two ways: through open seawater intakes or through wells drilled on the coast or beneath the ocean floor. While open intakes usually represent a lower investment cost, these expose the plant to a higher variability of intake water quality, caused by algae bloom or stormy weather, for example. This requires adaptability of the subsequent pretreatment stage in order to secure continuous operation, which in turn translates into tighter operational control and higher capital costs. Additionally, for larger desalination plants, the high feed water flow rate disturbs the local aquatic biota, and impingement of marine organisms on the intake screens, as well as entrainment further into the system, become serious concerns. Both these issues can be mitigated by situating the intake deeper, at the expense of higher capital costs (Gille, 2003). Subsurface intakes have the advantage of the intake water being naturally filtered through the permeable stratum, decreasing the need for a flexible pretreatment system. The limits imposed by local soil permeability of conventional vertical beach wells have typically disqualified these for large-scale operation. However, new techniques in horizontal drilling may prove to provide high capacity intake systems of constant water quality without disturbing the marine life and at reasonable costs (Peters et al., 2007). Ultimately, site-specific characteristics, such as meteorological, oceanographic and geological features and marine life determine the optimal intake system.

The purpose of pretreatment is to supply the RO-stage with feed water of even quality for continuous operation. The main mechanisms causing flux decline in SWRO, which need to be avoided or mitigated through pretreatment, can be summarized as fouling and scaling. Fouling includes the accumulation of particulates as well as growth of biological matter in the feed channel. Scaling, on the other hand, is the precipitation of super-saturated inorganic compounds on the membrane surface.

Protecting pipes and pumps in the intake system, coarse screens in the extraction stage serve as the first step in the physical pretreatment of the feed water. To further separate suspended solids and colloids from the feed, flocculation agents, e.g. iron or aluminum salts, are added to agglomerate suspended matter. These compounds are then removed through e.g. sedimentation or multimedia filtration followed by micro filtration to reach acceptable levels of clarity (Fritzmann et al., 2007; Bonnelye et al., 2004). Protecting the membranes from biofouling, oxidants such as chlorine or ozone are sometimes added to the feed water (Lee et al., 2009). These oxidative compounds need to be carefully removed before the

RO-stage to prevent membrane degradation.

With an overall increase in feed salinity along the RO-membrane channel as fresh water gets diverted, precipitation of inorganic compounds becomes more likely. These crystals are sometimes hard to remove and besides severely inhibiting flux they may also permanently damage the membranes. The more notable foulants in this case are calcium carbonate $CaCO_3$, calcium sulfate $CaSO_4$ and silica SiO_2 (Oh et al., 2009). Precipitation of these and other salts can be limited or prevented through a combination of adding chemical agents as antiscalants and controlling the pH.

Post-treatment of the brine is focused on the removal of chemicals, introduced either as antiscalants or for plant cleaning, before discharge. Also, depending on operational conditions and the sensitivity of the discharge site, pH adjustment in the form of raised alkalinity may be needed. With desalination plants typically operating at 40-50% recovery (Greenlee et al., 2009; Reddy and Ghaffour, 2007), one of the main environmental issues with SWRO is however the discharge of the high saline concentrate. The salinity of the brine dramatically differs from that of the natural water causing severe osmotic stresses to the marine life (Fritzmann et al., 2007; Dupavillon and Gillanders, 2009).

One of the most important operating parameters in SWRO desalination is the recovery rate, i.e. the fraction of fresh water produced from a given amount of intake water. From the outside, this parameter determines, for a given production capacity, the overall water circulation through the plant as well as the salinity of the effluent brine. Seen from the inside, altering the recovery rate potentially affects every process step, most importantly the RO-separation stage. According to many observers, achieving continued reductions in energy consumption and overall cost in SWRO hinges on the ability to operate at even higher recovery rates than today's, i.e. at 50% and above (Busch and Mickols, 2004; Kim et al., 2007, 2009; Semiat, 2000, 2008; Stover, 2007; Wilf and Klinko, 2001; Liang et al., 2009).

While increasing the recovery means processing a smaller volume of feed water, this feed water has to be pretreated more carefully to avoid scaling. Moreover, a higher recovery also requires increased operating pressures to overcome the raised osmotic pressures caused by the higher average salinities in the brine stream. Consequently, this route entails higher capital costs for stricter pretreatment stages of the feed water, tighter operational control and more sturdy equipment capable of withstanding the higher pressures and more corrosive environment.

Next to the recovery rate, the permeation rate, or the flux of produced water, also significantly impacts the energy required to desalinate seawater. Akin to Ohm's law, the permeation rate through a membrane is proportional to the pressure drop (hydrostatic minus osmotic) across the membrane. The specific energy consumption is therefore, to a first approximation, linear in the flux rate. In environments with low-cost power and high capital costs, the incentives are lined up to operate at the highest possible flux. However, if this cost structure is reversed, then the incentives would pull in the opposite direction. Two settings in which such a reversal could be a reality for seawater desalination are envisaged. First, renewable power today typically commands a premium over its fossil-derived counterpart. Desalinating water using electricity from e.g. solar and wind energy would therefore fall into this category. Second, along the theme of this thesis, transitioning to a paradigm of mass production of small-scale and modular units could offer additional benefits. Small-scale desalination systems exist today in niche markets, see e.g. (Spectra Watermakers, 2012), proving the fundamental validity of this approach. Mass-producing similar units could bring down investment cost to levels at, or below the cost of today's large-scale and centralized plants.

6.2 Transport and Energy Consumption in the RO-stage

The amount of energy required to produce one unit of potable water from a saline solution via reverse osmosis depends on several factors. Perhaps the most prominent locational factor is the salinity of the feed. Even when limiting the discussion to seawater RO, the salinities of different bodies of water typically range from 30,000 to 45,000 mg/L (Greenlee et al., 2009), which complicates comparisons of energy consumption of RO plants around the world. Other locational factors that influence energy consumption are general feed water quality and intake structures. For instance, open surface intakes require more pretreatment of the feed in the plant compared to beach wells or other subsurface intakes where the permeable stratum provides a natural filtration. In addition to these locational factors, two operational parameters substantially affect the specific energy consumption of any membrane-based desalination process: the recovery rate and permeate flux rate. These two parameters are controlled through the hydrostatic pressure and the volumetric flow rate of the feed entering the main separation stage. In order to find the specific energy consumption in the RO-stage, the local pressure of the feed and the permeation rates are modeled along the feed channel in the separation stage.

Before introducing the transport models the minimum necessary work required in desalination is briefly reviewed from a thermodynamic perspective. Given a volume of saline solution with known concentration c, the osmotic pressure π can be formulated as a function of the fraction χ of water that has been withdrawn. Considering the solution to be ideal, the osmotic pressure π is approximately linear in the concentration: $\pi = f_{os}c$, where f_{os} depends on the chemical properties of the solute. Then

$$\pi(\chi) = f_{os}\rho\frac{m_s}{m_s + (1-\chi)m_w} = f_{os}\rho\frac{r}{r + (1-\chi)}, \tag{6.1}$$

where the density ρ is assumed constant. The ratio, $r = m_s/m_w$, is the mass ratio of dissolved salts and water respectively per unit volume of initial solution. The ideal work required to separate 1 cubic meter of fresh water from $1/\alpha$ cubic meter of salty solution, i.e. the ideal specific work at a recovery rate α, is

$$W_{ideal} = \frac{1}{\alpha}\int_0^\alpha \pi(\chi)d\chi = \frac{f_{os}\rho}{\alpha} r \ln\left(\frac{r+1}{r+(1-\alpha)}\right). \tag{6.2}$$

In the limit $\alpha \to 0$, the work W_{ideal} is the same as the osmotic pressure of the initial solution. This limit is sometimes stated as the minimum theoretical amount of work needed to produce one cubic meter of freshwater from seawater. Novel approaches to desalination call for submarine operations at working depths such that the hydrostatic pressure overcomes the osmotic pressure (Charcosset et al., 2009; Pacenti et al., 1999). The only work required is that of pumping the permeate to the surface, which roughly corresponds the limit $\alpha \to 0$ in (6.2). For regular land-based desalination, this limit is not practically attainable due to a finite volumetric feed flow rate to the plant. Hence, the minimal work required in real RO operation is higher, depending, among other things, on the recovery rate α.

According to solution-diffusion theory, the permeation velocity, or flux, v_p is proportional to the difference between the fluid pressure difference and the osmotic pressure difference across the membrane (Baker, 2004; Mulder, 1991):

$$v_p = A\left[\Delta p - \Delta\pi\right] = A\left[(p - p_0) - f_{os}\left(c_m - c_p\right)\right]. \tag{6.3}$$

The coefficient A is the membrane water permeability. Crucially, the osmotic pressure difference across the membrane depends on the feed side concentration c_m immediately above the membrane surface. The concentration polarization that occurs when selectively extracting water from the feed will elevate the concentration close to the membrane compared to the rest of the feed. The permeate concentration c_p to c_m are related through a rejection coefficient R, typically used to characterize membrane performance:

$$c_m - c_p = Rc_m. \tag{6.4}$$

Additionally, the pressure on the permeate side, p_0 in (6.3), is assumed constant and equal to ambient pressure.

The most common assembly of membrane surface area in RO-desalination is the spiral-wound module. In such a module, the feed channel is a thin leaf, multiples of which are wound around a permeate collector tube. Because the thickness of each leaf (channel height) is much smaller than the radius of the assembly, each leaf can be approximated with a rectangular channel, (Schwinge et al., 2004; Geraldes et al., 2002). Below, the pressure and concentration profile for both a laminar flow and for a disturbed flow are modeled in such a channel, see Figures 6.2 and 6.3.

6.2.1 Laminar Flow

If either the feed channel is unobstructed or if the feed flow rates are sufficiently low, the flow can be considered laminar. Moreover, since the width of the channel is much greater than the channel height, the problem can be limited to two dimensions and a Poiseuille flow can be stipulated in the feed channel. With such a flow field, the continuity equation at steady state for the salt in the feed channel can be solved in conjunction with the expression for the local permeation rate in (6.3). Similar laminar models, which make some simplifying assumptions that circumvent solving the partial differential equation that is the continuity

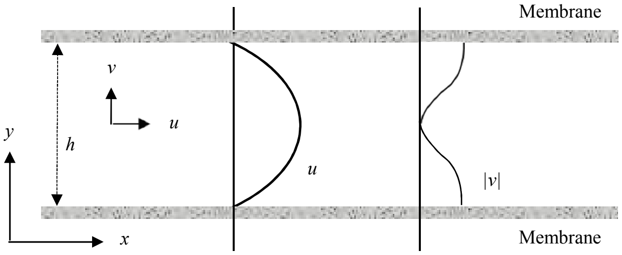

Figure 6.2: *An unobstructed feed channel with the same height as those found in the spiral-wound module would allow for a laminar feed flow. The longitudinal feed velocity component, u, is approximated with a parabolic Poiseuille profile onto which a vertical velocity component, v, obeying the continuity equation, is superimposed.*

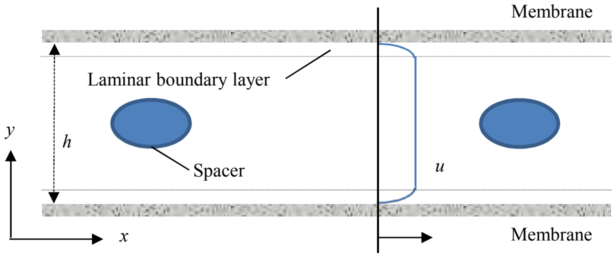

Figure 6.3: *The feed flow in a typical spiral-wound membrane module is disturbed by spacers to enhance mixing. At sufficient flow rates the flow can be considered turbulent. Concentration is assumed uniform across the height of the channel except in the laminar boundary layers. The superficial feed velocity is denoted u.*

equation, can be found in the literature, see e.g. (Elimelech and Phillip, 2011; Song and Elimelech, 1995; Song, 2010; Bouchard et al., 1994; Brian, 1965).

Consider a flow field $\vec{w}$ in the feed channel, where

$$\vec{w}(x,y) = \left(u(x,y), v(x,y)\right), \tag{6.5}$$

which is approximated by a parabolic longitudinal component $u(x,y)$ and a superimposed transverse velocity component $v(x,y)$ satisfying the continuity equation. The transgression made by simply superimposing a transverse velocity and thus violating the laws of momentum transport in the fluid is negligible since $|v|$ is typically four to five orders of magnitude smaller than the longitudinal velocity $|u|$ in reverse osmosis applications (Song, 2010; Bouchard et al., 1994). A parabolic profile in a channel of height h with vanishing velocities at the channel walls (both semi-permeable) yields

$$u(x,y) = 6\overline{u}(x)\frac{y}{h}\left(1 - \frac{y}{h}\right), \tag{6.6}$$

where $\overline{u}(x)$ is the average longitudinal velocity at x. Integrating the continuity equation at steady state, $\nabla \cdot \vec{w} = 0$, together with the symmetry condition $v(x, h/2) = 0$ results in

$$v(x,y) = \frac{d\overline{u}}{dx}\frac{h}{2}\left[1 - 6\left(\frac{y}{h}\right)^2 + 4\left(\frac{y}{h}\right)^3\right], \tag{6.7}$$

where the boundary condition is expressed with the use of the permeation rate in (6.3):

$$v(x,0) = \frac{h}{2}\frac{d\overline{u}}{dx} = v_p(x). \tag{6.8}$$

For a laminar flow between two planes, the pressure loss along the channel is given by

$$\frac{dp}{dx} = -\frac{12\eta\overline{u}}{h^2}, \tag{6.9}$$

where η is the dynamic viscosity of the solution. The assumption of a laminar flow in a clear channel holds for Reynolds numbers $Re = \rho\overline{u}h/\eta < 2100$ (Kim and Hoek, 2005).

In order to find the local permeation rates $v_p(x)$ from (6.3), which varies along the channel, the concentration $c_m(x)$ at the membrane wall is required. This is achieved by solving the general continuity equation at steady state for the salt in the channel:

$$\begin{aligned} \partial_t c &= -\nabla \cdot (c\vec{w}) - \nabla \cdot (-D\nabla c) \\ 0 &= -u\partial_x c + D\partial_{xx} c - v\partial_y c + D\partial_{yy} c. \end{aligned} \tag{6.10}$$

Both the density ρ of the total fluid and the diffusion constant D for the salt in the fluid are assumed to be spatially uniform in (6.10). Details regarding the implementation of a discretization of (6.10) and (6.9), as well as the longitudinally varying (6.3) can be found in Appendix B.

6.2.2 Turbulent Flow

The hydrodynamics of the feed flow in a commercial RO-separation stage is an area under significant current investigation, see (Guillen and Hoek, 2009; Kim and Hoek, 2005; Lyster and Cohen, 2007; Song et al., 2002; Lu et al., 2007; Kaghazchi et al., 2010). Short of CFD analyses of the flow with detailed information on the geometry of the feed channel, most turbulent models rely on thin-film theory to link concentration profiles and permeation rates. This theory suggests that outside a laminar boundary layer of thickness δ, the fluid is well mixed with a concentration c_b (bulk concentration). Furthermore, this theory presumes that inside this boundary layer, there is negligible longitudinal solute transport and the vertical velocity component is constant, $v = v_p$. Under these assumptions the continuity equation (6.10) in the boundary layer can be expressed as

$$0 = \partial_y \left(v_p c - D \partial_y c \right), \quad 0 \leq y \leq \delta. \tag{6.11}$$

Taking v_p to be a positive quantity (directed toward the membrane at $y = 0$) and integrating over the height δ of the boundary layer yields

$$v_p c + D \partial_y c = v_p (1 - R) c_m, \tag{6.12}$$

where $c_m = c(0)$ is the concentration next to the membrane and where $(1 - R)c_m$ is the permeate concentration. Solving this differential equation, with the boundary condition $c(\delta) = c_b$, which can be extended to the center of the channel, yields

$$c(y) = \begin{cases} \left(c_b - (1 - R) c_m \right) e^{(\delta - y) v_p / D} + (1 - R) c_m, & 0 \leq y \leq \delta, \\ c_b, \quad \delta \leq y \leq h/2. \end{cases} \tag{6.13}$$

Particularly, for $y = 0$, the commonly encountered expression for the concentration polarization in obstructed channels is retrieved,

$$\frac{Rc_m}{c_b - (1-R)c_m} = e^{v_p\delta/D} = e^{v_p/k}. \tag{6.14}$$

Note the introduction of the mass-transfer coefficient $k = D/\delta$.

