Resource pooling games

Resource Pooling Games

This thesis is number D173 of the thesis series of the Beta Research School for Operations Management and Logistics. The Beta Research School is a joint effort of the School of Industrial Engineering and the department of Mathematics and Computer Science at Eindhoven University of Technology, and the Center for Production, Logistics and Operations Management at the University of Twente.

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-3482-1

Printed by Proefschriftmaken.nl || Uitgeverij BOXPress

Cover design by Roy Lurken || BureauNobel

Resource Pooling Games

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op donderdag 28 november 2013 om 16.00 uur

door

Franciscus Jacobus Pierre Karsten

geboren te Heerlen

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. A.G.L. Romme
1e promotor:	prof.dr.ir. G.J.J.A.N. van Houtum
co-promotor:	dr. M. Slikker
leden:	prof.dr. P.E.M. Borm (Tilburg University)
	prof.dr. A.A. Scheller-Wolf (Carnegie Mellon University)
	prof.dr.ir. J.C. Fransoo
	prof.dr. J.S.H. van Leeuwaarden
adviseur:	dr.ir. R.J.I. Basten (Universiteit Twente)

—Being a graduate student is like becoming all of the Seven Dwarves. In the beginning you're Dopey and Bashful. In the middle, you are usually sick (Sneezy), tired (Sleepy), and irritable (Grumpy). But at the end, they call you Doc, and then you're Happy.

Ronald Azuma

Acknowledgements

This thesis is the result of a project that I have worked on for the last four years. Looking back, I can say that doing a PhD project is comparable to preparing for a Magic: the Gathering Pro Tour. Although the time frame is totally different, the process is similar: Exploring a new format, finding the best decks, enthusiastically tweaking them, feeling frustrated because they are all imperfect, but finally registering an acceptable one nevertheless—it's all very similar to the process of doing research and writing a PhD thesis. Or, to use a non-gaming analogy, a PhD project is similar to being locked inside a clock factory with some working clocks, lots of clock parts, and machines for building clocks, but with incomplete instructions (at best) on how to build new ones. It takes a bit of mental fortitude to overcome the challenges, but hearing the ticking sound of a newly constructed clock is a fulfilling experience.

Even though my PhD project expanded the boundaries of human knowledge by only a microscopic amount, I experienced a sense of wonder in the process of building a bit of new knowledge out of nothing. I am grateful for the opportunity to do fun, appealing research. Those last three words deserve some explanation.

It was fun to come up with new ideas, to sketch new formulations on the white-board, to complete a rigorous proof after many arduous explorations, to have coffee with colleagues and chat about interesting ideas, to give presentations, and to write research papers. I also had fun in coming up with unconventional examples to illustrate my ideas, such as the Black Lotus story on page 23 and the Ducktales story on page 138.

The appeal of my project was largely in bridging the mathematical fields of queueing theory and game theory. I could explain how these fields resonate with my affinity for randomness and interactive situations, but instead I will just mention that "queueing" is one of the few English words with five consecutive vowels, next to such words as miaouing (making a sound like a cat) and zoaeae (larvae of crabs and relatives). That's already more than enough appeal for now, and I'll leave it to the thesis to hopefully spark a genuine appreciation for its mathematical fields in the reader.

And finally, research. After four years of research, my perspective is that it is about creatively searching for truth and beauty via the tools of logic. Truth, beauty, and logic are all found in mathematical elegance, and that is what I have strived for in this thesis. Neat theorems for abstract models are collected in one cohesive document.

It would not have been possible to complete this document without the contributions of a number of people—people who helped, supported, and encouraged me over the course of my PhD trajectory. I consider myself lucky to have worked with a number of very pleasant coauthors and colleagues, and I am grateful for the support of my friends and family. I would like to use this opportunity to thank them all.

First and foremost, I would like to thank the person who deserves it the most: Marco Slikker, my co-promotor and daily advisor. I thank you for encouraging me to start this project, for your countless inspiring remarks and helpful ideas, for teaching me to be careful before saying "but that's obvious," and for always having time for me whenever I was stuck. I am grateful for the extensive and detailed comments on my papers and chapters, and for your guidance in writing everything down in a mathematically precise way. Your uncanny ability to find counterexamples and your mathematical insights are phenomenal, and I greatly enjoyed standing in front of the white-board to work out examples and proofs with you. Marco, I really appreciate the time and effort that you invested in me, and I've learned a lot.

Next, I wish to express my gratitude to my promotor Geert-Jan van Houtum. Your enthusiasm for the research and its practical relevance motivated me to keep going. You always managed to ask the right questions that helped me structure my thoughts, and I benefited from your academic experience, your friendly supervision, and high-level comments on my papers and chapters. You deserve to have a probability distribution named after you.

I would like to thank Rob Basten and Peter Borm for the joint research that, respectively, formed the basis for Chapters 4 and 8. Rob, it was a pleasure to work with you, and it was funny how we got the best ideas not when huddled over stacks of papers behind a desk, but on the day that we visited the Niagara Falls after a research conference. Peter, our discussions were always useful and inspiring, your numerous detailed comments on the work greatly improved it, and you managed to turn the research in the right direction. I am glad that Rob and Peter agreed to be part of my committee.

From January to April 2013, I visited Alan Scheller-Wolf and Mustafa Akan at Carnegie Mellon University. I thank them for their hospitality, and for our cooperation during this period that led to the research presented in Chapter 7. I appreciate our discussions on research, academia, and life, and I am glad that Alan agreed to be a part of my committee. I thank Chiel van Oosterom, with whom I shared an apartment in Pittsburgh, and the PhD students in the Tepper School for the enjoyable time that I had in Pittsburgh.

I would like to thank the other members of my committee, Johan van Leeuwaarden and Jan Fransoo, for the time they spent to read and evaluate the thesis and for providing valuable suggestions for improvement.

It has been a pleasure to work at the Eindhoven University of Technology, and to be part of the OPAC group. I thank the current and former colleagues in the group. I am especially grateful to Ben Vermeulen, Frank van den Heuvel, Joachim Arts, Maarten Driessen, Qiushi Zhu, and Sandra van Wijk for the enjoyable discussions and the pleasant atmosphere during our joint lunches, coffee breaks, wine hours, courses, occasional dinners, and conference visits. I thank Joachim in particular for his feedback on parts of this thesis. Furthermore, I am grateful to my successive office mates Dilay Çeleby, Rob Basten, Said Dabia, and Kristina Sharypova for their company and for putting up with my peculiar card shuffling habit.

I thank my friends for being interested in this project, but mostly for helping me take my mind off the project every now and then. Thanks in particular to Bas Melis, Jasper Blaas, Job Meertens, Menno Dolstra, Remco Smits, Roel van Heeswijk, Ron Cadier, Ruben Snijdewind, Stan van der Velden, and Victor van den Broek for the pleasant holiday trips, movie nights, game days, and other welcome distractions. I thank Roy Lurken for designing the cover of my thesis.

Finally, I thank my family for their love and support. Mom and dad, this thesis is in some way the end result of a lifetime of education, which I wouldn't have completed without you. Thanks for giving me the freedom to follow my own path in life.

Frank Karsten
Eindhoven, October 2013

Contents

1 Introduction **1**

1.1 Let's pool! . . . 1
1.2 Pooling of resources between independent players . . . 3
1.3 Research question . . . 5
1.4 Research methodology . . . 6
1.4.1 Quantitative modeling . . . 7
1.4.2 Queueing and inventory theory . . . 8
1.4.3 Cooperative game theory . . . 8
1.5 Illustration of the research approach . . . 9
1.6 Overview and contribution of the thesis . . . 12
1.6.1 The starter . . . 12
1.6.2 The main course . . . 13
1.6.3 The dessert . . . 14
1.6.4 Reading guide . . . 14

2 Preliminaries **17**

2.1 Introduction . . . 17
2.2 Terminology and notation . . . 17
2.2.1 Sets . . . 17
2.2.2 Functions . . . 19
2.3 Game theory . . . 21
2.3.1 Non-cooperative games . . . 21
2.3.2 Cooperative games . . . 22
2.3.3 Subadditivity and concavity . . . 22
2.3.4 The imputation set and the core . . . 23
2.3.5 Balancedness . . . 25
2.3.6 The Shapley value and the nucleolus . . . 27
2.3.7 Population monotonic allocation schemes . . . 28
2.3.8 Overview of implications . . . 29
2.3.9 We have to go stricter . . . 29

2.3.10 Single-attribute games . . . 33
2.4 Stochastic processes . . . 34
2.4.1 Probability theory and random variables . . . 34
2.4.2 Poisson distribution and exponential distribution . . . 35
2.4.3 Markov processes . . . 35
2.4.4 Queueing models . . . 37
2.4.5 Discussion of assumptions . . . 39

3 Literature review **41**
3.1 Introduction . . . 41
3.2 Cooperative inventory games . . . 42
3.2.1 EOQ games . . . 42
3.2.2 ELS games . . . 43
3.2.3 Newsvendor games . . . 44
3.2.4 Continuous-review games . . . 45
3.2.5 Spare parts games . . . 45
3.2.6 Contribution to the literature . . . 46
3.3 Cooperative queueing games . . . 46
3.3.1 $M/M/1$ games with fixed numbers of servers . . . 48
3.3.2 $M/M/1$ games with optimized numbers of servers . . . 49
3.3.3 Multi-server queueing games . . . 51
3.3.4 Contribution to the literature . . . 52
3.4 Pooling in queueing and inventory systems . . . 53

4 Erlang loss games **55**
4.1 Introduction . . . 55
4.2 Scalability of extensions of the Erlang loss function . . . 57
4.2.1 The classic Erlang loss function . . . 58
4.2.2 Extensions of the Erlang loss function . . . 61
4.2.3 Scalability of the linear interpolation . . . 63
4.3 Model description . . . 67
4.3.1 The basic setup . . . 67
4.3.2 Erlang loss situations . . . 72
4.4 Fixed numbers of servers . . . 74
4.4.1 Game . . . 74
4.4.2 Beneficial and adverse pooling effects . . . 75
4.4.3 Balancedness under symmetric penalty costs . . . 80
4.4.4 Cost allocation: stability and population monotonicity . . . 81
4.5 Optimized numbers of servers . . . 82

4.5.1 Games . . . 83
4.5.2 Relationship to the game with fixed numbers of servers . . . 85
4.5.3 The cost-based formulation . . . 88
4.5.4 The service-based formulation . . . 91
4.6 Optimal pooling, rationing, and transshipments . . . 93
4.7 Conclusion . . . 96

5 Erlang delay games **99**
5.1 Introduction . . . 99
5.2 The continuous extension of the Erlang delay function . . . 101
5.3 Model description . . . 106
5.4 Fixed numbers of servers . . . 108
5.4.1 Game . . . 108
5.4.2 Cost allocation: stability . . . 109
5.4.3 Cost allocation: population monotonicity . . . 111
5.5 Optimized numbers of servers . . . 113
5.5.1 Games . . . 114
5.5.2 Cost allocation: stability and population monotonicity . . . 116
5.5.3 Approximate stability and population monotonicity . . . 119
5.6 Conclusion . . . 121

6 Spare parts games with backordering **123**
6.1 Introduction . . . 123
6.2 Analysis of the $(S-1,S)$ model with backlogging . . . 125
6.2.1 The inventory model . . . 125
6.2.2 Behavior of the costs as a function of the demand rate . . . 127
6.3 Model description . . . 135
6.4 Fixed base stock levels . . . 136
6.4.1 Game . . . 136
6.4.2 A strict core allocation . . . 137
6.5 Optimized base stock levels . . . 142
6.5.1 Game . . . 142
6.5.2 A strictly population monotonic allocation scheme . . . 144
6.5.3 Who reaps the benefits? . . . 147
6.5.4 A truth-inducing implementation . . . 149
6.6 Conclusion . . . 151

7 Inventory pooling games with non-stationary, dynamic demands **153**
7.1 Introduction . . . 153
7.2 Model description . . . 156

7.3 Observation possibilities and games . . . 158
7.4 Balancedness . . . 161
7.5 Cost allocation . . . 165
7.6 Strategically joining late . . . 168
7.7 Conclusion . . . 174

8 Allocation rules on elastic single-attribute situations 175
8.1 Introduction . . . 175
8.2 An axiomatic characterization of the proportional rule . . . 177
8.3 An impossibility result . . . 181
8.4 The serial rule . . . 183
8.5 The benefit-proportional rule . . . 187
8.6 Concavicated marginal rules . . . 187
8.6.1 Concave single-attribute situations . . . 188
8.6.2 Concave functions under elastic functions . . . 192
8.6.3 Two concavicated rules . . . 197
8.7 Conclusion . . . 199

9 Conclusion 201
9.1 Game over? . . . 201
9.2 Description and illustration of the main results . . . 201
9.3 Lessons and surprises . . . 204
9.4 Game on! . . . 207

Bibliography 209

Summary 223

About the Author 227

—Humankind cannot gain anything without first giving something in return. To obtain, something of equal value must be lost. That is Alchemy's first law of equivalent exchange.

Alphonse Edric, Fullmetal Alchemist

1 Introduction

1.1 Let's pool!

Alchemists who are devoted to the philosophy of equivalent exchange believe that you can never get ahead without sacrificing something. These alchemists are wrong. Sure, we[1] cannot create gold out of thin air, but we can often gain something for free by organizing things more efficiently. One way to achieve this is via resource pooling: an arrangement in which a group of common resources (machines, inventories, servers, etc.) is shared between multiple customer streams. The idea is that when the customer arrival process is stochastic[2] and we don't know in advance which customer needs a given resource at a certain time, then resource flexibility can mitigate the uncertainty. It is well-known in the literature (e.g., Eppen, 1979; Smith and Whitt, 1981) that, compared to an arrangement in which each resource is dedicated to an individual customer stream, resource pooling arrangements typically result in reduced congestion and improved availability.

All of us regularly encounter pooled resources in daily life, whether we realize it or not. One example is when we place an order for tennis shoes at an internet retailer who fulfills the customer orders from all over Europe from a single, centralized warehouse. Here, resources (inventories) are pooled over multiple customer streams (countries). Another example is when we go to the post office and are served by an employee who can provide both mail and

[1] "We" refers to you, the reader, and me, the author. It is customary in the operations research literature to write in this "we" voice. However, any personal opinions will be written in the "I" voice.

[2] Stochastic is basically just another word for random or uncertain.

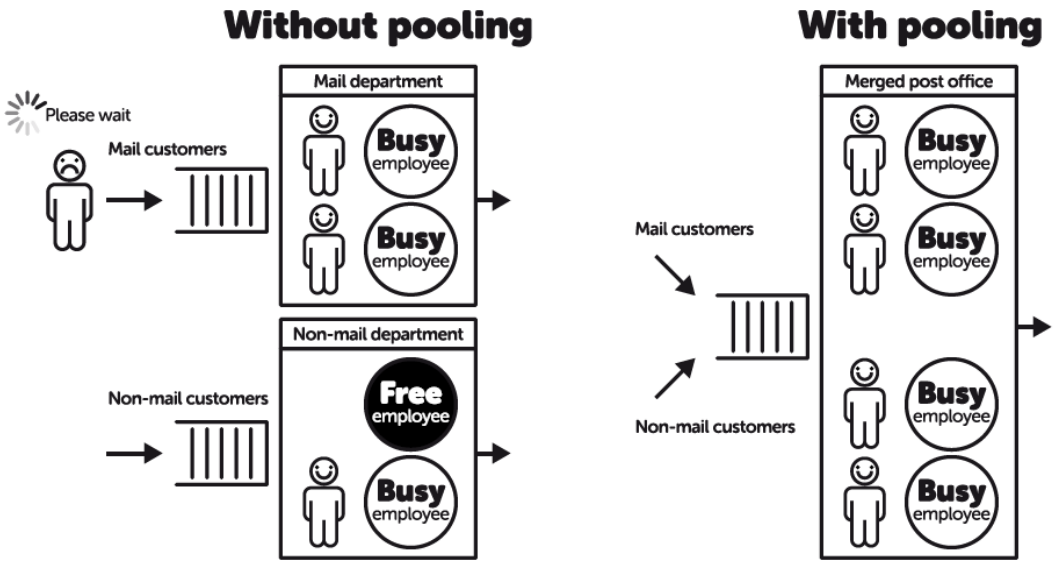

Figure 1.1: *In a post office without pooling (in which employees are dedicated to either mail or non-mail services) a customer who requires mail service has to wait in a queue when all mail servers are busy, even if a non-mail server is free. In a post office with pooling (in which employees are not dedicated to a specific service category) the same customer can immediately receive service.*

non-mail services. Here, resources (employees) are pooled over multiple customer streams (service categories). In both settings, pooling is beneficial. To understand why, suppose that the internet retailer would hold dedicated, separate inventories per country instead. Then, if demand in one country is higher than expected, even if there are extra units available in another country, there may still be a stock-out. In contrast, by pooling inventories over multiple countries, high demand in one country can be offset by lower demand in another country. The argument for the merging the mail and non-mail departments in a post office is analogous. Figures 1.1 and 1.2 provide an illustration.

These examples show that the efficiency benefits of resource pooling are already commonly exploited in case there is one service provider (e.g., an internet retailer or a post office) who owns all resources and who is responsible for serving all customer or demand streams. However, these benefits can also be obtained if there are multiple, self-interested players, each associated with a number of separate resources and customer streams. Pooling of the players' resources and customer streams can yield operational improvements, as described above, but the presence of multiple players—who care only about their own bottom line—does lead to all kinds of questions, such as who pays for the shared resources and how much. Such questions are not relevant when there is only one player, but they become essential when there are multiple players, and they form the main focus of this thesis.

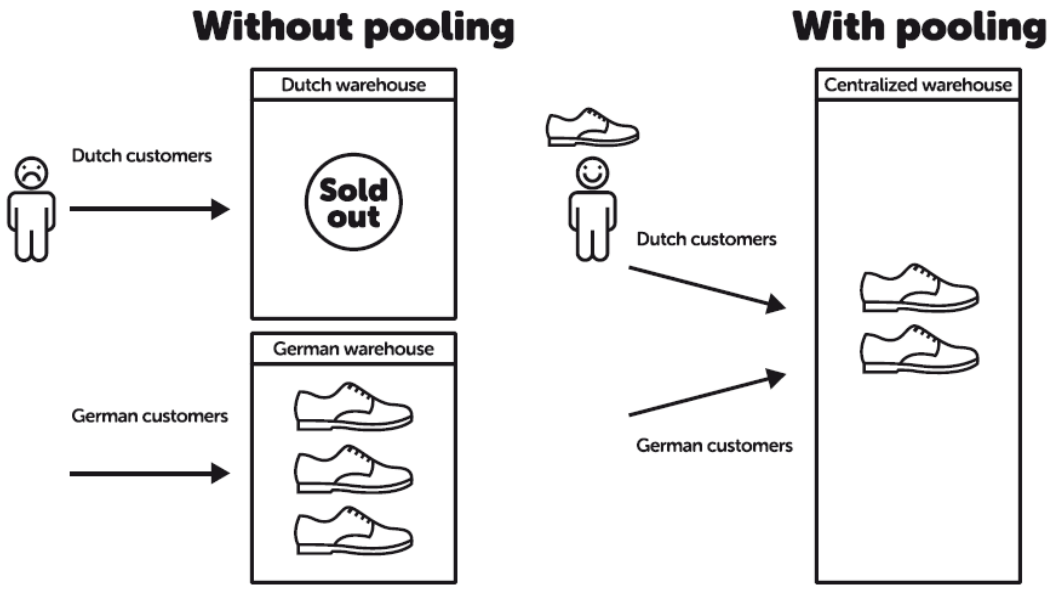

Figure 1.2: *For an internet retailer who does not pool inventories (i.e., with separate warehouses per country) a Dutch customer will not receive his desired tennis shoe if the Dutch warehouse is stocked out, even if the German warehouse is not. For an internet retailer who pools inventories (i.e., with one centralized warehouse) the same customer's demand can be immediately fulfilled.*

The remainder of this chapter is organized as follows. In Section 1.2, we start with examples of resource pooling between independent players. Subsequently, we get things rolling in Sections 1.3 and 1.4 with a more detailed description of our research questions and research methodology, respectively. We illustrate them via examples in Section 1.5. Finally, Section 1.6 presents an overview of the rest of this thesis.

1.2 Pooling of resources between independent players

Excellent examples of pooling between multiple decision makers, each with their own interests in mind, can be found in the capital goods industry. Capital goods such as power-generating plants, chemical production lines, naval vessels, MRI scanners, wind turbines, baggage handling systems, semiconductor fabs, construction equipment, and airplanes form the backbone of much of our society. To ensure operational continuity of these capital goods, maintenance activities are often needed. For the execution thereof, different types of costly resources are needed, such as spare parts and specialized service engineers. In the case of spare parts, it happens all too often that separate companies in the same geographical region keep their own parts on stock and do not share them with each other. Since the total demand for parts on a given day is driven by random component failures and therefore cannot be

predicted with certainty in advance, there is always the risk that demand is higher than expected. Hence, if companies do not pool their parts, even high safety stocks cannot always prevent the unfortunate situation in which one company has a shortage of inventory while another company has an excess of inventory. Such mismatches cannot occur if companies join forces and pool inventories. Moreover, pooling mitigates the risk of a shortage, which implies that the total safety stock can usually be reduced. Real-life cases of independent firms considering a spare parts pool are described in Guajardo et al. (2012) and Braglia and Frosolini (2013). Sharing (non-branded) service engineers makes sense for similar reasons.

The capital goods industry forms an important application because it embodies opportunities for massive monetary savings. The sale of spare parts and after-sales services in 2006 has been pegged at $1 trillion in the United States alone, representing 8% of its gross domestic product (Cohen et al., 2006, pp. 129-130, and references therein). At the same time, lacking a spare part or available service engineer can lead to downtime of capital goods, which is very expensive due to loss of operational continuity: an unresolved failure of a capital good can result in the standstill of a complete factory. For example, in the semiconductor industry, the opportunity costs for lost production are estimated to run into tens of thousands of euros per hour (Kranenburg, 2006, p. 17). Several studies (such as Kukreja et al., 2001, who looked at pooling possibilities between independently operating power-generating plants) showed that reductions in spare parts provisioning and maintenance costs in the order of 50% are easily possible when resources are pooled. Indeed, as stated by Cohen et al. (2006, p. 136): "the best way for companies to realize economies of scale is to pool spare parts."

While pooling of spare parts and service engineers is one of our main motivations for the research presented in this thesis, there is an abundance of resource pooling opportunities in other sectors as well. For instance, clinical departments in a hospital, each with their own patient populations, may share operating rooms, intensive care beds, medical staff, blood banks, diagnostic equipment, or medicine inventories. Here, the availability of the right resource at the right time can mean the difference between life and death. Less critical but still relevant, manufacturing facilities may share flexible production equipment between several job types, university faculties may share a high-performance computer cluster, and neighboring municipalities may pool stocks of snow salt (see van Outeren, 2010). Another example is that of pooling of mobility scooters between senior citizens living in a home for the elderly (see Pinedo, 2012). Other examples that come to mind are business units of a large insurance firm that share a common call center with cross-trained telephone agents, airline companies that pool check-in counters at an airport, and regional broadcasters that share internet capacity in case of a localized disaster (see ANP, 2011). Shared ownership of corporate jets also occurs in practice (see Keskinocak, 1999). Another example is found in the rental industry, where independent car rental agencies may collaborate by holding a common set of rental cars for all their customers. And as a final example, a group of banks may share tellers by directing customers at the branch locations to automated video screens that connect

to tellers who remotely authorize transactions and review checks from the headquarters (see Sidel, 2012).

We could go on to list many other examples, but hopefully, these examples already make the point that the world is teeming with resource pooling opportunities between independent players. These players could be business units of a single firm, or even competitors that collaborate on, e.g., back-end maintenance operations to improve their competitiveness against the rest of the world. In general, such collaborations enable more efficient use of resources and offer the opportunity to benefit from large (statistical) economies of scale: benefits aplenty! However, establishing a resource pool between independent, self-interested players is not easy—indeed, it raises all kinds of difficult questions, such as how to make sure that all parties benefit from collaboration. The following section deals with these questions.

1.3 Research question

When several players pool resources, the first issue to be addressed is what the best operating policy might be: How many resources would be needed in total if they are shared between multiple players? This question does not pose a big hurdle because it also comes up if all customer streams belong to just one service provider. Accordingly, the literature provides many models and solutions. By estimating probability distributions of uncertain quantities (such as the demand pattern for spare parts), quantifying the monetary consequences of our decisions (such as inventory holding and/or shortage costs), and subsequently applying a mathematical queueing or inventory model from the literature, we can find an adequate operating policy.

The next natural concern, which is more closely linked to our problem with separate players, is whether or not full pooling among all players is actually the arrangement with minimal cost. Would partial pooling or separate pools for smaller coalitions (subgroups of players) be more beneficial? If so, which coalitions will form? Furthermore, when multiple players are waiting for the same resource to become available, who should get it first? These questions are tricky when the customer streams are associated with different cost parameters or different service requirements, or in case transshipping a resource from one location to another comes with a substantial cost. This thesis will tackle some of these complexities to a certain extent. However, the majority of our work focuses on situations where all customers have identical costs parameters, identical service requirements, and negligible transshipment costs. Under these assumptions, the answer to the aforementioned questions is simple: full, complete pooling between all players on a first-come first-serve basis is optimal.

The more interesting question—indeed, the key question that this thesis is dedicated to—is the following:

> *When multiple service providers collaborate by sharing resources in a stochastic inventory or queueing system, is it possible to allocate the total costs of their pooled system amongst them in a stable way, and if so, can we identify a rule that results in stable cost allocations with appealing properties?*

Four clarifications on this research question are in order. First, "total costs of their pooled system" refers to the sum of the resource costs and shortage costs in the service facility set up by the grand coalition of all players together. Second, "a stable way" means a way ensuring that the cooperation doesn't break up into smaller coalitions in the long run. Third, "identify a rule ... with appealing properties" implies that we will be aiming to prove structural results, not trying to formulate an optimization program that picks the best cost allocation rule. Due to the plethora of appealing, yet conflicting, properties, there wouldn't be an unambiguous best. Fourth, we remark that although this question could also be studied from a profit allocation perspective, we will study this from the perspective of companies minimizing costs; from a stability perspective, this makes no difference.

The motivation for this research question is that a fair, transparent cost allocation mechanism is an essential prerequisite for a successful cooperation. Indeed, when independent decision makers join forces, they are not really interested in the benefits of pooling for society as a whole; they are mainly interested in the consequences for themselves. Without an equitable mechanism to allocate the total costs, or a guarantee that such a mechanism even exists, firms will not be convinced that participation in a resource pool is profitable. They may fear that some firms will end up paying to subsidize the others, and that new members may take more benefit out of the pool than they bring in. The construction of fair allocation mechanisms that do sustain support for the collaboration tends to be challenging—indeed, it is one of the more severe impediments for horizontal cooperation in the logistics industry (Cruijssen et al., 2007).

The existing literature, which we discuss in more detail later, does not provide satisfactory solutions to the cost allocation problems that arise in pooled stochastic inventory or queueing systems. This thesis takes a step towards a more comprehensive examination of the issue of resource pooling and cost allocation among independent parties, whether it arises in service, inventory, or manufacturing applications. We do this by proving novel, elegant theoretical results for mathematical models, as we describe next.

1.4 Research methodology

To answer the questions raised in the preceding section, we will follow the research methodology of quantitative modeling and draw upon the mathematical fields of queueing theory, inventory theory, and cooperative game theory. In this section, we briefly describe these approaches. Chapter 2 introduces the mathematical concepts in more detail.

1.4.1 Quantitative modeling

In this thesis, we formulate and analyze mathematical models. In the terminology of Bertrand and Fransoo (2002), we do quantitative, model-based, axiomatic research. This methodology uses equations, real numbers, sets, functions, probabilities, etcetera to provide a (simplified) abstraction of reality. The messy details of the real-world situation are purposefully disregarded to end up with a tractable model that allows us to study the aspects of interest in more detail.

A mathematical model for a real-world situation is similar to a geographical map. Indeed, in making a model, it is necessary to make many simplifying assumptions so as to render the mathematics tractable; consequently, we find ourselves in an imaginary world. To paraphrase Rubinstein (2012), models can be denounced for being simplistic and unrealistic, but they are still the best tool we have for clarifying concepts and acquiring insights. Indeed, no one would mistake a geographical map for physical reality, but it can still help us find our way. In similar fashion, a model is merely a simplified version of reality, but it can still help us grasp the world around us.

Analysis of a model can provide us with new knowledge in the form of mathematical statements or theorems. A theorem may state, for example, that a cost allocation rule always satisfies a certain fairness property in a specific model under certain assumptions. The truth or falsity of these theorems is demonstrated not by observations or experiments, as in the natural sciences, but by a rigorous mathematical proof that follows logically from the assumptions. The purpose of these theorems is to provide insights that will serve us when we return from the model to real life.

Models and theorems may sometimes be abstract. To make these abstractions more concrete, more vivid, and more specific, many illustrative examples will be presented. As humans tend to be good at inductive reasoning, I strongly believe in the value of simple examples; they allow us to grasp definitions, understand real-world applications, and prevent obfuscation. Obfuscation is not unlike a bewildered soliloquy on a non-isomorphic, thinly veiled non-theoremhood that results, considering several metaphysical and etymological ramifications, in an inability to understand as to what may be the original intention of a model or theorem, often inclusive of a subsemantic identity containing unnecessarily aggrandized words and a superfluous amount of clauses, deliberately constructed in a mythopoetic fashion to concoct confusion and verbal subterfuge for the express purpose of, indeed, escaping comprehension. I realize that, to a reader who is not well versed in the language of mathematics, an academic thesis can be just as puzzling as the previous sentence. Still, I hope to avoid obfuscation by explaining complex ideas via simple examples. I consider this to be an important part of quantitative modeling research.

1.4.2 Queueing and inventory theory

Queueing theory is the mathematical modeling and analysis of waiting lines. In queueing theory, a model is constructed wherein customers enter a service system according to a generally unpredictable arrival process, receive some kind of service, and subsequently leave the system. If all servers are occupied, customers either wait in a queue or leave without receiving service, depending on the model. The inherent uncertainty with respect to interarrival and service times is translated into a mathematical model via the use of probabilities and random variables. Using queueing models, steady-state performance measures, such as the expected time one has to wait for an available repairman, can be predicted. The foundations of queueing theory were laid by Erlang[3] about a century ago.