Applying (6.14) requires an approximation of the mass-transfer coefficient, or equivalently the boundary layer thickness. This is typically achieved with a Sherwood-number correlation:

$$k = Sh\frac{D}{d_h} = g(Re_t, Sc)\frac{D}{d_h}, \tag{6.15}$$

where d_h denotes the hydraulic diameter ($d_h \sim h$), and where $Re_t = \rho d_h u/\eta$ and $Sc = \eta/(\rho D)$ are the dimensionless Reynolds and Schmidt numbers respectively. Affixing the subscript to the notation of the turbulent Reynolds number highlights the dependence on the hydraulic diameter rather than channel height. The hydraulic diameter is a parameter that depends on the channel geometry, including the shape of the spacers. In addition to the approximation of the Sherwood number, a pressure drop,

$$\frac{dp}{dx} = -\frac{f}{d_h}\frac{\rho u^2}{2}, \tag{6.16}$$

is also inferred by an approximation of the friction factor f as a function of the Reynolds number. The multitude of different channel geometries available in RO applications, together with various flow regimes possible, has spawned a plethora of correlations g in (6.15) and of the friction factor f in (6.16), see e.g. (Guillen and Hoek, 2009; Lyster and Cohen, 2007; Gekas and Hallstrom, 1987; Schock and Miquel, 1987). The model used here is adopted from (Guillen and Hoek, 2009), which studies similar applications (SWRO at typical permeation rates in spiral-wound modules). In that work, the Sherwood number correlation and the

Description	Value
d_h (mm)	0.81
κ	0.46
λ	0.34
β	0.69
γ	244.17
ε	1.00

Table 6.1: *Parameter values used in the Sherwood number correlation and pressure drop approximation of the turbulent flow.*

friction factor approximation are given by:

$$\begin{aligned} Sh &= \kappa \left(Re_t Sc\right)^{\lambda}, \\ f &= \beta + \frac{\gamma}{Re_t^{\varepsilon}}, \end{aligned} \tag{6.17}$$

where the numerical values can be found in Table 6.1. The numerical solution scheme can be found in Appendix B.

The numerical values for the parameters used are only valid for a fixed channel geometry and a for limited range of Reynolds numbers. Particularly, for low Reynolds numbers the notion of turbulent flow is questionable. Use of this turbulent model is therefore limited to the conditions specified in (Guillen and Hoek, 2009), which corresponds well to the conditions in commercial SWRO operation.

6.3 Results and Discussion

The two main factors that impede flux, beyond thermodynamic limitations, are concentration polarization and longitudinal pressure drop in the feed channel. These factors both affect the permeation rate directly since this rate is proportional to the difference $\Delta p - \Delta\pi$, see (6.3). A more pronounced polarization raises the concentration next to the membrane, thereby also the osmotic pressure gradient $\Delta\pi$ across the membrane. On the other hand, pressure losses in the feed channel diminish the driving force that is the hydrostatic pressure p causing flux decline along the longitudinal direction of the channel. With limited mass transport in the transversal direction, the laminar flow is expected to yield greater levels of polarization. However, since this transport is correlated with transport of x-momentum in the y-direction, lower pressure drops for laminar flows are also expected. For a turbulent flow, the situation is reversed. With turbulent mixing across the major part of the channel height, concentration polarization is expected to decrease at the expense of a higher pressure drop.

These arguments are verified in Figure 6.4, where flow conditions of commercial SWRO operations are simulated, $L = 8$m, $\Delta p_0 = 65$ bar, $u_0 = 0.25$m/s ($Re = 113$.) As can be seen in Figure 6.5, the greater polarization of the laminar flow causes lower flux early in the channel compared to the turbulent flow. However, permeation rates decrease more quickly along the channel due to a greater pressure drop. Comparing the two flows, the turbulent channel yields slightly higher average permeation rates under these conditions ($\overline{v}_p = 3.7\mu$m/s versus $\overline{v}_p = 3.3\mu$m/s.)

Figures 6.4 and 6.5 detail the polarization, permeation velocity and pressure drop in a long channel. Given the feed flow rate and the operating pressure the flow early in the channel is unaffected by conditions downstream. Thus, these figures also reveal conditions in channels shorter than $L = 8$m. Importantly, the laminar model shows the intrinsic benefits of using a shorter channel. As seen in Figure 6.5 the permeation velocity, in contrast to the pressure, drops rapidly early in the channel. Using a massively parallel configuration of shorter channels rather than one long channel would therefore yield more permeate at the same pressure, which consequently results in lower energy consumption per unit fresh water produced.

The transversal transport in the feed channel can be described by the average Peclet number, $Pe = \overline{v}_p h/D$, which expresses the ratio between the rates of convective and diffusive transport. With a Peclet number less than unity diffusion is expected to dominate convection

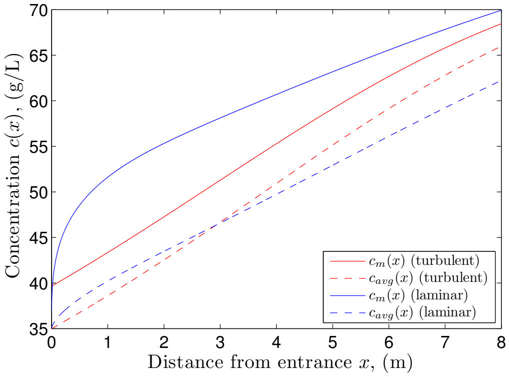

Figure 6.4: *The wall concentration $c_m(x)$ compared to the average concentration $c_{avg}(x)$ a distance x from the inlet. The turbulent flow exhibits lower polarization in both real terms (c_m) and relative terms c_m/c_{avg}.The feed flow rate and applied pressure were the same in both cases, $u_0 = 0.25m/s$ and $\Delta p_0 = 65$ bar.*

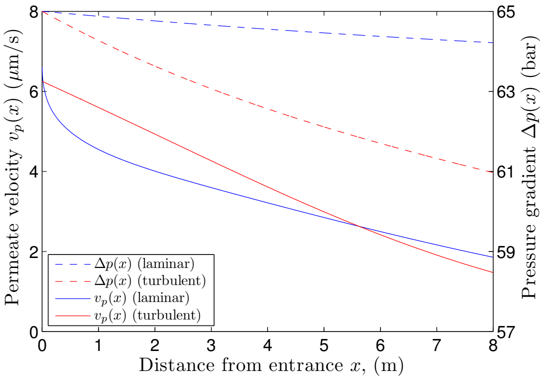

Figure 6.5: *Permeation rate $v_p(x)$ and transmembrane pressure gradient $\Delta p(x)$ for laminar and turbulent flows. A higher polarization for the laminar flow causes lower fluxes early in the channel. However, a higher pressure drop for the turbulent flow depresses flux rates more quickly than for the laminar flow.*

and thereby prohibit the build-up of any significant polarization in the channel. Conversely, if $Pe > 1$ then the convective transport is presumed to dominate diffusion and therefore yield a significant transversal gradient in concentration. This effect is verified in Figure 6.6, where the concentration profiles of two flows, with the same average permeation rate and the same Reynolds number, are given in two channels of height $2.5 \cdot 10^{-3}$m and $1 \cdot 10^{-4}$m respectively[1]. With $\overline{v}_p = 4\mu$m/s in both cases, the Peclet numbers are 6 and 0.2 respectively. While a smaller channel height reduces polarization, the flow suffers a greater pressure drop because of the inverse dependence on h^2 in (6.9). In order for this drop not to obscure the reduced polarization, a relatively short length of the channel, $L = 0.1$m, was used for illustration purposes. The two flows result in the same amount of permeate but, because of the substantial difference in concentration levels near the membrane, the difference in required pressure is equally substantial, $\Delta p_0 = 68$ bar and $\Delta p_0 = 51$ bar respectively. Since these flows have the same Reynolds numbers, they also have the same volumetric feed flow

[1] These values where chosen since they can be considered extremes compared to the height of the channel in typical spiral-wound modules ($0.5 - 0.7$mm).

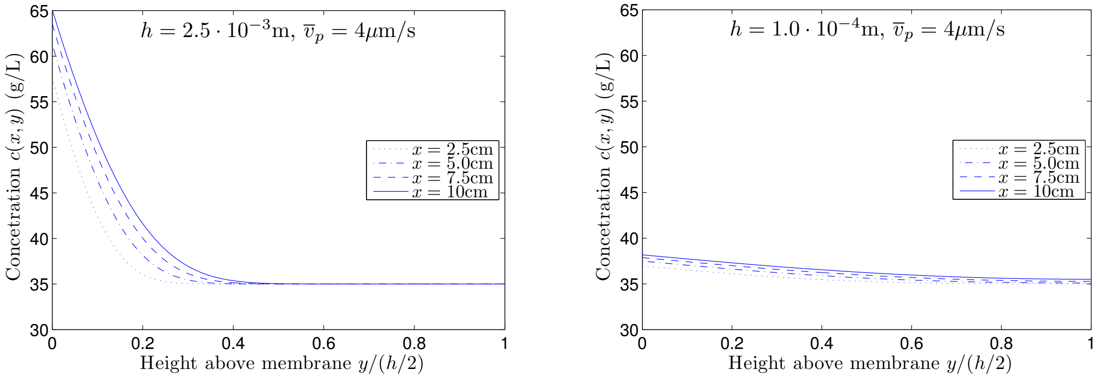

Figure 6.6: *Concetration profiles in channels of height* $2.5 \cdot 10^{-3}m$ *(right) and* $1 \cdot 10^{-4}m$ *(left) respectively. The average permeation rate is the same in both cases, as where the channel length* $L = 0.1m$ *and the Reynolds number* $Re = 25$. *In the channel with the smaller height, diffusion away from the membrane overcomes the convective transport of solute towards the membrane. The reduced polarization causes lower osmotic pressure at the membrane boundary and therefore requires a lower pressure gradient for the same flux.*

rate. The lower operating pressure would therefore directly correspond to a lower specific energy consumption in the thinner channel.

The concentration polarization is however not merely a function of the transversal Peclet number. From (6.10), it is expected that the longitudinal velocity u in the feed channel also influences the polarization. With a higher velocity u, retained particles are swept away in the longitudinal direction to a greater extent. Indeed, two flows with different flow speeds $\overline{u}_0$, expressed through the Reynolds number $Re = \rho h \overline{u}_0 / \eta$, but with the same average permeation rate in the same channel are displayed in Figure 6.7. The difference, evidenced in Figure 6.7, suggests that treating the concentration polarization as a boundary layer phenomenon in laminar flows and neglecting the transversal terms in (6.10), as done in Song (2010); Song and Yu (1999); Song and Elimelech (1995), has doubtful validity.

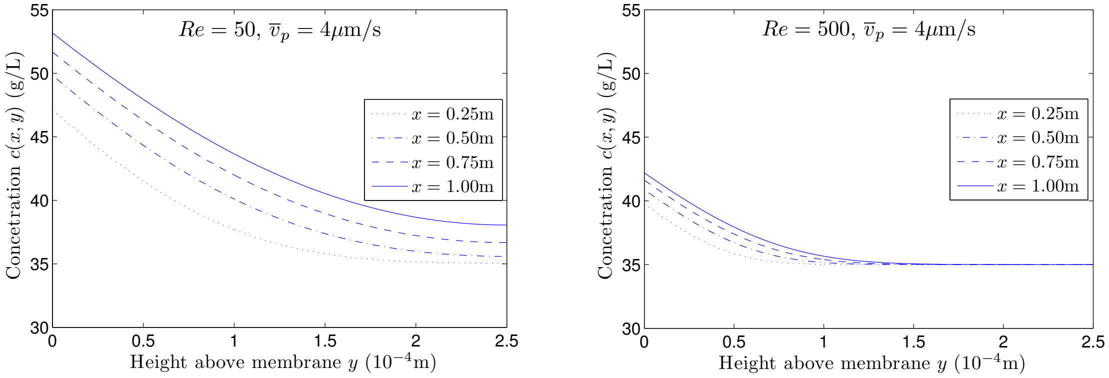

Figure 6.7: *Two flows in the same channel with the same transversal Peclet number $Pe = \overline{v}_p h/D$ but with different Reynolds numbers $Re = \rho\overline{u}_0 h/\eta$. Increasing the transversal velocity u also increases the importance of the transversal transport terms in (6.10), which depresses polarization and also decreases the necessary pressure to produce the flux.*

6.3.1 Specific Work

In a typical RO-plant up to 8 spiral-wound modules are placed in series in a single pressure vessel. These modules are typically manufactured with a length of 1 meter, making $L = 8\,\text{m}$ an appropriate length simulating the conditions in stadard practice RO. Morevover, standard modules, such as e.g. the SW30XLE-400 produced by FilmTec, has a feed channel height of approximately 0.7mm. However, to be consistent with the parameter values adopted from (Guillen and Hoek, 2009), a channel height $h = 0.5$mm is used. As to the energy recovery, there are several different implementations of such devises. The first generation let the pressurized brine drive a turbine, thereby generating electricity to power the main pumps. Later generations typically let the brine impart the pressure directly on a part of the feed in so-called pressure exchangers. This design has two pump systems, one main pump elevating one part of the feed to operating pressure, whereas the other part flows through the pressure exchanger followed by a booster pump to make up the difference. However, smaller and more compact RO-systems integrate the recovery directly, e.g. with a Clark pump (Spectra Watermakers, 2012). These newer recovery devices can achieve very high efficiencies, especially the small and integrated ones, of up to 98% (Thomson et al., 2003;

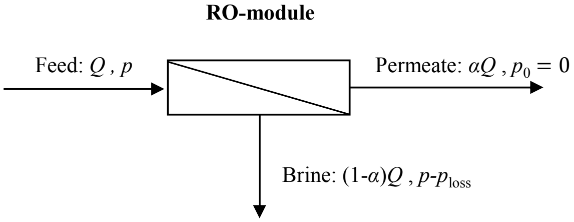
RO-module
Feed: Q , p
Permeate: αQ , $p_0 = 0$
Brine: $(1\text{-}\alpha)Q$, $p\text{-}p_{\text{loss}}$

Figure 6.8: *Schematic view of volumetric flow rates and pressures in the different streams in the RO-separation stage. All pressures are gauge pressures and a zero back-pressure is assumed on the permeate stream.*

Sun et al., 2009; Sanz et al., 2007). The subsequent analysis will adopt the integrated approach to energy recovery and use a recovery efficiency of $\zeta = 95\%$. Lastly, modern RO-desalination plants, such as those in Perth, Australia and Ashkelon, Israel, both operate their first pass membranes at an average permeate flux rate of around 4μm/s (Sanz et al., 2007; Sauvet-Goichon, 2007). This value will serve as a benchmark in this study. All membrane and fluid specific parameter values used can be found in Table 6.2.

The specific energy consumption in standard practice RO-desalination (turbulent flow model, $L = 8$m) is compared to the energy required in a smaller channel with a length of one meter, representing a small-scale and modular design. While theory, and simulation results, implies increasing specific energy consumption with channel length (due to greater pressure drop) the choice of one meter and not shorter for the small-scale design is motivated by the scale of currently manufactured spiral-wound modules. The laminar model is used for the flow in the shorter channel.

From a practical perspective, the specific energy consumption W_{RO} in the RO-stage, operating at a recovery rate α, is determined by the flow rates and pressures of the different streams, see Figure 6.8. With the given recovery efficiency, the specific energy consumption is given by

$$W_{RO} = \frac{Q\Delta p_1 - \zeta(1-\alpha)Q\Delta p_2}{\alpha Q} = \frac{\Delta p_1 - \zeta\Delta p_2}{\alpha} + \zeta\Delta p_2. \tag{6.18}$$

Notation	Description	Value	
A	Membrane permeability	$1.8 \cdot 10^{-12}$	(m s^{-1}Pa^{-1})
R	Membrane salt rejection	99%	
h	Channel height	$5.0 \cdot 10^{-4}$	(m)
L	Channel length	1, 8	(m)
c_0	Feed salinity	35	(kg m^{-3})
η	Feed dynamic viscosity	$1.1 \cdot 10^{-3}$	(kg m^{-1}s^{-1})
ρ	Feed density	$1.0 \cdot 10^{3}$	(kg m^{-3})
f_{os}	Osmotic pressure coefficient	78	(Pa m^{-3} kg^{-1})
D	Diffusivity	$1.6 \cdot 10^{-9}$	m^2/s
ζ	Recovery efficiency	95%	

Table 6.2: *Default parameter values characterizing membrane and feed properties. These values were, where appropriate, sourced in (Guillen and Hoek, 2009).*

The volumetric feed flow rate is denoted Q and $\Delta p_i = p_i - p_0$, $i = 1, 2$ are the hydrostatic heads of the feed and brine streams respectively relative the permeate stream. In the ideal limit, where $p_2 = p_1 = p$ and $\zeta = 1$, the limit in (6.2) is recovered with p being the osmotic pressure of the solution.

When operating the separation stage at a recovery α the volumetric flow rate of the permeate stream is given by αQ, where Q is the flow rate of the incoming feed. For a channel of length L, height h and width w this relation can be written as

$$\alpha \overline{u}_0 h w \quad = \quad 2 \overline{v}_p w L \quad \Leftrightarrow \quad \overline{u}_0 = \frac{2}{\alpha} \frac{L}{h} \overline{v}_p, \tag{6.19}$$

and the Reynolds number can be expressed as

$$Re \quad = \quad \frac{\rho \overline{u}_0 h}{\eta} = \frac{2 \rho L \overline{v}_p}{\alpha \eta}. \tag{6.20}$$

Given the recovery rate and the average permeate flux, the average feed flow speed is set by (6.19) and the required pressure is subsequently found using either of the two models introduced before. From (6.20) it can be seen that the Reynolds number of the flow, given α and $\overline{v}_p$, is proportional to the length of the channel. This suggests that the turbulent model is less appropriate for a shorter channel.

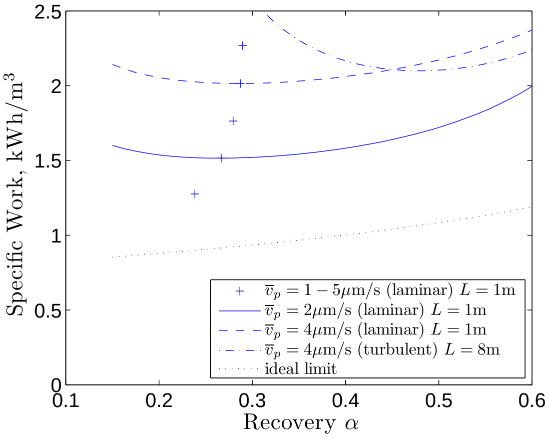

Figure 6.9: *At a fixed permeate flux rate the competing forces of high pressure drops (at lower recoveries) and increased polarization (at higher recoveries) give rise to an intermediate optimum in both the laminar and the turbulent channels. For the two permeation rates used, the shorter channel exhibits more than 25% lower energy consumption at respective optimal recoveries. The '+'-markers indicate the specific energy consumption at the optimal recovery rate for different flux rates* ($1 - 5\mu$*m/s) in the shorter, laminar channel.*

The specific energy consumption in the RO-separation stage is displayed in Figure 6.9 for several different flows. At lower recoveries the feed flow rates are higher causing greater pressure losses in the channel at fixed $\overline{v}_p$. On the other hand, lower flow rates (higher recoveries) results in higher concentration polarization. Additionally, the theoretical minimum, also shown in the same figure, increases with recovery. These competing forces explain the existence of an intermediate optimum recovery rate. The specific energy consumption in the longer and turbulent channel, at the standard permeate flux rate of $\overline{v}_p = 4\mu$m/s, exhibit a more pronounced optimum. While it is the same competing forces as in the laminar case, the penalty incurred by the pressure drop is much greater for turbulent flows at lower recoveries. This increases the optimal recovery rate compared to the shorter channel with a laminar flow. Moreover, the result of an optimal recovery around 50% supports the current trend in the industry of pursuing high recovery rates. Examining the laminar flows in the shorter channel it is noticed that the optimal recovery is much lower, less than 30%. At the same permeate flux rate, $\overline{v}_p = 4\mu$m/s, the shorter channel shows slightly improved energy

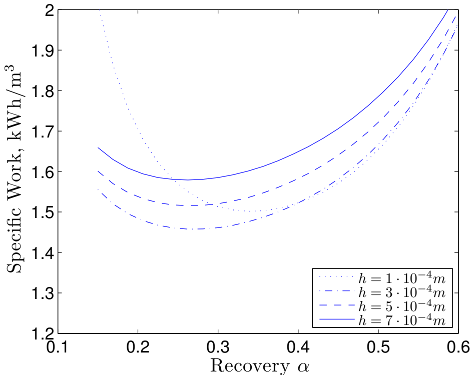

Figure 6.10: *Decreasing the height of the feed channel reduces polarization for laminar flows. For very thin channels the increased pressure loss overshadows this benefit. Decreasing the channel height in standard modules by half results in a 10% reduction in energy consumption at optimal recovery.*

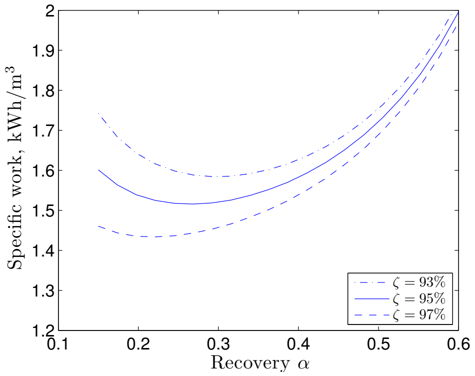

Figure 6.11: *The specific energy consumption is more sensitive to changes in recovery efficiency at lower recovery rates. At lower recovery rates the feed flow rates are higher, which in turn causes a greater pressure drop for a given average permeate flux rate. Consequently, we notice from (6.18) the increased energy consumption.*

consumption compared to the standard long channel. This suggests that the lower pressure drop and greater polarization in the laminar flow is balanced by the higher pressure drop but lower polarization in the turbulent channel.

Decreasing the permeate flux rate results in a significant reduction in specific energy consumption, as is expected. For instance, reducing the flux by half in the shorter channel, results in a 25% reduction in energy consumption. Furthermore, the results indicate that the optimal recovery increases slightly with the flux rate. This trend can be attributed to the influence of feed flow speeds on concentration polarization. Increasing the recovery rate at a fixed flux rate implies decreased longitudinal velocities in the feed channel. As seen in Figure 6.7, decreasing the longitudinal velocities at fixed average Peclet number enhances polarization. To achieve the same permeate flux, the operating pressure needs to be raised, which results in higher specific energy consumption.

In the previous section, it was shown that the height of the channel can significantly affect the concentration polarization for laminar flows. A thinner channel reduces polarization and consequently also reduces the required operating pressure to achieve the same flux. However, since a thinner channel implies higher volumetric flow rates at a given recovery and permeate flux rate, the pressure losses scale with $1/h^3$, see (6.9) and (6.19). This suggests the existence of an optimal channel height for laminar flows, see Figure 6.10. The results displayed in this figure indicate a 10% decrease in specific energy consumption by reducing the channel height in standard modules (0.7mm) with almost 50%.

The sensitivity of the energy consumption with respect to changes in the recovery efficiency is exhibited in Figure 6.11 for the laminar flow. Increasing the efficiency to 97% would decrease the optimal recovery rate for the laminar flow even further, down to almost 20%. The optimum of the turbulent flow is less sensitive (not shown here) to this efficiency. This difference in sensitivity is explained by the difference in pressure losses at different recoveries, see (6.18). Decreasing the membrane permeability (also not shown) shifts both curves down a similar amount.

The required operating pressures for the shorter channel and the longer channel are shown in Figure 6.12. At the same flux rate $\overline{v}_p = 4\mu\text{m/s}$, we observe that the laminar flow requires a greater pressure than the turbulent flow at higher recovery rates. This is again explained by the relatively greater concentration polarization in the laminar channel. For lower recovery rates (higher feed flow speeds) the pressure drop is significantly higher in the turbulent channel, which therefore requires a higher pressure to achieve the same average flux. Moreover, lowering the flux rate corresponds to significantly reduced pressures in the system. The difference in pressures, at respective optimal recovery (see Figure 6.9), is almost 20 bars.

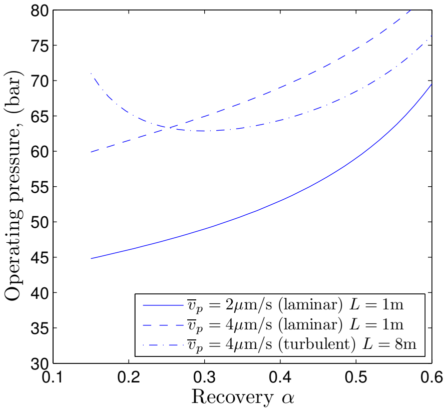

Figure 6.12: *Due to concentration polarization and increased average salinities in the feed channel, the required pressure increases with the recovery rate. For the turbulent channel, operating at lower recoveries incurs significant pressure losses along the channel, explaining the increased operating pressure to achieve the same flux.*

6.4 Summary

This analysis is focused on the RO-separation stage and the energy expenditures in the intake and pretreatment stages have not been included. Since these penalties scale with the overall volume circulated in the plant, the optimal recovery rate, viewed from an energy consumption perspective, would increase when including these steps in an overarching analysis. However, the magnitude of this increase is uncertain for several reasons. Locational factors influence the quality of the feed water and therefore also the level of pretreatment needed. Surface intakes require substantially more pretreatment compared to beach wells or deep sub-surface intakes. One case study on conventional pretreatment, a sand filter and a cartridge filter, measured a head loss of 1-2 bar in the pretreatment stage (Djebedjian et al., 2007). Another source cites head differences of 3.5 bar including seawater intake, while a yet another source indicates very minor head losses in a direct filtration pretreatment of around 0.25 bar (Semiat, 2008; Bonnelye et al., 2004). To put these numbers into a perspective of energy consumption, a plant operating at 25% recovery would incur an energy penalty, upstream of the separation stage, of less than 0.1 kWh/m^3 relative to a plant operating at 50% recovery, assuming the same level of treatment is required.

This study reveals that small-scale desalination system require slightly less specific energy in the separation stage. Moreover, optimality occur under much different conditions than those observed in the industry and suggested in the literature. A shorter channel results in optimal energy consumption at lower operating pressure and lower recovery rate. The former gives rise to the possibility of using less sturdy materials and thereby reducing cost. Operating at a lower recovery will require a reevaluation of all steps in the process. For instance, a lower recovery rate translates into lower salinities in the membrane modules, which reduces the risk of scaling. In current SWRO operation scaling is mitigated through pH-control and by adding chemical agents as antiscalants to the feed, a regiment which typically requires post-treatment of the brine before it can be discharged back into the ocean. Lower recovery will decrease the need for such treatment, perhaps alleviating the need for it altogether. Thus, in addition to slight reductions in energy consumption, scaling down RO desalination units has the potential to reduce costs of ancillary processes as well.

Also highlighted in this study is the effect of membrane productivity on energy consumption. While it makes sense under conditions of high capital costs and relatively low energy costs to maximize the permeation rates through the membranes, this strategy changes if these conditions are reversed. For instance, relying on more expensive renewable power, combined with cheaper mass-produced, small-scale units could render such a scenario possible.

Chapter 7

Future Work

The main benefits of a small-scale approach to physical capital in commodity-based industries will likely be realized in novel technology solutions. Nonetheless, there are some technologies that appear to be 'low-hanging fruit' in that they would benefit from being scaled down from both a physical and economic perspective. The technologies suggested below for further study from a scale perspective share two features. They are central to a modern economy and they have followed the traditional orthodoxy of upscaling.

7.1 Ammonia synthesis

Since its inception almost a century ago, the synthesis of ammonia through the Haber-Bosch process has provided the world with synthetic fertilizers, thereby dramatically altering Malthusian projections of population levels constrained by food supply. Except for minor alterations to the catalyst, the process has only changed appreciably in one regard – unit size. The first commercial plant, built by BASF in 1913, had a capacity of 30 metric tons per day (MTPD) (Jennings, 1991). This should be contrasted to the nominal capacity of 3,500 MTPD of the planned facility in Collie, Australia (Haldor Topsoe, 2009).

The synthesis of ammonia, as with most other catalytic processes, requires careful control of temperatures throughout the reactor in order to maintain favorable kinetics, thermodynamic equilibrium conditions and stability of the catalyst. The reduced surface area to volume ratio makes it harder to shed the heat generated by the exothermic process when scaling up in unit size. Indeed, modern ammonia synthesis reactors require internal heat exchangers to effectively manage process parameters. Moreover, operated at extremely high pressures, the handling of explosive gases also pose a substantial safety concern, in the event of critical failure, when scaling up in size.

Considering the synthesis reaction

$$\mathrm{N_2 + 3\,H_2 = 2\,NH_3} \qquad \Delta H = -50\mathrm{kJ/mol}, \tag{7.1}$$

from a heat management perspective reveals factors that benefit a small scale. Assuming a regular iron-based catalyst under typical operating conditions (T = 700K), an activity of $a = 10\mu\mathrm{mol/g{\cdot}s}$ is reasonable (Jennings, 1991). Assuming further that this iron catalyst is assembled at an effective density of $\rho = 3\mathrm{g/cm^3}$ in a cylindrical reactor with a volume $V = \pi r^2 h$ and surface area $A = 2\pi rh$. For the purposes here, the reactor is assumed to have an infinite internal heat conductivity. Lastly, the transfer of heat to the ambient (T = 300K) through free convection can be achieved with a heat transfer of $k = 10\,\mathrm{W/m^2K}$. Auto-thermal conditions arise when the generated hear $Q_{\mathrm{gen}} = V\rho a\Delta H$ is balanced by the heat $Q_{\mathrm{con}} = kA\Delta T$ conducted to the ambient air. This is achieved at a scale of

$$\frac{V}{A} = \frac{k\Delta T}{\rho a\Delta H} \qquad \Leftrightarrow \qquad r \approx 0.5\mathrm{cm}. \tag{7.2}$$

From the crude heat balance above, it can be concluded that a test tube-sized reactor can be operated auto-thermally, with little to no need for active removal of the process heat. Moreover, as shown in Chapter 2, *scaling down* a structure subjected to the same boundary conditions (pressure) typically require disproportionately less material. Thus, from a construction cost perspective, the ammonia synthesis reactor itself would benefit from a smaller scale.

The main use of ammonia is in fertilizers in the agricultural sector. Depending on crop type and climate, typical fertilizer use is on the order of (100kg N/ha y) (Metwally et al., 2011). A test tube-sized reactor (h = 10cm), with the same activity as discussed above, produces an output commensurate with the demand for one hectare of farmland. Moreover, the material inputs of nitrogen and hydrogen can be sourced directly from the ambient, e.g. through membrane air separation technologies and water electrolysis. These factors all point to the possibility of distributed operation.