Inventory theory is the mathematical modeling and analysis of inventory systems. In inventory theory, a model is constructed wherein stock is depleted by demands, which occur according to a stochastic process, and stock is replenished by orders. The key decisions that inventory theory can help with are how much to order and when to order. Uncertainties with respect to demands are again captured via probabilities and random variables. Inventory models allow us to calculate performance measures such as the expected number of items on hand or the probability that we face a stock-out. The 1950s and 1960s were banner years for the development of stochastic inventory models.[4]

There is a tight link between queueing theory and inventory theory; the central analogy is between the number of parts in the replenishment pipeline in inventory models and the number of busy servers in queueing models. However, in inventory theory the cost functions are typically different, as inventory theorists are interested in, e.g., the average inventory, not the average queueing time.

1.4.3 Cooperative game theory

Cooperative game theory is concerned primarily with the mathematical modeling and analysis of situations where coalitions (groups of players) may collaborate to save costs. To represent a cooperative game, we must list all the possible coalitions that can be formed, together with how much each coalition would have to pay when cooperation would be limited to just that

[3]Agner Krarup Erlang (1878-1929) was a Danish mathematician and telephone engineer. While working for the Copenhagen Telephone Company, Erlang was presented with the problem of determining how many circuits were needed to provide an acceptable telephone service. To answer this problem, he developed his classic formulae for loss and waiting time. As told in Brockmeyer et al. (1948), an American researcher even learnt Danish to be able to read Erlang's papers in the original language. A unit of telephone traffic and a probability distribution have been named in his honor.

[4]The rigorous analysis of multi-period inventory models with stochastic demand began with the seminal paper of Arrow et al. (1951). Most of the basic analytical methods of the field, such as dynamic programming and stationary analysis, were established in the 1950s and 1960s (see, e.g., the memoir of Scarf, 2002).

coalition. The basis of cooperative game theory was laid by von Neumann[5] during World War II, and it offers a natural paradigm to tackle cost allocation problems. In the model of a cooperative game, enforceable binding agreements and side payments are allowed, while players are assumed to coordinate their actions and share all information.

An important aspect in cooperative game theory is how the total costs when all players collaborate should be split equitably over the players. Cooperative game theory provides several solutions that convert information about smaller coalitions into fair payoffs to each individual player. A main set of solutions from cooperative game theory that we will often focus on is the core. The core is the set of cost allocations that are stable in the sense that no smaller coalition is given an incentive to split off and, in our context, set up a separate resource pool. Under such a stable allocation, each player will feel motivated to collaborate with everyone, and no group of players is paying to subsidize the others. Such an allocation can be seen as a bond that keeps the players from defecting. Yet, even if collaboration leads to an overall cost reduction, it is not always the case that a stable allocation exists. Moreover, stability is not the only interesting notion of fairness. Various alternatives will be considered in this thesis.

1.5 Illustration of the research approach

This section presents two examples to make the research questions and research methodology more concrete. The first example deals with pooling of spare parts when emergency procedures and optimized base stock levels are used, and it illustrates the *multiplicity* of stable cost allocations in a two-player setting. The second example (inspired by Curiel, 1997) describes pooling of an exogenously given number of repairmen to reduce waiting times, and it illustrates the problem of the *existence* of a core allocation in a setting with a lot of players.

Example 1.1. Consider two independently managed airlines, A and B, that are located at negligible distance from each other. Both use spare parts to replace failed components. The number of failures observed for some component is subject to a lot of uncertainty: one year a given component might not fail at all, while the next year it may suddenly fail multiple times. Since obtaining a new part from the regular replenishment channel takes several months, having enough spare parts on hand to guard against the uncertain demands is essential. Indeed, when a spare part is unavailable when needed, an airplane goes out of service for a while, and an emergency procedure is needed to obtain the part from an alternative channel.

[5]John von Neumann (1903-1957) was a Hungarian-born American mathematician who made major contributions to a vast range of fields, including set theory, quantum mechanics, game theory, computer science, and even the Manhattan Project. He can be rightfully called the founder of game theory, both cooperative and non-cooperative. His tome "Theory of Games and Economic Behavior," co-written with Oskar Morgenstern and published in 1944, introduced the concepts of a cooperative game in characteristic function form. For more of a historical flavor, see Poundstone (1993).

This costs a lot of money. However, if one attempts to reduce stock-out risk with excessively large stocks, then these stocks will tie up a lot of capital. So, there is a trade-off between inventory holding costs and shortage costs. But no matter how this trade-off is managed, the sum of holding and shortage costs for a single player is likely to be substantial.

To reduce these costs, the two players consider cooperating by setting up a joint stock point for spare parts. Their inventory planners meet up, share information, discuss the options, and figure out how to jointly set optimal base stock levels.[6] Let us focus on one specific repairable part, for which player A expects $\lambda_A = 0.2$ demands per year and B expects $\lambda_B = 0.3$ demands per year. To analyze the stock point, the inventory planners use a mathematical inventory model: the $(S-1, S)$ infinite-horizon model with lost sales under the assumption of Poisson demand processes. This allows them to estimate the expected yearly costs for both inventory holding and shortages under any base stock level. Subsequently, an optimal, cost-minimizing base stock level can be identified.

Let us suppose that it would be optimal for player A, when acting alone and in absence of pooling, to stock 1 part with expected total yearly costs of \$13K. For player B, when acting solitarily, it would be optimal to stock 1 part with expected total yearly costs of \$15K. (These numbers encompass costs for both inventory holding and shortages.) If the two players cooperate while maintaining 2 parts in total, then their collective costs would go down to \$22K as a result of pooling, but they can do even better by re-optimizing the base stock level: when they share only 1 part, minimal costs of \$17K in total are attained.

At first, player A has cold feet about this type of collaboration. "What if I need a part," player A might wonder, "when player B took it the day before?" However, after mulling it over, player A realizes that a single-minded focus on this worst-case scenario is not productive. The expected total yearly cost is what really matters, and it is reduced significantly by sharing 1 part.

But there is still an element of competition here: the two players are fighting to pay as little as possible. Many solutions are possible. They could agree on side payments such that the \$17K are split equally, proportional to their individual demand rates, or proportional to their individual costs. They could also focus on the benefits, \$13K+\$15K−\$17K=\$11K, rather than the costs, which can lead to a different allocation rule. And how about letting player A pay half the costs of a fictive pool with total demand rate $2\lambda_A$, and B the rest?

All these allocations may very well make both players better off, but it leaves us with the problem of picking the fairest or most reasonable one, whatever that even means. This problem can be cracked with the axiomatic approach of cooperative game theory. This approach involves first formalizing the vague, intuitive notion of fairness by putting down a set of natural properties that an allocation should satisfy, and then either deriving a solution

[6] The base stock level is the initial number of parts on hand. The terminology stems from the $(S-1, S)$ inventory model, in which a new part is immediately ordered upon any demand; consequently, the number of parts on hand plus the number of parts on order is kept at a fixed level: the base stock level.

(preferably a unique one) that satisfies these properties, or showing that these properties are incompatible.

Choosing a certain way to split the \$17K is not the end-all, as the players may quickly face additional challenges. First off, how to implement an allocation in expected terms for any demand realization, especially when the realized costs in any period of time may differ greatly from expected costs? Furthermore, what if a player believes that it is not in his best interest to disclose his demand rate truthfully and decides to lie about it? Finally, what if a third player shows up and joins the pool; will that make the existing players better off, and how does that depend on the cost allocation mechanism? All these questions may, of course, influence the choice for an allocation rule in the first place.

With so many questions, the player may very well keep on fighting over the cost allocation and never even come to an agreement, in which case the spare parts pool would not materialize. We need a fair, implementable cost allocation that all parties can trust. This thesis will provide that. ◇

Example 1.2. Consider a manufacturer of advanced office equipment, such as 3D printers and computer mainframes, who offers service contracts to customers that purchase a new product. Each of the equipment categories is managed by a separate department. Each department employs a number of repairmen and has its own demand rate for repair requests. The departments have separate, limited budgets and can decide for themselves how to run their service operations. Due to the random duration of a repair and the uncertainty in the number of repair requests that are placed every day, customers sometimes have to wait for a repairman to arrive.

To reduce waiting times, five departments (A, B, C, D, and E) consider cooperating by pooling all their repairmen. Repairmen are trained to repair all types of products. The departments do not want to hire or fire anyone, so the total number of repairmen will stay fixed at the prior staffing levels. Using a mathematical queueing model—specifically, the $M/M/s$ model—the expected waiting time of an arbitrary repair request can be estimated in the pooled service system for any coalition that might form. These waiting times can be translated into monetary terms via contractual penalties stipulated in the service contracts.

After a careful analysis of the situation, the five departments conclude that resource pooling would lead to expected yearly costs of \$20K for the five of them together. Of course, this amount has to be divided among them. At first, assigning \$4K to each department seems like a reasonable allocation. However, each department contributes a different demand rate and a different number of repairmen to the pool. After some additional analysis, D and E figure out that if they would work together without the other three, they would face yearly expected costs of \$7K, which is less than the \$8K that they would have to pay according to the allocation proposed above. An ad hoc solution might be to allocate \$2.5K to each of D and E and \$5K to each of A, B, and C. This settles the original problem, but only results in a new one: A, B, and D quickly figure out that setting up up a pool for only the three of them

results in costs \$11K, which is less than the the \$12.5K that they would have to pay under the new allocation. Stability is threatened once again.

This could go on for a while. In the meantime, the departments are losing precious pooling opportunities because no one wants to commit without knowing first how the costs will be divided. They need a more systematic way to analyze the situation.

Formulating the problem as a cooperative game does exactly that. Cooperative game theory can explain the departments' problem as the search for a core allocation. Recall that a core allocation guarantees that no coalition can do better by themselves: it not only keeps the individual players happy, but every possible coalition happy, too. However, the core might also be an empty set. If that would be the case, then no matter how hard you try, you can never find a core allocation.

So, the big question is: is there a certain *structure* in the game of our five departments, or more generally any game that results from pooling of $M/M/s$ queues, that somehow guarantees the existence of a core allocation? (And if so, how to determine such an allocation?) This thesis will investigate such fundamental existence questions for several classes of resource pooling games. ◇

1.6 Overview and contribution of the thesis

The setup of this thesis is like a 3-course restaurant meal. Chapters 2 and 3 are the starter; Chapters 4, 5, and 6 are the main course; and Chapters 7 and 8 are the dessert. We receive the bill in Chapter 9, which concludes with a summary of the main results and insights.

1.6.1 The starter

Chapters 2 and 3 lay the foundation for the rest of the dinner.

Chapter 2 describes mathematical preliminaries to make this thesis self-contained. This will equip us with all the mathematical concepts and formalisms required to formulate and analyze a variety of cooperative games that arise from resource pooling arrangements in inventory and queueing systems.

Chapter 3 provides an extensive review of the related literature. After all, we are not the first to study resource pooling between independent players from the perspective of cooperative game theory. However, as we will describe in Chapter 3, the existing game theoretical literature has mostly focused on simple newsvendor, lot sizing, and $M/M/1$ models. These works provide useful insights, but they are based on limiting assumptions such as a single period, deterministic demand, and/or a single server. Due to these limitations, they cannot be used directly in our motivating resource pooling settings. Chapter 3 also positions the work in this thesis in the existing literature: this thesis will extend the game theoretical analysis to more general models that describe the reality of our motivating applications

	Queueing problem	Inventory problem
No waiting allowed	Chapter 4	Chapter 4
Waiting possible	Chapter 5	Chapter 6

Table 1.1: *Classification of Chapters 4, 5, and 6 based on modeling assumptions.*

more accurately. Specifically, we will study inventory models with both *multiple* periods and *stochastic* demands, as well as queueing models with *multiple* servers in parallel.

1.6.2 The main course

The main content of our work is in Chapters 4, 5, and 6. In these chapters, we consider similar, yet disparate models because different situations require different modeling assumptions. When considering a resource pooling situation, we could be dealing with an inventory problem, as in Example 1.1, or with a queueing problem, as in Example 1.2. Likewise, we could have a situation in which waiting for an available resource is not possible, as in Example 1.1, or a situation in which waiting in a queue is allowed, as in Example 1.2. Every combination of choices leads to a different model, and conclusions that are drawn for one model need not remain valid when we change our assumptions. So, to cover a variety of practically relevant resource pooling arrangements, we will consider different models in different chapters.

In a queueing context, we will study games derived from pooling of Erlang loss systems (i.e., $M/G/s/s$ queues, in which waiting is not allowed) and of Erlang delay systems (i.e., $M/M/s$ queues, in which waiting in a queue is possible). In an inventory context, we will study resource pooling games in a model that is realistic for expensive, low-demand spare parts: the $(S-1,S)$ infinite-horizon model with Poisson demand processes, both under lost sales (waiting for a part is not allowed) and backlogging (waiting for a part is possible). Table 1.1 provides a chapter overview. For each of those queueing and inventory settings, we study two variants: one where the total number of resources for any coalition is optimized, as in Example 1.1, and one where each player brings a fixed number of resources to every coalition, as in Example 1.2. The typical approach for each model variant is to capture the relevant aspects of the problem in a cooperative game, subsequently provide sufficient conditions for the games to possess a core allocation, and finally identify a specific stable cost allocation, if possible.

Chapter 4 fills a dual role in that it analyzes Erlang loss games that simultaneously capture pooling of $M/G/s/s$ queues and pooling in the $(S-1,S)$ model with lost sales. The most important theorem in that chapter is that Erlang loss games always admit a stable allocation if players have symmetric shortage costs. We also derive new analytical properties of extensions of the classic Erlang loss function.

Chapter 5 deals with Erlang delay games in which $M/M/s$ queues join forces. The most

surprising result of that chapter is that Erlang delay games with optimized numbers of servers need not admit a stable allocation, even if players have symmetric shortage costs. We also derive new analytical properties of an extension of the classic Erlang delay function.

Chapter 6 treats games derived from the $(S-1,S)$ model with backlogging. The most important theorem in that chapter is that the resulting games always admit a stable allocation if players have symmetric shortage costs. We also derive new analytical properties of the cost function in the $(S-1,S)$ inventory model with backlogging.

1.6.3 The dessert

Chapters 7 and 8, like a cake, could be devoured as a separate, quick snack. Nevertheless, they taste best after the entire main course.

In Chapter 7, we relax an assumption that underlies the $(S-1,S)$ infinite-horizon model with Poisson demand processes. In that model, the stochastic demand process is stationary and independent over time, and the focus is on steady-state behavior. In Chapter 7, we analyze the resource pooling games that arise from a finite-horizon, periodic inventory model in which the demand processes are not only stochastic but also non-stationary and dynamic due to, e.g., seasonal effects or correlations over time. In the presence of such correlations, information on a past demand observation can matter. The key theorem for this model is that a stable cost allocation is still guaranteed to exist.

In Chapter 8, we study the appeal of several cost sharing mechanisms, including simple proportional ones, with regard to various fairness criteria. The analysis in that chapter does not take place in a specific queueing or inventory model but in a more general class of so-called elastic single-attribute games. This class includes, amongst others, the inventory games with optimized numbers of servers from Chapters 4 and 6 in which each player has one attribute only: his demand rate. We define six fairness properties and five allocation rules. After showing the impossibility of combining two of those fairness properties, we show for each of the five rules which of the six properties it satisfies.

1.6.4 Reading guide

This thesis can be read in several ways, and it is written with different audiences in mind. Students, operations managers, practitioners, my mom, and other interested readers can browse through it, looking especially at the game theory tutorial in Section 2.3, the model formulation in Section 4.3.1, the examples in each chapter, and Chapter 9. In this way, one can get a fairly good idea of what this thesis is about. Queueing or inventory researchers who are mainly interested in new contributions to the literature on queues or inventories may be particularly interested in Sections 4.2, 5.2, and 6.2, where new analytical properties of (extensions of) the classic Erlang loss and Erlang delay functions and of the cost function in the $(S-1,S)$ inventory model with backlogging are derived. Game theorists who are mainly

interested in new contributions to the literature on cooperative games might enjoy the new game theoretical results and allocation rules described in Chapters 2 and 8. These sections and chapters have been set up such that they can be read separately from the rest of the thesis. Finally, one can do a thorough reading of the thesis, which is of course the path that is wholeheartedly recommended to anyone doing research on the interface of cooperative games and stochastic operations management.

—If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.

John von Neumann

2 Preliminaries

2.1 Introduction

In this chapter, we start off with a brief review of the basic terminology and notation that will be used throughout this thesis. Afterwards, we define a number of elementary concepts of game theory and stochastic processes. Several examples will be used to illustrate the concepts. Any result for which a formal proof is worked out is a new contribution to the literature.

2.2 Terminology and notation

The beauty of mathematics is that rather complex concepts are built up from basic notions such as sets and functions. In this section, we briefly review these and related notions. Although the reader is probably familiar with many of the concepts that will be treated, this section serves to fix notations and definitions, to collect them in a single place, and to aid the reader who needs a quick refresher on the terminology of sets and functions. For more detail, we refer the reader to, e.g., Kosmala (2004).

2.2.1 Sets

The notion of a set is fundamental to mathematics. A *set* is a collection of distinct, well-defined objects. These objects can be anything: numbers, names of people, other sets, and so on. If a is an element of the set A, this is denoted as $a \in A$. The *cardinality* of a set A,

written $|A|$, is the number of elements of the set. The empty set $\emptyset$ is the unique set with zero cardinality.

There are several fundamental operations for constructing new sets from given sets. The *union* of two sets A and B, denoted $A \cup B$, is the set that contains all elements that belong to either A or B. The *intersection* of two sets A and B, written $A \cap B$, is the set that contains all elements of A that also belong to B. The *subtraction* of a set A from B, written $B \setminus A$, is the set of all elements that are in B but not in A. Finally, the *Cartesian product* of two sets A and B, denoted $A \times B$, is the set of all possible ordered pairs whose first component is a member of A and whose second component is a member of B.

Sets may contain a finite or an infinite number of elements. In line with the game theoretical orientation of this thesis, we call any finite, nonempty set a *player set*. Some of the infinite sets that we will encounter in this thesis are:

- $\mathbb{N} = \{1, 2, \ldots\}$, the set of natural numbers;
- $\mathbb{N}_0 = \mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$, the set of non-negative integers;
- $\mathbb{R} = (-\infty, +\infty)$, the set of real numbers;
- $\mathbb{R}_+ = [0, \infty)$, the set of non-negative real numbers;
- $\mathbb{R}_{++} = \mathbb{R}_+ \setminus \{0\} = (0, \infty)$, the set of positive real numbers;
- $\mathbb{Q}$, the set of rational numbers, i.e., numbers that can be expressed as a fraction of two integers.

A set A is said to be a *subset* of B, written $A \subseteq B$, if every element of A is also contained in B. Note that if $A = B$ then $A \subseteq B$ and $B \subseteq A$. If A is a *proper subset* of B, i.e., if $A \subseteq B$ and $A \neq B$, then we write $A \subset B$.

The *power set* of a finite set N, written 2^N, is the set of all subsets of N (including N itself and the empty set). Two variants are denoted by $2^N_- = 2^N \setminus \{\emptyset\} = \{M \mid M \subseteq N, M \neq \emptyset\}$ and $2^N_{--} = 2^N \setminus \{\emptyset, N\} = \{M \mid M \subseteq N, M \neq N, M \neq \emptyset\}$.

For $n \in \mathbb{N}$, we denote by $\mathbb{R}^n$ the set of real vectors of length n, where the coordinates correspond to $1, 2, \ldots, n$. So, an element $x \in \mathbb{R}^n$ can be represented by $(x_1, x_2, \ldots, x_n)$. For a player set N, we denote by $\mathbb{R}^N$ the set of real vectors of length $|N|$, where the coordinates correspond to the elements of N. So, an element $x \in \mathbb{R}^N$ can be represented by $(x_i)_{i \in N}$.

Given any finite set N, a set $A \subseteq \mathbb{R}^N$ is said to be *convex* if, for all $x, y \in A$ and all $t \in [0, 1]$, the point $(1 - t)x + ty$ is contained in A, i.e., if every point on the line segment connecting x and y is an element of A.

Given any finite set N, we call a set P of non-empty sets a *partition of* N if it is a collection of non-empty proper subsets of N such that every element of N is in exactly one of these subsets (i.e., if $\bigcup_{A \in P} A = N$) and if each two distinct elements of P are disjoint, i.e., if $A \cap B = \emptyset$ for all $A, B \in P$.

2.2.2 Functions

The notion of a function is also fundamental to mathematics. A function describes a relation (specifically, a set of ordered pairs) between a set of inputs, called the *domain*, and a set of permissible outputs, called the *codomain*, such that each element in the domain is paired to exactly one element in the codomain. We denote a function f with domain D and codomain C by $f : D \to C$. The set of elements in the codomain that are paired with at least one element in the domain is called the *image* of a function. In the literature, domains of functions are often not specified explicitly. This makes sense when the domain is understood from context, but in this thesis we will often focus on domain restrictions. Many of the functional relations that we will study turn out to exhibit different properties under different domains, so we will need to be more explicit with our domain specifications.

A function can have various interesting properties. First off, a function $f : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}$ is called *non-decreasing* (resp. *increasing*) if, for all $x, y \in D$ with $x < y$, it holds that $f(x) \leq f(y)$ (resp. $f(x) < f(y)$). The properties *non-increasing* and *decreasing* are defined analogously.

A real-valued function $f : A \to \mathbb{R}$ defined on a convex set $A \subseteq \mathbb{R}^N$ is called *convex* if, for any two points $x_1, x_2 \in A$ and any $t \in [0, 1]$,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2),$$

and is called *strictly convex* if, for any two points $x_1, x_2 \in A$ with $x_1 \neq x_2$ and any $t \in (0, 1)$,

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2).$$

So, the graph of a (strictly) convex function lies (strictly) below any line segment joining two points from its graph. The function f is called (*strictly*) *concave* if $-f$ is (strictly) convex.

We will often be interested in optimization. For a function $f : D \to \mathbb{R}$, the *minimum* $\min_{x \in D} f(x)$ is the smallest element in the image of f, if such an element exists, and the *set of minima* $\operatorname{argmin}_{x \in D} f(x)$ is comprised of all elements $x \in D$ for which $f(x)$ equals the minimum value. The *maximum* $\max_{x \in D} f(x)$ and the *set of maxima* $\operatorname{argmax}_{x \in D} f(x)$ are defined analogously. Note that $\operatorname{argmin}_{x \in A} f(x)$ and $\operatorname{argmax}_{x \in A} f(x)$ are sets. A minimum and a maximum are guaranteed to exist if D is finite and non-empty. Alternatively, a minimum and a maximum exists if f is continuous while D is a closed, bounded, non-empty set[7].

An important property signifying economies of scale is subhomogeneity of degree zero. A function $f : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}^N$ is called *subhomogeneous of degree zero* if $f(x) \geq f(tx)$ for all $x \in D$ and all $t \geq 1$ with $tx \in D$.[8] Subhomogeneity of degree zero of f says that scaling up all arguments with the same relative amount t leads to a value that is no greater than the initial value.

[7] Any interval containing its boundaries, i.e., of the form $[a, b]$, is closed and bounded.

[8] More generally, subhomogeneity of degree a is based on the inequality $f(x) \geq t^a f(tx)$. Of course, $t^0 = 1$ for all $t \geq 1$. This illustrates where the "zero" in subhomogeneity of degree zero comes from.

Example 2.1. Consider the function $f : D \to \mathbb{R}$ with $D = \mathbb{N} \times \mathbb{R}_{++}$ defined by $f(s,a) = 1/(sa)$ for all $(s,a) \in D$. This function is subhomogeneous of degree zero because for any $(s,a) \in D$ and any $t \geq 1$ with $(ts,ta) \in D$ (i.e., with $ts \in \mathbb{N}$) it holds that $f(s,a) = 1/(sa) \geq 1/(t^2sa) = f(ts,ta)$. ◇

Another property related to economies of scale is elasticity. A function $f : \mathbb{R}_+ \to \mathbb{R}$ is called *elastic* if $f(0) = 0$ and $f(x_1)/x_1 \geq f(x_2)/x_2$ for all $x_1, x_2 \in \mathbb{R}_{++}$ with $x_1 \leq x_2$. If $f(x)$ expresses the cost of, say, serving demand level x, then elasticity of f says that the per-demand cost is non-increasing in the total demand served.

Example 2.2. Consider the function $f : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$f(x) = \begin{cases} 0 & \text{if } x = 0; \\ 5 & \text{if } x \in (0,5]; \\ x & \text{if } x > 5. \end{cases}$$

This function is elastic because $f(0) = 0$ and because $f(x)/x$ is non-increasing for $x > 0$. Specifically, $f(x)/x$ is decreasing with minimal value 1 for x on $(0,5]$, while $f(x)/x = 1$ for all $x > 5$. ◇

The following theorem characterizes an elastic function as a function f for which any straight line segment drawn through a point $(a, f(a))$ on its graph and the origin lies completely below or on the graph of f. This has an obvious link to concavity. We omit the obvious proof.

Theorem 2.1. *Consider any function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(0) = 0$.*
(i) The function f is elastic if and only if $f(x) \geq f(a)x/a$ for all $a \in \mathbb{R}_{++}$ and all $x \in (0,a]$.
(ii) If f is concave, then f is elastic.

The converse of Part *(ii)* of this theorem is not true. In fact, many of the elastic functions that will be considered in this thesis are *not* concave. One example is illustrated in Figure 4.3 on page 90. All elastic functions that we consider in this thesis are, however, non-decreasing. This implies continuity, except possibly at 0, as shown in the following theorem.

Theorem 2.2. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing, elastic function. Then f is continuous on $\mathbb{R}_{++}$.*

Proof. Let $a \in \mathbb{R}_{++}$ and let $\epsilon > 0$. If $f(a) > 0$, then fix $\delta = \min\{a, \epsilon a/[2f(a)]\}$; otherwise, if $f(a) = 0$, then fix $\delta = a$. Let $x \in (a - \delta, a + \delta)$. If $x < a$, then fix $b = a - \delta$; otherwise, if $x \geq a$, then fix $b = a + \delta$. We obtain

$$|f(a) - f(x)| \leq |f(a) - f(b)| \leq \left| f(a) - f(a)\frac{b}{a} \right| = f(a)\frac{\delta}{a} \leq \epsilon/2 < \epsilon,$$

where the first inequality holds because f is non-decreasing and the second inequality holds by Part (i) of Theorem 2.1. We conclude that f is continuous at a. □

2.3 Game theory

Sets and functions, while abstract, can help to formalize situations of conflict or cooperation into mathematical models. Those kinds of models are studied in *game theory*. The field of game theory can roughly be divided into non-cooperative game theory and cooperative game theory.

Non-cooperative game theory primarily deals with the analysis of conflict situations. In a non-cooperative game, several rational decision makers (the players) independently have to choose a strategy that eventually results in a payoff for every player. This payoff depends not only on the player's own strategy but also on the other players' strategies, and the key question is which strategies the players will choose. Non-cooperative games in strategic form, in which every player makes a single decision simultaneously, are defined in Section 2.3.1.

In non-cooperative game theory, binding agreements between the players are not possible. If binding agreements and side payments are allowed, and if coordination of players' actions and sharing of information is possible, then it is more appropriate to study an interactive situation from the perspective of *cooperative* game theory.

In a cooperative game, the players still have their own best interest in mind, but they typically benefit from collaborating with others, i.e., the players become better off when they join forces. A cooperative game captures the synergy and cost sharing aspects in a way that conveniently enables us to work out how much every player should pay. We will solely focus on cooperative games with transferable utility, which means that players can measure their utilities in a common currency and can make monetary transfer payments to each other. As this thesis is crafted around cooperative games, we dedicate Sections 2.3.2 through 2.3.10 to the introduction of important concepts from cooperative game theory. A comprehensive reference on cooperative game theory is the book by Peleg and Sudhölter (2007).

2.3.1 Non-cooperative games

A *non-cooperative cost game in strategic form* is a triple $(N, (S_i)_{i\in N}, (f_i)_{i\in N})$, where

- N is a player set;
- S_i is a nonempty set containing the strategies for player $i \in N$;
- $f_i : \prod_{i\in N} S_i \to \mathbb{R}$ is a function that determines the costs inflicted on player $i \in N$.

Note that the cost functions f_i for player $i \in N$ depends not only on player i's own strategy but also on the other players' strategies. Now, a *strategy profile* is any vector $s \in \prod_{i\in N} S_i$ that contains an element from the strategy set S_i for each player $i \in N$. A strategy profile $s^* = (s_i^*)_{i\in N}$ is called a *Nash equilibrium* (cf. Nash, 1951) if no player has anything to gain by changing his own strategy unilaterally, i.e., if for all $i \in N$ it holds that $f_i(s^*) \leq f_i(\sigma)$ for every strategy profile σ with $\sigma_i \in S_i$ for player i and $\sigma_j = s_j^*$ for all other players $j \in N \setminus \{i\}$.

2.3.2 Cooperative games

A cooperative cost game with transferable utility, which we will simply refer to as *game*, is a pair (N, c) where

- N is a player set;
- $c : 2^N \to \mathbb{R}$ is a characteristic cost function with $c(\emptyset) = 0$.

A non-empty subset $M \in 2^N_-$ is called a *coalition*, the set N of all players is called the *grand coalition*, and any coalition other than the grand coalition is called a *subcoalition*. Note that the empty set is *not* a coalition.

The function c assigns to every subset $M \subseteq N$ the value $c(M)$, which is interpreted as the total costs of the joint cooperative effort if only the players in M are involved in it. In other words, $c(M)$ is determined solely by what M can accomplish on its own, without the participation of the other players. Because $c(\emptyset) = 0$ by definition, we will systematically refrain from listing $c(\emptyset) = 0$ when defining a characteristic cost function.

The costs of any coalition M can be measured in monetary terms and are freely transferable between the players of M. In particular, $c(N)$ represents the total costs for the grand coalition when all players agree to work together.

2.3.3 Subadditivity and concavity

Games can satisfy all kinds of interesting properties. A game (N, c) is called *subadditive* if for any two coalitions $M, L \in 2^N_-$ with $M \cap L = \emptyset$ it holds that $c(M) + c(L) \geq c(M \cup L)$. In a subadditive game, it is always beneficial to combine coalitions, and cooperation by the grand coalition is socially optimal.

Another interesting property that a game might satisfy is concavity. A game (N, c) is called *concave* (cf. Shapley, 1971) if any player's marginal cost contribution is smaller if he joins a larger subset, i.e., if for each $i \in N$ and for all $M, L \subseteq N \setminus \{i\}$ with $M \subseteq L$ it holds that $c(M \cup \{i\}) - c(M) \geq c(L \cup \{i\}) - c(L)$. The concavity property of set functions is sometimes referred to as submodularity.