Synthesizing ammonia from ambient nitrogen and water requires energy, and with the sub-processes mentioned in a distributed setting this energy would likely be in the form of electric power, perhaps from renewable sources. Compared to standard processes, which typically uses natural gas both as a hydrogen feedstock but also as an energy carrier, it is unlikely that distributed ammonia synthesis from renewable power can be cost competitive with today's prices. However, areas in developing countries far from the mainstream distribution network, this process could find a niche foothold in anticipation of cheaper renewable energy.

7.2 Liquid fuel synthesis

Technologies to synthesize liquid fuels have been known almost as long as the 'grandfather technology' of high-pressure chemistry in the Haber-Bosch process mentioned above. Limiting the whole host of technologies and process routes to those starting with synthesis gas, a

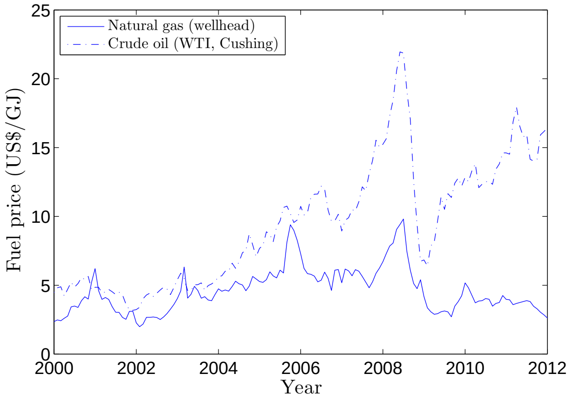

Figure 7.1: *The increase in unconventional natural gas reserves in the U.S. have effectively decoupled the domestic price of natural gas from oil. The conversion assumes* 1.1 *GJ/MMBTU and* 6.1 *GJ/bbl respectively. Data from EIA (2013).*

general reaction can be written as

$$\mathrm{CO} + 2\,\mathrm{H_2} = \text{liquid} + \text{heat}. \tag{7.3}$$

Thus, the same heat transfer argument as with ammonia synthesis holds that the reactor design can be greatly simplified at smaller sizes. Moreover, when synthesizing longer hydrocarbon chains, e.g. with Fischer-Tropsch synthesis, the distribution of the output depends strongly on process conditions, which in conjunction with heat management is easier to control at a smaller scale. These factors have generated interest in micro-channel synthesis reactors (Knobloch et al., 2013; LeViness et al., 2011).

The economic incentives to synthesize liquid fuels from natural gas have been greatly magnified as the price of this commodity has been decoupled from crude oil in recent years, see Figure 7.1. Moreover, much of the new resource base is in the form of shale gas, or other unconventional deposits. In comparison to conventional sources, these are generally much smaller. Converting the natural gas to a liquid on-site using modular and small-scale technologies reduces the need for a pipeline infrastructure, while at the same time yields a product which today commands a higher price. Similarly, gas that currently is flared at remote petroleum plays due to lack of transportation infrastructure could also be transformed into liquids and more easily stored and subsequently transported.

While the economic incentives exist today to transform natural gas to liquid fuels other carbon and energy feedstocks can be used. Mentioned by LeViness et al. (2011) are biomass and waste, which both typically are distributed and therefore require small-scale synthesis technologies. More interesting perhaps is the possibility of recycling atmospheric carbon as liquids, either for use in the transportation sector directly or as an intermediate energy storage. Either way, such capabilities are more tractable in small-scale and modular implementations than in rigid, large-scale installations.

7.3 Mining

Labor productivity is a key metric in evaluating mining operations since labor cost typically accounts for a large fraction of total costs. Therefore the most natural way to increase profitability of a given mine has been to scale up the size of individual process equipment such as loaders and haulers. Indeed, as seen in Figure 1.1 (right), the size of the largest available haulers has increased by a factor ten over the past 50 years. A general consequence of this trend is that smaller mines which preclude the use of larger equipment become less profitable, and mining operations tend to be more concentrated on large mines (Bozorgebrahimi et al., 2003). This trend was documented by Bartos (2007) in the case of the global copper industry from 1975 to 2000.

In a report on the future of the mining industry published by the Rand corporation in 2001 (Petersen et al., 2001), the opinion among key industry insiders was divided as to whether there would be a continued increase in truck sizes or if the economies of scale had reached a peak. In hindsight, knowing that truck sizes have indeed stagnated and the very largest trucks have stalled in the 400 ton class (as can be observed by considering the current portfolio of Caterpillar), the latter viewpoint seems to have prevailed. Bozorgebrahimi et al. (2003) illustrate some of the possible factors behind this apparent trend. For instance, auxiliary civil works, e.g. roads and bridges, to accommodate larger trucks going in and out of a mine become more costly. Moreover, larger equipment diminishes the possibility of selective mining techniques, thus resulting in the transportation of lower grade ores for further processing. The complexity of larger machinery also increases markedly at the largest end of the spectrum and hence requires additional training of operators and repair crews as well and larger (and more expensive) maintenance facilities.

With these issues in mind as well as considering workers' safety, automating various mining processes is a potentially attractive means for future cost reductions. Moreover, with mining typically being a remote operation, avoiding the additional infrastructure that has to be provided in order to accommodate on-site labor further increases the attractiveness of automation. According to Bellamy and Pravica (2011), the cost of one mining truck driver in remote areas of Australia amounts to $150,000 per year[1], of which more than $36,000 goes to auxiliary support such as transportation, accommodation and food. Operating in three shifts, this translates into $450,000 per year per truck in labor costs which does not include personnel in training. Assuming that the investment required for a single truck is on the order of $5 million, the capital charges at 10% interest are almost on par with labor costs, making the potential benefits of automation apparent.

Even though tests have been performed recently on operating retro-fitted autonomous mining trucks in Australian mines (Bellamy and Pravica, 2011), such technology has not yet caught on. With non-stationary and interacting robotic systems making progress by the day, as manifested by Google's autonomous car (Folsom, 2011) and 'Junior', the driver less vehicle

[1] Numbers are given in $ Australian but with an exchange rate of roughly 1:1, the use of USD is roughly equivalent

developed by Volkswagen and Stanford through DARPA (Stanek et al., 2010), it is only a matter of time before such technology becomes viable in an isolated area such as a mine. When the technology does become available, there is little reason automation should proceed with the ultra-large-size equipment of today and not with much smaller units, perhaps in the 1-10 ton class. In addition to the flexibility arguments raised in previous example favoring small unit size, smaller automated units can make smaller mines economical alongside large ones, thus increasing the total resource base. Also, from a physical perspective, the dead weight to payload ratio is likely to decrease with smaller trucks, suggesting potential fuel savings as well.

Chapter 8

Conclusion

In the discussion on economies of scale it has here been emphasized that this concept should not be viewed as synonymous with large-scale individual units of production. Many of the same benefits can be obtained in a paradigm of small-scale and modular units. For instance, external economies are to a greater extent functions of aggregate firm input and output rather than granularity of capacity. Similarly, the reduction in overheads and other indivisibilities on a firm level that are achieved through centralization can be garnered in exactly the same way by centralizing many small units in one location. On the other hand, only by scaling down in size and up in numbers can benefits arising from distributed operation be attained.

The common assertion that the often observed 'two-thirds-law' in engineering cost estimates is inextricably linked to material consumption in the production stage of physical capital has been refuted on physical grounds. Were this argument inherently true, any strategy that relies on providing large-scale aggregate capacity through mass-production of small-scale units would potentially be disqualified by resource constraints. It should be noted that the veracity of relative materials reductions observed in various scale-up enterprises are not being challenged. However, if these observations have been made where the scale-up was truly (or close to) uniform, then the original small-scale pilot unit was not optimally designed to begin with from a structural perspective.

The notion that cost tend to decline as cumulative production grows has been studied from a unit-scale perspective. Importantly, with a clear distinction between large and small-scale technologies, it was concluded that the latter, on average, exhibit rates 10 percentage-points higher learning rates. The suggested explanation to this statistical result is that the mass-production process allows for continuous improvements, product and process alterations and incorporation of exogenous technological improvement. Such effects arguably have a lower return in the generally highly customized construction process of large-scale installations.

Comparing the estimated cost reductions of scaling up in size and scaling up in numbers reveals that for typical values of respective parameters the two strategies result in roughly the same investment cost. Based on a case study of the operational cost in the four main electricity generating technologies in the U.S., the only factor that gives rise to significant operational economies of unit scale is increased labor productivity. Thus, in an environment where automation technologies are not only low-cost but increasingly reliable and capable the proposition of scaling down and possibly distributing capacity is presents many opportunities for reduced cost and increased utility by diffusing the technology to a wider group of users.

The observation that a paradigm marked by mass-production of small-scale units, as opposed to customized large-scale installations, likely will entail shorter lifetime and lead time of capital has financial ramifications. To capture the increased flexibility of engaging and disengaging markets a real options model is introduced as an optimal multiple stopping problem with lifetime and lead time as explicit parameters. In addition to a sensitivity analysis with respect to relevant parameters, it was shown that the investment cost of capital lasting 2.5 years need only be roughly half (per unit capacity) of an investment lasting 25 years in order to be competitive when given the opportunity of multiple consecutive investments.

In addition to a brief discussion on three technologies that in their current implementation would benefit from scaling down, a more detailed case study was performed on the actual separation stage in reverse osmosis desalination. With the backdrop of the current trend of operating at ever higher recovery rates and longer feed channels it was shown that shorter channels has an intrinsic advantage based on transport phenomena. In addition to the possibility of mass-producing small-scale desalination system that compare favorably to large-scale systems from an energy consumption, the insights gathered are also valuable for future unconventional approaches to desalination, e.g. submarine operation.

In conclusion, this thesis has provided several reasons why a *small unit scale* should be strongly considered in any technology development process that a priory does not have a natural scale. Rather than predicating continued development on a positive response to the question “Does the technology scale up?”, the question that should be posed is “Does the technology scale up in size or in numbers?”. With the unrelenting progress in automation technologies it is the prediction of this observer that firms that cling to the largely outdated mantra of “bigger-is-better” will find them self outmaneuvered by nimbler and modular technologies.

Bibliography

Ackerman, F., Biewald, B., White, D., Woolf, T., and Moomaw, W. (1999). Grandfathering and coal plant emissions: the cost of cleaning up the clean air act. *Energy Policy*, 27(15):929–940.

Anderson, K. P. (1973). Residential demand for electricity: Econometric estimates for California and the United States. *The Journal of Business*, 46(4):526–553.

Argote, L. and Epple, D. (1990). Learning Curves in Manufacturing. *Science*, 247(4945):920–924.

Arrow, K. (1962). The economic implications of learning by doing. *The Review of Economic Studies*, 29(3):155–173.

Baily, M. N., Chakrabarti, A. K., and Levin, R. C. (1985). Innovation and productivity in U.S. industry. *Brookings Papers on Economic Activity*, 1985(2):609–639.

Baker, R. W. (2004). *Membrane Technology and Application, second edition.* John Wiley & Sons ltd.

Bartos, P. J. (2007). Is mining a high-tech industry? Investigations into innovation and productivity advance. *Resources Policy*, 32(4):149–158.

Bellamy, D. and Pravica, L. (2011). Assessing the impact of driverless haul trucks in Australian surface mining. *Resources Policy*, 36(2):149–158.

Bender, C. (2011). Dual pricing of multi-exercise options under volume constraints. *Finance and Stochastics*, 15:1–26.

Bernard, A. B., Redding, S. J., and Schott, P. K. (2006). Multi-product firms and product switching. Technical report, National Bureau of Economic Research.

Bonnelye, V., Sanz, M. A., Durand, J.-P., Plasse, L., Gueguen, F., and Mazounie, P. (2004). Reverse osmosis on open intake seawater: pre-treatment strategy. *Desalination*, 167:191–200.

Bouchard, C. R., Carreau, P. J., Matsuura, T., and Sourirajan, S. (1994). Modeling of ultrafiltration: Predictions of concentration polarization effects. *Journal of Membrane Science*, 97:215–229.

Bozorgebrahimi, E., Hall, R. A., and Blackwell, G. H. (2003). Sizing equipment for open pit mining - a review of critical parameters. *Mining Technology*, 112(3):171–179.

Brealey, R. A., Myers, S. C., and Allen, F. (2008). *Principles of Corporate Finance*. McGraw Hill/Irwin.

Brian, P. L. T. (1965). Concentration polarization in reverse osmosis desalination with variable flux and incomplete salt rejection. *Industrial & Engineering Chemistry Fundamentals*, 4(4):439–445.

Busch, M. and Mickols, W. (2004). Reducing energy consumption in seawater desalination. *Desalination*, 165:299–312. Desalination Strategies in South Mediterranean Countries.

Carelli, M., Garrone, P., Locatelli, G., Mancini, M., Mycoff, C., Trucco, P., and Ricotti, M. (2010). Economic features of integral, modular, small-to-medium size reactors. *Progress in Nuclear Energy*, 52(4):403–414.

Carlaw, K. I. (2004). Returns to scale generated from uncertainty and complementarity. *Journal of Economic Behavior & Organization*, 53(2):261–282.

Carley, S. (2011). Historical analysis of U.S. electricity markets: Reassessing carbon lock-in. *Energy Policy*, 39(2):720–732. Special Section on Offshore wind power planning, economics and environment.

Carley, S. and Andrews, R. N. (2012). Creating a sustainable us electricity sector: the question of scale. *Policy Sciences*, 45(2):97–121.

Carmona, R. and Dayanik, S. (2008). Optimal multiple stopping of linear diffusions. *Mathematics of Operations Research*, 33(2):446–460.

Carmona, R. and Touzi, N. (2008). Optimal multiple stopping and valuation of swing options. *Mathematical Finance*, 18(2):239–268.

Castelnuovo, E., Galeotti, M., Gambarelli, G., and Vergalli, S. (2005). Learning-by-doing vs. learning by researching in a model of climate change policy analysis. *Ecological Economics*, 54(2-3):261–276. Technological Change and the Environment.

Cerci, Y. (2002). Exergy analysis of a reverse osmosis desalination plant in California. *Desalination*, 142(3):257–266.

Charcosset, C., Falconet, C., and Combe, M. (2009). Hydrostatic pressure plants for desalination via reverse osmosis. *Renewable Energy*, 34(12):2878–2882.

Chiara, N., Garvin, M., and Vecer, J. (2007). Valuing simple multiple-exercise real options in infrastructure projects. *Journal of Infrastructure Systems*, 13(2):97–104.

Christensen, L. R. and Greene, W. H. (1976). Economies of Scale in U.S. Electric Power Generation. *Journal of Political Economy*, 84(4):655–676.

Cunningham, J. A. (1980). Management: Using the learning curve as a management tool: The learning curve can help in preparing cost reduction programs, pricing forecasts, and product development goals. *Spectrum, IEEE*, 17(6):45–48.

Dahlgren, E., Göçmen, C., Lackner, K., and van Ryzin, G. (2013). Small modular infrastructure. *The Engineering Economist*, 58:231–266.

Dahlgren, E. and Lackner, K. (2012). Questioning a simple explanation to the two-thirds rule in scaling equipment cost. *Working Paper*.

Dahlgren, E. and Lackner, K. (2013). The Potential of Small and Modular Reverse Osmosis Desalination Units. *Working Paper*.

Dahlgren, E. and Leung, T. (2013). An optimal multiple stopping approach to infrastructure investment decisions. *Journal of Economic Dynamics and Control*.

Deng, S. and Oren, S. (2006). Electricity derivatives and risk management. *Energy*, 31(6-7):940–953. Electricity Market Reform and Deregulation.

Deutch, J., Moniz, E. J., Ansolabehere, S., Driscoll, M., Gray, P., Holdren, J., Joskow, P., Lester, R., and Todreas, N. (2003). The future of nuclear power. *MIT*.

Dixit, A. and Pindyck, R. (1994). *Investment Under Uncertainty*. Princeton University Press.

Djebedjian, B., Gad, H., Khaled, I., and Rayan, M. A. (2007). Reverse osmosis desalination plant in Nuweiba City (case study). *Proceedings of IWTC11*, pages 315–330.