Example 2.3. Consider three service providers (named 1, 2, and 3) who wish to set up a service system. If they set up separate systems, then player 1 has to pay 10 euros, player 2 has to pay 20 euros, and player 3 has to pay 30 euros to adequately satisfy all their demands. If the individuals decide to work together and set up a joint system, then (due to the peculiarities of the synergies between the player's demands) a saving of 1 euro can be obtained if the system is set up to serve player 1 and exactly one other player, and a saving of 10 euros can be obtained if the system is set up to serve at least players 2 and 3. This situation may be

modeled by a game (N, c) with player set $N = \{1, 2, 3\}$ and characteristic cost function

$$c(M) = \begin{cases} 10 & \text{if } M = \{1\}; \\ 20 & \text{if } M = \{2\}; \\ 30 & \text{if } M = \{3\}; \\ 29 & \text{if } M = \{1, 2\}; \\ 39 & \text{if } M = \{1, 3\}; \\ 40 & \text{if } M = \{2, 3\}; \\ 50 & \text{if } M = N. \end{cases}$$

This game is subadditive because combining coalitions never increases their collective costs. This game is not concave because $c(\{1, 2\}) - c(\{2\}) = 9 < 10 = c(\{1, 2, 3\}) - c(\{2, 3\})$. ◇

Example 2.4. Consider four wizards (Jon, Kai, Gabriel, and Olivier) who each wish to cast the powerful spell Lightning Helix. Olivier vengefully even wants to cast not just one, but two copies of this spell. To cast any Lightning Helix, a certain amount of magical fuel is required: one red mana and one white mana. The wizards can obtain this mana by buying Black Lotuses at the bargain price of 5 euros apiece; each Black Lotus generates either three red mana or three white mana. The wizards can collaborate by buying several Black Lotuses jointly and pooling the generated mana.

For example, if Jon and Olivier collaborate, then they can obtain the required 3 red mana and 3 white mana for the 3 spells they wish to cast together from two copies of Black Lotus, at a total cost of 10 euros. If Kai or Gabriel would join, then two additional copies of Black Lotus would be needed. This situation may be modeled by a game (N, c) with player set $N = \{J, K, G, O\}$ and characteristic cost function

$$c(M) = \begin{cases} 20 & \text{if } |M| \geq 3 \text{ and } O \in M; \\ 10 & \text{otherwise.} \end{cases}$$

This game is subadditive because combining coalitions never increases costs. This game is not concave because $c(\{J, K\}) - c(\{J\}) = 0 < 10 = c(\{J, K, O\}) - c(\{J, O\})$. ◇

2.3.4 The imputation set and the core

A central problem in any game (N, c) is to allocate $c(N)$ to the individual players in a fair way. Formally, an *allocation* for a game (N, c) is a vector $x = (x_i)_{i \in N} \in \mathbb{R}^N$ satisfying $\sum_{i \in N} x_i = c(N)$. The requirement $\sum_{i \in N} x_i = c(N)$ is often called *efficiency*. The value x_i should be interpreted as the costs assigned to player i. An allocation x for a game (N, c) is called *individually rational* if $x_i \leq c(\{i\})$ for all players $i \in N$. Individual rationality means that every player is allocated no more costs than what he would face by staying alone. The set of all individually rational allocations of a game (N, c) is called the *imputation set*, denoted as $\mathcal{I}(N, c)$.

An allocation x for a game (N, c) is called *stable* if $\sum_{i \in M} x_i \leq c(M)$ for all $M \in 2^N_-$. Stability extends individual rationality to all coalitions. Under a stable allocation, each group of players has to pay no more collectively than what they would face by acting independently as a group. Hence, if the costs of the grand coalition are assigned according to a stable allocation, then no subcoalition has an incentive to split off and establish cooperation on its own. The set of all stable allocations is called the *core*, proposed[9] by Gillies (1959). So, the core $\mathcal{C}(N, c)$ of a game (N, c) is

$$\mathcal{C}(N, c) = \{x \in \mathbb{R}^N \mid \textstyle\sum_{i \in N} x_i = c(N) \text{ and } \sum_{i \in M} x_i \leq c(M) \text{ for all } M \in 2^N_-\}.$$

Note that the core is a convex set. A stable allocation is also sometimes called a *core allocation*; the two terms bear the same meaning. The following two examples illustrate individual rationality and stability. The second example additionally shows that a game may have an empty core, even if it is subadditive.

Example 2.5. Reconsider the game (N, c) described in Example 2.3. The vector $\hat{x} = (9, 20, 21)$ is an individually rational allocation, but it is not stable because $\hat{x}_2 + \hat{x}_3 = 41 > 40 = c(\{2, 3\})$. The allocation $\bar{x} = (10, 19, 21)$, however, is both individually rational and stable.

To describe the core of this game, note that for any $x \in \mathcal{C}(N, c)$, we must have $x_1 = 10$ because $c(\{1\}) = 10$, $c(\{2, 3\}) = 40$, and $c(N) = 50$. Combining this observation with the values of $c(\{1, 2\})$ and $c(\{1, 3\})$, we obtain that the core of this game is described as the set of all convex combinations of the points $(10, 11, 29)$ and $(10, 19, 21)$, i.e., $\mathcal{C}(N, c) = \{t(10, 11, 29) + (1 - t)(10, 19, 21) \mid 0 \leq t \leq 1\}$. ◇

Example 2.6. Reconsider the game (N, c) described in Example 2.4. The imputation set is given by $\mathcal{I}(N, c) = \{x \in \mathbb{R}^N \mid \sum_{i \in N} x_i = 20, x_i \leq 10 \text{ for all } i \in N\}$. For example, the allocation $x \in \mathbb{R}^N$ that assigns $x_i = 5$ to each player $i \in N$ is individually rational. But this allocation x is not stable because $x_J + x_K + x_G = 15 > 10 = c(\{J, K, G\})$. In fact, there is no stable allocation. To see this, assume to the contrary that $\hat{x}$ is an element of $\mathcal{C}(N, c)$. Then, the stability conditions $\hat{x}_J + \hat{x}_K + \hat{x}_G \leq 10$ and $\hat{x}_O \leq 10$ together with efficiency imply that $\hat{x}_O = 10$. At the same time, these conditions imply that $\hat{x}_i$ must be positive for at least one $i \in \{J, K, G\}$; suppose without loss of generality that Jon is the unlucky one, i.e., $\hat{x}_J > 0$. Then, $\hat{x}_O + \hat{x}_J > 10 = c(\{O, J\})$, which means that it would be beneficial for Jon and Olivier together to split off and only buy the required Black Lotuses for the two of them. Hence, our assumption that $\hat{x}$ is a core allocation must be wrong. We conclude that, even though (N, c) is subadditive, $\mathcal{C}(N, c) = \emptyset$. So, any proposed allocation is unstable. ◇

[9] The idea of the core already appeared in Von Neumann and Morgenstern (1944), for instance in footnote 3 on page 41, but they rejected it as a solution concept because it may be empty. The core is attributed to Gillies because he did not let the possible emptiness get in his way: he revived and analyzed the concept in his PhD thesis (Gillies, 1953) before publishing his work in 1959.

Although picking an allocation from the core may, at first, have seemed like an unshakably good idea, the preceding examples emphasize two of the major challenging problems with this. First, as shown in Example 2.6, the core may be empty. In Section 2.3.5, we provide a set of necessary and sufficient conditions for a game to possess a core allocation. Second, as shown in Example 2.5, the core may contain (infinitely) many allocations, which raises the important question of which allocation to choose. In Section 2.3.6, we present several rules that prescribe an allocation to any game.

2.3.5 Balancedness

Though in a small example it is not difficult to determine whether a game has an empty core or not, we are often interested in the derivation of general results for classes of games. To prove such results, it can be helpful to use the notion of a balanced game, introduced by Bondareva (1963) and Shapley (1967). Before describing balancedness of a game, we first need to define balanced maps and balanced collections for any player set N. A map $\kappa : 2^N_- \to [0,1]$, which assigns a number from the unit interval to any coalition, is called *balanced for* N if $\sum_{M \in 2^N_-: i \in M} \kappa(M) = 1$ for all $i \in N$. Such a map may be interpreted as describing an artificial arrangement wherein each player $i \in N$ divides 1 unit (of, say, time/effort) over the coalitions that include i, with $\kappa(M)$ the fraction dedicated to coalition M. For each balanced map κ for N, we define $\mathbb{B}(\kappa) = \{M \in 2^N_- \mid \kappa(M) > 0\}$ to be the set of coalitions with positive coefficients.

Example 2.7. Consider player set $N = \{1,2,3\}$ and map $\kappa : 2^N_- \to [0,1]$ defined by

$$\kappa(M) = \begin{cases} 1/2 & \text{if } M \in \{\{1,2\},\{1,3\},\{2,3\}\}; \\ 0 & \text{otherwise.} \end{cases}$$

This map[10] is balanced because for player 1 we have $\sum_{M \in 2^N_-: 1 \in M} \kappa(M) = \kappa(\{1,2\}) + \kappa(\{1,3\} = 1$ with similar equations for players 2 and 3. Accordingly, $\mathbb{B}(\kappa) = \{\{1,2\}, \{1,3\}, \{2,3\}\}$. Observe that balanced maps may be fractional-valued. ◇

Example 2.8. Consider player set $N = \{J,K,G,O\}$ and map $\kappa : 2^N_- \to [0,1]$ defined by

$$\kappa(M) = \begin{cases} 2/3 & \text{if } M = \{J,K,G\}; \\ 1/3 & \text{if } M \in \{\{J,O\},\{K,O\},\{G,O\}\}; \\ 0 & \text{otherwise.} \end{cases}$$

This map is balanced because for player J we have $\sum_{M \in 2^N_-: J \in M} \kappa(M) = \kappa(\{J,K,G\}) + \kappa(\{J,O\}) = 1$ with similar equations for the other players. Accordingly, $\mathbb{B}(\kappa) = \{\{J,K,G\}, \{J,O\}, \{K,O\}, \{G,O\}\}$. ◇

[10] The arrangement described by this balanced map is artificial because in a real partitioning of the players in which players work the same hours every day, player 3 should be in coalition $\{3\}$ whenever players 1 and 2 are in coalition $\{1,2\}$.

Any collection $\mathbb{B} \subseteq 2^N_-$ of coalitions is called *balanced for* N if there is a balanced map κ for N such that $\mathbb{B} = \mathbb{B}(\kappa)$. Every partition of N is balanced for N. A balanced collection $\mathbb{B}$ is called *minimally balanced for* N if $\mathbb{B} \neq \{N\}$ and there does not exist another balanced collection for N that is a proper subset of $\mathbb{B}$. Correspondingly, a balanced map κ is called *minimally balanced for* N if $\mathbb{B}(\kappa)$ is a minimally balanced collection for N. Each minimally balanced collection is associated with a unique minimally balanced map (Shapley, 1967). We let $\mathcal{W}^N$ denote the set of minimally balanced maps for N.

It follows immediately from the definition of a balanced map that every $\kappa \in \mathcal{W}^N$ satisfies $\sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \cdot \sum_{i \in M} f(i) = \sum_{i \in N} f(i)$ for all functions $f : N \to \mathbb{R}$. This property will prove pivotal in the analysis of resource pooling games later on.

Example 2.9. Consider player set $N = \{1, 2, 3\}$. The collection $\mathbb{B} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ is balanced for N because its elements are the coalitions with positive coefficients in the balanced map κ of Example 2.7, i.e., $\mathbb{B}(\kappa) = \mathbb{B}$. The collection $\mathbb{B}$ is minimally balanced because $\mathbb{B} \neq \{N\}$ and every proper subset of $\mathbb{B}$ is not balanced on N. The (only) other minimally balanced collections on N are the partitions of N.

If we now consider the function $f : N \to \mathbb{R}$ defined by $f(i) = 2i$ for all $i \in N$, then we obtain that $\sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \cdot \sum_{i \in M} f(i) = \frac{1}{2}(2+4) + \frac{1}{2}(2+6) + \frac{1}{2}(4+10) = 12 = \sum_{i \in N} f(i)$. $\diamond$

Example 2.10. Consider player set $N = \{J, K, G, O\}$ and collection $\mathbb{B}' = \{\{J, K, G\}, \{J, K\}, \{J, O\}, \{K, O\}, \{G, O\}\}$. It is balanced for N because the map $\kappa' : 2^N_- \to [0, 1]$, described by

$$\kappa'(M) = \begin{cases} 1/3 & \text{if } M = \{J, K, G\}; \\ 1/2 & \text{if } M = \{J, K\}; \\ 1/6 & \text{if } M \in \{\{J, O\}, \{K, O\}\}; \\ 2/3 & \text{if } M = \{G, O\}; \\ 0 & \text{otherwise.} \end{cases}$$

is a balanced map satisfying $\mathbb{B}'(\kappa') = \mathbb{B}'$. However, the collection $\mathbb{B}'$ is not minimally balanced for N because the collection $\mathbb{B}(\kappa)$ described in Example 2.8 is balanced for N whilst being a proper subset of $\mathbb{B}'$. $\diamond$

We are now ready to define balancedness of a game. A game (N, c) is called *balanced* if for every minimally balanced map $\kappa \in \mathcal{W}^N$ it holds that $\sum_{M \in 2^N_-} \kappa(M) c(M) \geq c(N)$. The following theorem was derived independently[11] by Bondareva (1963) and Shapley (1967).

Theorem 2.3. *Let (N, c) be a game. The core of (N, c) is non-empty if and only if (N, c) is balanced.*

[11] The seminal contribution of Bondareva was written in Russian and appeared in a rather obscure source. It therefore remained hidden for the game theorists in the West (Gilles, 2010), while Shapley independently found the same fundamental existence theorem, which has since been known as the Bondareva-Shapley Theorem. Bondareva's work was discovered before Shapley's work was published, however. Indeed, Shapley (1967) refers to Bondareva (1963) and refines her results by eliminating redundant balanced collections.

A game (N, c) is called *totally balanced* (cf. Shapley and Shubik, 1969) if, for each coalition $L \in 2^N_-$, the *subgame* (L, c^L), defined by $c^L(M) = c(M)$ for all $M \subseteq L$, is balanced. It is well-known that any concave game is totally balanced and that any totally balanced game is subadditive.

Example 2.11. Reconsider the game (N, c) described in Example 2.4 and the minimally balanced map κ described in Example 2.8. Observe that $\sum_{M \in 2^N_-} \kappa(M)c(M) = 16\frac{2}{3}$. Since $16\frac{2}{3} < 20 = c(N)$, the notion of balancedness provides an alternative way to prove that the game of Example 2.4 has an empty core. However, the subgame $(\{K, G, J\}, c^{\{K,G,J\}})$ is easily seen to be balanced. ◇

2.3.6 The Shapley value and the nucleolus

As mentioned earlier, a central question in cooperative game theory is how to allocate the costs of the grand coalition to the individual players. Cooperative game theory offers various solutions or allocation rules that provide an answer to this question. An *allocation rule* is a function f that assigns to any game (N, c) a vector $f(N, c) \in \mathbb{R}^N$ satisfying $\sum_{i \in N} f_i(N, c) = c(N)$. Two well-known allocation rules are the *Shapley value* (introduced in Shapley, 1953) and the *nucleolus* (introduced in Schmeidler, 1969).

To describe the Shapley value, we first need to define orderings and marginal allocations. An *ordering* on player set N is a bijection[12] $\sigma : N \to \{1, \ldots, n\}$, which should be interpreted as saying that player i is in position $\sigma(i)$. We let $\Pi(N)$ denote the set of all orderings on N. For an ordering $\sigma \in \Pi(N)$, we let $\sigma^{-1}(j)$ denote the player in position $j \in \{1, \ldots, |N|\}$ and we let $P^\sigma_i = \{j \in N | \sigma(j) < \sigma(i)\}$ describe the set of players that precede i.

The marginal contribution of player i according to ordering σ in a game (N, c) is denoted by $m^\sigma_i(N, c) = c(P^\sigma_i \cup \{i\}) - c(P^\sigma_i)$, i.e., the cost difference when player i joins his predecessors. The vector $m^\sigma(N, c) = (m^\sigma_i(N, c))_{i \in N}$ is called the *marginal allocation* according to σ. The Shapley value $\Phi(N, c)$ of a game (N, c) is then defined as the average of all marginal allocations, i.e.,

$$\Phi_i(N, c) = \frac{1}{|N|!} \sum_{\sigma \in \Pi(N)} m^\sigma_i(N, c) \qquad \text{for all } i \in N.$$

An alternative description is that

$$\Phi_i(N, c) = \sum_{M \subseteq N \setminus \{i\}} \frac{(|M|)!(|N| - |M| - 1)!}{|N|!} \cdot [c(M \cup \{i\}) - c(M)] \qquad \text{for all } i \in N.$$

As shown in Shapley (1971), concavity of a game (N, c) implies that $m^\sigma(N, c) \in \mathcal{C}(N, c)$ for any $\sigma \in \Pi(N)$. Because $\mathcal{C}(N, c)$ is a convex set, $\Phi(N, c) \in \mathcal{C}(N, c)$ as well if (N, c) is concave.

[12] A function $f : D \to C$ is called a *bijection* if every element of D is paired with exactly one element of C, and vice versa.

In the absence of concavity, however, $\Phi(N,c)$ may fall outside the core, even if the core is non-empty.

The nucleolus is defined for games with a non-empty imputation set only. To define the nucleolus, we first need to introduce satisfactions and the lexicographic ordering. Consider a game (N,c).[13] The *satisfaction* of a coalition $M \subseteq N$ under an allocation x is given by $c(M) - \sum_{i\in M} x_i$. Given an allocation x, we let $\theta(x) \in \mathbb{R}^{2^{|N|}-1}$ represent the vector that takes the corresponding satisfactions of all $2^{|N|}-1$ different coalitions and arranges them in non-decreasing order, so that $\theta_a(x) \leq \theta_b(x)$ for all $a,b \in \{1,\ldots,2^{|N|}-1\}$ with $a \leq b$. Given two allocations x and $\hat{x}$, we say that $\theta(x)$ is *lexicographically larger* than $\theta(\hat{x})$ if there exists an index $b \in \{1,\ldots,2^{|N|}-1\}$ such that $\theta_a(x) = \theta_a(\hat{x})$ for all $a \in \{1,\ldots,b-1\}$ and such that $\theta_b(x) > \theta_b(\hat{x})$.

The *nucleolus* $\nu(N,c)$ of a game (N,c) with $\mathcal{I}(N,c) \neq \emptyset$ is then defined as the (unique) allocation in $\mathcal{I}(N,c)$ whose θ is lexicographically maximal. In words, the nucleolus considers how satisfied each coalition is with any proposed allocation and then selects the allocation that maximizes the minimal satisfaction. The nucleolus always picks an allocation in the core whenever it is nonempty. Moreover, it satisfies appealing fairness properties (see Snijders, 1995).

Example 2.12. Reconsider the game (N,c) described in Example 2.3. The Shapley value $\Phi(N,c)$ is $(9\frac{2}{3}, 15\frac{1}{6}, 25\frac{1}{6})$ and the nucleolus $\nu(N,c)$ is $(10,15,25)$. The nucleolus is stable, but the Shapley value is not. $\diamond$

2.3.7 Population monotonic allocation schemes

Population monotonicity (cf. Sprumont, 1990) refines the concept of stability. Under a population monotonic allocation scheme, adding extra players to an existing coalition does not make anyone worse off. An *allocation scheme* for a game (N,c) is a vector $y = (y_{i,M})_{i\in M, M\in 2^N_-}$ with $\sum_{i\in M} y_{i,M} = c(M)$ for any $M \in 2^N_-$, which specifies how to allocate the costs of every coalition to its members. This scheme is called a *population monotonic allocation scheme* (PMAS) if the amount that a player has to pay does not increase when the coalition to which he belongs grows. That is, $y_{i,M} \geq y_{i,L}$ for all coalitions $M, L \in 2^N_-$ with $M \subset L$ and $i \in M$. If a game (N,c) admits a PMAS, say y, then it is totally balanced and $(y_{i,N})_{i\in N}$ is an element of its core. Not every totally balanced game admits a PMAS, though. Norde and Reijnierse (2002) provide a set of necessary and sufficient conditions for a game to admit a PMAS. One class of games that always admits a PMAS is the class of concave games.

Example 2.13. Reconsider the game (N,c) that was introduced in Example 2.3. For convenience, it is described in Table 2.1. The vector y as described in Table 2.1 is an allocation scheme because $\sum_{i\in M} y_{i,M} = c(M)$ for all coalitions $M \in 2^N_-$. To see that y is population

[13] Although Schmeidler (1969) defined the nucleolus for games in profit rather than cost terms, it carries over straightforwardly by simply reversing all inequalities involved. We tacitly did the same for the core.

Coalition M	$c(M)$	$y_{1,M}$	$y_{2,M}$	$y_{3,M}$
$\{1\}$	10	10	*	*
$\{2\}$	20	*	20	*
$\{3\}$	30	*	*	30
$\{1,2\}$	29	10	19	*
$\{1,3\}$	39	10	*	29
$\{2,3\}$	40	*	15	25
N	50	10	15	25

Table 2.1: *The game and the (population monotonic) allocation scheme of Example 2.13.*

monotonic, observe that if player 2 would join player 1 to form coalition $\{1,2\}$, player 2's cost reduces from $y_{2,\{2\}} = 20$ to $y_{2,\{1,2\}} = 19$. This population monotonicity can be verified for the members of all other nested pairs of coalitions as well, implying that y is population monotonic. We remark that that the allocation $(10, 15, 25)$ is an element of the core of (N, c), which we described in Example 2.5. ◇

2.3.8 Overview of implications

Figure 2.1 summarizes the relationships among the various properties that we considered so far. The reverse directions in Figure 2.1 do not hold in general.

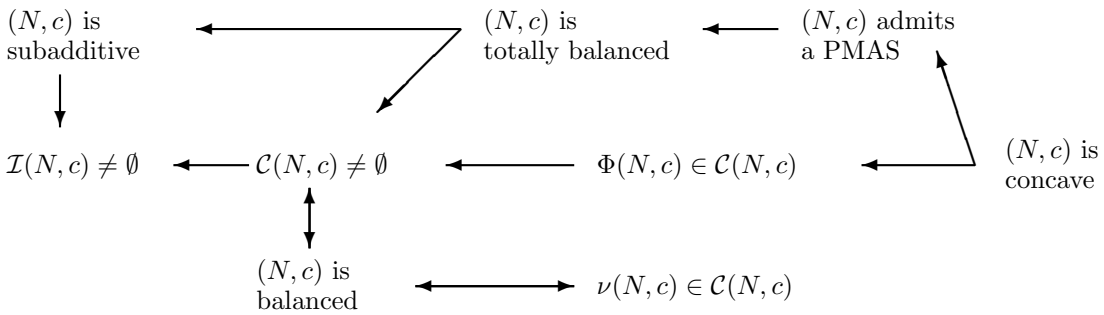

Figure 2.1: *Implications between properties for any game* (N, c).

2.3.9 We have to go stricter

In this section, we strengthen the notions of stability, population monotonicity, and so on by changing non-strict inequalities to strict inequalities. We propose the stronger concepts for two reasons.

First, in practice, people are usually interested in allocations that make each coalition *strictly* better off when cooperating. Indeed, an allocation x under which some group of players is indifferent between cooperating or not (i.e., if $\sum_{i\in M} x_i = c(M)$ for some subcoalition $M \in 2^N_{--}$, while perhaps $\sum_{i\in L} x_i < c(L)$ for another subcoalition $L \in 2^N_{--}$) may be hard to defend in practice because coalition M may decide to stay "alone" if they do not strictly benefit from the collaboration, out of spite. This is avoided if every coalition reaps a strict cost saving.

Second, when we analyze a real-life cooperation via a mathematical model, then some aspects of real life are invariably left out of the model, and the cost parameters will almost always be based on imprecise, judgemental accounting standards (as Dror et al., 2012, also point out). Consequently, the characteristic cost function is only an approximation of the true cost figures. So, even if players never hold a grudge, coalition M may still be hesitant to collaborate under the allocation x that was described above. This hesitance now stems from M's worry that even a very small change of its characteristic costs (as a result of a more precise analysis, modeling, or measurement) may imply that cooperation is strictly worse than staying alone. Such worries can be alleviated, at least to some extent, if each coalition reaps a strictly positive cost saving.

The issues described above are rarely addressed in the literature on cooperative game theory. An exception is Zhao (2001), who introduces and characterizes the relative interior of the core. We will instead consider the *strict core*, which is defined as

$$\mathcal{C}_+(N,c) = \left\{x \in \mathbb{R}^N \mid \textstyle\sum_{i\in N} x_i = c(N) \text{ and } \sum_{i\in M} x_i < c(M) \text{ for all } M \in 2^N_{--}\right\}.$$

Note that the strict core is a convex set. Given a game (N,c), we call any element of $\mathcal{C}_+(N,c)$ a *strictly stable* allocation. By definition, such allocations remain stable after small perturbations in the costs of subcoalitions.

We next define three more strict concepts. Firstly, a game (N,c) is called *strictly subadditive* if for any two coalitions $M, L \in 2^N_-$ with $M \cap L = \emptyset$ it holds that $c(M) + c(L) > c(M \cup L)$. Secondly, an allocation scheme y for a game (N,c) is called *strictly population monotonic* (an SPMAS) if $y_{i,M} > y_{i,L}$ for all coalitions $M, L \in 2^N_-$ with $M \subset L$ and $i \in M$. Thirdly, the *strict imputation set* $\mathcal{I}_+(N,c)$ is defined to be $\left\{x \in \mathbb{R}^N \mid \sum_{i\in N} x_i = c(N) \text{ and } x_i < c(\{i\}) \text{ for all } i \in N\right\}$; any element of this set is called a *strictly individually rational* allocation. Although strict concavity could be defined in similar fashion, we disregard this concept because it need not be satisfied by the (generally non-concave) resource pooling games encountered in this thesis.

Example 2.14. Reconsider the 3-player game (N,c) from Example 2.3 and its allocation scheme y from Table 2.1. Recall that y is a PMAS, which implies that $(10, 15, 25) \in \mathcal{C}(N,c)$ and that (N,c) is subadditive. However, these properties do *not* extend to the strict concepts: the allocation scheme y is *not* strictly population monotonic because $y_{2,\{2,3\}} = y_{2,\{1,2,3\}}$, the allocation $(10, 15, 25)$ is *not* strictly stable because $10 = c(\{1\})$, and the game (N,c) is *not*

strictly subadditive because $c(\{1\}) + c(\{2,3\}) = 50 = c(\{1,2,3\})$. One strict concept still applies, though: the strict imputation set $\mathcal{I}_+(N,c)$ is not empty because it contains, e.g., the allocation $(8,16,26)$. $\diamond$

The class of games with a non-empty strict core can be characterized via balanced maps, in line with the characterization result of Bondareva (1963) and Shapley (1967) for games with a non-empty "traditional" core. We call a game (N,c) *strictly balanced* if for every minimally balanced map $\kappa \in \mathcal{W}^N$ it holds that $c(N) < \sum_{M\in\mathbb{B}(\kappa)} \kappa(M)c(M)$. Moreover, this game is called *strictly totally balanced* if for each coalition $L \in 2^N_-$ the *subgame* (L, c^L) is strictly balanced.

The following theorem deals with the relationships between these strict concepts; as it turns out, all corresponding implications portrayed in Figure 2.1 extend to the strict concepts if there are two or more players. Part (ii) of the following theorem pertains to the multiplicity of (strictly) stable allocations; there is no analogue in Figure 2.1.

Theorem 2.4. *Let (N,c) be a game with two or more players.*

(i) *The strict core $\mathcal{C}_+(N,c)$ is non-empty if and only if (N,c) is strictly balanced.*

(ii) *If (N,c) is strictly balanced, then (N,c) has infinitely many (strict) core allocations.*

(iii) *If (N,c) admits an SPMAS, say y, then $(y_{i,N})_{i\in N} \in \mathcal{C}_+(N,c)$ and (N,c) is strictly totally balanced.*

(iv) *If (N,c) is strictly totally balanced, then it is strictly subadditive.*

(v) *If (N,c) is strictly subadditive, then its strict imputation set $\mathcal{I}_+(N,c)$ is non-empty.*

(vi) *If $\mathcal{C}_+(N,c) \neq \emptyset$, then $\mathcal{I}_+(N,c) \neq \emptyset$.*

(vii) *The strict core $\mathcal{C}_+(N,c)$ is non-empty if and only if $\nu(N,c) \in \mathcal{C}_+(N,c)$.*

Proof. We first simultaneously prove the $\Leftarrow$ direction of Part (i) and Part (ii) because both proofs require the same setup. We will return to the $\Rightarrow$ direction of Part (i) afterwards.

Suppose that $c(N) < \sum_{M\in\mathbb{B}(\kappa)} \kappa(M)c(M)$ for each $\kappa \in \mathcal{W}^N$. Then, let $\kappa^* \in \operatorname{argmin}_{\kappa\in\mathcal{W}^N} \sum_{M\in\mathbb{B}(\kappa)} \kappa(M)c(M)$. Such a κ^* exists because the number of minimally balanced collections for N is finite (this number is no larger than the number of subsets of 2^N_{--}) and because each minimally balanced collection is associated with a unique minimally balanced map. Let $\epsilon^* = \sum_{M\in\mathbb{B}(\kappa^*)} \kappa^*(M)c(M) - c(N)$; note that $\epsilon^* > 0$. We will aim to identify two different strictly stable allocations for the game (N,c). To this end, we define the auxiliary game (N,c^*) by $c^*(M) = c(M)$ for all proper subsets $M \subset N$ and $c^*(N) = c(N) + \epsilon^*$. Then, for any $\kappa \in \mathcal{W}^N$, we obtain

$$c^*(N) = c(N) + \epsilon^* = \sum_{M\in\mathbb{B}(\kappa^*)} \kappa^*(M)c(M) \leq \sum_{M\in\mathbb{B}(\kappa)} \kappa(M)c(M) = \sum_{M\in\mathbb{B}(\kappa)} \kappa(M)c^*(M),$$

where the second inequality holds by definition of ϵ^* and the inequality holds by choice of κ^* as a minimizer. Hence, the game (N,c^*) has a non-empty core as well, which implies that the nucleolus $\nu(N,c^*)$ is in the core of (N,c^*).