Dupavillon, J. L. and Gillanders, B. M. (2009). Impacts of seawater desalination on the giant Australian cuttlefish Sepia apama in the upper Spencer Gulf, South Australia. *Marine Environmental Research*, 67(4-5):207–218.

Dutton, J. M. and Thomas, A. (1984). Treating progress functions as a managerial opportunity. *Academy of Management Review*, pages 235–247.

Edwards, B. K. and Starr, R. M. (1987). A note on indivisibilities, specialization, and economies of scale. *The American Economic Review*, pages 192–194.

EIA (1998). Challenges of electric power industry restructuring for fuel suppliers. `eia.gov`.

EIA (2000). The changing structure of the electric power industry 2000: An update. `eia.gov`.

EIA (2003). Annual Energy Outlook 2003. `eia.gov`.

EIA (2004). Form 860, Annual Electric Generator Report 2004. `eia.gov`.

EIA (2010). Updated Capital Cost Estimates for Electricity Generation Plants. `eia.gov`. Table 1. p. 7.

EIA (2011). Form 860, Annual Electric Generator Report 2011. `eia.gov`.

EIA (2013). Price data. `eia.gov`.

Elimelech, M. and Phillip, W. (2011). The future of seawater desalination: Energy, technology, and the environment. *Science*, 333(6043):712–717.

Esty, B. C. (1999). Improved techniques for valuing large-scale projects. *The Journal of Structured Finance*, 5(1):9–25.

Euzen, J., Trambouze, P., and Wauquier, J. (1993). *Scale-up Methodology for Chemical Processes*. Gulf Publishing Company.

Fan, L., Hobbs, B. F., and Norman, C. S. (2010). Risk aversion and CO2 regulatory uncertainty in power generation investment: Policy and modeling implications. *Journal of Environmental Economics and Management*, 60(3):193–208.

FERC (2010). FERC Form No. 1 for the year 2010. `ferc.gov`. Accessed on 2012.10.5.

Ferioli, F. and van der Zwaan, B. (2009). Learning in times of change: A dynamic explanation for technological progress. *Environmental Science & Technology*, 43:4002–4008.

Fight, A. (2005). *Introduction to project finance*. Butterworth-Heinemann.

Finnerty, J. D. (2007). *Project financing: asset-based financial engineering*, volume 386. Wiley.

Folsom, T. (2011). Social ramifications of autonomous urban land vehicles. In *IEEE International Symposium on Technology and Society, May 2011, Chicago*.

Frayer, J. and Uludere, N. (2001). What is it worth? Application of real options theory to the valuation of generation assets. *The Electricity Journal*, 14(8):40–51.

Fritzmann, C., Lwenberg, J., Wintgens, T., and Melin, T. (2007). State-of-the-art of Reverse Osmosis Desalination. *Desalination*, 216(1-3):1–76.

Funk, J. (2010). Exponential improvements and increasing returns to scale. In *Management of Innovation and Technology (ICMIT), 2010 IEEE International Conference on*, pages 506–511. IEEE.

Garcia, P., Knights, P. F., and Tilton, J. E. (2001). Labor productivity and comparative advantage in mining:: the copper industry in chile. *Resources Policy*, 27(2):97–105.

Garvin, M. J. and Cheah, C. Y. (2004). Valuation techniques for infrastructure investment decisions. *Construction Management and Economics*, 22(4):373–383.

Gekas, V. and Hallstrom, B. (1987). Mass transfer in the membrane concentration polarization layer under turbulent cross flow : I. critical literature review and adaptation of existing sherwood correlations to membrane operations. *Journal of Membrane Science*, 30(2):153–170.

Geraldes, V., Semiao, V., and de Pinho, M. N. (2002). The effect of the ladder-type spacers configuration in nf spiral-wound modules on the concentration boundary layers disruption. *Desalination*, 146(1-3):187–194.

Gille, D. (2003). Seawater intakes for desalination plants. *Desalination*, 156(1-3):249–256.

Grasselli, M. and Henderson, V. (2009). Risk aversion and block exercise of executive stock options. *Journal of Economic Dynamics and Control*, 33(1):109–127.

Greenlee, L., Lawler, D., Freeman, B., B.Marrot, and P.Moulin (2009). Reverse osmosis desalination: Water sources, technology, and today's challenges. *Water Research*, 43(9):2317–2348.

Guillen, G. and Hoek, E. M. (2009). Modeling the impacts of feed spacer geometry on reverse osmosis and nanofiltration processes. *Chemical Engineering Journal*, 149(1-3):221–231.

Haldi, J. and Whitcomb, D. (1967). Economies of scale in industrial plants. *Journal of Political Economy*, 75(4):373–385.

Haldor Topsoe (2009). Press release: Topsoe contracts the world's largest ammonia plant. `http://www.topsoe.com/news/News/2009/020709.aspx`. Accessed in Dec 2010.

Hamelinck, C. and Faaij, A. (2002). Future prospects for production of methanol and hydrogen from biomass. *Journal of Power Sources*, 111(1):1–22.

Harmon, C. (2000). Experience curves of photovoltaic technology. *Laxenburg, IIASA*, 17.

Harrison, M. (1985). *Brownian Motion and Stochastic Flow Systems*. New York, John Wiley & Sons.

Henderson, V. (2007). Valuing the option to invest in an incomplete market. *Mathematics and Financial Economics*, 1(2):103–128.

Herder, P. M., de Joode, J., Ligtvoet, A., Schenk, S., and Taneja, P. (2011). Buying real options - valuing uncertainty in infrastructure planning. *Futures*, 43(9):961–969.

Hettinga, W., Junginger, H., Dekker, S., Hoogwijk, M., McAloon, A., and Hicks, K. B. (2009). Understanding the reductions in US corn ethanol production costs: An experience curve approach. *Energy Policy*, 37(1):190–203.

Hisnanick, J. and Kymn, K. O. (1999). Modeling economies of scale: the case of US electric power companies. *Energy Economics*, 21(3):225–237.

Hreinsson, E. (1987). Hydroelectric project sequencing using heuristic techniques and dynamic programming. Presented at the 9th. Power Systems Computation Conference.

Humphreys, K. and Katell, S. (1981). *Basic cost engineering*. New York : M. Dekker.

Husan, R. (1997). The continuing importance of economies of scale in the automotive industry. *European Business Review*, 97(1):38–42.

Irwin, D. A. and Klenow, P. J. (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of Political economy*, pages 1200–1227.

Jack, M. (2009). Scaling laws and technology development strategies for biorefineries and bioenergy plants. *Bioresource Technology*, 100(24):6324–6330.

Jaillet, P., Ronn, E. I., and Tompaidis, S. (2004). Valuation of commodity-based swing options. *Management Science*, 50(7):909–921.

Jaimungal, S. (2011). Irreversible investments and ambiguity aversion. Working Paper, University of Toronto.

Jamasb, T. (2006). Technical change theory and learning curves: patterns of progress in energy technologies.

Jenkins, B. M. (1997). A comment on the optimal sizing of a biomass utilization facility under constant and variable cost scaling. *Biomass and Bioenergy*, 13(1-2):1–9.

Jennings, J. (1991). *Catalytic ammonia conversion*. Plenum Press, New York.

Joskow, P. L. (2005). Markets for power in the United States: An interim assessment. *The American Economic Review*. AEI-Brookings Joint Center Working Paper No. 05-20.

Joskow, P. L. and Baughman, M. L. (1976). The future of the U.S. nuclear energy industry. *The Bell Journal of Economics*, 7(1):3–32.

Junginger, M., de Visser, E., Hjort-Gregersen, K., Koornneef, J., Raven, R., Faaij, A., and Turkenburg, W. (2006). Technological learning in bioenergy systems. *Energy Policy*, 34(18):4024–4041.

Junginger, M., Faaij, A., and Turkenburg, W. C. (2005). Global experience curves for wind farms. *Energy policy*, 33(2):133–150.

Kaghazchi, T., Mehri, M., Ravanchi, M. T., and Kargari, A. (2010). A mathematical modeling of two industrial seawater desalination plants in the Persian Gulf region. *Desalination*, 252(1-3):135–142.

Kahouli-Brahmi, S. (2008). Technological learning in energy–environment–economy modelling: A survey. *Energy Policy*, 36(1):138–162.

Kaslow, T. and Pindyck, R. (1994). Valuing flexibility in utility planning. *The Electricity Journal*, 7(2):60–65.

Keeney, R. L. and Sicherman, A. (January/February 1983). Illustrative comparison of one utility's coal and nuclear choices. *Operations Research*, 31(1):50–83.

Kim, D. W. and Chang, H. J. (2012). Experience curve analysis on South Korean nuclear technology and comparative analysis with South Korean renewable technologies. *Energy Policy*, 40(0):361–373. Strategic Choices for Renewable Energy Investment.

Kim, S. and Hoek, E. M. (2005). Modeling concentration polarization in reverse osmosis processes. *Desalination*, 186(1-3):111–128.

Kim, S.-H., Lee, S.-H., Yoon, J.-S., Moon, S.-Y., and Yoon, C.-H. (2007). Pilot plant demonstration of energy reduction for RO seawater desalination through a recovery increase. *Desalination*, 203(1-3):153–159.

Kim, Y. M., Kim, S. J., Kim, Y. S., Lee, S., Kim, I. S., and Kim, J. H. (2009). Overview of systems engineering approaches for a large-scale seawater desalination plant with a reverse osmosis network. *Desalination*, 238(1-3):312–332.

Knobloch, C., Güttel, R., and Turek, T. (2013). Holdup and pressure drop in micro packed-bed reactors for fischer-tropsch synthesis. *Chemie Ingenieur Technik*.

Koellner, W., Brown, G., Rodriguez, J., Pontt, J., Cortes, P., and Miranda, H. (2004). Recent advances in mining haul trucks. *Industrial Electronics, IEEE Transactions on*, 51(2):321–329.

Lai, W. M., Rubin, D., and Krempl, E. (1993). *Introduction to Continuum Mechanics*. Elsevier, 3 edition.

Laitner, J. A. and Sanstad, A. H. (2004). Learning-by-doing on both the demand and the supply sides: implications for electric utility investments in a heuristic model. *International Journal of Energy Technology and Policy*, 2(3):142–152.

Langevin, A., Mbaraga, P., and Campbell, J. F. (1996). Continuous approximation models in freight distribution: An overview. *Transportation Research Part B: Methodological*, 30(3):163–188.

Larminie, J. and Dick, A. (2003). *Fuel Cell Systems Explained.* John Wiley & Sons Ltd.

Lazonick, W. and Brush, T. (1985). The 'Horndal effect'in early US manufacturing. *Explorations in Economic History*, 22(1):53–96.

Lee, L. Y. et al. (2009). Ozone-biological activated carbon as a pretreatment process for reverse osmosis brine treatment and recovery. *Water Research*, In Press, Corrected Proof:–.

Leung, T. and Sircar, R. (2009). Accounting for risk aversion, vesting, job termination risk and multiple exercises in valuation of empoloyee stock options. *Mathematical Finance*, 19(1):99–128.

Levendis, J., Block, W., and Morrel, J. (2006). Nuclear power. *Journal of business ethics*, 67(1):37–49.

Levin, R. (1977). Technical Change and Optimal Scale: Some Evidence and Implications. *Southern Economic Journal*, 44:208–211.

LeViness, S., Tonkovich, A., Jarosch, K., Fitzgerald, S., Yang, B., and McDaniel, J. (2011). Improved Fischer-Tropsch economics enabled by microchannel technology. *White Paper generated by Velocys.*

Liang, S., Liu, C., and Song, L. (2009). Two-step optimization of pressure and recovery of reverse osmosis desalination process. *Environmental Science & Technology*, 43(9):3272–3277.

Lieberman, M. (1987). Market Growth, Economies of Scale, and Plant Size in the Chemical Processing Industries. *The Journal of Industrial Economics*, 36:175–191.

Lindman, Å. and Söderholm, P. (2012). Wind power learning rates: A conceptual review and meta-analysis. *Energy Economics*, 34(3):754–761.

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T. (2005). *Economic Transformations.* Oxford University Press.

Locatelli, G. and Mancini, M. (2010). Small–medium sized nuclear coal and gas power plant: A probabilistic analysis of their financial performances and influence of CO2 cost. *Energy Policy*, 38(10):6360–6374.

Lovins, A. B., Datta, E. K., Feiler, T., Rabago, K. R., Swisher, J. N., Lehmann, A., and Wicker, K. (2002). *Small is Profitable.* Rocky Mountain Institute.

Lu, Y.-Y., Hu, Y.-D., Zhang, X.-L., Wu, L.-Y., and Liu, Q.-Z. (2007). Optimum design of reverse osmosis system under different feed concentration and product specification. *Journal of Membrane Science*, 287(2):219–229.

Ludkovski, M. (2008). Financial hedging of operational flexibility. *International Journal of Theoretical and Applied Finance*, 11(8):799–839.

Luettel, T., Himmelsbach, M., and Wuensche, H. J. (13). Autonomous ground vehicles x2014; concepts and a path to the future. *Proceedings of the IEEE*, 100(Special Centennial Issue):1831–1839.

Lumley, R. and Zervos, M. (2001). A model for investments in the natural resource industry with switching costs. *Mathematics of Operations Research*, 26(4):637–653.

Lundberg, E. (1961). *Produktivitet och räntabilitet: studier i kapitalets betydelse inom svenskt näringsliv.* Studieförbundet Näringsliv och samhälle.

Luo, R. C. and Chang, C.-C. (2012). Multisensor fusion and integration: A review on approaches and its applications in mechatronics. *Industrial Informatics, IEEE Transactions on*, 8(1):49–60.

Lüthi, S. and Prässler, T. (2011). Analyzing policy support instruments and regulatory risk factors for wind energy deployment – a developers' perspective. *Energy Policy*, 39(9):4876–4892.

Lyster, E. and Cohen, Y. (2007). Numerical study of concentration polarization in a rectangular reverse osmosis membrane channel: Permeate flux variation and hydrodynamic end effects. *Journal of Membrane Science*, 303(1-2):140–153.

Manne, A. S. (1961). Capacity expansion and probabilistic growth. *Econometrica: Journal of the Econometric Society*, pages 632–649.

Martinot, E., Dienst, C., Weiliang, L., and Qimin, C. (2007). Renewable energy futures: Targets, scenarios, and pathways. *Annu. Rev. Environ. Resour.*, 32:205–239.

McDonald, A. and Schrattenholzer, L. (2001). Learning rates for energy technologies. *Energy Policy*, 29(4):255–261.

McDonald, R. and Siegel, D. (1985). Investment and the valuation of firms when there is an option to shut down. *International Economic Review*, 26(2):331–349.

McDonald, R. and Siegel, D. (1986). The value of waiting to invest. *The Quarterly Journal of Economics*, 101(4):707–728.

Meinshausen, N. and Hambly, B. M. (2004). Monte Carlo methods for the valuation of multiple-exercise options. *Mathematical Finance*, 14(4):557–583.

Metwally, T., Gewaily, E., and Naeem, S. (2011). Nitrogen response curve and nitrogen use efficiency of Egyptian hybrid rice. *J, Agric, res, kafer EL-Sheikh univ*, 37(1).

Mohan, Y. and Ponnambalam, S. (2009). An extensive review of research in swarm robotics. In *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*, pages 140–145. IEEE.

Mulder, M. (1991). *Basic Principles of Membrane Technology*. Kluwer Academic Publishers.

National Academy of Engineering (2012). Grand Challenges of Engineering. engineeringchallenges.com. Accessed on 2012.12.28.

Neij, L. (1997). Use of experience curves to analyse the prospects for diffusion and adoption of renewable energy technology. *Energy policy*, 25(13):1099–1107.

Neij, L. (2008). Cost development of future technologies for power generation – A study based on experience curves and complementary bottom-up assessments. *Energy Policy*, 36(6):2200–2211.