For notational ease, let $n = |N|$. Now, we let $i, j \in N$ with $i \neq j$ be two different players, and we define two allocations for the game (N, c): $x^{(i)}$ and $x^{(j)}$. The first allocation, $x^{(i)}$, is defined by $x_k^{(i)} = \nu_k(N, c^*) - \epsilon^*/((n+1)(n-1))$ for all $k \in N \backslash \{i\}$ and $x_i^{(i)} = \nu_i(N, c^*) - \epsilon^* n/(n+1)$. The second allocation, $x^{(j)}$, is defined analogously, now reducing player j's allocation by $\epsilon^* n/(n+1)$ instead. Since $n \in \{2, 3, \ldots\}$, it holds that $n/(n+1) \neq 1/((n+1)(n-1))$; hence, $x^{(i)} \neq x^{(j)}$. Note that $x^{(i)}$ is an efficient allocation for (N, c) since

$$\sum_{k \in N} x_k^{(i)} = \left(\sum_{k \in N \backslash \{i\}} \nu_k(N, c^*) \right) - \frac{\epsilon^*}{n+1} + \nu_i(N, c^*) - \frac{\epsilon^* n}{n+1} = c^*(N) - \epsilon^* = c(N),$$

where the second equality holds because $\nu(N, c^*)$ is an efficient allocation for the (N, c^*). Moreover, $x^{(i)}$ is a strictly stable allocation for game (N, c) since, for any $M \in 2^N_{--}$, it holds that

$$\sum_{k \in M} x_k^{(i)} < \sum_{k \in M} \nu_k(N, c^*) \leq c^*(M) = c(M),$$

where the first inequality holds because $\epsilon^* > 0$ and the second inequality is valid because $\nu(N, c^*)$ was a stable allocation for the game (N, c^*). We conclude that $x^{(i)}$ is a strictly stable allocation for (N, c). The argument for strict stability of $x^{(j)}$ goes analogously. Thus, the game (N, c) has a non-empty strict core, and the $\Leftarrow$ direction of Part (i) is proven. Finally, as the (strict) core is a convex set, the existence of two different (strict) core allocations implies that there are infinitely many (strict) core allocations. This completes the proof of Part (ii).

We now return to the $\Rightarrow$ direction of Part (i). Suppose that (N, c) has a non-empty strict core. Consider any $x \in \mathcal{C}_+(N, c)$ and $\kappa \in \mathcal{W}^N$. Then, by definition of a balanced map,

$$c(N) = \sum_{i \in N} x_i = \sum_{i \in N} \sum_{M \in \mathbb{B}(\kappa): i \in M} \kappa(M) x_i = \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \sum_{i \in M} x_i < \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) c(M),$$

where the inequality holds because x is a strictly stable allocation for the game (N, c).

(iii). Let y be an SPMAS. By definition, y satisfies $\sum_{i \in M} y_{i,N} < \sum_{i \in M} y_{i,M} = c(M)$ for every $M \in 2^N_{--}$ and $\sum_{i \in N} y_{i,N} = c(N)$. Hence, $(y_{i,N})_{i \in N}$ is a strictly stable allocation. To show strict total balancedness, let $L \in 2^N_-$ and consider the subgame (L, c^L). By definition, y satisfies $\sum_{i \in M} y_{i,L} < \sum_{i \in M} y_{i,M} = c(M)$ for every $M \in 2^L_{--}$ and $\sum_{i \in L} y_{i,L} = c(L)$. Hence, $(y_{i,N})_{i \in N}$ is a strictly stable allocation, and thus $\mathcal{C}_+(L, c^L) \neq \emptyset$. By the $\Leftarrow$ direction of Part (i), this means that (L, c^L) is balanced. We conclude that (N, c) is strictly totally balanced.

(iv). Assume that (N, c) is strictly totally balanced. Let $M, L \in 2^N_-$ with $M \cap L = \emptyset$. Consider the subgame $(M \cup L, c^{M \cup L})$, and take the map $\kappa : 2^{M \cup L} \to [0, 1]$ with $\kappa(M) = \kappa(L) = 1$ and $\kappa(J) = 0$ for all other coalitions $J \in 2^{M \cup L} \setminus \{M, L\}$. Note that this map is balanced for $M \cup L$ because M and L are disjoint sets by assumption. Then, $c(M) + c(L) = \sum_{J \in \mathbb{B}(\kappa)} \kappa(J) c(J) > c(M \cup L)$, where the inequality follows from strict total balancedness of (N, c). We conclude that (N, c) is strictly subadditive.

(v). Assume that (N, c) is strictly subadditive. This immediately implies that $\sum_{i \in N} c(\{i\}) > c(N)$. Hence, the allocation x defined by $x_i = c(\{i\}) - \big(\sum_{i \in N} c(\{i\}) - c(N)\big)/|N|$ is strictly individually rational, and thus $\mathcal{I}_+(N, c) \neq \emptyset$.

(vi). This readily follows from the observation that $\mathcal{C}_+(N, c) \subseteq \mathcal{I}_+(N, c)$.

(vii). The $\Leftarrow$ direction is trivial. The $\Rightarrow$ direction follows from the observation that the nucleolus maximizes the minimal satisfaction of subcoalitions. Indeed, when $\mathcal{C}_+(N, c) \neq \emptyset$, an allocation with a positive satisfaction for every subcoalition exists, which implies that the minimal satisfaction of a subcoalition under $\nu(N, c)$ is positive. Hence, $\nu(N, c) \in \mathcal{C}_+(N, c) \neq \emptyset$. □

We remark that, obviously, any strict concept immediately implies its non-strict counterpart. The reverse of Part (ii) of Theorem 2.4 is not true if the parenthesized word "strict" is removed. Specifically, as shown in the following example, a game with infinitely many core allocations need not admit a strictly stable allocation.

Example 2.15. Reconsider the game (N, c) that was described in Example 2.3. As derived in Example 2.5, $\mathcal{C}(N, c) = \{\alpha(10, 11, 29) + (1 - \alpha)(10, 19, 21) \mid 0 \leq \alpha \leq 1\}$, i.e., this game has infinitely many core allocations. However, since $x_1 = 10 = c(\{1\})$ for all $x \in \mathcal{C}(N, c)$, this game does not possess a *strictly* stable allocation. ◇

2.3.10 Single-attribute games

In this section, we consider games that arise from situations in which every player is associated with a certain resource endowment: his *attribute*, described by a positive real number. The players can pool their endowments to attain potential cost savings. To describe this, we introduce a *single-attribute situation* as a triple $\varphi = (N, \tilde{K}, \lambda)$, where

- N is a player set;
- $\tilde{K}$ is a non-decreasing function mapping $\mathbb{R}_+$ to $\mathbb{R}_+$ with $\tilde{K}(0) = 0$;
- $\lambda = (\lambda_i)_{i \in N}$ is an element of $\mathbb{R}^N_{++}$.

Here, λ should be interpreted as a vector of attributes, with λ_i the attribute of player $i \in N$. We will systematically write $\lambda_M = \sum_{i \in M} \lambda_i$ for any $M \subseteq N$. The function $\tilde{K}$ expresses the costs to serve any level of attributes. The non-decreasing requirement on $\tilde{K}$ is imposed to highlight the applications that we have in mind: in these applications, $\tilde{K}$ expresses the costs to serve any level of demand in a resource pooling setting and such a cost does not shrink as demand increases.

We remark that single-attribute situations have the same mathematical structure as the cost sharing situations considered in, e.g., Hamlen et al. (1977), Moulin and Shenker (1992), and Sudhölter (1998). We will return to this correspondence in Section 8.4. Our current interpretation and terminology is based on Özen et al. (2011).

A single-attribute situation naturally leads to a corresponding game. Given a single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$, we call the game (N, c^φ) that is defined by $c^\varphi(M) = \tilde{K}(\lambda_M)$ for all $M \in 2^N_-$ the *associated single-attribute game*.

Let $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\tilde{K}(0) = 0$. A game (N, c) is called a *single-attribute game embedded in* $\tilde{K}$ if there exists a vector $\lambda \in \mathbb{R}^N_{++}$ such that $c(M) = \tilde{K}(\lambda_M)$ for all $M \in 2^N_-$. We remark that the set of *all* single-attribute games embedded in $\tilde{K}$ contains games with two players, games with three players, etc.

The following theorem is based on Özen et al. (2011).[14] Part (i) of the following theorem also follows from Hamlen et al., 1977, p. 621.

Theorem 2.5. *Let $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\tilde{K}(0) = 0$.*

- (i) *If $\tilde{K}$ is elastic, then for any single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ the allocation scheme assigning $\tilde{K}(\lambda_M) \cdot \lambda_i / \lambda_M$ to any $i \in M$ and $M \in 2^N_-$ is a PMAS for the associated single-attribute game (N, c^φ).*
- (ii) *If $\tilde{K}$ is not elastic, then there exists a single-attribute game embedded in $\tilde{K}$ that has an empty core.*
- (iii) *If all single-attribute games embedded in $\tilde{K}$ have a non-empty core, then $\tilde{K}$ is elastic.*

Theorem 2.5 indicates that elasticity of K is of crucial importance. Fortunately, the cost functions underlying many of models to be studied in this thesis will turn out to be elastic.

2.4 Stochastic processes

In this thesis, we are interested in phenomena that are not predictable with certainty in advance but, rather, exhibit an inherent variation. One example is the number of customers in a service system, which may evolve over time in an apparently random fashion. To model such quantities, we employ notions from probability theory and queueing theory. In this section, we briefly introduce several elementary concepts from those fields. For a more extensive treatment, see the book by Ross (2007).

2.4.1 Probability theory and random variables

Probability theory provides a mathematical way to model random phenomena or experiments that can produce a number of outcomes, each with a specified likelihood of happening. Although the reader is probably familiar with basic notions from probability theory, we will provide a quick refresher on the terminology. For brevity, in this section we only recall the formal definitions for finite or countable sample spaces.

A *sample space* Ω is a set that contains all possible scenarios (possible futures, experimental outcomes, etcetera). For finite or countable Ω, a *probability measure* $\mathbb{P} : \Omega \to$

[14] Although Özen et al. (2011) focus on games in profit rather than cost terms, their results carry over straightforwardly by simply reversing all inequalities involved.

$[0,1]$ is a function with $\sum_{\omega\in\Omega}\mathbb{P}(\omega)=1$. This function describes the occurrence probability of the various scenarios.

We are frequently interested in functions that assign a real number to each scenario, as opposed to the scenario itself. These functions are known as *random variables*. So, a random variable has domain Ω and the probability of taking on a certain value is described by $\mathbb{P}$. The *expected value* or *mean* of a random variable X, written $\mathbb{E}[X]$, is the average, weighted by $\mathbb{P}$, of the possible values that X can take on.

For finite or countable Ω, two random variables $X_1:\Omega\to C_1$ and $X_2:\Omega\to C_2$ are called *independent* if observing the value of one does not affect the probability of observing any value of the other, i.e., if for every $y_1\in C_1$ that can occur with positive probability, it holds for all $y_2\in C_2$ that

$$\sum_{x\in\Omega:X_2(x)=y_2,X_1(x)=y_1}\mathbb{P}(x)=\sum_{x\in\Omega:X_2(x)=y_2}\mathbb{P}(x)\cdot\sum_{x\in\Omega:X_1(x)=y_1}\mathbb{P}(x).$$

For uncountable Ω, notions such as random variables and their independence have analogous, well-known meanings. As we do not use the formal, technical definitions for uncountable Ω in this thesis, we omit them for brevity.

2.4.2 Poisson distribution and exponential distribution

In this thesis, we will often consider random variables with Poisson or exponential distributions. These distributions are relatively easy to work with and are often a good approximation of natural processes.

An integer-valued random variable X is said to be *Poisson distributed* with parameter $\lambda>0$, written $X\sim Poi(\lambda)$, if for every $y\in\mathbb{N}_0$ the probability that X takes on the value y, written $\mathbb{P}[X=y]$, is equal to $\lambda^y e^{-\lambda}/y!$. The mean of a Poisson distributed random variable with parameter λ is equal to λ.

A real-valued random variable X is said to be *exponentially distributed* with parameter $\lambda>0$ if for every $y\in\mathbb{R}_+$ the probability that X takes on any value smaller than or equal to y, written $\mathbb{P}[X\leq y]$, is equal to $1-e^{-\lambda y}$. The mean of an exponentially distributed random variable with parameter λ is equal to $1/\lambda$.

2.4.3 Markov processes

Given any set T, often interpreted as a set of *time* indices, and any set S, often referred to as the *state space*, we call a collection of indexed random variables $(X^t)_{t\in T}$, all taking values in S, a *stochastic process*.

A stochastic process $(X^t)_{t\in\mathbb{R}_+}$ taking values in $\mathbb{N}_0$ is said to be a *Poisson process with rate* $\lambda>0$ if it counts the number of events that have occurred in a system where the time between each pair of successive events is independently and identically distributed (i.i.d.)

according to an exponential distribution with parameter λ.[15] The exponential interarrival time distribution is *memoryless*, which means that the probability of having to wait until at least time $x + y$ for the first arrival, given that the first arrival has not happened yet after time y, is equal to the initial probability of having to wait until at least time x. Merging of independent Poisson processes results in another Poisson process with the sum of the rates. That is, if $(X_1^t)_{t\in T}$ and $(X_2^t)_{t\in T}$ are two independent Poisson processes with respective rates λ_1 and λ_2, then the superposition $(X^t)_{t\in T}$ defined by $X^t = X_1^t + X_2^t$ for all $t \in T$ is also a Poisson process with rate $\lambda_1 + \lambda_2$.

A stochastic process $(X^t)_{t\in\mathbb{R}_+}$ taking values in S is said to be a (stationary, continuous-time, non-absorbing) *Markov process* if there exists a non-negative transition rate vector $q = (q_{i,j})_{i,j\in S, i\neq j}$ satisfying $\sum_{j\in S:j\neq i} q_{i,j} > 0$ for all $i \in S$ such that each time the process enters a state $i \in S$, the amount of time it spends in that state before making a transition to a different state is exponentially distributed with parameter $\sum_{j\in S:j\neq i} q_{i,j}$, and when the process leaves state i, it next enters a different state j with probability $q_{i,j}/\sum_{k\in S:k\neq i} q_{i,k}$. This means that the future is independent of the past and present. A Markov process can be graphically represented as a directed graph with set of nodes S, where an arc is drawn from node i to a different node j if $q_{i,j} > 0$, with the value $q_{i,j}$ placed on that arc.

A simple example of a Markov process is a Poisson process. The following example provides another example of a Markov process and simultaneously illustrates the merging property of independent Poisson processes.

Example 2.16. Suppose that broken machines of type A (resp. B) arrive at a repairman (i.e., a single-server service system) in accordance with a Poisson process with rate λ_A (resp. λ_B). The two arrival processes are independent. This means, by the merging property of independent Poisson processes, that the times between successive arrivals of an arbitrary machine (i.e., a customer) are exponential i.i.d. with parameter $\lambda = \lambda_A + \lambda_B$.

Upon arrival, any machine directly goes into repair (i.e., enters service) if the repairman is free; if not, then the machine joins the waiting line. When the repairman finishes serving a machine, that machine leaves the system and the next in line, if there are any waiting, goes into repair. The successive repair times (i.e., service times) are exponential i.i.d. with parameter $\mu > \lambda$ for both types A and B.

The corresponding stochastic process $(X^t)_{t\in\mathbb{R}_+}$, where X^t represents the total number of customers in the system at time t, is the Markov process represented in Figure 2.2. ◇

Under suitable regularity conditions, which are satisfied by all Markov processes considered in this thesis, the probability of observing a Markov process $(X^t)_{t\in\mathbb{R}_+}$ in state $i \in S$ at time

[15] Alternatively, we may say that $(X^t)_{t\in\mathbb{R}_+}$ is a Poisson process with rate $\lambda > 0$ if for all $s \geq 0$ and all $t > 0$ it holds that $X^{s+t} - X^s$ has a Poisson distribution with parameter λt, where $X^{s+t} - X^s$ is interpreted as the number of events that occur in the interval $(s, t]$, and if it holds that the number of events occurring in disjoint time intervals are independent.

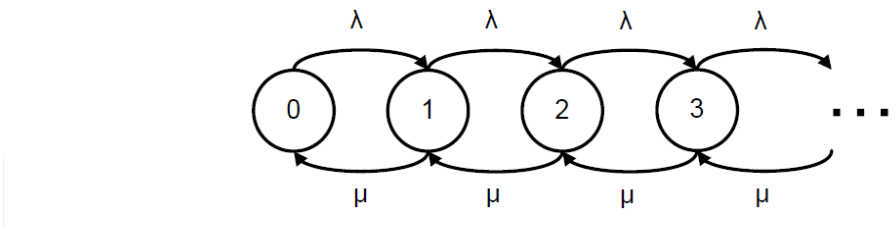

Figure 2.2: *The Markov process (an $M/M/1$ queueing model) of Examples 2.16 and 2.17.*

t converges, as t tends to infinity, to a limiting value. We call this value π_i the *steady-state probability* of state i, and it should be interpreted as the long-run proportion of time that the process is in state i.

2.4.4 Queueing models

A *queueing model*, at least in this thesis, is a stochastic process taking values in $\mathbb{N}_0$ that describes the evolution over time of the number of customers in a service system where customers randomly arrive, possibly wait in a queue, and leave after being served. Kendall's $A/B/C/D$ notation is the standard way to classify a queueing model when the service discipline is first-come first-serve, as in all queueing models considered in this thesis. Here, A is a code for the interarrival time distribution, B is a code for the service time distribution, C is the number of servers, and D is the maximum number of customers allowed in the system including those in service. The D entry is often omitted if there is infinite waiting space. Common codes for distributions (A and B) are M, which represents the memoryless exponential distributions, and G, which indicates a general or arbitrary independent distribution. The G code encompasses deterministic times.

Example 2.17. Reconsider the (merged) process described in Example 2.16. Clearly, this is a queueing model. In this model, both the interarrival and service times are exponentially distributed, there is a single server, and there is infinite waiting space. Accordingly, in Kendall's notation, this system is an $M/M/1(/\infty)$ queue. This is the most elementary of queueing models and it has been analyzed extensively in the literature. Two well-known results are that the steady-state probabilities are given by $\pi_i = (\lambda/\mu)^i(1-\lambda/\mu)$ for all $i \in \mathbb{N}_0$ and that the expected time spent in the system (i.e., time waiting in the queue and in service) by an arbitrary customer in steady-state is $1/(\mu-\lambda)$. ◇

The queueing models underlying the various resource pooling games analyzed in this thesis are the $M/G/s/s$ model, the $M/M/s$ model, and the $M/G/\infty$ model. We will briefly describe these three models. Firstly, the $M/G/s/s$ queue, also known as *Erlang loss system*, is a model where interarrival times follow an exponential distribution (with parameter λ), service times follow a general distribution (with mean $1/\mu$), there are s servers, each arriving

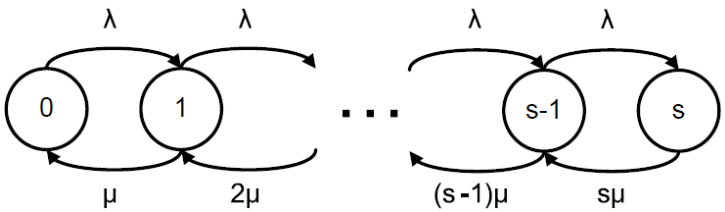

Figure 2.3: *The $M/M/s/s$ queue.*

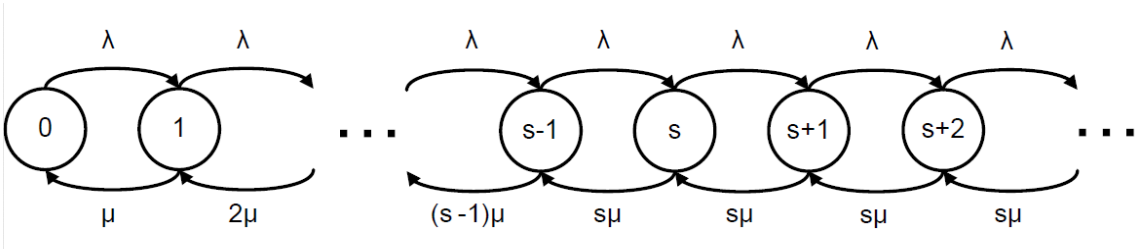

Figure 2.4: *The $M/M/s$ queue.*

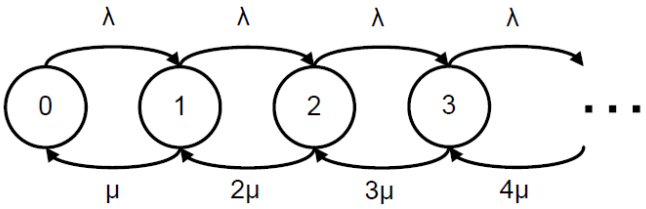

Figure 2.5: *The $M/M/\infty$ queue.*

customer immediately goes into service if there is an unoccupied server available, and an arriving customer that finds no free server does not enter but rather is lost and never served. Such customers are called *blocked* customers. If the service time distribution in an Erlang loss system is exponential, then the corresponding stochastic process is a Markov process, as pictured in Figure 2.3. Secondly, the $M/M/s$ queue, also known as *Erlang delay system*, is a model where waiting *is* allowed; it features exponential i.i.d. interarrival and service times, s servers, and infinite waiting space. Figure 2.4 represents the corresponding Markov process. Finally, the $M/G/\infty$ queue is a model where interarrival times follow an exponential distribution, service times follow a general distribution, and there is an unlimited number of servers. If the service time distribution is exponential, then the corresponding stochastic process is a Markov process, as shown in Figure 2.5. In any of these models, we will call the product of the arrival rate and the mean service time the *offered load*.

We now make a connection with infinite-horizon inventory models. Suppose that initially there are $S \in \mathbb{N}_0$ items on stock, that demands occur according to a Poisson process with rate λ, and that a one-for-one replenishment strategy is followed under i.i.d. lead times with mean $\tau > 0$. This setting is known as the $(S-1, S)$ model. If demands are *lost* in case of a stock-out, then the process for the number of items in the replenishment pipeline is identical to the process for the number of busy servers in an $M/G/S/S$ system with arrival rate λ and mean service time τ. If demands are *backlogged* in case of a stock-out, then the process for the number of items in the replenishment pipeline is identical to the process for the number of busy servers in an $M/G/\infty$ system with arrival rate λ and mean service time τ.

2.4.5 Discussion of assumptions

An appealing feature of the $M/G/s/s$, the $M/M/s$, and $M/G/\infty$ queueing models described above is that closed form expressions for their steady-state probabilities are available. From a steady-state distribution π, performance measures such as the mean time spent waiting in the queue can be easily obtained. Although technically π only describes the system when it has been running for an infinite amount of time, we are motivated by structural collaborations over long periods of time. For the context of pooling spare parts or repairmen in particular, the machines that the spare parts or repairmen aim to service typically have lifetimes of several decades, which is long enough for π to reasonably approximate the state of the system at an arbitrary point in time during the collaboration, assuming that the installed base stays constant over time. Research on queueing models and infinite-horizon inventory models customarily focuses on the steady-state distributions, due to its analytical tractability.

All above-mentioned queueing models feature a Poisson arrival process. The assumption of a Poisson arrival process is reasonable when customers arrive independently from a large population. One example of this in a service operations setting where a customer arrival represents a component failure. Components of high-tech machines typically have close to

exponentially distributed lifetimes (cf. Wong et al., 2006). As such, the failure process of a single machine is close to Poisson, and a merged stream of component failure processes — as faced by a stock point for spare parts or a group of repairmen that serve a large number of machines — corresponds very well to a Poisson process. The assumption of a Poisson arrival process may even be justified when a customer arrival represents the arrival of a human who strategically chooses his arrival time to avoid congestion: as shown in Lariviere and van Mieghem (2004) with Nash equilibirum concepts, the assumption of Poisson arrivals is still acceptable in such a setting.

Besides its ability to resemble real-world phenomena, a Poisson arrival process also enables tractable mathematics. One important reason for this is that Poisson arrivals see time averages—a property known under the tasty acronym PASTA, which means that the fraction of arriving customers who observe n other customers in the system upon arrival is equal to π_n, i.e., the steady-state probability of having n customers in the system (Wolff, 1982).

—If I have seen further, it is only by standing on the shoulders of giants.

Isaac Newton

3 Literature review

3.1 Introduction

In the previous chapter, we took the characteristic cost function of a game as a given and explored concepts such as the core. We did not pay much attention to the *origin* of a characteristic cost function, i.e., the underlying situation from which a game actually stems. Often, a game stems from a situation where several players face a joint optimisation or operations research problem. There is a growing stream of literature that considers various operations research (OR) problems from the perspective of cooperative game theory. Such research typically studies the general properties (e.g., balancedness or convexity) of all games arising from that specific type of OR problem and/or the properties of context-specific cost allocation rules defined on the parameters of that specific type of problem.

The applications of cooperative game theory to operations research problems were surveyed at the start of this century by Borm et al. (2001). They mainly focused on games stemming from problems of a combinatorial nature, such as connection, routing, scheduling, and production. Since then, various researchers have applied cooperative game theory to inventory and queueing problems. In this chapter, we survey the corresponding games whilst simultaneously positioning the work of this thesis in the existing literature. We describe inventory games in Section 3.2 and queueing games in Section 3.3. In Section 3.4, we briefly describe the literature on pooling in stochastic inventory and queueing systems with only a single decision maker.

3.2 Cooperative inventory games

There is a growing stream of literature on cooperative games in inventory systems. In two recent papers, Fiestras-Janeiro et al. (2011) and Dror and Hartman (2011) provide detailed reviews of this stream of literature. This section provides a briefer overview with three aims: acquainting the reader with the various lines of inventory game research, formally defining concepts that we will use later, and pinpointing the literature gaps that this thesis aims to fill. We distinguish five lines of inventory game research: economic order quantity (EOQ) games, economic lot sizing (ELS) games, newsvendor games, continuous-review games, and spare parts games. We next review each line separately.

3.2.1 EOQ games

The first line of research is on EOQ games in which players face deterministic demand at a constant rate. For any order that is placed, a fixed ordering cost is incurred. Holding inventory in stock is also costly. The players may cooperate by ordering jointly and sharing the fixed ordering costs. The optimal replenishment policy for any coalition is determined via the basic Economic Order Quantity model (see, e.g., Zipkin, 2000). Formally, an *EOQ situation* is a triple $(N, (m_i^2)_{i\in N}, a)$ where

- N is the set of players;
- $m_i^2 > 0$ is the square of the optimal number of orders per time unit for player $i \in N$ if ordering alone;
- $a > 0$ is the fixed (major) ordering cost, say of leasing a truck.

The *EOQ game* (N, c) corresponding to such a situation is described by

$$c(M) = 2a\sqrt{\sum_{i\in M} m_i^2}$$

for each coalition $M \in 2^N_{-}$. Meca et al. (2004) show that EOQ games are concave. Moreover, they show that the allocation assigning $c(N) \cdot m_i^2 / \sum_{j\in N} m_j^2$ to each player $i \in N$ is stable and reachable through a PMAS. Mosquera et al. (2008) axiomatically characterize this proportional cost allocation rule. Their characterization uses an axiom saying that the rule should be immune to coalitional manipulation via artificial splitting or merging of the players.

Meca et al. (2003) show that a wider class of Economic Production Quantity situations with shortages lead to exactly the same class of games, and Meca et al. (2007) analyze a further extension with temporary price discounts.

Anily and Haviv (2007), Dror and Hartman (2007), and Zhang (2009) study the balancedness of generalized EOQ games in which, in addition to the fixed (major) ordering costs, there is also an individual (minor) delivery cost for every player that is incurred only

when that player participates in an order. Dror et al. (2012) numerically study how frequently these generalized EOQ games are concave and numerically study the performance of various cost allocation rules.

Fiestras-Janeiro et al. (2012) analyze a variant where the (minor) delivery costs are based on the distance to the supplier and the players are located on a line route, such that only the (minor) delivery cost of the player at maximal distance to the supplier has to be paid. He et al. (2012) analyze a variant where the joint ordering cost can be any nondecreasing, submodular function.

3.2.2 ELS games

Economic Lot Sizing (ELS) situations are similar to EOQ situations in that there are deterministic demands, but instead of having demands coming in at a constant rate over an infinite time horizon, each player now has discrete demands over a finite number of periods. There are possible economies of scale in the production. Inventory may be held, at a holding cost, to satisfy demands in future periods. The players may cooperate by constructing and executing an optimal joint production plan. The optimal plan may be determined via the Wagner-Whitin algorithm (see, e.g., Zipkin, 2000). Formally, an *ELS situation* is a tuple[16] $(N, T, \{(d_i^t)_{i\in N}, h^t, b^t, p^t\}_{t=1}^T)$ where

- N is the set of players;
- $T \in \mathbb{N}$ is the number of periods in the planning horizon;
- $d_i^t \geq 0$ is the deterministic demand for player $i \in N$ in period t;
- $h^t \geq 0$ is the unit inventory holding costs for period t;
- $b^t \geq 0$ is the unit backlogging costs for period t;
- $p^t : \mathbb{R}_+ \to \mathbb{R}_+$ is the concave, nondecreasing production cost function for period t.

The *ELS game* corresponding to such a situation is described by the total costs of each coalition's optimal production plan. For details, see Chen and Zhang (2008).

Van den Heuvel et al. (2007) focus on a specific subclass of ELS games in which backlogging is not allowed (i.e., $b^t = \infty$ for all periods t) and in which p^t is the sum of a variable cost component that is linear in the number of units produced and a fixed setup cost component that is incurred any time production is initiated. They show that the corresponding games are balanced. Chen and Zhang (2008) prove the stronger, more general result that all ELS games as described above are balanced.

[16] A tuple is an ordered list of elements. In contrast to a set, a tuple is ordered and can contain repeat elements. A tuple with two elements is called a pair, and a tuple with three elements is called a triple.

Xu and Yang (2009) and Gopaladesikan et al. (2012) propose cost-sharing methods for ELS games. Zeng et al. (2011) consider an extension of ELS games with perishable inventory, and Drechsel and Kimms (2011) consider an extension of ELS games with transshipment costs and scarce capacities.

Guardiola et al. (2008, 2009) study a variant of ELS games, called production-inventory games, where p^t is a linear function for all periods t. So, they disregard setup costs. However, their variant has different cost parameters for every player, and they allow any coalition to use the cheapest cost parameters among its members.

3.2.3 Newsvendor games

Next up is the vast line of research on newsvendor games. In this setting, each player faces a stochastic demand for the same item in a single period. There is a trade-off between the costs of ordering too much and the costs of ordering an insufficient amount. The players may cooperate by coordinating their orders and pooling their inventories after the demand realization is known. The optimal order quantity is determined via the basic newsvendor model (see, e.g., Zipkin, 2000). Formally, a *newsvendor situation* is a tuple $(N, (D_i)_{i\in N}, c_o, c_u)$ where

- N is the set of players;
- D_i is the stochastic demand (a real-valued random variable with finite mean) of player $i \in N$;
- $c_o \geq 0$ is the oversupply cost;
- $c_u \geq 0$ is the undersupply cost.

The corresponding *newsvendor game* (N, c) is defined by

$$c(M) = \min_{s\geq 0} \left(c_o \cdot \mathbb{E} \max\Big\{ s - \sum_{i\in M} D_i, 0 \Big\} + c_u \cdot \mathbb{E} \max\Big\{ \sum_{i\in M} D_i - s, 0 \Big\} \right)$$

for all coalitions $M \in 2^N_-$.

Müller et al. (2002) and Slikker et al. (2001) independently derived that every newsvendor game is balanced. Later, several authors introduced and analyzed various games associated with generalizations of the basic newsvendor model. We start our overview of these generalizations with Slikker et al. (2005), who enrich the basic model with transshipment costs and asymmetric (i.e., possibly player-dependent) oversupply and undersupply costs. Many other generalizations have been studied as well: Montrucchio and Scarsini (2007) consider an infinite number of newsvendors, Özen et al. (2008) and Chen and Zhang (2009) consider supplying via warehouses, Chen (2009) consider price-dependent demands and quantity

discounts, Kemahlıoğlu-Ziya and Bartholdi III (2011) consider a separate supplier who enables pooling, Özen et al. (2012a) consider lower bounds on the amount of inventory that each player must receive, and Özen et al. (2012b) consider demand forecast updates. Most papers show that the corresponding games remain balanced. The properties of newsvendor games and their generalizations are reviewed in more detail in Montrucchio et al. (2012).

Alternative model formulations, convexity, and truthful information disclosure for newsvendor situations have also stimulated research. Anupindi et al. (2001) and Granot and Sošić (2003) study a setting where the ordering decisions are made individually and only the transshipment decision is taken cooperatively. Dror and Hartman (2005) and Dror et al. (2008) consider games based on demand realizations rather than expectations. Özen et al. (2011) tackle the convexity of newsvendor games. Finally, Norde et al. (2011) propose mechanisms for truthful revelation of private demand information.[17]

3.2.4 Continuous-review games

The fourth line of research is on inventory centralization games in a continuous-review setting where several players face stochastic demand for the same item and penalty costs have to be paid per backorder occurrence independent of duration. The players can cooperate by pooling inventory, thereby reducing their total required safety stocks. Gerchak and Gupta (1991), Robinson (1993) and Hartman and Dror (1996) examined the problem of allocating the joint costs and showed balancedness of games formulated via an approximation of the total costs rather than an exact evaluation.

3.2.5 Spare parts games

The final, relatively scarce line of inventory game research is on spare parts games. Though similar to continuous-review games, spare parts games are based on Poisson demand processes, which permits exact Markov chain analysis. Wong et al. (2007) are the first to study a multi-location, continuous-review, infinite-horizon model that is motivated by spare parts applications. In their model, every player has a number of machines, each containing a critical component whose operational lifetime and repair time is exponentially distributed. The players may cooperate by pooling their spare parts via costly lateral transshipments. Several cost allocation mechanisms are proposed and numerically illustrated in a 3-player example. We also mention Kilpi et al. (2009), who specify a framework of cooperative strategies, each

[17]The newsvendor model has also formed the backdrop for several papers on strategic competition and bargaining in supply chains. Often, such papers analyze interactions between suppliers and retailers with non-cooperative game theory and look for a Nash equilibrium (see the review of Cachon and Netessine, 2004) or study the coalition formation process under farsightedness (as reviewed in Nagarajan and Sošić, 2008). In contrast to those works, we are interested in settings where cooperation via a binding pooling agreement is possible. We abstract away the intricacies of negotiation and coalition formation and, in doing so, we model our settings as cooperative games. This allows us to focus on the cost allocation problem in more detail.

with a different degree of contractual integration, for the availability of repairable aircraft components. They do not explicitly formulate a game, but they do propose mechanisms for sharing the pooling benefits in the spirit of the game theoretical approach. Their mechanisms are illustrated numerically. Neither of these two papers investigate analytical properties of the underlying game or stability of cost allocations in general; this is left as a future research direction.

3.2.6 Contribution to the literature

This literature review has shown that most of the inventory game research has dealt with deterministic demands (the EOQ games and ELS games) or stochastic demands in a single-period model with no further inventory replenishment or inventory carry-over (the newsvendor game). However, in our motivating spare parts pooling applications, we have stochastic demands over a long time horizon. So, any reasonable inventory model for this application should feature multiple periods (possibly an infinite number of periods, i.e., a steady-state setting) with inventory carry-over and a plethora of replenishment opportunities. The importance of extending newsvendor games to multiple periods has also been pointed out by Chen (2009).

Although some research has been done on games derived from inventory models with these features (the continuous-review and spare parts games), this existing research has its limitations. In particular, the continuous-review games feature approximate rather than exact cost expressions and, moreover, assume that penalty costs have to be paid per backorder occurrence independent of duration. In our motivating applications, either such costs are paid for each unit of time an item is backordered (as we will assume in Chapter 6 and 7) or backordering is not an acceptable option and stock-outs are dealt with via emergency procedures instead (as we will assume in Chapter 4). The spare parts games, on the other hand, have only received ad hoc analysis; structural results are lacking.

We conclude that there is a lack of knowledge on structural properties of games and cost allocation mechanisms for pooled spare parts. Indeed, not even the elementary $(S-1, S)$ infinite-horizon inventory model with Poisson demand processes has been adequately studied from the perspective of cooperative game theory, in spite of its practical relevance for spare parts inventory management. Moreover, the inventory game literature hasn't come close to touching a variant with non-stationary, stochastic demands and correlations over time. These are precisely the voids that we aim to fill in Chapters 4, 6, and 7: We will formulate these games and prove, amongst other results, existence of a stable allocation.

3.3 Cooperative queueing games

In this section, we direct our attention to games in queueing systems. The existing literature that applies cooperative game theory to analyze resource pooling in queueing models can be

	Optimized service capacity	Fixed service capacity
Waiting in queue	González and Herrero (2004) García-Sanz et al. (2008) Yu et al. (2009)	Anily and Haviv (2010) Timmer and Scheinhardt (2010) Anily and Haviv (2011)

Table 3.1: *Classification of literature on **single**-server cooperative queueing games*

	Optimized numbers of servers	Fixed numbers of servers
Waiting in queue	Yu et al. (2007) Chapter 5	Chapter 5
No waiting allowed	Özen et al. (2011) McLean and Sharkey (1993) Chapter 4	Chapter 4

Table 3.2: *Classification of literature on **multi**-server cooperative queueing games*

classified along several dimensions: Tables 3.1 and 3.2 position the existing literature and the queueing chapters in this thesis according to various modeling assumptions. (Although the model in Chapter 6 has queueing characteristics, its game is based on inventory analysis; accordingly, Chapter 6 is not included in the tables.)

As can be seen in the tables, the literature on cooperative queueing games is scarce. This scarcity is surprising because related problems have been studied extensively. Indeed, cooperative games stemming from sequencing or scheduling problems, which consider a finite population of customers and are not concerned with steady-state congestion as we are, have been fruitfully analyzed; see Curiel et al. (2002) for a survey. Likewise, competition among service providers and customer equilibrium behavior in queues, analyzed from a non-cooperative perspective, has been a rich research domain; see Hassin and Haviv (2003) for a survey. However, as revealed in Tables 3.1 and 3.2, only a few papers have investigated queueing models from a cooperative point of view.

In this section, we survey the various categories of queueing games separately. We first treat single-server queueing games where the service capacity can be varied continuously and where the service capacity can be easily consolidated into a single server (e.g., by choice of material or technology). Afterwards, we treat multi-server queueing games where the service rate of a server is given exogenously and where service capacity can only be adjusted by changing the number of servers (e.g., by hiring additional repairmen). Finally, we pinpoint the gaps in the existing literature on queueing games that this thesis aim to fill.

3.3.1 $M/M/1$ games with fixed numbers of servers

Anily and Haviv (2010) study a model with several service providers. Each service provider corresponds to a player. Each player is associated with a Poisson customer arrival stream and a capacity endowment. The work content for each customer is exponential i.i.d. Any coalition operates an $M/M/1$ queue that serves the union of its members' arrival streams with a single exponential server whose service rate is the sum of the capacity endowment of the coalition members. Such collaboration leads to reduced congestion, as measured by the total number of customers in the system. To describe the model considered by Anily and Haviv (2010), we define an *M/M/1 situation with fixed service capacity* as a tuple $(N, (\lambda_i)_{i\in N}, (\mu_i)_{i\in N}, (d_i)_{i\in N})$, where

- N is the set of players;
- $\lambda_i > 0$ is the arrival rate of customers that belong to player i;
- $\mu_i > \lambda_i$ is the capacity endowment of player i;
- $d_i > 0$ is the delay cost incurred for each unit of time a customer of player $i \in N$ spends in the system (i.e., time waiting in the queue and in service).

As mentioned, any coalition M operates an $M/M/1$ queue with total customer arrival rate $\lambda_M = \sum_{i\in M} \lambda_i$ and uses a single exponential server under first-come first-served discipline with service rate $\mu_M = \sum_{i\in M} \mu_i$.[18] In this system, the expected sojourn time (time waiting in the queue and in service) of an arbitrary customer is equal to $1/(\mu_M - \lambda_M)$. By defining the costs of any coalition M as the long-term average delay costs per unit time in the queueing system that only involves the players in M, we obtain the corresponding game (N, c). Since customers for player $i \in M$ arrive at rate λ_i and have to pay d_i per unit of time spent in the system, the corresponding game (N, c) is defined by $c(M) = \sum_{i\in M} \lambda_i d_i/(\mu_M - \lambda_M)$ for all coalitions $M \in 2^N_-$. Note that we used the PASTA property here.

Anily and Haviv (2010) study this setting under symmetric unit delay cost parameters, i.e., $d_i = 1$ for all $i \in N$. They show that, even though the game (N, c) need not be concave, it is balanced: the allocation x that assigns $x_i = \lambda_i/(\mu_N - \lambda_N)$ to each player $i \in N$ is in the core. Subsequently, they investigate the multiplicity and non-negativity of core allocations. They show that, if $|N| > 1$, there are infinitely many core allocations and there always exists a core allocation where all entries are non-negative.

[18] Anily and Haviv (2010) consider a more general model featuring a measure $\alpha \in (0, 1]$ that indicates whether the actual service rate is closer to the arrival rate or the capacity endowment. The service rate for any coalition M in this more general model is given by the geometric mean $\lambda_M^{1-\alpha}\mu_M^{\alpha}$. The measure α, however, lacks a natural interpretation: it is unclear where it would come from and why it would be the same for all coalitions. It also does not provide additional insights. Our formulation corresponds to $\alpha = 1$, which Anily and Haviv (2010) tout as an important special case. It allows a more natural interpretation and a closer fit with the model described in the next subsection.

The stability result of Anily and Haviv (2010) is easily generalized to strict population monotonicity and to asymmetric delay cost parameters. Indeed, if we consider the allocation scheme y that assigns $y_{i,M} = \lambda_i d_i/(\mu_M - \lambda_M)$ to each player $i \in M$ in any coalition M, then for all coalitions $M, L \in 2^N_-$ with $M \subset L$ and $i \in M$ it holds that

$$y_{i,M} = \frac{\lambda_i d_i}{\mu_M - \lambda_M} > \frac{\lambda_i d_i}{\mu_M - \lambda_M + \mu_{L\setminus M} - \lambda_{L\setminus M}} = y_{i,L},$$

where the inequality is valid because, by assumption, $\mu_j > \lambda_j$ for each $j \in L \setminus M$.[19]

We conclude this section with some comments on more complex, alternative single-server queueing games with fixed numbers of servers. Timmer and Scheinhardt (2010) and Anily and Haviv (2011) analyze models that feature several single-server stations in a certain network structure. In these models, the network structure is kept intact (as opposed to pooling capacities into a single station) and the total network capacity is predetermined. The various stations may cooperate by redistributing their combined service capacity or by re-routing arrivals, resulting in a network of $M/M/1$ queues.

3.3.2 $M/M/1$ games with optimized numbers of servers

Yu et al. (2009) study a model that is similar to the one introduced by Anily and Haviv (2010). As before, each player is associated with a customer arrival stream, and every coalition operates an $M/M/1$ queue which serves the union of its members' arrival streams. However, instead of letting the service capacity for each coalition be determined exogenously, Yu et al. (2009) assume that the service rate is optimized. Of course, any coalition desires both low capacity costs and low sojourn times, but these are conflicting objectives. To manage this trade-off, there are two main approaches:

1. Measure the costs for capacity and delay in monetary terms and choose a service rate that minimizes the sum of delay and capacity costs;

2. Stipulate a sojourn time constraint and choose the lowest service rate that satisfies this constraint.

Yu et al. (2009) combine both approaches[20] in a single, overarching model. That is, they

[19] Although this result is new, it is nevertheless placed in the literature review chapter because it is straightforward to derive, because its proof may help understand the original stability result of Anily and Haviv (2010), and because we do not analyze $M/M/1$ situations with fixed service capacity elsewhere in this thesis.

[20] Which of the two approaches should be selected in practice? This often depends on how easy it is to quantify the delay costs. Delay costs may be easy to quantify when customers are external to the service provider (in which case delay costs may correspond to contractually agreed penalties) and when customers are machines or employees that are internal to a service provider (in which case delay costs may be measured in terms of idle time or lost productivity). However, in the absence of such contractual agreements or internal customers, delays may be indirect and hard to quantify. In those cases, service constraints may be more readily adoptable.

consider an *M/M/1 situation with optimized service capacity*, which we will represent as a tuple $(N, (\lambda_i)_{i\in N}, (d_i)_{i\in N}, (\alpha_i)_{i\in N}, (w_i)_{i\in N}, k)$, where

- N is the set of players;
- $\lambda_i > 0$ is the arrival rate of customers that belong to player i;
- $d_i \geq 0$ is the delay cost incurred for each unit of time a customer of player $i \in N$ spends in the system (i.e., time waiting in the queue and in service);
- $\alpha_i \in [0, 1)$ and $w_i > 0$ together make up a sojourn time constraint saying that the time an arbitrary customer of player $i \in N$ spends in the system cannot be larger than w_i time units with probability α_i;
- $k > 0$ is the linear cost rate for service capacity.

The special case of $\alpha_i = 0$ for all $i \in N$ corresponds to a pure cost-based formulation. The special case of $d_i = 0$ for all $i \in N$ corresponds to a pure service constraint model. The case where both $d_i = 0$ and $\alpha_i = 0$ for all $i \in N$ is *not* allowed.

As mentioned, any coalition M operates an $M/M/1$ queue with total customer arrival rate $\lambda_M = \sum_{i\in M} \lambda_i$ and uses a single exponential server under first-come first-serve discipline with an adjustable service rate μ. For any particular choice of this service rate $\mu > \lambda_M$, the expected costs per unit time in steady state are $K_M(\mu) = k\mu + \sum_{i\in M} d_i\lambda_i/(\mu - \lambda_M)$ and the time spent in the system by an arbitrary customer in the queueing system of coalition M is exponentially distributed with parameter $\mu - \lambda_M$. The service rate must satisfy the sojourn time constraint for all members of M, which is done by guaranteeing it for the most "urgent" stream. Straightforward analysis then reveals that a cost minimizing service rate that adheres to the sojourn time constraints is given by

$$\mu_M^* = \lambda_M + \max\left\{\max_{i\in S} \frac{-\ln(1-\alpha_i)}{w_i}, \sqrt{\frac{\sum_{i\in M} d_i\lambda_i}{k}}\right\}.$$

If the first term in the maximization is larger than the second, then the service level constraint is more stringent than the cost considerations. The corresponding game (N, c) is defined by $c(M) = K_M(\mu_M^*)$, which represents the long-term average delay and capacity costs per unit time, for all coalitions $M \in 2_-^N$.

Yu et al. (2009) study these games under the assumption that $w_i = w_j$ for all $i, j \in N$, i.e., they consider a setting where the maximal sojourn time is symmetric. They show that this subclass of games is always balanced, and they identify a cost allocation rule that is in the core. Subsequently, they investigate settings where the delay costs are private information and design a cost allocation mechanism that induces all players to truthfully report their private information.

García-Sanz et al. (2008) study these games under the assumption that $d_i = 0$ for all $i \in N$, i.e., they consider a pure service constraint model. They show that this subclass of games is always concave. Their model is a generalization of the model considered in González and Herrero (2004), who were motivated by a cost allocation problem arising from sharing a medical service in the Spanish health system: the players in their problem correspond to medical procedures. González and Herrero (2004) assume, in addition to $d_i = 0$ for all $i \in N$, that $1 - \alpha_i = 1/e$ for all $i \in N$, which implies (by the exponentially distributed sojourn times) that they consider a model with a constraint on the *mean* sojourn time of an arbitrary customer.

García-Sanz et al. (2008) also study an alternative model that allows mixed[21] preemptive priority disciplines. Their motivation is that the first-come first-served discipline is generally not optimal if the players have specified different sojourn time constraints. To analyze their alternative model, they assume $1 - \alpha_i = 1/e$ for all $i \in N$ (cf. González and Herrero, 2004) and show that the allocation rule which assigns the total costs in proportion to players' arrival rates accomplishes core allocations. They uniquely characterize this rule via several axioms. One of their axioms says that the rule should be such that artificial splitting or merging of the players is non-advantageous. This axiom is defined on a domain that we will refer to as *symmetric M/M/1 situations*, where each element can be represented as a tuple (N, λ, w), with all elements as described before. As all players in such a situation have symmetric sojourn time constraints, the *M/M/1 game* corresponding to such a situation is described by $c(M) = \lambda_M + 1/w$ for each coalition $M \in 2^N_-$.

We conclude this section by pointing out a connection with single-attribute games. For a given sojourn time constraint $w > 0$, consider the non-decreasing function $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\tilde{K}(0) = 0$ and $\tilde{K}(\ell) = \ell + 1/w$ for each $\ell > 0$. Note that $\tilde{K}$ is concave, even affine on $\mathbb{R}_{++}$, and thus elastic. Any single-attribute game (N, c) embedded in $\tilde{K}$, say via attribute vector λ, is an $M/M/1$ game corresponding to symmetric $M/M/1$ situation (N, λ, w). Hence, Theorem 2.5 provides an alternative way to show stability of the proportional allocation and additionally shows that it can be reached through a PMAS.[22]

3.3.3 Multi-server queueing games

The preceding two subsections showed that most of the research on cooperative queueing games has been carried out in the context of the $M/M/1$ model. This model is applicable when service capacity can be easily consolidated into a single server whose service rate can be set at arbitrary levels. The $M/M/1$ model is not appropriate, however, when there are

[21] A *mixed* priority discipline consists of multiplexing a finite set of priority disciplines in such a way that each of them will operate during a desired percentage of time. This makes it possible to further reduce the required service capacity and to meet specific service constraints.

[22] Although the population monotonicity result is new, it is nevertheless placed in the literature review chapter for the same reasons as in Footnote 19.

multiple servers whose service speeds are given. For settings where additional servers have to be acquired to increase the service capacity, the $M/M/s$ model or the $M/G/s/s$ model is more appropriate. The main difference between the $M/M/1$ and the $M/M/s$ model is that the latter has possibly more than one server in parallel. Accordingly, the behavior of the $M/M/1$ model is fundamentally different from the behavior of the $M/M/s$ model; a detailed comparison is made in Section 5.2. The $M/G/s/s$ model differs from the $M/M/1$ model in two additional ways: it disallows waiting in a queue (e.g., because waiting for service is not an acceptable option), while it allows general service time distributions.

Many of the real-life examples described in Section 1.2 can be accurately modeled as Erlang delay or Erlang loss systems, whereas an $M/M/1$ model would not fit. However, almost no research has been carried out on games where coalitions operate a queueing system with possibly more than one parallel server.

Two exceptions are McLean and Sharkey (1993) and Yu et al. (2007). McLean and Sharkey (1993) consider a multi-channel telecommunication model that contains the $M/G/s/s$ model as a special case. They describe several cost allocation mechanisms, including the Shapley value, and illustrate them with numerical examples. However, they do not investigate analytical properties of the underlying game or stability of allocation rules in general. Yu et al. (2007), a previous version of Yu et al. (2009), briefly consider a setting in which each coalition operates an $M/M/s$ queue with an adjustable number of servers. They show balancedness of a game formulated via heavy traffic limit formulas under the assumption that the number of servers is chosen via the square-root safety staffing principe (Borst et al., 2004), a close-to-optimal rule of thumb. However, their results do not apply to settings with non-heavy traffic and/or settings where the number of servers is optimized according to exact formulas.

3.3.4 Contribution to the literature

Based on the preceding subsections, we conclude that there is a lack of knowledge on structural properties of games and cost allocations corresponding to resource pooling in multi-server queueing systems. This is exactly the void that we aim to fill in Chapters 4 and 5, in which we formulate and analyze Erlang loss games and Erlang delay games, respectively. We study these settings under both fixed numbers of servers (the analogue of the model by Anily and Haviv, 2010, as reviewed in Section 3.3.1) and optimized numbers of servers (the analogue of the model by Yu et al., 2009, as reviewed in Section 3.3.2).

While the research that led to these chapters was carried out, Özen et al. (2011) independently formulated and analyzed some of these games as well. Specifically, they use the framework of single-attribute games, which we reviewed in Section 2.3.10, to study balancedness of various cooperative games arising from optimization of capacity or optimization of the amount of demand to serve in Erlang loss or delay systems.

Our games with fixed numbers of servers differ from their games because we endow each

player with two attributes (arrival rate and number of servers), whereas the framework of Özen et al. (2011) calls for a single attribute per player. Erlang loss games with optimized numbers of servers, which we analyze in Chapter 4, are also studied by Özen et al. (2011). Under linear resource costs, they independently derived the same conclusion as we did: the allocation proportional to players' arrival rates is stable. We generalize this result, however, by showing that it remains valid for concave resource costs and for a service constraint formulation. Erlang delay games with optimized numbers of servers, for which in Chapter 5 we prove that they are not balanced in general, are disregarded in the analysis of Özen et al. (2011); they focus on alternative models instead.

3.4 Pooling in queueing and inventory systems

There is a rich literature on resource pooling in queueing and stochastic inventory systems, under the assumption that the system is owned by a single entity who decides whether or not to pool. We will only mention a few works and do not aim to give an exhaustive review.

Considering pooling in queueing systems, Smith and Whitt (1981) were the first to prove that sharing the servers of multiple Erlang loss or Erlang delay systems with identical service time distributions into an aggregate system is always beneficial. For the Erlang delay system, Calabrese (1992) found that combining servers into larger groups, while keeping server utilization constant, leads to reduced congestion, and Benjaafar (1995) provides performance bounds on the effectiveness of resource pooling. A good survey on the analysis, design, and control of queueing systems in general appears in Stidham Jr. (2002). In more practically oriented works, Papier and Thonemann (2008) and De Bruin et al. (2010) use (generalizations of) Erlang loss models to study merging of rental fleets and hospital wards, respectively.

As for pooling in inventory models, Eppen (1979) was the first to explore the benefits of pooling inventories in a simple newsvendor model with normally distributed demands. He shows that pooling leads to a decrease in the required safety stock and that the savings are higher when demands are negatively correlated. Eppen and Schrage (1981) and Corbett and Rajaram (2006) generalize some of the results in Eppen (1979) to positive lead times and arbitrary (non-normal) distributions, respectively.

Inventory pooling has also been considered in infinite-horizon continuous-review inventory models with Poisson demand processes for spare parts under one-for-one replenishments. Kukreja et al. (2001) and Wong et al. (2005) study such a model and show that if there are multiple stock points at one echelon level, it is generally worthwhile to use lateral transshipments between these stock points in order to reduce costs or increase the service level. Alfredsson and Verrijdt (1999), Grahovac and Chakravarty (2001), Wong et al. (2006), and Kranenburg and van Houtum (2009) analyze multi-echelon variants with inventory pooling via lateral transshipments. Wong et al. (2006) and Paterson et al. (2011) give more extensive overviews of the literature on pooling in spare parts inventory models.

All of the above-described papers consider models with a *single* decision maker, i.e., a single player who owns all the resources and is responsible for all the demands or customers. As opposed to this literature, we will consider resource pooling arrangements between independent service providers, each with their own interests, and explicitly address the issue of fair cost allocation.

—It's only after we've lost everything that we're free to do anything.

Tyler Durden

4

Erlang loss games

4.1 Introduction

In this chapter, based on Karsten et al. (2011a) and Karsten et al. (2012), we formulate and analyze games—called Erlang loss games—wherein an arbitrary number of service providers face exogenous Poisson streams of customer arrivals for whom waiting is not an option. The service providers are allowed to collaborate by completely sharing their servers, combining their individual customer streams into a pooled resource system whose steady-state behavior is equivalent to that of an Erlang loss system. The main benefit of such resource pooling is a reduction in the number of blocked customers that find no free server upon arrival (Smith and Whitt, 1981). As we describe in more detail in Subsection 4.3.1, the Erlang loss model has applications to a broad range of service systems in practice. In a queueing context, one may think of clinical departments that pool hospital beds or car rental agencies that pool rental cars. In an inventory context, one may think of high-tech companies that pool low-demand, expensive spare parts whose stock-outs are dealt with via emergency procedures.

Costs consist of penalty costs for blocked customers and concave resource costs for servers. We consider two settings: one where players' numbers of servers are exogenously given and one in which every coalition can pick an optimal number of servers. The former setting leads to a corresponding game with fixed numbers of servers. The latter setting leads to two games: one based on a cost formulation and the other on a service formulation.[23] As we will see,

[23]These two games correspond to the two main ways to manage the trade-off between resource and penalty costs, as discussed in Subsection 3.3.2.

there is no obvious relationship between the (cost-based) games of the two different settings. Nevertheless, we are able to prove interesting results under the assumption that penalty costs (or naturally chosen service level constraints) are symmetric. First and foremost, we show that all three games have a non-empty core. Moreover, when the number of servers may be optimized, we identify a specific allocation in the core: the division of the total system-wide costs proportional to the players' arrival rates.

One important aspect is that the number of servers has to be an integer at all times. Indeed, resources such as spare parts or service engineers can only be varied in discrete amounts. Yet, the difference between placing one or two spare parts in a warehouse or between employing one or two service engineers in a repair shop can be massive, and as a result, the integrality requirement can be a pivotal obstacle. Not from an optimization perspective—we can simply choose the integer number of servers that yields minimal costs. But from a game analysis perspective—when comparing coalitions, it could happen that, say, one spare part is sufficient for two companies together while two spare parts are needed for three companies. This discrete number of resources, which may differ across coalitions, results in analytical complications. Essentially, structural properties such as the stability of a cost allocation rule come down to inequalities involving the classic Erlang loss function (which represents the steady-state blocking probability in an Erlang loss system, i.e., the probability that an arriving customer finds no free server), and while proving such inequalities is challenging within the confines of integrality, it becomes doable when non-integral numbers of servers are allowed.

For these technical reasons, we focus on *extensions* of the Erlang loss function (i.e., functions that extend the domain of the Erlang loss function to non-integral numbers of servers) and derive structural properties of such extensions. In particular, we prove that several such extensions satisfy a property that we call *scalability*. Scalability means that the blocking probability as described by the extension becomes no larger when the offered load a and the number of servers s are scaled by the same relative amount, even when scaling up from integral s to non-integral s. Most importantly, we prove scalability of the linear interpolation, which signifies that the blocking probability in a system with $a = 2$ and $s = 3$ is higher than the blocking probability in a system with $a = 3$ that uses 4 servers half of the time and 5 servers the other half, assuming steady-state conditions are instantaneously achieved under both, with zero mixing times. These results may be relevant beyond the context of our games, as extensions of the Erlang loss function have seen applications in other fields as well. For example, loss systems with overflow layers come to mind. Indeed, techniques such as the equivalent random method (Wilkinson, 1956) and Hayward's approximation (Fredericks, 1980) for performance approximations of such systems also employ extensions of the Erlang loss function. However, they mainly allow us to prove positive results on balancedness and stability.

These positive results are valid under symmetric penalty costs. If penalty costs are *not*

symmetric and full pooling of all servers is mandated, then the corresponding games are no longer even guaranteed to be subadditive. The reason for this, as we will show, lies in our assumption of full pooling: Full pooling under asymmetric penalty costs has two, possibly conflicting, effects. On the beneficial side, full pooling always reduces the number of customers that find all servers occupied upon arrival. On the adverse side, full pooling may allow a low-penalty customer to occupy a server that could have been saved for a high-penalty one. We disentangle these two effects to provide general conditions for balancedness of Erlang loss games with fixed numbers of servers. Although the effects of fully pooling customer streams with different *service times* into a single Erlang loss system have been well studied (see, e.g., Smith and Whitt, 1981; Papier and Thonemann, 2008), to our knowledge the effects of fully pooling customer streams with different penalty costs have not been investigated before, let alone in a cooperative framework. At the end of this chapter, we briefly consider partial pooling approaches in which some servers are held back for more "critical" customer classes. As it turns out, a game that was not balanced under full pooling may become balanced under an optimal "partial" pooling policy.

Altogether, the model, results, and analysis we present in this chapter make three primary contributions. First, we derive several new analytical properties of extensions of the classic Erlang loss function. Second, we introduce Erlang loss games with fixed numbers of servers and prove that they are totally balanced if the penalty costs are symmetric among players. Third, we show that Erlang loss games with optimized numbers of servers lack an obvious relationship to the ones with fixed numbers of servers, but nevertheless admit a proportional core allocation if penalty costs are symmetric among players; this holds for both the cost-based and the service-based variant.

The structure of this chapter is as follows. We start in Section 4.2 by analyzing the extensions of the Erlang loss function. Then, in Section 4.3, we describe our basic model. Sections 4.4 and 4.5 focus on the games with fixed and optimized numbers of servers, respectively. In Section 4.6, we describe a more general model that allows for optimization of the pooling policy. Finally, we draw conclusions and suggest directions for future research in Section 4.7.

4.2 Scalability of extensions of the Erlang loss function

In this section, we define the classic Erlang loss function, subsequently introduce three extensions (including linear interpolation) that extend its domain to non-integral numbers of servers, and finally show that all three extensions satisfy the scalability property.