Norberg-Bohm, V. (2000). Creating incentives for environmentally enhancing technological change: Lessons from 30 years of U.S. energy technology policy. *Technological Forecasting and Social Change*, 65(2):125–148.

Nordhaus, W. D. (2007). A review of the "Stern Review on the Economics of Climate Change". *Journal of Economic Literature*, 45(3):686–702.

OECD/IEA (2010). Projected costs of generating electricity: 2010 edition. oecd-nea.org. Accessed on 2013.02.10.

Oh, H.-J. et al. (2009). Scale formation in reverse osmosis desalination: model development. *Desalination*, 238(1-3):333–346.

Pacenti, P. et al. (1999). Submarine seawater reverse osmosis desalination system. *Desalination*, 126(1-3):213–218. European Conference on Desalination and the Environment.

Panzar, J. C. and Willig, R. D. (1977). Economies of scale in multi-output production. *The Quarterly Journal of Economics*, 91(3):481–493.

Pepermans, G., Driesen, J., Haeseldonckx, D., Belmans, R., and D'haeseleer, W. (2005). Distributed generation: definition, benefits and issues. *Energy Policy*, 33(6):787–798.

Peters, T., Pint, D., and Pint, E. (2007). Improved seawater intake and pre-treatment system based on neodren technology. *Desalination*, 203(1-3):134–140. EuroMed 2006 - Conference on Desalination Strategies in South Mediterranean Countries.

Petersen, D. J., LaTourrette, T., and Bartis, J. T. (2001). http://www.rand.org/pubs/monograph_reports/MR1324. New Forces at Work in Mining: Industry Views of Critical Technologies.

Petrina, A. (2011). Advances in robotics (review). *Automatic Documentation and Mathematical Linguistics*, 45(2):43–57.

Pindyck, R. S. (1986). Irreversible investment, capacity choice, and the value of the firm.

Reddy, K. and Ghaffour, N. (2007). Overview of the cost of desalinated water and costing methodologies. *Desalination*, 205(1-3):340–353.

Riahi, K., Rubin, E. S., Taylor, M. R., Schrattenholzer, L., and Hounshell, D. (2004). Technological learning for carbon capture and sequestration technologies. *Energy Economics*, 26(4):539–564.

Rivera-Tinoco, R., Schoots, K., and van der Zwaan, B. (2012). Learning curves for solid oxide fuel cells. *Energy Conversion and Management*, 57:86–96.

Rubin, E. S., Yeh, S., Antes, M., Berkenpas, M., and Davison, J. (2007). Use of experience curves to estimate the future cost of power plants with CO2 capture. *International Journal of Greenhouse Gas Control*, 1(2):188 – 197.

Sahal, D. (1985). Technological guideposts and innovation avenues. *Research Policy*, 14(2):61 – 82.

Sanz, M. A., Stover, R. L., Degrémont, S., and Recovery, E. (2007). Low energy consumption in the Perth seawater desalination plant. In *Proceedings of the International Desalination Association Congress, Maspalomas, Gran Canaria, Spain*.

Sauvet-Goichon, B. (2007). Ashkelon desalination plant – a successful challenge. *Desalination*, 203(1-3):75–81. EuroMed 2006 - Conference on Desalination Strategies in South Mediterranean Countries.

Schock, G. and Miquel, A. (1987). Mass transfer and pressure loss in spiral wound modules. *Desalination*, 64(0):339–352.

Schoots, K., Ferioli, F., Kramer, G. J., and van der Zwaan, B. C. (2008). Learning curves for hydrogen production technology: an assessment of observed cost reductions. *International Journal of Hydrogen Energy*, 33(11):2630–2645.

Schoots, K., Kramer, G., and Van Der Zwaan, B. (2010). Technology learning for fuel cells: An assessment of past and potential cost reductions. *Energy policy*, 38(6):2887–2897.

Schwinge, J., Neal, P. R., Wiley, D. E., Fletcher, D. F., and Fane, A. G. (2004). Spiral wound modules and spacers: Review and analysis. *Journal of Membrane Science*, 242(1-2):129 – 153. Membrane Engineering Special Issue.

Semiat, R. (2000). Desalination: Present and future. *Water International*, 25(1):54–65.

Semiat, R. (2008). Energy issues in desalination processes. *Environmental Science & Technology*, 42:8193–8201.

Shreve, S. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models.* New York, NY: Springer.

Sick, G. and Gamba, A. (2010). Some important issues involving real options: An overview. *Multinational Finance Journal*, 14(1/2):73–123.

Smith, A. (2000). *The wealth of nations.* New York: Modern Library, 2000.

Solow, R. M., Tobin, J., von Weizsäcker, C. C., and Yaari, M. (1966). Neoclassical growth with fixed factor proportions. *The Review of Economic Studies*, 33(2):79–115.

Song, L. (2010). Concentration polarization in a narrow reverse osmosis membrane channel. *AIChE Journal*, 56(1):143–149.

Song, L. and Elimelech, M. (1995). Theory of concentration polarization in crossflow filtration. *J. Chem. Soc., Faraday Trans.*, 91:3389–3398.

Song, L., Hong, S., Hu, J. Y., Ong, S. L., and Ng, W. J. (2002). Simulations of full-scale reverse osmosis membrane process. *Journal of Environmental Engineering*, 128(10):960–966.

Song, L. and Yu, S. (1999). Concentration polarization in cross-flow reverse osmosis. *AIChE Journal*, 45(5):921–928.

Spectra Watermakers (2012). Marine Products, Catalina 300 Mk II. `http://www.spectrawatermakers.com/`.

Srikanth, N. and Funk, J. L. (2011). Geometric scaling and long-run reductions in cost: The case of wind turbines. In *Technology Management Conference (ITMC), 2011 IEEE International*, pages 691–696. IEEE.

Stanek, G., Langer, D., Müller-Bessler, B., and Huhnke, B. (2010). Junior 3: A test platform for advanced driver assistance systems. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 143–149.

Stern, N. et al. (2007). The economics of climate change: the Stern report. *Cambridge, UK.*

Stover, R. L. (2007). Seawater reverse osmosis with isobaric energy recovery devices. *Desalination*, 203(1-3):168–175. EuroMed 2006 - Conference on Desalination Strategies in South Mediterranean Countries.

Sun, J., Wang, Y., Xu, S., Wang, S., and Wang, Y. (2009). Performance prediction of hydraulic energy recovery (HER) device with novel mechanics for small-scale swro desalination system. *Desalination*, 249(2):667–671.

Thomson, M., Miranda, M. S., and Infield, D. (2003). A small-scale seawater reverse-osmosis system with excellent energy efficiency over a wide operating range,. *Desalination*, 153(1-3):229–236.

Tone, K. and Sahoo, B. K. (2003). Scale, indivisibilities and production function in data envelopment analysis. *International Journal of Production Economics*, 84(2):165–192.

Tribe, M. and Alpine, R. (1986). Scale economies and the '0.6 rule'. *Engineering Costs and Production Economics*, 10(1):271–278.

U.S. Census Bureau (2007). 2007 economic census. `http://www.census.gov/econ/census07/`. Accessed on 2012.10.12.

Van Den Wall Bake, J., Junginger, M., Faaij, A., Poot, T., and Walter, A. (2009). Explaining the experience curve: Cost reductions of Brazilian ethanol from sugarcane. *Biomass and Bioenergy*, 33(4):644–658.

van Mieghem, J. A. (2008). *Operations Strategy: Principles and Practice*. Dynamic Ideas, Llc.

Vrba, P. (2013). Review of industrial applications of multi-agent technologies. In *Service Orientation in Holonic and Multi Agent Manufacturing and Robotics*, pages 327–338. Springer.

Weaver, R. D. (2008). Collaborative pull innovation: origins and adoption in the new economy. *Agribusiness*, 24(3):388–402.

Weiss, M., Junginger, H., and Patel, M. K. (2008). Learning energy efficiency: experience curves for household appliances and space heating, cooling, and lighting technologies.

Weiss, M., Junginger, M., Patel, M. K., and Blok, K. (2010a). A review of experience curve analyses for energy demand technologies. *Technological Forecasting and Social Change*, 77(3):411–428.

Weiss, M., Patel, M. K., Junginger, M., and Blok, K. (2010b). Analyzing price and efficiency dynamics of large appliances with the experience curve approach. *Energy Policy*, 38(2):770–783.

Westner, G. and Madlener, R. (2012). Investment in new power generation under uncertainty: Benefits of CHP vs. condensing plants in a copula-based analysis. *Energy Economics*, 34(1):31–44.

Whitaker, R. (1998). Investment in volume building: the virtuous cycle in pafc. *Journal of Power Sources*, 71(1-2):71–74.

White, C., Steeper, R., and Lutz, A. (2006). The hydrogen-fueled internal combustion engine: a technical review. *International Journal of Hydrogen Energy*, 31(10):1292–1305.

Wilf, M. and Klinko, K. (2001). Optimization of seawater RO systems design. *Desalination*, 138(1-3):299–306.

Williams, J. H., DeBenedictis, A., Ghanadan, R., Mahone, A., Moore, J., Morrow, W. R., Price, S., and Torn, M. S. (2012). The technology path to deep greenhouse gas emissions cuts by 2050: the pivotal role of electricity. *Science*, 335(6064):53–59.

Winkler, H., Hughes, A., and Haw, M. (2009). Technology learning for renewable energy: Implications for South Africa's long-term mitigation scenarios. *Energy Policy*, 37(11):4987–4996.

Wittholz, M. K., O'Neill, B. K., Colby, C. B., and Lewis, D. (2008). Estimating the cost of desalination plants using a cost database. *Desalination*, 229(1-3):10–20.

Wright, T. (1936). Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences*, 3:122–128.

Yeh, S. and Rubin, E. S. (2007). A centurial history of technological change and learning curves for pulverized coal-fired utility boilers. *Energy*, 32(10):1996–2005.

Yeh, S., Rubin, E. S., Taylor, M. R., and Hounshell, D. A. (2005). Technology innovations and experience curves for nitrogen oxides control technologies. *Journal of the Air & Waste Management Association*, 55(12):1827–1838.

Appendix A: Proofs of Chapter 5

Some fundamental properties of asset pricing with a log-normal underlying process (geometric Brownian motion) are briefly reviewed before the statements in Chapter 5 are proved.

A complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is assumed to be equipped with a filtration $\mathbb{F} = (\mathcal{F})_{t \geq 0}$ associated with the Brownian motion B_t, which by definition satisfies $B_t - B_s \in \mathcal{N}(0, t-s)$ for $t - s > 0$. The uncertain output price process X_t is modeled by a geometric Brownian motion through the stochastic differential equation in (5.4), which is repeated here for convenience:

$$dX_t = \alpha X_t \, dt + \sigma X_t \, dB_t, \qquad X_0 \in (0, \infty). \tag{A.1}$$

Given a sufficiently smooth function $f(x,t)$, the Itô-Doeblin formula, see e.g. (Shreve, 2004, p.146), can be stated as

$$\begin{aligned} df(t, X_t) &= \partial_t f(t, X_t) dt + \partial_x f(t, X_t) dX_t + \frac{1}{2} f(t, X_t) dX_t dX_t \\ &= \left(\partial_t f(t, X_t) + \alpha X_t \partial_x f(t, X_t) + \frac{1}{2} \sigma^2 X_t^2 \partial_{xx} f(t, X_t) \right) dt \\ & \quad + \sigma X_t f(t, X_t) dB_t, \end{aligned} \tag{A.2}$$

where $dB_t^2 = dt$ and all higher order terms in dt have been neglected. If the dt-term vanishes it can be shown that $(f(t, X_t))_{t \geq 0}$ is a martingale, i.e. $\mathbb{E}\left\{ f(t, X_t) | \mathcal{F}_s \right\} = f(s, X_s), \ s < t$. Assuming that the value v of an asset (or investment opportunity) depends only on the price

X_t means that v can be determined by requiring that the discounted present value process $\left(e^{-rt}v\left(X_t\right)\right)_{t\geq 0}$ is a martingale. That is, v has to satisfy the differential equation from (A.2)

$$-rv + \alpha x v' + \frac{1}{2}\sigma^2 x^2 v'' = 0\ ,$$

which, through the ansatz $v = Av^\gamma$, gives rise to the polynomial equation

$$\frac{1}{2}\sigma^2\gamma(\gamma-1) + \alpha\gamma - r = 0\ . \tag{A.3}$$

Thus, any functions of the form

$$v(x) = Ax^{\gamma_-} + Bx^{\gamma_+}, \tag{A.4}$$

where $\gamma_\pm$ are the two solutions to (A.3), render the process $\left(e^{-rt}v\left(X_t\right)\right)_{t\geq 0}$ a martingale. With a positive discount rate r and with the requirement that the discount rate exceeds the drift rate, $r > \alpha$, it can be seen that one of the solutions, γ_-, to the equation above is negative and the other, γ_+, is greater than one. Demanding that the value $v(x)$ is bounded for finite x forces $A = 0$. The positive root, simply referred to as γ, is given by

$$\gamma = \frac{1}{2} - \frac{\alpha}{\sigma^2} + \sqrt{\left(\frac{1}{2} - \frac{\alpha}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}}\,.$$

For further background to stochastic calculus see e.g. Shreve (2004); Harrison (1985).

Proposition 1. *The value function $v^{(k)}(x)$, $k \geq 1$, satisfies,*

$$v^{(k)}(x) = \sup_{\vec{\tau}\in\mathcal{S}^k} \mathbb{E}\left\{\sum_{i=1}^{k} e^{-r\tau_i}\psi^+(X^{0,x}_{\tau_i})\right\}, \tag{A.5}$$

with $\psi^+(x) = \max\{0, \psi(x)\}$.

Proof: For every fixed stopping rule $\vec{\tau} = (\tau_i)_{i=1}^k \in \mathcal{S}^k$, we define a subsequence $(\hat{\tau}_j)_{j=1}^s$, $s \leq k$, of $(\tau_i)_{i=1}^k$ by recording only those stopping times at which the reward is non-negative:

$$\{\hat{\tau}_j\} = \left\{\tau_i \,|\, \psi(X_{\tau_i}^{0,x}) \geq 0\,;\, i = 1, \ldots, k\right\}.$$

In the case of finite k we can append the subsequence $(\hat{\tau}_j)_{j=1}^s$ with $k - s$ infinite stopping times, $\hat{\tau}_{k-s+1} = \cdots = \hat{\tau}_k = \infty$ to create the full sequence $\vec{\hat{\tau}} = (\hat{\tau}_j)_{j=1}^k$. Since the stopping times within $\vec{\tau}$ are refracted with at least the constant time T, the same is true for those in $\vec{\hat{\tau}}$ by construction, and therefore $\vec{\hat{\tau}} \in \mathcal{S}^k$.