4.2.1 The classic Erlang loss function

The *Erlang loss function* $\hat{B} : \mathbb{N}_0 \times \mathbb{R}_{++} \to [0,1]$ is a classic in queueing theory. It was first published by Erlang (1917) and is defined by

$$\hat{B}(s,a) = \frac{a^s/s!}{\sum_{y=0}^{s} a^y/y!} \qquad \text{for any } s \in \mathbb{N}_0 \text{ and } a \in \mathbb{R}_{++}. \tag{4.1}$$

The value $\hat{B}(s,a)$ may be interpreted as the steady-state probability that an arriving customer finds no free server in an Erlang loss system with s servers and offered load a. Recall that, in such a system, the service times are i.i.d. according to some general distribution function with finite and positive mean, say τ, and that customers arrive according to a Poisson process. The arrival rate will typically be denoted by λ. Given λ and τ, the offered load a is equal to $\lambda\tau$. Since customers that find no free server upon arrival are often called *blocked* customers, $\hat{B}(s,a)$ is also sometimes referred to as the *blocking probability*.

The Erlang loss function satisfies several useful properties. The following three theorems collect several properties that will be important in our analysis.

Theorem 4.1. *Let $s, s' \in \mathbb{N}$ and $a, \tau, \lambda, \lambda' \in \mathbb{R}_{++}$. Then,*

(i) $\hat{B}(s, \lambda/\mu)$ *is decreasing and convex in* μ *for* μ *on* $\mathbb{R}_{++}$.
(ii) $D_2\hat{B}(s,a) = [\hat{B}(s,a) - 1 + s/a] \cdot \hat{B}(s,a)$.[24]
(iii) $\hat{B}(s,a) = a\hat{B}(s-1,a)/[a\hat{B}(s-1,a) + s]$.
(iv) If $s < s'$, *then* $\hat{B}(s,a) > \hat{B}(s',a)$.
(v) $\hat{B}(s+s', (\lambda+\lambda')\tau) \cdot (\lambda+\lambda') \leq \hat{B}(s,\lambda\tau)\cdot\lambda + \hat{B}(s',\lambda'\tau)\cdot\lambda'$.
(vi) $\hat{B}(s, \bar{\lambda}\tau)\bar{\lambda}$ *is increasing and convex in* $\bar{\lambda}$ *for* $\bar{\lambda}$ *on* $\mathbb{R}_{++}$.

Property (i) corresponds to Proposition 3 in Harel (1990). The partial derivative with respect to the load, (ii), is due to Theorem 15 in Jagerman (1974). The recursive relation of property (iii) is well-known; see, e.g., Jagerman (1974, p. 531). Property (iv) is also well-known; see, e.g., Whitt (2002). The subadditivity property, (v), is due to Theorem 1 in Smith and Whitt (1981). Finally, property (vi) is due to Krishnan (1990).

Theorem 4.2. *For all $s \in \mathbb{N}$ and $a \in \mathbb{R}_{++}$, $a[\hat{B}(s,a)]^2 - 1 - (a-s-1)\hat{B}(s,a) \leq 0$.*[25]

Proof. Let $S \in \mathbb{N}$ and $a \in \mathbb{R}_{++}$. By Equation (4.1), what we aim to show is that

$$a\left[\frac{a^S/S!}{\sum_{y=0}^{S} a^y/y!}\right]^2 - 1 - (a-S-1)\frac{a^S/S!}{\sum_{y=0}^{S} a^y/y!} \leq 0. \tag{4.2}$$

[24] The operator D_2 represents differentiation with respect to the second argument. We avoid the more standard notation $\delta/\delta a$ because it may lead to confusion when we swap arguments in a proof later on.

[25] This theorem was independently derived by Özen et al. (2011): it corresponds to their Inequality (19). Their proof is by contradiction, whereas our alternative proof is direct.

Re-arranging, combining all terms into a single fraction, and multiplying both sides of the resulting inequality with $-\left(\sum_{y=0}^{S} a^y/y!\right)^2 < 0$, we obtain that Inequality (4.2) is equivalent to

$$\left(\sum_{y=0}^{S} \frac{a^y}{y!}\right)^2 - a \cdot \left(\frac{a^S}{S!}\right)^2 + (a - S - 1) \cdot \left(\sum_{y=0}^{S} \frac{a^y}{y!}\right) \cdot \frac{a^S}{S!} \geq 0. \tag{4.3}$$

For all $s \in \mathbb{N}$, define

$$f(s) = \left(\sum_{y=0}^{s} \frac{a^y}{y!}\right)^2 - \frac{a^{2s+1}}{s! \cdot s!} + \sum_{y=0}^{s} \frac{a^{y+s} \cdot (a - s - 1)}{y! \cdot s!}. \tag{4.4}$$

Notice that Inequality (4.3) corresponds to $f(S) \geq 0$. To complete the proof, we show that $f(s) \geq 0$ for all $s \in \mathbb{N}$ by induction. Firstly,

$$f(1) = (1+a)^2 - a^3 + (a + a^2)(a - 2) = (1+a)^2 - a^3 + a^2 - 2a + a^3 - 2a^2 = 1 \geq 0.$$

To avoid empty summations later on, it is convenient to treat the case $s = 2$ separately:

$$\begin{aligned} f(2) &= \left(\frac{1}{2}a^2 + a + 1\right)^2 - \frac{1}{4}a^5 + \left(\frac{1}{4}a^4 + \frac{1}{2}a^3 + \frac{1}{2}a^2\right) \cdot (a - 3) \\ &= \frac{1}{4}a^4 + a^3 + 2a^2 + 2a + 1 - \frac{1}{4}a^5 + \frac{1}{4}a^5 + \frac{1}{2}a^4 + \frac{1}{2}a^3 - \frac{3}{4}a^4 - \frac{3}{2}a^3 - \frac{3}{2}a^2 \\ &= \frac{1}{2}a^2 + 2a + 1 \geq 0. \end{aligned}$$

For the induction step, let $s \in \{3, 4, \ldots\}$ and assume that $f(s-1) \geq 0$. Then,

$$\begin{aligned} f(s) &= \left(\sum_{y=0}^{s} \frac{a^y}{y!}\right)^2 - \frac{a^{2s+1}}{s! \cdot s!} + \sum_{y=0}^{s} \frac{a^{y+s} \cdot (a - s - 1)}{y! \cdot s!} \\ &= \left(\sum_{y=0}^{s} \frac{a^y}{y!}\right)^2 - \frac{a^{2s+1}}{s! \cdot s!} + \frac{a^{2s+1}}{s! \cdot s!} - \frac{a^{2s} \cdot s}{s! \cdot s!} - \frac{a^{2s}}{s! \cdot s!} + \sum_{y=0}^{s-1} \frac{a^{y+s} \cdot (a - s - 1)}{y! \cdot s!} \\ &= \frac{a^{2s}}{s! \cdot s!} + 2 \cdot \sum_{y=0}^{s-1} \frac{a^{y+s}}{y! \cdot s!} + \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 - \frac{a^{2s} \cdot s}{s! \cdot s!} - \frac{a^{2s}}{s! \cdot s!} + \sum_{y=0}^{s-1} \frac{a^{y+s} \cdot (a - s - 1)}{y! \cdot s!} \\ &= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 - \frac{a^{2s} \cdot s}{s! \cdot s!} + \sum_{y=0}^{s-1} \frac{a^{y+s} \cdot (a - s + 1)}{y! \cdot s!} \\ &= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[-\frac{a^s \cdot s}{s!} + \sum_{y=1}^{s} \frac{a^y}{(y-1)!} + \sum_{y=0}^{s-1} \frac{a^y \cdot (-s+1)}{y!}\right] \cdot \frac{a^s}{s!} \\ &= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[\sum_{y=1}^{s-1} \frac{a^y}{(y-1)!} + \sum_{y=0}^{s-1} \frac{a^y \cdot (-s+1)}{y!}\right] \cdot \frac{a^s}{s!} \end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[\sum_{y=1}^{s-2} \frac{a^y}{(y-1)!} + \sum_{y=0}^{s-2} \frac{a^y \cdot (-s+1)}{y!}\right] \cdot \frac{a^s}{s!} \\
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[\left(\sum_{y=1}^{s-2} \frac{a^y}{y!} \cdot (y-s+1)\right) - s + 1\right] \cdot \frac{a^s}{s!} \\
&\geq \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[\sum_{y=1}^{s-2} \frac{a^y}{y!} \cdot \left(\frac{-s^2}{y+1} + s\right) - s + 1\right] \cdot \frac{a^s}{s!} \\
&\geq \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[\sum_{y=1}^{s-2} \frac{a^y}{y!} \cdot \left(\frac{-s^2}{y+1} + s\right) - s^2 + s - \frac{s^2}{a}\right] \cdot \frac{a^s}{s!} \\
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[\sum_{y=0}^{s-2} \frac{a^y}{y!} \cdot \left(\frac{-s^2}{y+1} + s\right) - \frac{s^2}{a}\right] \cdot \frac{a^s}{s!} \\
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[-\frac{a^{s-1} \cdot s}{(s-1)!} - \sum_{y=0}^{s-2} \frac{a^y \cdot s^2}{(y+1)!} - \frac{s^2}{a} + \sum_{y=0}^{s-2} \frac{a^{y-1} \cdot a \cdot s}{y!} + \frac{a^{s-1} \cdot s}{(s-1)!}\right] \cdot \frac{a^s}{s!} \\
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[-\frac{a^{s-1} \cdot s}{(s-1)!} - \sum_{y=1}^{s-1} \frac{a^{y-1} \cdot s^2}{y!} - \frac{s^2}{a} + \sum_{y=0}^{s-1} \frac{a^{y-1} \cdot a \cdot s}{y!}\right] \cdot \frac{a^s}{s!} \\
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 + \left[-\frac{a^{s-1} \cdot s}{(s-1)!} + \sum_{y=0}^{s-1} \frac{a^{y-1} \cdot (a-s) \cdot s}{y!}\right] \cdot \frac{a^s}{s!} \\
&= \left(\sum_{y=0}^{s-1} \frac{a^y}{y!}\right)^2 - \frac{a^{2(s-1)+1}}{(s-1)! \cdot (s-1)!} + \sum_{y=0}^{s-1} \frac{a^{y+s-1} \cdot (a-(s-1)-1)}{y!(s-1)!} \\
&= f(s-1) \geq 0.
\end{aligned}
$$

In the first couple of steps, we split up summations, split off terms from summations, and cancel out common terms. The first inequality is valid because, for all $y \in \mathbb{N}_0$, it holds that $s^2 - 2s(y+1) + (y+1)^2 = (s-(y+1))^2 \geq 0$, and thus $y - s + 1 \geq -s^2/(y+1) + s$. The second inequality holds because $-s + 1 \geq -s^2 + s \geq -s^2 + s - s^2/a$, since $s > 1$. In the final steps, we rearrange our expression to show that it is equal to $f(s-1)$, which was non-negative by the induction hypothesis. Hence, by the principle of mathematical induction we have for all $s \in \mathbb{N}$ that $f(s) \geq 0$. This completes the proof. □

Theorem 4.3. *The Erlang loss function is subhomogeneous of degree zero.*

Theorem 4.3 is attributed to Paul Burke in the appendix of Smith and Whitt (1981). Subhomogeneity of degree zero was defined in Section 2.2.2; it means that for any $s \in \mathbb{N}$ and $a > 0$, $\hat{B}(ts, ta)$ is decreasing in t for $t \in \{1/s, 2/s, \ldots\}$. This captures the economies of scale

in Erlang loss systems: when we increase the offered load a and the number of servers s with the same relative amount t, the blocking probability always decreases. This scaling occurs, for instance, when we combine the servers and arrival streams of two Erlang loss systems with the same a/s value into a single joint system. Note that this is only meaningful when the scaling factor t is chosen such that the number of servers in the scaled system, ts, is an integer number. As an example, the blocking probability in a system with $a = 2$ and $s = 3$ is higher than in a system with $a = 4$ and $s = 6$.

To prove that various Erlang loss games have a non-empty core, we will frequently use the notion of balancedness. In doing so, we encounter fractional-valued balanced maps and, accordingly, we want to be able to construct Erlang loss systems with possibly fractional numbers of servers. In other words, it will be convenient to be able to scale up to an Erlang loss system with, say, $a = 3$ and $s = 4.5$. But 4.5 is not an admissible number of servers. Indeed, the Erlang loss function is not defined for non-integral numbers of servers. A seemingly obvious way to deal with this problem would be to interpret the blocking probability via an *extension*, i.e., a function that extends the domain of the Erlang loss function to non-integral numbers of servers and that coincides with the Erlang loss function whenever the number of servers is an integer. However, for such an extension to be helpful, it needs to satisfy several properties, such as scalability or convexity in the number of servers. These properties are not satisfied by all extensions.

4.2.2 Extensions of the Erlang loss function

As mentioned, we are interested in functions that extend the domain of the Erlang loss function to non-integral numbers of servers. Formally, we call any function $E : \mathbb{R}_+ \times \mathbb{R}_{++} \to [0, 1]$ an *extension of the Erlang loss function* (or *extension* for short) if $E(s, a) = \hat{B}(s, a)$ for all $s \in \mathbb{N}_0$ and $a \in \mathbb{R}_{++}$. We next introduce three different extensions; all are depicted in Figure 4.1.

First, the continuous extension $B : \mathbb{R}_+ \times \mathbb{R}_{++} \to [0, 1]$ is defined by

$$B(s, a) = \left(a \int_0^\infty e^{-ax}(1 + x)^s dx \right)^{-1} \quad \text{for all } s \in \mathbb{R}_+ \text{ and } a \in \mathbb{R}_{++}. \tag{4.5}$$

This function, which is related to the Erlang loss function via the Gamma function, is indeed an extension (Jagerman, 1974). It is one of the most commonly used extensions in the literature (Fredericks, 1980).

The extension B satisfies several useful properties. First, as shown by Jagers and van Doorn (1986), it is convex in the number of servers.

Theorem 4.4. *For each fixed $a > 0$, $B(s, a)$ is a convex function of s for s on $\mathbb{R}_+$.*

As stated in the next theorem, B is subhomogeneous of degree zero, which may be viewed as a generalization of Theorem 4.3. The proof of this theorem is based on an argument in the

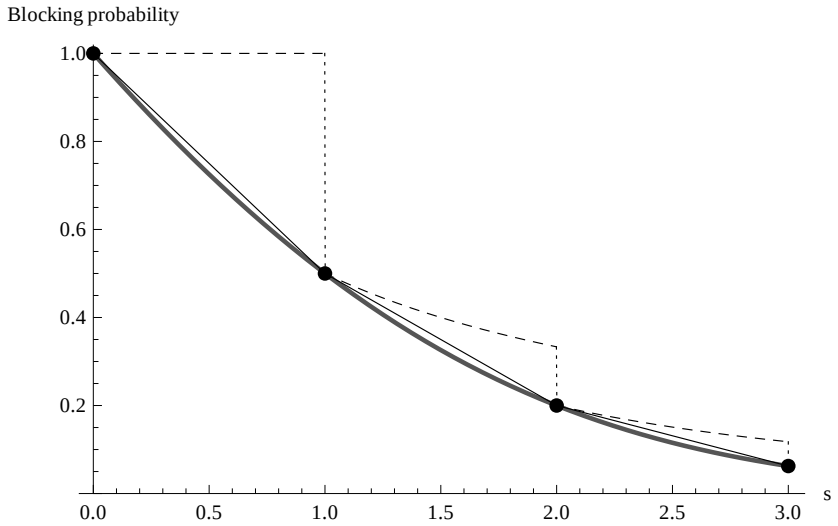

Figure 4.1: *The extensions $X(s,a)$ (dashed), $L(s,a)$ (middle), and $\hat{B}(s,a)$ (thick) for $a=1$ fixed.*

appendix of Smith and Whitt (1981), simultaneously taking out some of their inaccuracies w.r.t. strictness of inequalities.

Theorem 4.5. *The extension B is subhomogeneous of degree zero. More specifically, for any $a \in \mathbb{R}_{++}$ and $s \in \mathbb{R}_{+}$, the following holds: if $s > 0$, then $B(ts,ta)$ is decreasing in t for t on $\mathbb{R}_{++}$; if $s = 0$, then $B(ts,ta) = 1$ for each $t > 0$.*

Proof. Let $a > 0$. First, let $s \in (0,\infty)$. By letting $w = tax$ in Equation (4.5), we obtain for any $t > 0$ that

$$\begin{aligned} B(ts,ta) &= \left(ta \int_0^\infty e^{-tax}(1+x)^{ts} dx \right)^{-1} \\ &= \left(\int_0^\infty e^{-w} \left(1 + \frac{w}{ta}\right)^{ts} dw \right)^{-1}. \end{aligned} \tag{4.6}$$

For $w > 0$ and $t > 0$, it holds that

$$\begin{aligned} \frac{d}{dt}\left(1+\frac{w}{ta}\right)^{ts} &= \frac{d}{dt} e^{ts\cdot\ln(1+w/(ta))} \\ &= e^{ts\cdot\ln(1+w/(ta))} \cdot \left[s\ln\left(1+\frac{w}{ta}\right) + ts \cdot \frac{1}{1+w/(ta)} \cdot \frac{-w}{t^2 a} \right] \\ &= \left(1+\frac{w}{ta}\right)^{ts} \cdot s \cdot \left[\ln\left(1+\frac{w}{ta}\right) - \frac{w}{ta+w} \right] > 0, \end{aligned}$$

where the inequality is valid because, for all $h > 0$,

$$\ln(1+h) = \int_1^{1+h} \frac{1}{k} dk = \int_0^h \frac{1}{1+k} dk > \int_0^h \frac{1}{1+h} dk = \frac{h}{1+h}.$$

We conclude that for t on $\mathbb{R}_{++}$, the integral in Equation (4.6) is increasing in t, which implies that $B(ts, ta)$ is decreasing in t.

Next, let $s = 0$. Then, for any $t > 0$, it holds that $B(ts, ta) = B(0, ta) = \hat{B}(0, ta) = 1$ since B is an extension. □

Another extension is obtained by using linear interpolation (cf. Kortanek et al., 1981). To be more precise, this piecewise linear function $L : \mathbb{R}_+ \times \mathbb{R}_{++} \to [0, 1]$ is defined by

$$L(s, a) = (1 - (s - \lfloor s \rfloor)) \cdot \hat{B}(\lfloor s \rfloor, a) + (s - \lfloor s \rfloor) \cdot \hat{B}(\lceil s \rceil, a) \quad \text{for all } s \in \mathbb{R}_+ \text{ and } a \in \mathbb{R}_{++} \tag{4.7}$$

where $\lceil s \rceil$ denotes the smallest integer larger than or equal to s, and $\lfloor s \rfloor$ denotes the largest integer smaller than or equal to s. By virtue of being a linear interpolation, this function is obviously an extension of the Erlang loss function.

The question of whether or not an extension satisfies various properties is particularly interesting for the linear interpolation L, as this interpolation represents a long-run blocking probability that is actually achievable by "mixing" between two consecutive integer numbers of servers. That is, operating under each of those two numbers of servers during a desired percentage of time. This "mixing" (cf. van Houtum and Zijm, 2000) sometimes occurs in practice to meet a service constraint exactly. The scalability of L is investigated in the next subsection.

Finally, we introduce a new extension of the Erlang loss function $X : \mathbb{R}_+ \times \mathbb{R}_{++} \to [0, 1]$, which is defined by

$$X(s, a) = \begin{cases} \hat{B}(\lfloor s \rfloor, a \cdot \lfloor s \rfloor / s) & \text{if } s \geq 1 \text{ and } a \in \mathbb{R}_{++}; \\ 1 & \text{if } s \in [0, 1) \text{ and } a \in \mathbb{R}_{++}. \end{cases} \tag{4.8}$$

This function is not continuous, but it is clearly an extension by definition. We use this extension to prove scalability of L in the next section. As illustrated in Figure 4.2, for any $s \in \mathbb{R}_+$ and $a \in \mathbb{R}_{++}$, it holds that $X(ts, ta)$ as a function of t is stepwise constant and non-increasing for t on $\mathbb{R}_{++}$; indeed, if $s > 0$, then for t between two successive values $t^- \in \{1/s, 2/s, \ldots\}$ and $t^+ = t^- + 1/s$, $X(ts, ta)$ equals $\hat{B}(st^-, at^-)$. Accordingly, it is easy to prove that X satisfies subhomogeneity of degree zero.

Theorem 4.6. *The extension X is subhomogeneous of degree zero.*

Proof. Let $s \in \mathbb{R}_+$, $a \in \mathbb{R}_{++}$, and $t > 1$. If $s \in [0, 1)$, then $X(s, a) = 1 \geq X(ts, ta)$. Otherwise, if $s \geq 1$, then $X(s, a) = \hat{B}(\lfloor s \rfloor, a \cdot \lfloor s \rfloor / s) \geq \hat{B}(\lfloor ts \rfloor, a \cdot \lfloor ts \rfloor / s) = X(ts, ta)$, where the inequality holds by Theorem 4.3. □

4.2.3 Scalability of the linear interpolation

In this subsection, we aim to show that the linear interpolation L satisfies the scalability property. This property will prove instrumental in analyzing Erlang loss games later on. We start by defining it formally.

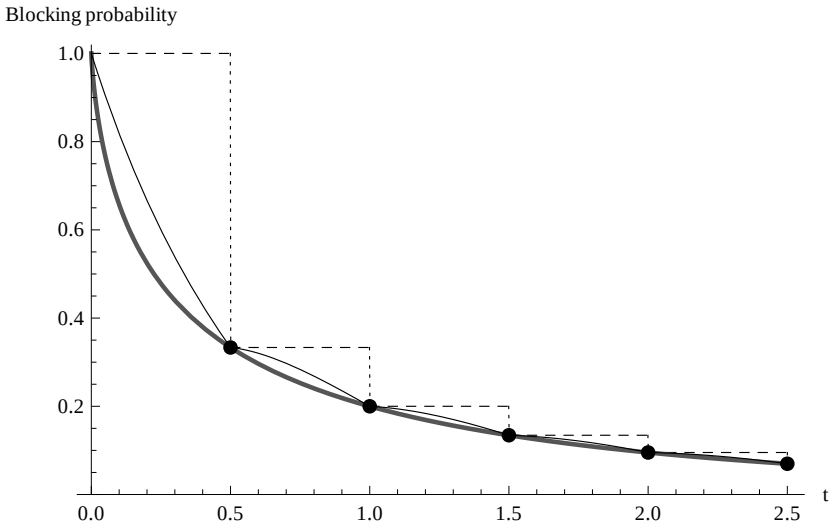

Figure 4.2: *For $s = 2$ and $a = 1$ fixed, the extensions $X(ts, ta)$ (dashed), $L(ts, ta)$ (middle), and $\hat{B}(ts, ta)$ (thick) as functions of t.*

Definition 4.1. An extension E is said to be *scalable* if, for each $s \in \mathbb{N}_0$ and $a \in \mathbb{R}_{++}$, $E(ts, ta) \leq E(s, a)$ for all $t \in (1, \infty)$ and $E(ts, ta) \geq E(s, a)$ for all $t \in (0, 1)$.

In words, scalability means that if we take an Erlang loss system with an *integer* number of servers as our starting point, an increase (resp. decrease) of the offered load and the number of servers with the same relative amount will result in a blocking probability that is no larger (resp. no smaller) than before. If this scaling would result in a system with a *nonintegral* number of servers, then the blocking probability is described by the extension.

Not every extension is scalable. A sufficient condition for scalability of an extension E is subhomogeneity of degree zero. (This goes beyond scalability by additionally comparing two systems that *both* have a real number of servers.) So, scalability of the extensions B and X follow as straightforward corollaries to Theorems 4.5 and 4.6, respectively.

To prove scalability of L, it will be convenient to show that the graph of L never dips below the graph of B and never jumps above the graph of X, as illustrated in Figures 4.1 and 4.2. The next theorem tackles the first of these two properties.

Theorem 4.7. *$B(s, a) \leq L(s, a)$ for all $s \in \mathbb{R}_+$ and $a \in \mathbb{R}_{++}$.*

Proof. Let $a \in \mathbb{R}_{++}$. Since, by Theorem 4.4, $B(s, a)$ as a function of s is convex for s on $\mathbb{R}_+$, whereas $L(s, a)$ linearly interpolates between points on the graph of $\hat{B}(s, a)$ at which $\hat{B}(s, a) = B(s, a)$, the desired inequality follows immediately from the definition of convexity. □

In the process of proving that the graph of L always lies at or below the graph of X, we will present two lemmas that consider X and L as a function of the number of servers on a

domain restricted to an interval between two consecutive integers. To describe this formally, it will be convenient to introduce two restricted functions: for any fixed $r = (S, a) \in \mathbb{N} \times \mathbb{R}_{++}$, we define the functions L_r and X_r, both mapping $[S, S+1)$ to $[0, 1]$, by $L_r(s) = L(s, a)$ and $X_r(s) = X(s, a)$ for all $s \in [S, S+1)$. So, these functions are described on their domain by

$$X_r(s) = \hat{B}(S, aS/s); \tag{4.9}$$

$$L_r(s) = (1 + S - s) \cdot \hat{B}(S, a) + (s - S) \cdot \hat{B}(S + 1, a). \tag{4.10}$$

Figure 4.1 on page 62 provides an illustration: there, the graph of $X_{(1,1)}$ corresponds to the dashed curve on $[1, 2)$ and the graph of $L_{(1,1)}$ corresponds to the middle curve on $[1, 2)$. So, X_r and L_r allow us to consider the functions X and L when the number of servers may only vary between two consecutive integers S and $S + 1$. The following lemma states two useful properties of X.

Lemma 4.8. *Let $r = (S, a) \in \mathbb{N} \times \mathbb{R}_{++}$. Then X_r is decreasing and convex on its domain $[S, S + 1)$.*

Proof. By Part (i) in Theorem 4.1, it holds for each fixed $\hat{s} \in \mathbb{N}$ and $\lambda \in \mathbb{R}_{++}$ that $\hat{B}(\hat{s}, \lambda/\mu)$ is decreasing and convex in μ for $\mu \in [S, S + 1)$. By substituting $\hat{s} = S$, $\lambda = aS$, and $\mu = s$, we conclude that $X_r(s) = \hat{B}(S, aS/s)$ is decreasing and convex in s on $[S, S + 1)$. □

In contrast to X and L, the functions X_r and L_r are differentiable, which allows us to compare their derivatives, evaluated at S, in the following lemma.

Lemma 4.9. *Let $r = (S, a) \in \mathbb{N} \times \mathbb{R}_{++}$. Then $L'_r(S) \leq X'_r(S)$.*

Proof. First of all, the derivative of L_r for any $s \in [S, S + 1)$ is

$$L'_r(s) = \hat{B}(S + 1, a) - \hat{B}(S, a). \tag{4.11}$$

To obtain the derivative of X_r, we combine Part (ii) of Theorem 4.1 with Equation (4.9) to derive that for any $s \in [S, S + 1)$,

$$\begin{aligned} X'_r(s) &= [\hat{B}(S, aS/s) - 1 + S/(aS/s)] \cdot \hat{B}(S, aS/s) \cdot \left(-aS/s^2\right) \\ &= [\hat{B}(S, aS/s) - 1 + s/a] \cdot \hat{B}(S, aS/s) \cdot \left(-aS/s^2\right). \end{aligned} \tag{4.12}$$

For notational ease, let $\mathfrak{B} = \hat{B}(S, a)$. Evaluating the derivatives (4.11) and (4.12) at $s = S$,

we obtain

$$\begin{aligned} L_r'(S) - X_r'(S) &= \hat{B}(S+1,a) - \mathfrak{B} - \left[\mathfrak{B} - 1 + \frac{S}{a}\right] \cdot \mathfrak{B} \cdot \frac{-a}{S} \\ &= \frac{a\mathfrak{B}}{a\mathfrak{B}+S+1} - \mathfrak{B} - \left[\mathfrak{B} - 1 + \frac{S}{a}\right] \cdot \mathfrak{B} \cdot \frac{-a}{S} \\ &= \mathfrak{B} \cdot \left[\frac{a}{a\mathfrak{B}+S+1} + (\mathfrak{B}-1) \cdot \frac{a}{S}\right] \\ &= \frac{a\mathfrak{B}}{S(a\mathfrak{B}+S+1)} \cdot \Big[S + (\mathfrak{B}-1) \cdot (a\mathfrak{B}+S+1)\Big] \\ &= \frac{a\mathfrak{B}}{S(a\mathfrak{B}+S+1)} \cdot \Big[a\mathfrak{B}^2 - 1 - (a-S-1)\mathfrak{B}\Big] \\ &\leq 0. \end{aligned}$$

The second equality holds by Part (*iii*) of Theorem 4.1. The other equalities hold by rewriting. The inequality holds because $\mathfrak{B} > 0$, $a > 0$, $S(a\mathfrak{B}+S+1) > 0$, and $a\mathfrak{B}^2 - 1 - (a-S-1)\mathfrak{B} \leq 0$, where the last-mentioned inequality holds by Theorem 4.2. We conclude that $L_r'(S) \leq X_r'(S)$. □

We use these lemmas to prove that the graph of L never jumps above the graph of X.

Theorem 4.10. *$X(s,a) \geq L(s,a)$ for all $s \in \mathbb{R}_+$ and $a \in \mathbb{R}_{++}$.*

Proof. Let $s \in \mathbb{R}_+$ and $a \in \mathbb{R}_{++}$. We distinguish two cases.

Case 1: $s < 1$. Then, by definition $X(s,a) = 1$, whereas $L(s,a) \leq 1$.

Case 2: $s \geq 1$. Then, we denote $S = \lfloor s \rfloor$ and consider the functions $X_{(S,a)}$ and $L_{(S,a)}$, which are described on their domain $[S, S+1)$ by Equations (4.9) and (4.10). First of all, observe that $X_{(S,a)}(S) = L_{(S,a)}(S)$ since both X and L are extensions of the Erlang loss function. Secondly, by Lemma 4.9, we observe that the derivative of $L_{(S,a)}$ at S does not exceed the derivative of $X_{(S,a)}$ at S. Thirdly, $X_{(S,a)}$ is convex by Lemma 4.8, whereas $L_{(S,a)}$ is by definition a linear function, which (together with the second observation) implies that $X'_{(S,a)} \geq L'_{(S,a)}$ on $[S, S+1)$. Combining these three observations yields $X_{(S,a)}(s) \geq L_{(S,a)}(s)$. We conclude that $X(s,a) \geq L(s,a)$. □

With the various theorems derived so far, we can now prove the following theorem.

Theorem 4.11. *The extension L is scalable.*

Proof. Let $s \in \mathbb{N}_0$ and $a \in \mathbb{R}_{++}$. For any $t \in (1, \infty)$, we find that

$$L(ts,ta) \leq X(ts,ta) \leq X(s,a) = L(s,a),$$

where the first inequality holds by Theorem 4.10, the second inequality holds by Theorem 4.6, and the equality holds because both X and L are extensions. Analogously, for any $t \in (0,1)$, we find that

$$L(ts,ta) \geq B(ts,ta) \geq B(s,a) = L(s,a),$$

where the first inequality holds by Theorem 4.7, the second inequality holds by Theorem 4.5, and the equality holds because both B and L are extensions. Hence, L is scalable. □

We remark that it remains an open question whether or not L satisfies the stronger property of subhomogeneity of degree zero. We have not been able to find a counterexample, but at the same time our proof approach for Theorem 4.11 is not readily adaptable for subhomogeneity of degree zero. Nevertheless, scalability is sufficient for the purpose of proving results in Section 4.5.

4.3 Model description

In this section, we introduce our model for sharing an Erlang loss system between a number of players. In Section 4.3.1, we start with an preliminary model which features a simplified cost structure—kept simple to allow illustration of the modeling approach with simple examples. In Section 4.3.2, we describe the final model which features a more general cost structure.

4.3.1 The basic setup

Consider a number of players who require certain costly servers for their customer populations. Let N denote the set of players. Each player's customers arrive according to mutually independent Poisson processes, with rate $\lambda_i > 0$ for player i. Service times (for an arbitrary customer of any player) are i.i.d. according to some general distribution function with finite, positive mean τ. Each newly arriving customer immediately goes into service if there is an unoccupied server available. Conversely, if that customer finds no free server, he is lost and penalty costs $p > 0$ have to be paid. Besides these penalty costs, there are resource costs for servers: $h > 0$ per server per unit of time. We assume throughout that players are interested in their long-term average costs per time unit.

Any coalition M may collaborate by operating an Erlang loss system that serves the union of the arrival streams of players in M. Since the superposition of independent Poisson processes is also a Poisson process, the Erlang loss system operated by this coalition faces a Poisson arrival process with merged rate $\lambda_M = \sum_{i \in M} \lambda_i$. The shared servers are able to handle all types of customers with equal ease, and all customers can effortlessly access the joint service facility. By Part (v) of Theorem 4.1, if the total number of servers remains unchanged when pooling, then such cooperation always leads to a reduction in the expected costs per unit time.

To derive an expression for the long-run average costs per unit time faced by any coalition M, we first need to know how many servers will be used in total. Let us suppose, for the time being, that this coalition uses $s \in \mathbb{N}_0$ servers. We will return later to the question of where this number s comes from. Given s, coalition M faces holding costs of hs per unit time and, by PASTA (as explained in Section 2.4.5), expected penalty costs of $\lambda_M p$ per unit time when

Servers	Players	Customers	Service time	Penalty costs	Res. costs
Hospital beds	Clinical departments in hospital	Patients	Length of stay	Diversion cost to other hospital or government fines	Costs of a bed
Call center agents	Call centers	Callers	Call duration	Goodwill loss when caller gets busy signal	Salary cost
Rental cars	Rental car agencies	Drivers	Rental period	Revenue loss due to inability of renting out a car	Capital and maintenance costs
Spare parts	Spare parts stock points	Failures	Repair / order lead time	Emergency procedure costs	Inventory holding costs
Fire trucks or helicopters	Townships or chemical factories	Fires	Time required to extinguish a fire	Damage while fire truck arrives from elsewhere	Capital and maintenance costs
Repairmen	Manufacturing departments	Repair requests	Repair lead time	Contractual penalties	Salary cost

Table 4.1: *Motivating examples.*

the system is in the state where all servers are occupied, which occurs with a steady-state probability of $\hat{B}(s, \lambda_M \tau)$. Accordingly, the expected costs per unit time in steady state for coalition M are given by $K_M(s) = hs + \hat{B}(s, \lambda_M \tau) \cdot \lambda_M p$.

Now that we have a general expression for these costs under an arbitrary integer number of servers s, we turn to the origin of s. In practice, two different arrangements may occur: fixed and optimized numbers of servers. These two settings will lead to two different games. In the game with fixed numbers of servers, which will be analyzed in Section 4.4, each player $i \in N$ owns a predetermined number of servers, denoted by S_i, and any coalition M uses $\sum_{i \in M} S_i$ servers. This leads to costs $K_M(\sum_{i \in M} S_i)$. In the game with optimized numbers of servers, which will be analyzed in Section 4.5, any group of players—including singletons—may benefit by (re-)optimizing the number of servers in their joint system. This leads to costs $\min_{s \in \mathbb{N}_0} K_M(s)$ for any coalition M. We will briefly treat an alternative service constraint formulation in Section 4.5.4. For clarity, however, we focus solely on the cost-based formulation in the current section.

A broad range of stochastic systems behave as Erlang loss systems. Emergency medical service providers (Restrepo et al., 2009), hospital wards (De Bruin et al., 2010), call centers (Gans et al., 2003), rental agencies (Papier and Thonemann, 2008), spare parts inventory systems (Kranenburg, 2006), or fire departments (Chiu and Larson, 1985) come to mind. In

Table 4.1, we phrase several real-life applications in the terminology of the model framework.

We next provide two extensive examples that illustrate in more detail how real-life settings fit into this modeling framework. The two examples illustrate two different application domains. The first example deals with the inventory setting of pooling of repairable spare parts, reminiscent of Example 1.1 from the introduction. The second example deals with the queueing setting of pooling of service technicians, reminiscent of Example 1.2 from the introduction. These examples serve to justify the *combination* of our modeling assumptions for these specific settings and to illustrate the difference between the settings with fixed and optimized numbers of servers.

Example 4.1. Consider three air carrier companies: Company A, located in Amsterdam; Company B, located in Brussels; and Company E, located in Eindhoven. Each company owns a large fleet of airplanes, all of the same model type. Each aircraft consists of a number of components that are subject to failures. We only consider one such component: a gearbox for an airplane's auxiliary power unit. For each company, failures of this component occur according to a Poisson process.[26] Based on historical data, A expects 4 failures per year, B expects 4 failures per year, and E expects 8 failures per year. The companies are not competitors to each other as they serve different geographical regions; hence, their failure processes are independent.

An aircraft with a failed gearbox has to be grounded due to safety regulations.[27] It is difficult to attach a monetary cost to this, but a reasonable estimate is that a grounded aircraft costs 500,000 euro per day. As all companies operate in the same industry, this downtime cost is the same for all of them. To avoid this downtime cost, each company can stock a number of spare gearboxes that can be used to replace failed gearboxes. Currently, the companies operate separate stockpoints. Company A owns 1 spare gearbox, Company B owns 0 gearboxes, and Company E owns 1 spare gearbox. Spare gearboxes are quite expensive. Holding costs, which mainly reflect the opportunity cost of capital tied up in stock that could have been used for another purpose, are incurred at a rate of 10,000 euro per spare gearbox per year. These costs are incurred when spares are in the on-hand inventory as well as when they are in repair.

Gearboxes are repairable.[28] The so-called Fix-It Corporation has a repair facility that

[26] A failure doesn't mean that an airplane crashes as a result—it just means that, during engineering inspection, the condition of the component was not in accordance with safety norms, in which case the component is termed "failed".

[27] We remark that when an airplane is grounded, no failures occur. However, the fraction of time during which an airplane is down is negligible since failure rates are low and, due to the use of emergency procedures, downtimes are never long. Additionally, the work for a grounded airplane may be taken over by a functional airplane (albeit while incurring the penalty costs due to the inconvenience). This implies that the total component failure rate remains close to constant.

[28] The assumption that a failed part can always be repaired is not critical: our model is also applicable for spare parts under condemnation in case one immediately procures a new part if a failed part appears to be

specializes in repairs of this gearbox. Therefore, all three companies have a standard contract with the Fix-It Corporation. This standard contract stipulates a fixed repair lead time of 3 months for any gearbox. Hence, repair lead times are i.i.d. for all players. Since the costs to send a shipment of parts to the Fix-It Corporation are small relative to the price of the part, there is no point in waiting to send a batch: Company A, B, and E always *immediately* send any failed gearbox to the Fix-It Corporation. Failed gearboxes that complete their repair are added back to the stock of spares.

If a demand for a spare gearbox cannot be immediately fulfilled from stock, an emergency procedure is instigated: a spare part is leased for the duration of the repair time from the Lease-It Corporation, an exogenous infinite supplier that always has gearboxes available.[29] It takes one day for such a leased part to arrive, during which a plane is grounded at a cost of 500,000 euro. Additionally, there are direct costs for the emergency shipment and the lease, which total 20,000 euro. So, the total costs incurred for such a stock-out are 520,000 euro. The probability that another part completes its repair and returns to the stock point before the arrival of the leased part is small, and we disregard it. The component that was responsible for this emergency procedure is still sent to the Fix-It Corporation for repair. When it completes repair, it is used to replace the leased part, and the leased part is sent back to the Lease-It Corporation. Therefore, a demand during a stock-out may be considered as a lost sale, and the inventory position for company A (resp. B and E) remains at a constant level: the base stock level 1 (resp. 0 and 1). These given base stock levels need not be optimal.

It follows that the stochastic process for the number of spare parts on hand is identical to the process for the number of free servers in an Erlang loss system. We can represent this situation in our model by letting $N = \{A, B, E\}$, $S_A = 1$, $S_B = 0$, $S_E = 1$, $\lambda_A = 4$, $\lambda_B = 4$, $\lambda_E = 8$, $\tau = 0.25$, $h = 10$, and $p = 520$, where we measure time in years and money in thousands of euros.

To avoid downtimes that occur when, e.g., a plane is down at Company A at a point in time when Company B has at least one part on hand, the three companies consider cooperating by fully pooling their spare parts inventories. That is, they consider setting up a single, jointly operated stock point (the "pool") near Breda where they will store their shared spare parts. This pool would face a total demand rate of $\lambda_{\{A,B,E\}} = 16$.

This setup will lead to some additional transportation time of about 1-2 hours to ship a part from Breda to either Amsterdam, Brussels, or Eindhoven. This, however, does not pose a problem: gearboxes are heavy components, so substantial time is often needed to remove a failed gearbox and to prepare the airplane to receive the spare part. During this

unrepairable, and, similarly, for consumable spare parts in case one-for-one procurements are applied.

[29] Such an arrangement is common in the airline industry when hourly downtime costs are very high (Wong et al., 2006) and it is in line with the assumptions made in Alfredsson and Verrijdt (1999) and Kranenburg and van Houtum (2009). Our model is also applicable if, instead of having an exogenous infinite supplier, the Fix-It Corporation would offer an expedited one-day repair at an additional charge of 20,000 euro.

time, the transshipment can take place. So, the additional transportation time does not lead to additional downtime, and thus we can neglect it. We also neglect any additional transportation costs, as these are small relative to the holding and downtime costs.

We now turn to the question of what the base stock level for the pool will be. As mentioned, before considering collaboration, Company A (resp. B and E) owns 1 (resp. 0 and 1) spare gearbox. After collaboration, the pool either takes over this entire stock of repairable parts in full, leaving the total base stock level of 2 unchanged (the case with fixed numbers of servers), or the pool picks a cost-minimizing base stock level (the case with optimized numbers of servers).

The assumption of fixed base stock levels might apply when, e.g., the airplanes were accompanied by a set of repairable gearboxes upon purchase, which cannot be easily produced or sold afterwards due to their specificity. Although players might reduce their inventory position by not repairing a failed part, this decision would not strongly affect interest costs on the initial capital investment (i.e., would not lead to a strong decrease in holding costs); hence, we disregard this possibility. Stock levels might also be unadjustable if governmental safety regulations specify a fixed base stock level.

The assumption of optimized base stock levels is appropriate when, e.g., the given stocks of players can be easily adjusted. This may happen if spare parts are still in production and/or if there is still a market to sell the parts. Stock levels would also be adjustable if the initial stocking levels of spare parts would not have been determined yet before considering pooling. ◇

Example 4.2. Consider a manufacturer of advanced office equipment such as 3D printers and computer mainframes, who offers service contracts for repairs to customers that purchase a new product. Any such service contract stipulates that, in case of a failure, a repairman must be immediately available. If the manufacturer cannot immediately send in a repairman, then the repair is taken over by a separate repair organization. Irrespective of equipment category, this outside repair comes at a cost of 3000 euro to the manufacturer. Each of the equipment categories is managed by a separate department. The departments have separate, limited budgets and can decide for themselves whether or not they wish to run their service operations separately or pool repairmen with other departments.

Let us consider three departments, named 1, 2, and 3. For each of them, requests for repair occur according to a Poisson process. Department 1 estimates that it faces 0.1 requests per day, Departments 2 estimates that it faces 0.1 requests per day, and Department 3 estimates that it faces 0.8 requests per day. Since these requests originate from failures of different equipment, the Poisson processes are independent. The time required by any repairman to solve a problem is the same for each equipment categories, but it is highly variable. The mean total time to fully execute a repair, including travel time to the customer and back, is equal to 8 hours. We assume that there are 20 working days per month and that a working day consists of 8 hours. Equipment failures only occur during those hours because the office

equipment is not used during weekends or evenings.

Currently, the three departments run their service operations independently. Department 1 employs 1 repairman, Department 2 employs 1 repairman, and Department 3 employs 3 repairmen. The salary of a repairman, including benefits, is 4000 euro per month. We can represent this situation in our model by letting $N = \{1, 2, 3\}$, $S_1 = 1$, $S_2 = 1$, $S_3 = 3$, $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.8$, $\tau = 1$, $h = 4000/20 = 200$, and $p = 3000$, where we measure time in working days and money in euros.

The three departments consider cooperating by fully pooling their repairmen. This pool would face a total repair request rate of $\lambda_{\{1,2,3\}} = 1$. Additional training may be needed to give the repairmen the knowledge to repair all types of products. This will come with some fixed costs. However, these costs are transient and negligible in the long run.

We now turn to the issue of how many repairmen the three departments will employ. As mentioned, before considering collaboration, Department 1 (resp. 2 and 3) employs 1 (resp. 1 and 3) repairmen. After collaboration, the pool either employs the 5 repairmen in total (the case with a fixed number of servers), or the pool picks a cost-minimizing number of repairmen (the case with an optimized number of servers).

The assumption of a fixed number of servers is appropriate when, e.g., repairmen cannot be easily hired or fired. The number might also be unadjustable if the departments do not want to enter a deep, long-term integration with a jointly optimized workforce right from the start, but rather prefer to build trust by starting with a short-term trial in which only their current employees are pooled.

The assumption of an optimized number of servers is appropriate when, e.g., repairmen *can* be easily hired or fired. If adjustment of the number of servers is allowed, then a natural trade-off arises between resource costs and penalty costs. Accordingly, the associated cooperative game will require a substantially different analysis than the game with fixed resource levels. ◇

4.3.2 Erlang loss situations

We now formally describe all input parameters of the situation described in the preceding subsection, while simultaneously generalizing the cost structure. After providing the definition, we discuss the modeling choices related to this generalization.

Definition 4.2. An *Erlang loss situation* is a tuple $(N, S, \lambda, \tau, H, p)$, where

- N is the non-empty, finite set of players;
- $S \in \mathbb{N}_0^N$ is the vector of numbers of servers, where S_i describes the number of servers possessed a priori by player $i \in N$;[30]

[30] For the game with optimized numbers of servers, S is in fact superfluous in the tuple, but we leave it in for consistency and to allow formal comparisons between the two games.

- $\lambda \in \mathbb{R}^N_{++}$ is the vector of arrival rates, where λ_i describes the rate of the Poisson process generating arrivals of customers that belong to player $i \in N$;
- $\tau \in \mathbb{R}_{++}$ is the mean service time for an arbitrary customer of any player;
- $H : \mathbb{N}_0 \to \mathbb{R}_+$ is a concave, non-decreasing function with $H(0) = 0$ specifying that the resource costs for holding s servers are $H(s)$ per unit time;
- $p \in \mathbb{R}^N_{++}$ is the vector of penalty costs, where p_i describes the expected penalty costs that are incurred whenever a customer of player $i \in N$ is blocked.

For notational ease, we write $S_M = \sum_{i \in M} S_i$ and $\lambda_M = \sum_{i \in M} \lambda_i$ for any coalition $M \in 2^N_-$.

Note that we now allow for concave resource costs. Concavity represents additional economies of scale that may be exploited by acquiring and maintaining resources collaboratively. Linear costs are an important special case. For clarity, all examples in this chapter will feature linear resource costs. For notational convenience, we call any Erlang loss situation $(N, S, \lambda, \tau, H, p)$ a *linear* Erlang loss situation if there exists an $h \in \mathbb{R}_+$ such that $H(s) = hs$ for all $s \in \mathbb{N}_0$, and we will represent such a situation directly via the tuple $(N, S, \lambda, \tau, h, p)$.

Having the same concave resource cost function H for every coalition is not the only way to capture synergy due to pooling. We could also have introduced linear resource cost rates $(h_M)_{M \in 2^N_-}$ that may differ across coalitions such that the resource cost rate does not increase as a coalition grows. In fact, we do exactly this in Chapter 5; it is an alternative way to model economies of scale in resource costs. This alternative formulation captures the ability of larger coalitions to acquire and maintain servers at a reduced cost rate because more players gives stronger negotiation or buying power. The concave formulation, in contrast, captures the ability of any coalition to acquire and maintain servers at a reduced cost rate because a larger total number of servers are bought or maintained simultaneously. Including both perspectives in the same model would lead to a muddled formulation. We choose to focus on the concave, yet symmetric cost function H in this chapter because it permits interesting analysis from a game theoretical perspective.

For our analysis, it will be convenient to extend the domain of the resource cost function H to all nonnegative reals by a linear interpolation: we let H^{lin} denote the function $H^{lin} : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$H^{lin}(s) = (1 - (s - \lfloor s \rfloor)) \cdot H(\lfloor s \rfloor) + (s - \lfloor s \rfloor) \cdot H(\lceil s \rceil) \qquad \text{for all } s \in \mathbb{R}_+.$$

Note that because H is a concave non-negative function, the same holds for its linear interpolation H^{lin}.

We next discuss the generalization from symmetric penalty costs to asymmetric penalty costs. This asymmetry allows us to capture the difference in downtime costs or monetary compensations for blocked customers across the various customer streams. At the same time,

it raises the question of which pooling policy is followed. While it is well-known that full pooling (i.e., each arriving customer is always admitted if there is a free server available and only denied when there is no free server) is the optimal policy if penalty costs are symmetric, a partial pooling policy in which some servers are held back for more "critical" customer classes may be better if penalty costs are asymmetric. We assume full pooling in Sections 4.4 and 4.5, and we consider partial pooling in Section 4.6.

Full pooling is a simple policy that is easy to operationalize in practice. Indeed, partial pooling can be more difficult to implement because saying "no" to an arriving customer when he can see that there is still a free server available could make that customer very upset. Furthermore, full pooling leads to a tractable model in which the blocking probability for any customer class is described by the Erlang loss function. We also remark that the assumption of full pooling is common in the spare parts literature (see, for example, Paterson et al., 2011).

4.4 Fixed numbers of servers

In this section, we analyze the setting with fixed numbers of servers. We first define the characteristic cost function and consider concavity in Subsection 4.4.1. Subsequently, we break down the effects of full pooling under asymmetric penalty costs into two separate, possibly conflicting, forces in Subsection 4.4.2. Specifically, we show that full pooling always reduces the expected number of blocked customers per unit time, but that it may increase the expected penalty costs of an arbitrary blocked customer. We disentangle these two effects to provide sufficient conditions for balancedness. In Subsection 4.4.3, we prove that symmetric penalty costs always leads to a balanced game. Finally, in Subsection 4.4.4, we show population monotonicity of a simple allocation scheme.

4.4.1 Game

We start by formulating our game.

Definition 4.3. For any Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$, the game (N, c^{φ}) with characteristic costs

$$c^{\varphi}(M) = H(S_M) + B(S_M, \lambda_M \tau) \cdot \textstyle\sum_{i \in M} \lambda_i p_i \tag{4.13}$$

for all $M \in 2^N_-$ is called the associated *Erlang loss game with fixed numbers of servers*.

First and foremost, we are interested in the existence of a stable cost allocation for this class of games. If Erlang loss games with fixed numbers of servers would always be concave, then this existence would have been guaranteed. The following example, however, shows that such games need not be concave, even if cost parameters are symmetric.

Example 4.3. Reconsider the air carrier companies from Example 4.1 and suppose that base stock levels are fixed. This results in the 3-player linear Erlang loss situation $\varphi =$

Coalition M	S_M	$\lambda_M\tau$	$\hat{B}(S_M, \lambda_M\tau)$	$c^\varphi(M)$
{A}	1	1	$\frac{1}{2}$	1050
{B}	0	1	1	2080
{E}	1	2	$\frac{2}{3}$	$2783\frac{1}{3}$
{A,B}	1	2	$\frac{2}{3}$	$2783\frac{1}{3}$
{A,E}	2	3	$\frac{9}{17}$	$3323\frac{9}{17}$
{B,E}	1	3	$\frac{3}{4}$	4690
{A,B,E}	2	4	$\frac{8}{13}$	5140

Table 4.2: *The games and parameter values of Example 4.3.*

$(N, S, \lambda, \tau, h, p)$ with $N = \{A, B, E\}$, $S_A = 1$, $S_B = 0$, $S_E = 1$, $\lambda_A = 4$, $\lambda_B = 4$, $\lambda_E = 8$, $p_i = 520$ for all $i \in N$, $\tau = 0.25$, and $h = 10$. The associated Erlang loss game with fixed numbers of servers is described in Table 4.2.

This game is balanced: for example, the allocation x that assigns $x_A = 700$, $x_B = 2000$, $x_C = 2440$ is a core element. The nucleolus, equal to $(642\frac{28}{51}, 1948\frac{4}{17}, 2549\frac{11}{51})$, is also in the core. In fact, the game is totally balanced.[31]

However, it is not concave as, e.g., $c(\{1,3\}) - c(\{1\}) = 2273\frac{9}{17} < 2356\frac{2}{3} = c(\{1,2,3\}) - c(\{1,2\})$. In other words, a player's marginal cost contribution may increase if he joins a larger coalition. ◇

4.4.2 Beneficial and adverse pooling effects

In this subsection, we consider the performance of full pooling arrangements in terms of the expected number of blocked customers per unit time and the expected costs of an arbitrary blocked customer. We are interested in the effects of establishing the full pooling arrangement of the grand coalition, compared to a balanced combination of full pooling arrangements between smaller coalitions. With "balanced combination" we mean a combination in which coalitions are assigned relative weights according to a balanced map. Balanced maps were introduced in Section 2.3.5.

To formally describe the two effects of full pooling, which may be conflicting when penalty costs are asymmetric, we introduce the following notation for any Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$. For a coalition $M \in 2^N_-$, the expected number of blocked customers per unit

[31] Total balancedness implies that the game—or equivalently and more conveniently, its cost savings game (N, v^φ), with $v^\varphi(M) = \sum_{i \in M} c^\varphi(\{i\}) - c^\varphi(M)$ for all $M \in 2^N_-$—is a linear production game (Owen, 1975) associated with its direct linear production situation (as described in, e.g., Borm et al., 2001, Theorem 6.2). Owen (1975) offers a set of resulting stable allocations, but unfortunately they exhaust the core (Borm et al., 2001, Proposition 6.3). So, this connection with linear production games is interesting, but does not directly aid in selecting a core allocation.

time is

$$U^{\varphi}(M) = \hat{B}(S_M, \lambda_M \tau) \cdot \lambda_M. \tag{4.14}$$

For a balanced map $\kappa \in \mathcal{W}^N$, the *weighted number of blocked customers per unit time* is

$$W_U^{\varphi}(\kappa) = \sum_{M \in 2_-^N} \kappa(M) \cdot U^{\varphi}(M). \tag{4.15}$$

Here, $W_U^{\varphi}(\kappa)$ describes the weighted average of the number of blocked customers per unit time for a number of sub-coalitions combined, where each sub-coalition is weighted according to κ. We will later compare such values to $U^{\varphi}(N)$, which will help our understanding of whether full pooling among all players is better or worse than full pooling among smaller coalitions.

Next, for a coalition $M \in 2_-^N$, the expected penalty costs of an arbitrary blocked customer are

$$E^{\varphi}(M) = \sum_{i \in M} \lambda_i p_i / \lambda_M. \tag{4.16}$$

For a balanced map $\kappa \in \mathcal{W}^N$, the *weighted penalty costs of an arbitrary blocked customer* are

$$W_E^{\varphi}(\kappa) = \frac{\sum_{M \in 2_-^N} \kappa(M) \cdot U^{\varphi}(M) E^{\varphi}(M)}{\sum_{M \in 2_-^N} \kappa(M) \cdot U^{\varphi}(M)}. \tag{4.17}$$

To grasp the meaning of $W_E^{\varphi}(\kappa)$, consider a balanced combination of pooling arrangements between sub-coalitions, specified by κ. For such an arrangement, the total penalty costs incurred over an infinite horizon, divided by the total number of blocked customers over that horizon, is equal to $W_E^{\varphi}(\kappa)$.

Now, the effects of establishing the full pooling arrangement of the grand coalition are twofold:

Effect 1: Full pooling changes the weighted number of blocked customers per unit time.

Effect 2: Full pooling changes the weighted costs of an arbitrary blocked customer.

Effect 1 is always beneficial, i.e., it leads to less frequent blocking, as we will show in Lemma 4.12. Effect 2 may be beneficial or adverse, depending on the underlying parameters of the players. as we will show in Example 4.4.

Lemma 4.12. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation. Then, for all minimally balanced maps $\kappa \in \mathcal{W}^N$, it holds that $W_U^{\varphi}(\kappa) \geq U^{\varphi}(N)$.*

Coalition M	S_M	λ_M	$\hat{B}(S_M, \lambda_M)$	$c^{\varphi}(M)$	$c^{\hat{\varphi}}(M)$
{1}	1	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{30}$	$33\frac{1}{3}$
{2}	1	$\frac{4}{5}$	$\frac{4}{9}$	$355\frac{5}{9}$	$\frac{16}{45}$
{1,2}	2	1	$\frac{1}{5}$	$160\frac{1}{25}$	$40\frac{4}{25}$

Table 4.3: *The games and parameter values of Example 4.4.*

Proof. Let $\kappa \in \mathcal{W}^N$. Then,

$$\begin{aligned} W_L^{\varphi}(\kappa) &= \sum_{M \in 2_-^N} \kappa(M) \cdot B\left(S_M, \lambda_M \tau\right) \cdot \lambda_M \\ &= \sum_{M \in 2_-^N} \kappa(M) \frac{\lambda_M}{\lambda_N} \cdot B\left(S_M, \lambda_M \tau\right) \cdot \lambda_N \\ &\geq \sum_{M \in 2_-^N} \kappa(M) \frac{\lambda_M}{\lambda_N} \cdot B\left(S_M \cdot \frac{\lambda_N}{\lambda_M}, \lambda_N \tau\right) \cdot \lambda_N \\ &\geq B\left(\sum_{M \in 2_-^N} \kappa(M) \frac{\lambda_M}{\lambda_N} \cdot S_M \cdot \frac{\lambda_N}{\lambda_M} \cdot, \lambda_N \tau\right) \cdot \lambda_N \\ &= B\left(S_N, \lambda_N \tau\right) \cdot \lambda_N = U^{\varphi}(N). \end{aligned}$$

The first and last equalities hold by Equation (4.14) and because B is an extension. The first inequality holds by Theorem 4.5. The second inequality is valid by the combination of the convexity in the number of servers, as stated in Theorem 4.4, and the fact that $\sum_{M \in 2_-^N} \kappa(M)\lambda_M = \lambda_N$; that is, we employ this convexity property of B by using that the nonnegative convex weights $\kappa(M)\lambda_M/\lambda_N$ add up to 1. The penultimate equality holds because $\sum_{M \in 2_-^N} \kappa(M) S_M = S_N$. □

Example 4.4. Consider the linear Erlang loss situations $\varphi = (N, S, \lambda, \tau, h, p)$ and $\hat{\varphi} = (N, S, \lambda, \tau, h, \hat{p})$. Both have $N = \{1, 2\}$, $S_1 = S_2 = 1$, $\lambda_1 = \frac{1}{5}$, $\lambda_2 = \frac{4}{5}$, $\tau = 1$, and $h = 0$, but the two situations differ in their penalty costs. For φ, $p_1 = 1$ and $p_2 = 1000$. For $\hat{\varphi}$, these costs are swapped, i.e., $\hat{p}_1 = 1000$ and $\hat{p}_2 = 1$. The associated Erlang loss games with fixed numbers of servers are described in Table 4.3.

Although (N, c^{φ}) has a non-empty core, the core of $(N, c^{\hat{\varphi}})$ is empty. In fact, $(N, c^{\hat{\varphi}})$ has an empty imputation set. The reason for this is that $(N, c^{\hat{\varphi}})$ is not subadditive.

To explain this, we compare the full pooling arrangement for the grand coalition to the no-pooling arrangement where all players are acting alone, by considering the balanced map $\kappa : 2_-^N \to [0, 1]$ with $\kappa(\{1\}) = \kappa(\{2\}) = 1$ and $\kappa(N) = 0$. Effect 1 of full pooling (the reduction in the weighted number of blocked customers per unit time) is equally beneficial in both situations. That is, $W_U^{\varphi}(\kappa) = W_U^{\hat{\varphi}}(\kappa) = \frac{1}{6} \cdot \frac{1}{5} + \frac{4}{9} \cdot \frac{4}{5} = \frac{7}{18}$ and $U^{\varphi}(N) = U^{\hat{\varphi}}(N) = \frac{1}{5} \cdot 1 = \frac{1}{5}$.

Effect 2, however, affects the situations in a vastly different way. For situation φ, the weighted penalty costs of an arbitrary blocked customer decrease from $W_E^{\varphi}(\kappa) = \frac{1}{6} \cdot \frac{1}{5} \cdot 1 + \frac{4}{9} \cdot \frac{4}{5} \cdot 1000 = 355\frac{53}{90}$ to $E^{\varphi}(N) = 1 \cdot \frac{1}{5} \cdot \left(\frac{1}{5} \cdot 1 + \frac{4}{5} \cdot 1000\right) = 160\frac{1}{25}$. In contrast, for situation $\hat{\varphi}$, the weighted penalty costs of an arbitrary blocked customer increase from $W_E^{\hat{\varphi}}(\kappa) = \frac{1}{6} \cdot \frac{1}{5} \cdot 1000 + \frac{4}{9} \cdot \frac{4}{5} \cdot 1 = 33\frac{31}{45}$ to $E^{\hat{\varphi}}(N) = 1 \cdot \frac{1}{5} \cdot \left(\frac{1}{5} \cdot 1000 + \frac{4}{5} \cdot 1\right) = 40\frac{4}{25}$. $\diamond$

This example illustrates Effect 2: compared to a no-pooling arrangement, full pooling may decrease or increase the expected penalty costs of an arbitrary blocked customer. Intuitively, a decrease in costs is possible if a customer stream with high penalty costs joins an existing pooling group whose members own excessively many servers that are infrequently used. In that case, full pooling not only leads to fewer blocked customers (Effect 1) but also to less expected penalty costs.