By avoiding those stopping times with a negative reward, the total discounted reward:

$$g_k(x; \vec{\tau}, \psi) := \sum_{i=1}^{k} e^{-r\tau_i} \psi(X_{\tau_i}^{0,x}), \quad \vec{\tau} \in \mathcal{S}^k,$$

is dominated by $g_k(x; \vec{\hat{\tau}}, \psi)$ in expectation, namely,

$$\mathbb{E}\left\{g_k(x; \vec{\tau}, \psi)\right\} \leq \mathbb{E}\left\{g_k(x; \vec{\hat{\tau}}, \psi)\right\}. \tag{A.6}$$

In addition, choosing $\vec{\tau}$ to be $\vec{\hat{\tau}}$ results in the equality $g_k(x; \vec{\hat{\tau}}, \psi) = g_k(x; \vec{\hat{\tau}}, \psi^+)$, almost surely. That means that maximizing over the stopping rules $\vec{\hat{\tau}}$ for the original problem $v^{(k)}(x)$ will achieve the upper bound (RHS of (A.5)) with the non-negative reward ψ^+. In summary,

$$v^{(k)}(x) \equiv \sup_{\vec{\tau} \in \mathcal{S}^k} \mathbb{E}\left\{g_k(x; \vec{\tau}, \psi)\right\} = \sup_{\vec{\hat{\tau}} \in \mathcal{S}^k} \mathbb{E}\left\{g_k(x; \vec{\hat{\tau}}, \psi^+)\right\} = \sup_{\vec{\tau} \in \mathcal{S}^k} \mathbb{E}\left\{g_k(x; \vec{\tau}, \psi^+)\right\}. \quad \square$$

The following refers to the single optimal stopping problem

$$v^{(1)}(x) = \sup_{\tau_1 \in \mathcal{S}} \mathbb{E}\left\{e^{-r\tau_1} \psi\left(X_{\tau_1}^{0,x}\right)\right\}\ . \tag{A.7}$$

Also, repeated reference will be made to the operator Λ, which is defined by

$$\Lambda = \gamma - x\frac{d}{dx} \, .$$

Lemma 1. *Let $\psi : \mathbb{R}^+ \to \mathbb{R}$ be a reward function in the single stopping problem (A.7). If x_1^* is a global maximum for $\psi(x)/x^\gamma$ on $\mathbb{R}^+$ and if*

$$\frac{d}{dx}\Lambda\psi(x) \equiv (\gamma-1)\psi'(x) - x\psi''(x) \geq 0, \quad x \geq x_1^*, \tag{A.8}$$

then

$$v^{(1)}(x) = \psi(x \vee x_1^*)\left[1 \wedge \left(\frac{x}{x_1^*}\right)^\gamma\right], \quad x \in \mathbb{R}^+,$$

where $v^{(1)}(x)$ is continuous on $\mathbb{R}^+$.

Proof: From the Laplace transform of the first passage time to x_1^* the function $\hat{v}(x)$, where

$$\hat{v}(x) = \begin{cases} \psi(x_1^*)\left(\frac{x}{x_1^*}\right)^\gamma, & x < x_1^*, \\ \psi(x), & x \geq x_1^*, \end{cases} \tag{A.9}$$

is a candidate for the solution. From the conditions on $\psi(x)$, it follows that, for $x \geq x_1^*$,

$$\begin{aligned} \Lambda\psi(x) &= \gamma\psi(x) - x\psi'(x) \geq 0, && \text{(A.10)} \\ \frac{d}{dx}\Lambda\psi(x) &= (\gamma-1)\psi'(x) - x\psi''(x) \geq 0. && \text{(A.11)} \end{aligned}$$

Looking at the drift term of $e^{-rt}\hat{v}(X_t)$,

$$\begin{aligned} \mathbb{E}\left\{de^{-rt}\hat{v}(X_t)\right\} &= e^{-rt}\frac{\psi(x_1^*)}{(x_1^*)^\gamma}\left[-r + \alpha\gamma + \frac{1}{2}\sigma^2\gamma(\gamma-1)\right]1_{\{X_t \leq x_1^*\}}dt \\ &+ e^{-rt}\left[-r\psi(X_t) + \alpha X_t\psi'(X_t) + \frac{1}{2}\sigma^2 X_t^2\psi''(X_t)\right]1_{\{X_t \geq x_1^*\}}dt, \quad \text{(A.12)} \end{aligned}$$

it is observed that first term vanishes identically, and the second term is non-positive following from (A.10) and (A.11). Hence, $(e^{-rt}\hat{v}(X_t))_{t\geq 0}$ is a supermartingale, which implies that

$$\hat{v}(x) = \mathbb{E}\left\{e^{-r(0\wedge\tau)}\hat{v}(X^{0,x}_{0\wedge\tau})\right\} \geq \mathbb{E}\left\{e^{-r(t\wedge\tau)}\hat{v}(X^{0,x}_{t\wedge\tau})\right\}, \quad \tau \in \mathcal{S}. \tag{A.13}$$

The assumption of a linear bound on $\psi(x)$ implies that $e^{-r(t\wedge\tau)}\hat{v}(X^{0,x}_{t\wedge\tau})$ is integrable. Also, taking the limit $t \to \infty$ in (A.13) and maximizing over τ yields:

$$\hat{v}(x) \geq \sup_{\tau\in\mathcal{S}} \mathbb{E}\left\{e^{-r\tau}\psi(X^{0,x}_{\tau})\right\}. \tag{A.14}$$

Conversely, choosing the specific stopping time $\tau = \tau_{x_1^*}$, the process $\left(e^{-r(t\wedge\tau_{x_1^*})}\hat{v}(X_{t\wedge\tau_{x_1^*}})\right)_{t\geq 0}$ is a martingale by construction and therefore

$$\begin{aligned}\hat{v}(x) &= \mathbb{E}\left\{e^{-r(t\wedge\tau_{x_1^*})}\hat{v}(X_{t\wedge\tau_{x_1^*}})\right\} = \mathbb{E}\left\{e^{-r\tau_{x_1^*}}\hat{v}(x^*)\right\} \\ &= \mathbb{E}\left\{e^{-r\tau_{x_1^*}}\psi(x^*)\right\} \leq \sup_{\tau\in\mathcal{S}} \mathbb{E}\left\{e^{-r\tau}\psi(X^{0,x}_{\tau})\right\}.\end{aligned} \tag{A.15}$$

The expressions in (A.14) and (A.15) together give the desired result,

$$\hat{v}(x) = v(x) = \sup_{\tau\in\mathcal{S}} \mathbb{E}\left\{e^{-r\tau}\psi(X^{0,x}_{\tau})\right\}. \quad \square$$

A.1 Optimal Multiple Stopping Problem

With Lemma 1 and Proposition 1 the main result can be stated and proved.

Theorem 1. *Let $\psi : \mathbb{R}^+ \to \mathbb{R}$ be a reward function with a break-even point x_0. If $\Lambda\psi(x)$ is convex for $x \in (x_0, \infty)$, with $\Lambda\psi(x)$ increasing for large x, then, for every $k \geq 1$, there exists* *an $x_k^* > x_0$ such that*

$$v^{(k)}(x) = \psi^{(k)}(x \vee x_k^*) \left[1 \wedge \left(\frac{x}{x_k^*}\right)^\gamma\right], \quad k \geq 1, \tag{A.16}$$

where

$$\psi^{(k)}(x) = \psi(x) + e^{-rT}\mathbb{E}\left\{v^{(k-1)}(X_T^{0,x})\right\}. \tag{A.17}$$

Moreover, the sequence $(x_k^*)_{k\geq 1}$ is strictly decreasing, and $\left(v^{(k)}\right)_{k\geq 1}$ is a strictly increasing sequence of continuous functions on $\mathbb{R}^+$. Also, for any bounded subset $D \in \mathbb{R}^+$ there exists a constant K_D, such that $v^{(k)}(x) \leq K_D$, for $x \in D$ and $k \geq 1$.

Proof: Be the definition of the reward function with a break-even point x_0, we know that $\psi(x) < 0$ and $\psi'(x) \geq 0$ for $x < x_0$, and $\psi'(x_0) > 0$. This implies that

$$\Lambda\psi(x) = \gamma\psi(x) - x\psi'(x) < 0, \quad x \leq x_0.$$

Particularly, since $\Lambda\psi(x)$ is convex on (x_0, ∞) and increasing for large x, there is exactly one solution, x_1^*, to $\Lambda\psi(x_1^*) = 0$, and furthermore, $(d/dx)\Lambda\psi(x_1^*) > 0$ for $x \geq x_1$. Recall from Lemma 1 that

$$v^{(1)}(x) = \psi^{(1)}(x \vee x_1^*) \left[1 \wedge \left(\frac{x}{x_1^*}\right)^\gamma\right], \tag{A.18}$$

where $\psi^{(1)} \equiv \psi$. In addition,

$$\psi^{(2)}(x) = \psi^{(1)}(x) + e^{-rT}\mathbb{E}\left\{v^{(1)}(X_T^{0,x})\right\}. \tag{A.19}$$

Since $\Lambda\mathbb{E}\left\{g(X_t)\right\} = \mathbb{E}\left\{\Lambda g(X_t)\right\}$ for any integrable function g, we apply (A.19) to get

$$\begin{aligned} \Lambda\psi^{(2)}(x) &= \Lambda\psi^{(1)}(x) + e^{-rT}\mathbb{E}\left\{\Lambda v^{(1)}(X_T^{0,x})\right\} \\ &= \Lambda\psi^{(1)}(x) + e^{-rT}\mathbb{E}\left\{\Lambda\psi^{(1)}(X_T^{0,x})1_{\{X_t \geq x_1^*\}}\right\}, \end{aligned} \tag{A.20}$$

where in the second step we have used the fact that $\Lambda v^{(1)}(x)$ vanishes in the continuation region of $v^{(1)}(x)$, i.e. for $x < x_1^*$. Since $\Lambda\psi^{(1)}(x)$ is assumed convex on (x_0, ∞), the expectation $\mathbb{E}\left\{\Lambda\psi^{(1)}(X_t^{0,x})1_{\{X_t \geq x_1^*\}}\right\}$ is also convex on (x_0, ∞). Being the sum of two convex functions, $\Lambda\psi^{(2)}(x)$ is also convex on (x_0, ∞). Moreover, since $\Lambda\psi^{(1)}(x)1_{\{x \geq x_1^*\}}$ is an increasing function (and strictly positive for $x > x_1^*$), (A.20) implies that $\Lambda\psi^{(2)}(x)$ is increasing for large enough x.

It is observed from (A.18) and (A.19) that $\psi^{(2)}(x)$ is a continuously differentiable increasing function with $\lim_{x\to 0}\psi^{(2)}(x) = \lim_{x\to 0}\psi^{(1)}(x) < 0$. Furthermore, if $\psi^{(1)}(x)$ is bounded by $f(x) = ax$, for some $a > 0$, then one can show that $v^{(1)}(x) \leq ax$, and therefore

$$\psi^{(2)}(x) \leq ax + e^{-rT}\mathbb{E}\left\{aX_T^{0,x}\right\} \leq a\left(1 + e^{-(r-\alpha)T}\right)x. \tag{A.21}$$

This ensures the existence of a maximum at $x = x_2^*$ to the function $\psi^{(2)}(x)/x^\gamma$, and consequently also the existence of a solution to $\Lambda\psi^{(2)}(x) = 0$. Also, Proposition 1 implies that $x_2^* \geq x_0$ as it is never optimal to exercise with negative payoff. The convexity of $\Lambda\psi^{(2)}(x)$ on (x_0, ∞) ensures that there is exactly one such maximum, uniquely defined by

$$\Lambda\psi^{(2)}(x_2^*) = 0, \quad \text{and} \quad \frac{d}{dx}\Lambda\psi^{(2)}(x) > 0, \quad x \geq x_2^*.$$

Hence, by applying Lemma 1 again one obtains

$$v^{(2)}(x) = \psi^{(2)}(x \vee x_2^*)\left[1 \wedge \left(\frac{x}{x_2^*}\right)^\gamma\right].$$

Also, from the bound on $\psi^{(2)}(x)$ in (A.21) we infer that $v^{(2)}(x) \leq a\left(1 + e^{-(r-\alpha)T}\right)x$.

Repeating the argument above, one can similarly show the existence and uniqueness of a stopping boundary x_k^*, for every $k \geq 1$, and also derive an upper bound for each $v^{(k)}$:

$$v^{(k)}(x) \leq ax \left(\sum_{i=0}^{k-1} e^{-(r-\alpha)iT} \right). \tag{A.22}$$

Proving by induction that $\Lambda\psi^{(k+1)} > \Lambda\psi^{(k)}$, would together with the previous part of the proof imply that $x_{k+1}^* < x_k^*$. As previously remarked, $\Lambda\psi^{(1)}(x)1_{\{x \geq x_1^*\}}$ is positive for $x > x_1^*$. From (A.20) it follows that

$$\Lambda\psi^{(2)}(x) - \Lambda\psi^{(1)}(x) = e^{-rT}\mathbb{E}\left\{\Lambda\psi^{(1)}(X_T^{0,x})1_{\{X_t \geq x_1^*\}}\right\} > 0,$$

Now, assume that $\Lambda\psi^{(k)}(x) > \Lambda\psi^{(k-1)}(x)$. From the definition of $\psi^{(k)}(x)$ in (A.17) one has

$$\begin{aligned}
\Lambda\psi^{(k+1)}(x) - \Lambda\psi^{(k)}(x) &= e^{-rT}\mathbb{E}\left\{\Lambda v^{(k)}(X_T^{0,x}) - \Lambda v^{(k-1)}(X_T^{0,x})\right\} \\
&= e^{-rT}\mathbb{E}\left\{\Lambda\psi^{(k)}(X_T^{0,x})1_{\{x_k^* \leq X_t \leq x_{k-1}^*\}}\right\} \\
&\quad + e^{-rT}\mathbb{E}\left\{\left(\Lambda\psi^{(k)}(X_T^{0,x}) - \Lambda\psi^{(k-1)}(X_T^{0,x})\right)1_{\{x_{k-1}^* \leq X_t\}}\right\} \\
&> 0.
\end{aligned}$$

As for the monotonicity of the sequence $(v^{(k)})_{k \geq 1}$, we first show that $v^{(2)} > v^{(1)}$. Since $v^{(1)} > 0$ from Lemma 1, it follows from (A.17) that

$$\psi^{(2)}(x) - \psi^{(1)}(x) = e^{-rT}\mathbb{E}\left\{v^{(1)}(X_T^{0,x})\right\} > 0. \tag{A.23}$$

By the fact that $x_2^* < x_1^*$ and $\psi^{(2)}(x) > \psi^{(1)}(x)$, the inequality

$$v^{(2)} = \psi^{(2)}(x) > \psi^{(1)}(x) = v^{(1)}$$

holds for $x \geq x_1^*$. Moreover, since $\psi^{(2)}(x)/x^\gamma$ is maximized at $x = x_2^*$, it can be seen that

$$v^{(2)}(x) \;\; = \;\; \psi^{(2)}(x) \geq \psi^{(2)}(x_1^*)\left(\frac{x}{x_1^*}\right)^\gamma > \psi^{(1)}(x_1^*)\left(\frac{x}{x_1^*}\right)^\gamma = v^{(1)}(x), \quad x \in (x_2^*, x_1^*).$$

Similarly, for $x \leq x_2^*$,

$$v^{(2)}(x) = \psi^{(2)}(x_2^*)\left(\frac{x}{x_2^*}\right)^\gamma \geq \psi^{(2)}(x_1^*)\left(\frac{x}{x_1^*}\right)^\gamma > \psi^{(1)}(x_1^*)\left(\frac{x}{x_1^*}\right)^\gamma = v^{(1)}(x), \quad x < x_2^*. \quad \text{(A.24)}$$

Hence, it is proven that $v^{(2)} > v^{(1)}$. By induction, one obtains

$$\psi^{(k)}(x) - \psi^{(k-1)}(x) = e^{-rT}\mathbb{E}\left\{v^{(k-1)}(X_T^{0,x}) - v^{(k-2)}(X_T^{0,x})\right\} > 0.$$

With this inequality the steps from (A.23) to (A.24) can be followed to arrive at $v^{(k)}(x) > v^{(k-1)}(x)$. Finally, the inequality in (A.22) implies that the value function $v^{(k)}$ admits the following bound for any k:

$$v^{(k)}(x) \leq \frac{aM}{1 - e^{-(r-\alpha)T}}, \quad x < M. \qquad \square$$

It now remains to consider the behavior of the solution to the optimal multiple stopping problem in the limit of infinitely many exercise rights. That is, the following value function is to be investigated

$$v^{(\infty)}(x) = \sup_{\vec{\tau} \in \mathcal{S}^\infty} \mathbb{E}\left\{\sum_{n \geq 1} e^{-r\tau_n}\psi(X_{\tau_n}^{0,x})\right\}. \quad \text{(A.25)}$$

Since $(e^{-rt}aX_t)_{t \geq 0}$, where $ax > ax - b \geq \psi(x)$, is a supermartingale for all X_t and since the refracted stopping times $(\tau_n)_{(n \geq 1)}$ satisfy $\tau_n \geq (n-1)T$, it follows that

$$v^{(\infty)}(x) \leq \mathbb{E}\left\{\sum_{n=1}^{\infty} e^{-r\tau_n}aX_{\tau_n}^{0,x}\right\} \leq \sum_{i=0}^{\infty}\mathbb{E}\left\{e^{-riT}aX_{iT}^{0,x}\right\} = \frac{ax}{1 - e^{-(r-\alpha)T}}. \quad \text{(A.26)}$$

That is, $v^{(\infty)}(x)$ is bounded on every bounded subset of $\mathbb{R}^+$. Defined through an integral in (A.25), this bound ensures continuity of $v^{(\infty)}(x)$ on every compact subset of $\mathbb{R}^+$. Finally, the following convergence result is proved.

Proposition 2. *The sequence $v^{(k)}(x)$ converges uniformly to $v^{(\infty)}(x)$ on every compact subset of $\mathbb{R}^+$.*

Proof: Without loss of substantial generality, one can assume that $x \leq M$. Let $(\tau^*_{n,\infty})_{n\geq 1}$ is an optimal stopping rule for the value function $v^{(\infty)}(x)$ in (A.25). Employing a similar argument as in (A.26), with $f(x) = ax$ bounding $\psi(x)$, one obtains

$$\begin{aligned} v^{(\infty)}(x) &= \mathbb{E}\left\{\sum_{n=1}^{k} e^{-r\tau^*_{n,\infty}}\psi(X^{0,x}_{\tau^*_{n,\infty}}) + \sum_{n=k+1}^{\infty} e^{-r\tau^*_{n,\infty}}\psi(X^{0,x}_{\tau^*_{n,\infty}})\right\} \\ &\leq v^{(k)}(x) + \mathbb{E}\left\{\sum_{n=k+1}^{\infty} e^{-r\tau^*_{n,\infty}} f(X^{0,x}_{\tau^*_{n,\infty}})\right\} \\ &\leq v^{(k)}(x) + aM\frac{e^{-(r-\alpha)kT}}{1-e^{-(r-\alpha)T}}. \end{aligned} \tag{A.27}$$

The fact that $(\tau^*_{n,\infty})^k_{n=1}$ is an admissible, but not necessarily optimal, stopping rule for $v^{(k)}(x)$ is used in the second step. With $x \leq M$ it follows from Theorem 1 that $(v^{(k)})_{k\geq 1}$ is strictly increasing and bounded, and hence convergent on $[0, M]$. The uniform convergence of $v^{(k)} \to v^{(\infty)}$ on $x \in [0, M]$ then follows from (A.27). □

A.2 Numerical Implementation

The expressions for $u^{(k)}(x)$ and $v^{(k)}(x)$ in (5.15) and (5.16) form the basis of our numerical algorithm, whereby $u^{(k)}(x)$, $v^{(k)}(x)$, and x^*_k are computed iteratively for $k = 1, 2, 3, \ldots$ The calculations were carried out on a grid of 500 points regularly spaced between 0 and $x_{\max}$, where the latter was determined by the process and model parameters. The computation of the expectation in (5.17) is complicated by the fact that $u^{(k-1)}(x)$ is not bounded on $\mathbb{R}^+$. However, since the choice of reward function, $\psi(x)$ in (5.23), is approximately affine for large x, so are $\Lambda\psi(x)$ and $u^{(k-1)}(x)$. Rather than truncating the distribution we can instead linearly extrapolate $u^{(k-1)}(x)$ outside the given grid. Since $(x_k^*)_{k\geq 1}$ is decreasing, the necessary range of the grid depends mainly on x_1^*. Specifically, $x_{\max}$ was chosen to be the upper bound on a two-sided 99.9% confidence interval around x_1^*,

$$x_{\max} = \exp\left(\ln(x_1^*) + \left(\alpha - \frac{1}{2}\sigma^2\right)T + 3.29\sigma\sqrt{T}\right).$$

This choice of $x_{\max}$, together with the number of gridpoints, was an acceptable compromise between having a large enough grid to ensure a linear behavior of $u^{(k-1)}(x)$ without undue computational complexity. When comparing different scenarios, i.e. different values for T, ν and I, the largest grid was used throughout.

Using a trapezoidal method, the expectation in (5.17) was calculated as

$$\mathbb{E}\left\{u^{(k-1)}(X_T^{0,x})\right\} = \int_0^{x_{\max}} u^{(k-1)}(z)g(z;x,T,\alpha,\sigma)dz + \int_{x_{\max}}^{x'} (kz+m)g(z;x,T,\alpha,\sigma)dz, \tag{A.28}$$

where $kz+m$ is the linear extrapolation of $u^{(k-1)}$ for $x > x_{\max}$ and where $g(z;x,T,\alpha,\sigma)$ is the density function of the lognormal distribution. The upper limit x' in (A.28) was chosen so that the support of the distribution contained a two-sided confidence interval around x, for every $x \leq x_{\max}$.

The boundary points x_k^* were found using a simple bisection method on the convex function in (5.18). The calculations of $u^{(k)}(x)$, and therefore also of x_k^* and $v^{(k)}(x)$, were terminated once a tolerance ε, defined by

$$\varepsilon = \frac{|u^{(k)} - u^{(k-1)}|}{|u^{(k)}|}, \tag{A.29}$$

had reached $\varepsilon \leq 10^{-3}$. Such a tolerance yields a solution $v^{(k)}(x)$ that is within 0.1% of $v^{(\infty)}(x)$, an accuracy that is likely good enough considering reasonable errors in the estimation of the underlying parameters. In the results, x^*_∞ and $v^{(\infty)}(x)$ denote the stopping boundary and the value function respectively at the termination of the algorithm according to the tolerance in (A.29).

Figure A.1 demonstrates the convergence of the algorithm for the value function $v^{(k)}(x)$ and the stopping boundary x^*_k respectively. Default parameter values used in the calculations below are given in Table 5.1. According to Theorem 1, $\left(v^{(k)}(x)\right)_{k\geq 1}$ is strictly increasing and the sequence of stopping boundaries $(x^*_k)_{k\geq 1}$ is strictly decreasing. In Figure A.1 (left), we see that the value function increases monotonically with each iteration. After 50 iterations the tolerance ε, defined above, was less than 10^{-5}. Moreover, we notice that the value function appears linear for large x. In Figure A.1 (right), the stopping boundary x^*_k decreases rapidly from 0.85 to 0.44, where it should be mentioned that the break-even point was $x_0 = 0.33$. The convergence is clear even after 20 iterations. In particular, the stopping boundary $x^*_1 = 0.85$ from the first iteration helps define the upper bound $x_{\max}$ of the grid.

The stopping boundary x^*_k is the price level at or above which the first investment should be made, with in mind the option to make $k-1$ more investments later. All investments have to be separated in time by at least the lifetime T. In Figure A.1, the number of exercises (iterations) is $k = 50$, which together with the lifetime of $T = 5$ years, gives a time horizon of ≥ 250 years. Such a time horizon is practically infinite under reasonable circumstances. On the other hand, with a large k, the number of remaining investment opportunities should have a smaller impact on the first investment timing. This is evidenced by the convergence of x^*_k to a constant level as k increases.

From numerical tests, it was found that the refraction time (lifetime) T strongly influences the speed of convergence. This is intuitive due to the discount factor e^{-rT} in the definition of

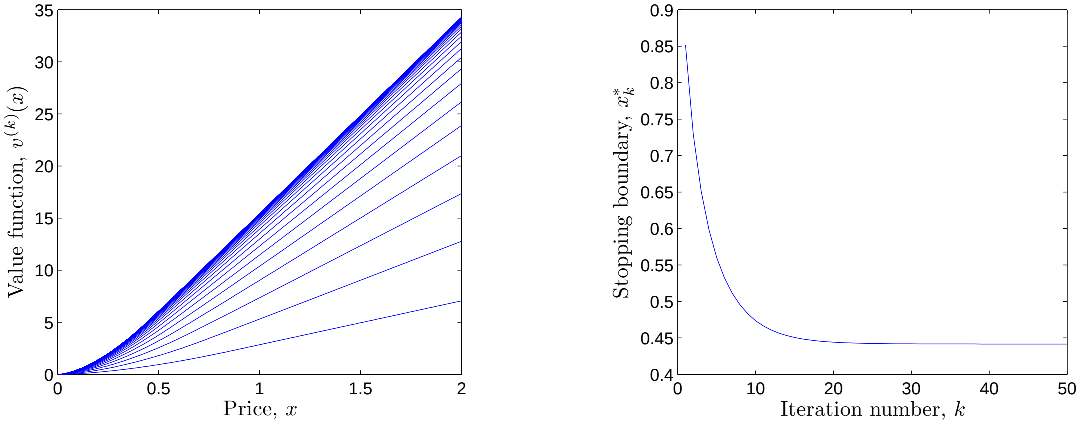

Figure A.1: (Left) Convergence of the value function, $v^{(k)}(x)$, for $k = 1, \ldots, 50$. (Right) The stopping boundary x^*_k decreases rapidly over iterations.

$u^{(k)}(x)$ in (5.17), reducing the differences over iterations. On the other hand, the lead time ν has much less bearing on the rate of convergence since it only affects the first investment timing.

Appendix B: Numerical Algorithm – RO Desalination

B.1 Laminar Channel

A flow field $\vec{w} = (u(x,y), v(x,y))$ is assumed in the laminar channel, where

$$u(x,y) = 6\overline{u}(x)\frac{y}{h}\left(1 - \frac{y}{h}\right) \ , \tag{B.1}$$

where $\overline{u}$ is the local channel average longitudinal velocity, and where

$$v(x,y) = \frac{d\overline{u}}{dx}\frac{h}{2}\left[1 - 6\left(\frac{y}{h}\right)^2 + 4\left(\frac{y}{h}\right)^3\right] \ . \tag{B.2}$$

The boundary condition on the vertical component v is given by

$$v(x,0) = \frac{h}{2}\frac{d\overline{u}}{dx} = v_p(x) \ , \tag{B.3}$$

where the permeation rate v_p is determined by the local pressure and concentration levels

$$v_p = A\left[\Delta p - \Delta\pi\right] = A\left[(p - p_0) - f_{os}(c_m - c_p)\right] \ . \tag{B.4}$$

For a laminar flow between two planes, the pressure loss along the channel is given by

$$\frac{dp}{dx} = -\frac{12\eta\overline{u}(x)}{h^2} \; . \tag{B.5}$$

Local concentration levels are found by solving the steady state continuity equation for salts in the feed channel:

$$0 = u\partial_x c + D\partial_{xx}c - v\partial_y c + D\partial_{yy}c \; , \tag{B.6}$$

subject to the following boundary conditions:

$$\begin{cases} c(x,y) = c_0, \quad x < 0, \\ \partial_y c(x,y)|_{y=\frac{h}{2}} = 0, \\ v_p(x)c(x,0) - D\partial_y c(x,y)|_{y=0} = v_p(x)c(x,0)(1-R) \; . \end{cases} \tag{B.7}$$

The first condition in (B.7) corresponds to the assumption of a uniform concentration profile of the feed before it enters the channel. The second condition follows from the symmetry of the problem across the entire height of the channel. A solute mass balance on the membrane surface, where longitudinal diffusion has been neglected, yields the third condition, where R is the rejection coefficient of the membrane. This last boundary condition also couples the steady state continuity equation to the permeation equations (B.3) and (B.4).

The equations (B.3) and (B.5) are discretized by

$$\overline{u}_i \;=\; \overline{u}_{i-1} - \frac{2v_{p,i-1}}{h}k_x, \quad v_{p,0} = 0, \;\; \overline{u}_0 = \overline{u}(0) \;\; i = 1, \ldots N_x \; , \tag{B.8}$$

$$p_i \;=\; p_{i-1} - \frac{12\eta\overline{u}_i}{h^2}k_x, \quad p_0 = p(0), \quad i = 1, \ldots N_x \; . \tag{B.9}$$

The continuity equation is discretized with a finite difference scheme

$$\begin{aligned} \partial_x c_{i,j} &= \frac{c_{i,j} - c_{i-1,j}}{k_x}, \quad \partial_{xx} c_{i,j} = \frac{c_{i-2,j} + 2c_{i-1,j} - c_{i,j}}{k_x^2} \ , && \text{(B.10)} \\ \partial_y c_{i,j} &= \frac{c_{i,j+1} - c_{i,j}}{k_y}, \quad \partial_{yy} c_{i,j} = \frac{c_{i,j-1} + 2c_{i,j} - c_{i,j+1}}{k_y^2} \ . && \text{(B.11)} \end{aligned}$$

The backward difference in the horizontal direction in (B.10) allows for the coupled equations (B.4) and (B.6) to be solved progressively along the channel given the initial applied pressure $p(0)$ and the feed flow rate, expressed via the average velocity $\overline{u}(0)$. The number of grid points N_x was chosen to be 500 (the same number of vertical grid points in the discretization of (B.6)). The numerical program was verified by performing an overall mass-balance of the streams entering and exiting the channel which gave an accuracy within 1%.

B.2 Turbulent Channel

The equation governing the flux through the membrane is the same as in the laminar channel (B.4). However, the pressure loss in the turbulent channel is modeled by

$$\frac{dp}{dx} = -\frac{f}{d_h}\frac{\rho u^2}{2} \ , \tag{B.12}$$

where the friction factor f and the hydraulic diameter d_h are estimated parameters. Also, as opposed to stipulating a flow field and solving the steady-state continuity equation for the salt in the channel the turbulent flow is assumed to exhibit a polarization layer of thickness δ according to

$$c(y) = \begin{cases} (c_b - (1-R)c_m)e^{(\delta - y)v_p/D} + (1-R)c_m, \quad 0 \leq y \leq \delta, \\ \\ c_b, \quad \delta \leq y \leq h/2, \end{cases} \tag{B.13}$$

where c_b and c_m are the concentrations in the bulk and at the membrane respectively. Evaluating (B.13) at the membrane wall gives the needed second relation between c_m and v_p, in addition to (B.4),

$$\frac{Rc_m}{c_b - (1-R)c_m} = e^{v_p/k} \ , \tag{B.14}$$

where the mass transfer coefficient k is also a numerically estimated parameter.

The longitudinal velocity and the pressure in the channel are calculated iteratively through

$$\begin{aligned} \overline{u}_i &= \overline{u}_{i-1} - \frac{2v_{p,i-1}}{h}k_x, \quad i = 1, \dots N_x & \text{(B.15)} \\ p_i &= p_{i-1} - \frac{f}{d_h}\frac{\rho u_i^2}{2}k_x, \quad i = 1, \dots N_x \ , & \text{(B.16)} \end{aligned}$$

where the subscript on f indicates a dependence on the local Reynolds number (feed velocity). Moreover, the superficial longitudinal velocity in the feed channel is discretized similarly as well

$$\overline{u}_i = \overline{u}_i - \frac{2v_{p,i-1}}{h}k_x, \quad i = 1, \dots N_x \ . \tag{B.17}$$

A solute mass balance,

$$\overline{c}_i = \left(2(1-R)v_{p,i-1}\frac{k_x}{h} + \overline{c}_{i-1}\overline{u}_{i-1}\right)/\overline{u}_i, \quad \overline{c}_0 = c_0, \quad i = 1, \dots, N_x,$$

allows for the calculation of the bulk concentration $c_{b,i}$, used in (B.14), from the average concentration

$$\overline{c}_i = \frac{1}{h/2}\int_0^{h/2} c_i(y)dy \ ,$$

with $c_i(y)$ from (B.13).

These iterative steps make possible the solution of the couple equations (B.4) and (B.14), giving $v_{p,i}$ and $c_{m,i}$. The same longitudinal grid was used as in the laminar case ($N_x =$

$500, k_x = L/N_x$.) The numerical program was verified with a solvent mass balance with an accuracy within 1%.