The intuition behind a cost increase, on the other hand, is that full pooling allow customers with low penalty costs to occupy servers that would have better been saved to guard against the high penalty costs of another customer stream. The resulting increase in costs may actually overshadow the reduction in the number of blocked customers (Effect 1), as is the case for the game $(N, c^{\hat{\varphi}})$ in Example 4.4, which has an empty imputation set. In this game, no pooling outperforms full pooling. This precludes cooperation.

Example 4.4 showed the existence of non-subadditive, non-balanced Erlang loss games with fixed numbers of servers, as well as the existence of a subadditive, balanced one. A natural question is whether subadditivity is sufficient for balancedness. The following example answer this question negatively by describing a subadditive, non-balanced Erlang loss game with fixed numbers of servers.

Example 4.5. Consider $\varphi = (N, S, \lambda, \tau, h, p)$, a 3-player linear Erlang loss situation with $N = \{1, 2, 3\}$, $S_1 = 7$, $S_2 = 1$, $S_3 = 1$, $\lambda_1 = \lambda_2 = \lambda_3 = 10$, $\tau = 1$, $h = 0$, $p_1 = 500$, $p_2 = 400$, and $p_3 = 400$. So, player 1 has many servers and high penalty costs, whereas players 2 and 3 have low, identical numbers of servers and low, identical penalty costs.

The characteristic cost function c^{φ} of the associated game (N, c^{φ}) is represented in Table 4.4. It is readily verified that this game is subadditive. Thus, there exists an individually rational allocation. However, this game is not balanced. To see this, consider the balanced map $\kappa : 2_{-}^{N} \to [0, 1]$ with $\kappa(M) = \frac{1}{2}$ if $|M| = 2$ and $\kappa(M) = 0$ otherwise. For this balanced map, $W_U^{\varphi}(\kappa) \approx 21.589$ and $W_E^{\varphi}(\kappa) \approx 429.041$. Hence, $\sum_{M \in 2_{-}^{N}} \kappa(M) c^{\varphi}(M) < 9265 < c^{\varphi}(N)$. Non-balancedness may be observed more clearly in the cost savings game (N, v^{φ}), with $v^{\varphi}(M) = \sum_{i \in M} c^{\varphi}(\{i\}) - c^{\varphi}(M)$ for all $M \in 2_{-}^{N}$, as provided in Table 4.4. $\diamond$

This example shows that, even if the costs of full pooling for the grand coalition are lower than the sum of the costs under any partitioning of players, there need not be a stable cost allocation. This result emphasizes the importance of taking the coalitional costs into account. We remark, however, that the players in Example 4.5 were vastly asymmetric. The

Coalition M	S_M	λ_M	$\hat{B}(S_M, \lambda_M)$	$U^\varphi(M)$	$E^\varphi(M)$	$c^\varphi(M)$	$v^\varphi(M)$
{1}	7	10	0.4090	4.0904	500	2045.204	0
{2}, {3}	1	10	0.9091	9.0909	400	3636.364	0
{1,2}, {1,3}	8	20	0.6270	12.5396	450	5642.817	38.751
{2,3}	2	20	0.9050	18.0995	400	7239.819	32.908
{1,2,3}	9	30	0.7127	21.3810	433.3333	9265.106	52.825

Table 4.4: *The game and parameter values of Example 4.5 (Values are rounded).*

following lemma provides sufficient conditions on the underlying parameters of the players for balancedness.

Lemma 4.13. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation. If $U^\varphi(N)/W_U^\varphi(\kappa) \leq W_E^\varphi(\kappa)/E^\varphi(N)$ for all $\kappa \in \mathcal{W}^N$, then the associated Erlang loss game with fixed numbers of servers (N, c^φ) has a nonempty core.*

Proof. If $S_N = 0$, then there are no servers to pool and the allocation x which assigns $x_i = \lambda_i p_i$ to each $i \in N$ is trivially stable. If $S_N > 0$, then we let $\kappa \in \mathcal{W}^N$, assume that $U^\varphi(N)/W_U^\varphi(\kappa) \leq W_E^\varphi(\kappa)/E^\varphi(N)$, and obtain

$$\begin{aligned}
\sum_{M \in 2_-^N} \kappa(M) c^\varphi(M) &= \sum_{M \in 2_-^N} \kappa(M) H(S_M) + \sum_{M \in 2_-^N} \kappa(M) U^\varphi(M) E^\varphi(M) \\
&= \sum_{M \in 2_-^N} \kappa(M) H(S_M) + W_E^\varphi(\kappa) W_U^\varphi(\kappa) \\
&\geq \sum_{M \in 2_-^N} \kappa(M) H(S_M) + U^\varphi(N) E^\varphi(N) \\
&= \sum_{M \in 2_-^N : S_M > 0} \kappa(M) S_M \frac{H^{lin}(S_M)}{S_M} + U^\varphi(N) E^\varphi(N) \\
&\geq \sum_{M \in 2_-^N : S_M > 0} \kappa(M) S_M \frac{H^{lin}(S_N)}{S_N} + U^\varphi(N) E^\varphi(N) \\
&= S_N \cdot \frac{H^{lin}(S_N)}{S_N} + \hat{B}(S_N, \lambda_N) \cdot \sum_{i \in N} \lambda_i p_i = c^\varphi(N).
\end{aligned}$$

The first inequality holds by assumption. The second inequality holds because H^{lin} is concave with $H^{lin}(0) = 0$, which by Part (ii) of Theorem 2.1 implies that H^{lin} is elastic. □

Intuitively, the condition on the balanced maps in Lemma 4.13 takes the impact of Effect 2 (the possible change in the weighted costs of an arbitrary blocked customer) as a given and states that the result of Effect 1 (the reduction in the number of blocked customers for the

grand coalition) should be sufficient to compensate for the possibly adverse impact of Effect 2. Under this condition, a stable cost allocation will exist. For instance, the nucleolus, which always yields core allocations for any game with a non-empty core, would accomplishes a stable cost allocation.

4.4.3 Balancedness under symmetric penalty costs

The condition in Lemma 4.13 immediately implies that if Effect 2 is not adverse, then—due to Lemma 4.12—the game will be balanced. This insight can be used to obtain our main result on Erlang loss games with fixed numbers of servers.

Theorem 4.14. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation with $p_i = p_j$ for all $i, j \in N$. Then, the associated Erlang loss game with fixed numbers of servers (N, c^φ) is totally balanced.*

Proof. Let $\kappa \in \mathcal{W}^N$ be a balanced map. As all players have equal penalty costs, say P, Effect 2 of full pooling is neutral, i.e., $E^\varphi(M) = P$ for all coalitions $M \in 2^N_-$. Hence, $W^\varphi_E(\kappa) = P$ as well. So,

$$U^\varphi(N) \leq W^\varphi_U(\kappa) = W^\varphi_U(\kappa) \cdot \frac{E^\varphi(N)}{E^\varphi(N)} = W^\varphi_U(\kappa) \cdot \frac{W^\varphi_E(N)}{E^\varphi(N)},$$

where the inequality holds by Lemma 4.12. Hence, the condition in Lemma 4.13 holds, which implies that (N, c^φ) is balanced. It is totally balanced because every sub-game of (N, c^φ) is a game associated with a Erlang loss situation with symmetric penalty costs itself. □

Theorem 4.14 states that if all players symmetric equal penalty costs, then the corresponding Erlang loss game with fixed numbers of servers has a nonempty core. Note that our extensive Examples 4.1 and 4.2 featured symmetric penalty costs. At the same time, the player in these examples had different demand rates and base stock levels; these are precisely the types of asymmetry that are allowed in Theorem 4.14.

The following theorem indicates that, from a theoretical perspective, the symmetry condition of Theorem 4.14 is tight. In other words, even if the penalty costs of the players differ only slightly, the associated Erlang loss game with fixed numbers of servers may be non-balanced.

Theorem 4.15. *For every $\epsilon > 0$, there exists an Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ with $\max_{i \in N} p_i - \min_{i \in N} p_i \leq \epsilon$ for which the associated Erlang loss game with fixed numbers of servers, (N, c^φ), is not balanced.*

Proof. Let $\epsilon > 0$. We now construct $\varphi = (N, S, \lambda, \tau, h, p)$, a 2-player linear Erlang loss situation with $N = \{1, 2\}$, $S_1 = 0$, $S_2 = 1$, $\lambda_1 = \lambda_2 = \Lambda = 1 + 1/\epsilon$, $\tau = 1$, $h = 0$, $p_1 = 1$, and

$p_2 = 1 + \epsilon$. Clearly, $\max_{i \in N} p_i - \min_{i \in N} p_i \leq \epsilon$, as desired. Then,

$$\begin{aligned}
c^{\varphi}(\{1\}) + c^{\varphi}(\{2\}) &= \hat{B}(0, \Lambda) \cdot \Lambda \cdot 1 + \hat{B}(1, \Lambda) \cdot \Lambda \cdot (1 + \epsilon) \\
&= \Lambda + \frac{\Lambda}{\Lambda + 1} \Lambda (1 + \epsilon) \\
&= \frac{\Lambda}{\Lambda + 1} [\Lambda + 1 + \Lambda (1 + \epsilon)] \\
&= \frac{\Lambda}{(\Lambda + 1)(2\Lambda + 1)} (2\Lambda + 1)[\Lambda + 1 + \Lambda (1 + \epsilon)] \\
&= \Lambda \frac{4\Lambda^2 + 2\Lambda^2 \epsilon + 4\Lambda + 1 + \Lambda \epsilon}{(\Lambda + 1)(2\Lambda + 1)} \\
&< \Lambda \frac{4\Lambda^2 + 2\Lambda^2 \epsilon + 4\Lambda + 2\Lambda \epsilon}{(\Lambda + 1)(2\Lambda + 1)} \\
&= \Lambda \frac{2\Lambda (2 + \epsilon)(\Lambda + 1)}{(\Lambda + 1)(2\Lambda + 1)} \\
&= \Lambda \frac{2\Lambda (2 + \epsilon)}{2\Lambda + 1} \\
&= \frac{2\Lambda (2 + \epsilon)}{2\Lambda + 1} [\Lambda \cdot 1 + \Lambda \cdot (1 + \epsilon)] \\
&= \hat{B}(1, 2\Lambda) \cdot [\Lambda \cdot 1 + \Lambda \cdot (1 + \epsilon)] = c^{\varphi}(N),
\end{aligned}$$

where the inequality holds because, by choice of arrival rates, $\Lambda \epsilon > 1$. We conclude that (N, c^{φ}) is not balanced. □

4.4.4 Cost allocation: stability and population monotonicity

In this section, we prove that Erlang loss games with fixed numbers of servers admit a PMAS under a reasonable symmetry condition. To describe this condition, let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation. For any player $i \in N$, we say that S_i/λ_i is his *server-demand ratio*. We now consider situations where all players have equal server-demand ratios. We do allow for different penalty costs among players.

This equality in server-demand ratios is clearly satisfied when players constitute equally sized business units, all with the same numbers of servers and arrival rates. It is also satisfied in the spare parts setting when any player $i \in N$ has $n_i \in \mathbb{N}_0$ machines, each with an individual failure rate of $\bar{\lambda} \in \mathbb{R}_{++}$ and a corresponding stock of $\bar{S} \in \mathbb{N}_0$ spare parts, possibly based on a per-machine recommendation of the Original Equipment Manufacturer (OEM). In this case, any player $i \in N$ will have a base stock level of $S_i = \bar{S} \cdot n_i$ and a total demand rate of $\lambda_i = \bar{\lambda} \cdot n_i$, which implies that players have a symmetric server-demand ratio of $\bar{S}/\bar{\lambda}$.

For Erlang loss situation φ, we now define a proportional allocation scheme. This scheme is easy to compute and easy to administer for any particular demand realization because it does not require transfer payments of penalty costs—each player simply incurs the penalty costs incurred for his own blocked customers. Formally, we define this allocation scheme

$\mathcal{P}^{FIX}(\varphi)$ for all $M \in 2^N_-$ and all $i \in M$ by

$$\mathcal{P}^{FIX}_{i,M}(\varphi) = \frac{S_i}{S_M} H(S_M) + \hat{B}\left(S_M, \lambda_M \tau\right) \cdot \lambda_i p_i. \tag{4.18}$$

The following theorem states that, when all players have the same server-demand ratio, this allocation scheme is population monotonic. The intuition behind this result is that, under this symmetry condition, the blocking probability does not increase as a coalition grows larger.

Theorem 4.16. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation with $S_i/\lambda_i = S_j/\lambda_j$ for all $i, j \in N$. Then, $\mathcal{P}^{FIX}(\varphi)$ is a PMAS for the associated Erlang loss game with fixed numbers of servers (N, c^φ).*

Proof. Let $M, L \in 2^N_-$ with $M \subseteq L$ and let $i \in M$. If the symmetric server-demand ratio is zero, then no player has any servers, which implies that $\mathcal{P}^{FIX}_{i,M}(\varphi) = \lambda_i p_i = \mathcal{P}^{FIX}_{i,L}(\varphi)$, and the proof is complete. Otherwise, if the symmetric server-demand ratio is positive, then we obtain

$$\begin{aligned}
\mathcal{P}^{FIX}_{i,L}(\varphi) &= S_i \frac{H^{lin}(S_L)}{S_L} + \hat{B}(S_L, \lambda_L \tau) \cdot \lambda_i p_i \\
&\leq S_i \frac{H^{lin}(S_L)}{S_L} + \hat{B}(S_M, \lambda_L \tau S_M / S_L) \cdot \lambda_i p_i \\
&= S_i \frac{H^{lin}(S_L)}{S_L} + \hat{B}(S_M, \lambda_M \tau) \cdot \lambda_i p_i \\
&\leq S_i \frac{H^{lin}(S_M)}{S_M} + \hat{B}(S_M, \lambda_M \tau) \cdot \lambda_i p_i = \mathcal{P}^{FIX}_{i,M}(\varphi).
\end{aligned}$$

The first inequality follows because the Erlang loss function is subhomogeneous of degree zero by Theorem 4.3. The subsequent equality holds because S_M/λ_M and S_L/λ_L coincide. The second inequality holds because H^{lin} is concave with $H^{lin}(0) = 0$, which by Part (ii) of Theorem 2.1 implies that H^{lin} is elastic. We conclude that $y(\varphi)$ is indeed a PMAS. □

As described in Section 2.3.7, the existence of a PMAS has several implications. Accordingly, we immediately obtain the following corollary from Theorem 4.16.

Corollary 4.17. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation with $S_i/\lambda_i = S_j/\lambda_j$ for all $i, j \in N$. Then, the associated game (N, c^φ) is totally balanced and the allocation assigning $\mathcal{P}^{FIX}_{i,N}(\varphi)$ to all $i \in N$ is an element of its core.*

4.5 Optimized numbers of servers

In this section, we analyze the setting where all coalitions optimize their numbers of servers. We consider two games: one with a service constraint and one that is purely cost-based. We first define the characteristic cost functions of both games and consider their concavity in Subsection 4.5.1. Subsequently, in Subsection 4.5.2, we show that there is no obvious relation

between the setting with fixed numbers of servers and the setting with optimized numbers of servers. In Subsection 4.5.3, we show that a simple, proportional allocation scheme is population monotonic for the cost-based formulation in case of symmetric penalty costs. Finally, in Subsection 4.5.4, we show that a similar proportional allocation (though not a scheme) is stable for the service constraint formulation under certain conditions.

4.5.1 Games

We first treat the cost-based formulation. Consider any Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ with unbounded[32] H and suppose that each coalition picks a cost minimizing number of servers. Accordingly, the initial vector S has become superfluous as an input parameter, but we leave the tuple structure intact for consistency reasons. For any particular choice of the number of servers $s \in \mathbb{N}_0$ in the joint system of a coalition $M \in 2^N_-$, the expected relevant costs per unit time in steady state are given by

$$K_M(s) = H(s) + \hat{B}(s, \lambda_M \tau) \cdot \textstyle\sum_{i \in M} \lambda_i p_i. \tag{4.19}$$

Since the image of the Erlang loss function is the interval $(0, 1]$, whereas the resource cost function H increases unboundedly, there exists an optimal number of servers, which can be found by, e.g., an enumerative search procedure. In case of a linear resource cost rate h, such an optimizer is given by the smallest $s_M \in \mathbb{N}_0$ satisfying $h \geq [\hat{B}(s_M + 1, \lambda_M \tau) - B(s_M, \lambda_M \tau)] \cdot \sum_{i \in M} \lambda_i p_i$. This follows from the convexity property of Theorem 4.4.

For any coalition $M \in 2^N_-$, there may be multiple cost minimizing numbers of servers, so to avoid ambiguity we define a specific optimal number of servers by[33]

$$S^*_M = \min\{s \in \mathbb{N}_0 \mid K_M(s) \leq K_M(\check{S}) \text{ for all } \check{S} \in \mathbb{N}_0\}. \tag{4.20}$$

We can now formulate the corresponding game.

Definition 4.4. For any Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ with unbounded H, the game (N, d^φ) with characteristic costs

$$d^\varphi(M) = K_M(S^*_M) \tag{4.21}$$

for all $M \in 2^N_-$ is called the associated *Erlang loss game with optimized numbers of servers.*

The following example shows that Erlang loss games with optimized numbers of servers need not be concave, even if all players have symmetric cost parameters.

Example 4.6. Reconsider the office equipment departments from Example 4.2 and suppose that the number of servers is optimized. Scaling the resource cost rate to 1, this results in

[32] Any non-decreasing function $f : \mathbb{N}_0 \to R_+$ is called *unbounded* if there does not exist an $a \in \mathbb{R}_+$ for which $f(x) \leq a$ for all $x \in \mathbb{N}_0$.

[33] We suppress the dependence of K_M and S^*_M on φ to avoid notational baggage.

Coalition M	λ_M	$\hat{B}(1,\lambda_M)$	$\hat{B}(2,\lambda_M)$	$\hat{B}(3,\lambda_M)$	S^*_M	$d^\varphi(M)$	$d^\varphi_\beta(M)$
{1},{2}	$\frac{1}{10}$	$\frac{1}{11}$	$\frac{1}{221}$	$\frac{1}{6631}$	1	$1\frac{3}{22}$	$1\frac{221}{672}$
{3}	$\frac{8}{10}$	$\frac{4}{9}$	$\frac{8}{53}$	$\frac{32}{827}$	3	$3\frac{384}{827}$	$2\frac{4135}{5248}$
{1,2}	$\frac{2}{10}$	$\frac{1}{6}$	$\frac{1}{61}$	$\frac{1}{916}$	1	$1\frac{1}{2}$	$1\frac{61}{88}$
{1,3}, {2,3}	$\frac{9}{10}$	$\frac{9}{19}$	$\frac{81}{461}$	$\frac{243}{4853}$	3	$3\frac{6561}{9706}$	$2\frac{810451}{899424}$
{1,2,3}	1	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{16}$	3	$3\frac{15}{16}$	3

Table 4.5: *The Erlang loss game with optimized numbers of servers (N, d^φ) for Example 4.6 and the Erlang loss game with service constraints (N, d^φ_β) for Example 4.7.*

the 3-player linear Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ with $N = \{1,2,3\}$, $\lambda_1 = \frac{1}{10}$, $\lambda_2 = \frac{1}{10}$, $\lambda_3 = \frac{8}{10}$, $\tau = 1$, $h = 1$, and $p_i = 15$ for all $i \in N$, and arbitrary S. For player 1 alone, it is optimal to use a single server because $K_{\{1\}}(0) = \hat{B}(0, \frac{1}{10}) \cdot \frac{1}{10} \cdot 15 = 1.5$, $K_{\{1\}}(1) = h + \hat{B}(1, \frac{1}{10}) \cdot \frac{1}{10} \cdot 15 = 1\frac{3}{22}$, and $K_{\{1\}}(2) > 2h = 2$. In similar fashion, we can find the optimal numbers of servers and the characteristic costs of all other coalitions. See Table 4.5.

Note that this game is balanced: for example, the allocation proportional to players' arrival rates, which assigns $\frac{1}{10}d^\varphi(N) = \frac{63}{160}$ to both players 1 and 2 and $\frac{8}{10}d^\varphi(N) = 3\frac{24}{160}$ to player 3, is a core element. However, this game is not concave, since $d^\varphi(\{1,3\}) - d^\varphi(\{3\}) = 3\frac{6561}{9706} - 3\frac{384}{827} < 0.212 < 0.261 < 3\frac{15}{16} - 3\frac{6561}{9706} = d^\varphi(\{1,2,3\}) - d^\varphi(\{2,3\})$. ◇

Next up is the service constraint formulation, which is suitable when penalty costs are hard to quantify. Our approach is as follows: We will let each player incur the penalty costs for its own blocked customers, and we will look for an allocation of *only* the resource costs that ensures that no subcoalition can improve (achieve the same blocking probability at lower resource cost) by splitting off. We do not consider transfer payments of penalty costs because the blocking probability for every coalition will be set at the same number; hence, the penalty costs will represent an additive term, unaffected by collaboration. Moreover, when penalty costs are hard to quantify, enforcing monetary transfer payments may be difficult in practice.

The blocking probability that each coalition will achieve is exogenously set at a number $\beta \in (0,1)$, the *service constraint*. This number indicates the maximal blocking probability for any arriving customer. We assume that all players impose the same service constraint for their customers; this is analogous to assuming that all players have identical penalty costs.

A game will be associated with the combination of an Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ and a service constraint β. Determining an adequate β is not trivial: Should it be 5%, 1%, or something else? It may be easier to think in monetary terms by estimating that a blocked customer costs a certain number of euros. Even if these penalty costs are hard to quantify, a good estimate for the penalty costs can induce an adequate service constraint. We call the service constraint $\beta^\varphi = \hat{B}(S^*_N, \lambda_N)$, which describes the blocking probability for

the grand coalition under the cost-minimizing number of servers, *natural for* φ. "Natural" refers to the equivalence between penalty costs and service level via Lagrange relaxation, as described in van Houtum and Zijm (2000).

To formulate the service constraint game, we assume that any coalition $M \in 2^N_-$ will pick the (possibly "mixed", cf. van Houtum and Zijm, 2000) number of servers that yields a blocking probability of exactly β.[34] That is, M picks the unique $s \in \mathbb{R}_{++}$ such that $L(s, \lambda_M \tau) = \beta$, where L is the linear interpolation as defined in Equation (4.7). We will denote this number of servers by σ_M. It is well-defined because $L(s, \lambda_M \tau)$ as a function of s on $\mathbb{R}_+$ is strictly decreasing and continuous, with image $(0, 1]$. These properties hold by Part (iv) of Theorem 4.1 and by nature of being a linear interpolation.

Definition 4.5. For the combination of an Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ and an service level constraint $\beta \in (0, 1)$, the game (N, d^φ_β) with characteristic costs

$$d^\varphi_\beta(M) = H^{lin}(\sigma_M) \tag{4.22}$$

for all $M \in 2^N_-$ is called the associated *Erlang loss game with service constraints.*

The following example provides an illustration.

Example 4.7. Reconsider the 3-player linear Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ from Example 4.6. Take the natural service constraint $\beta = \hat{B}(S^*_N, \lambda_N) = \frac{1}{16}$. The associated Erlang loss game with service constraints (N, d^φ_β) is given in Table 4.5. For all $M \in 2^N_-$, it holds that

$$L(\sigma_M, \lambda_M \tau) = (1 - (\sigma_M - \lfloor \sigma_M \rfloor)) \cdot \hat{B}(\lfloor \sigma_M \rfloor, \lambda_M \tau) + (\sigma_M - \lfloor \sigma_M \rfloor) \cdot \hat{B}(\lceil \sigma_M \rceil, \lambda_M \tau) = \frac{1}{16},$$

as is easily verified from Table 4.5 by using that $d^\varphi_\beta(M) = h\sigma_M = \sigma_M$.

Note that (N, d^φ_β) is concave, in contrast to the Erlang loss game with optimized numbers of servers (N, d^φ) for Example 4.6. Furthermore, the allocation proportional to players' arrival rates, which assigns $\frac{1}{10} d^\varphi_\beta(N) = \frac{3}{10}$ to both players 1 and 2 and $\frac{8}{10} d^\varphi_\beta(N) = 2\frac{2}{5}$ to player 3, is a core element. ◇

4.5.2 Relationship to the game with fixed numbers of servers

In this section, we investigate the relationship between games with fixed numbers of servers and games with optimized numbers of servers. Because the analogue between the two is clearer for the cost-based formulation, we do not consider the service level formulation in this section. To obtain examples with non-balanced games, we consider situations with asymmetric penalty costs. However, the main conclusions that we will draw, regarding shrinking cores and the

[34] This approach is similar to the multiplexing of priority disciplines by García-Sanz et al. (2008) in such a way that each of them will operate during a desired percentage of time, as mentioned in Footnote 21 on page 51.

Coalition M	λ_M	$\hat{B}(1,\lambda_M)$	$\hat{B}(2,\lambda_M)$	S_M	$c^\varphi(M)$	S^*_M	$d^\varphi(M)$
{1}	1	$\frac{1}{2}$	$\frac{1}{5}$	0	$1\frac{3}{10}$	0	$1\frac{3}{10}$
{2}	1	$\frac{1}{2}$	$\frac{1}{5}$	0	$1\frac{1}{10}$	0	$1\frac{1}{10}$
{3}	1	$\frac{1}{2}$	$\frac{1}{5}$	1	$2\frac{1}{5}$	1	$2\frac{1}{5}$
{1,2}	2	$\frac{2}{3}$	$\frac{2}{5}$	0	$2\frac{2}{5}$	0	$2\frac{2}{5}$
{1,3}	2	$\frac{2}{3}$	$\frac{2}{5}$	1	$3\frac{7}{15}$	1	$3\frac{7}{15}$
{2,3}	2	$\frac{2}{3}$	$\frac{2}{5}$	1	$3\frac{1}{3}$	1	$3\frac{1}{3}$
{1,2,3}	3	$\frac{3}{4}$	$\frac{9}{17}$	1	$4\frac{3}{5}$	2	$4\frac{46}{85}$

Table 4.6: *The parameter values, blocking probabilities, and games for Example 4.8.*

lack of an obvious relationship between the two classes of games, would remain valid if we would limit ourselves to symmetric penalty costs.

We start with an example of an Erlang loss situation for which the associated game with fixed numbers of servers has an empty core, while the associated game with optimized numbers of servers has a non-empty core. The intuition behind this is that, if adjustment of the number of servers is allowed, then the grand coalition will face no higher costs than in the case with fixed numbers of servers. Such a reduction in the costs of the grand coalition can annul the core's emptiness if all other coalitions retain their original costs.

Example 4.8. Consider the 3-player linear Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ with $N = \{1,2,3\}$, $S_1 = S_2 = 0$, $S_3 = 1$, $\lambda_i = 1$ for all $i \in N$, $\tau = 1$, $h = 1$, $p_1 = 1\frac{3}{10}$, $p_2 = 1\frac{1}{10}$, and $p_3 = 2\frac{2}{5}$. The associated game with fixed numbers of servers, (N, c^φ), and the associated game with fixed numbers of servers, (N, d^φ), are given in Table 4.6.

Note that (N, c^φ) has an empty core because $c^\varphi(N) = 4\frac{3}{5} > 4\frac{17}{30} = c^\varphi(\{2\}) + c^\varphi(\{1,3\})$. Also note that $c^\varphi(M) = d^\varphi(M)$ for all $M \subset N$ and that $c^\varphi(N) > d^\varphi(N)$. Indeed, the (sum of the) given numbers of servers are optimal for each singleton and two-player coalition. But the grand coalition can improve by re-optimizing its number of servers: the grand coalition's costs are minimized by two servers instead of one. The game (N, d^φ) has a non-empty core: e.g., the allocation $(1\frac{3}{10}, 1\frac{1}{10}, 2\frac{12}{85})$ is stable. ◇

The following example shows that the converse may also occur: a game with fixed numbers of servers may have a non-empty core while the corresponding game with optimized numbers of servers has an empty core. The intuitive explanation for this is that if sub-coalitions can choose a cost-minimizing number of servers, then they might reduce their costs. The core will shrink as the result of any such cost reduction if the grand coalition's cost stay the same. This example simultaneously shows that optimal numbers of servers need not be monotonic.

Example 4.9. Consider the 3-player linear Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ with $N = \{1,2,3\}$, $S_1 = S_2 = S_3 = 0$, $\lambda_i = 1$ for all $i \in N$, $\tau = 1$, $h = 1$, $p_1 = 1\frac{1}{2}$, $p_2 = \frac{1}{10}$, and

Coalition M	λ_M	$\hat{B}(1,\lambda_M)$	S_M	$c^{\varphi}(M)$	S_M^*	$d^{\varphi}(M)$
{1}	1	$\frac{1}{2}$	0	$1\frac{1}{2}$	0	$1\frac{1}{2}$
{2}	1	$\frac{1}{2}$	0	$\frac{1}{10}$	0	$\frac{1}{10}$
{3}	1	$\frac{1}{2}$	0	$1\frac{9}{10}$	0	$1\frac{9}{10}$
{1,2}	2	$\frac{2}{3}$	0	$2\frac{3}{5}$	0	$2\frac{3}{5}$
{1,3}	2	$\frac{2}{3}$	0	$3\frac{2}{5}$	1	$3\frac{4}{15}$
{2,3}	2	$\frac{2}{3}$	0	2	0	2
{1,2,3}	3	$\frac{3}{4}$	0	$3\frac{1}{2}$	0	$3\frac{1}{2}$

Table 4.7: *The parameter values, blocking probabilities, and games for Example 4.9.*

$p_3 = 1\frac{9}{10}$. The associated game with fixed numbers of servers, (N, c^{φ}), and the associated game with fixed numbers of servers, (N, d^{φ}), are given in Table 4.7.

Since no player possesses a server, there are no gains of cooperation if the number of servers is unadjustable. Accordingly, (N, c^{φ}) has a non-empty core: e.g., the allocation $(1\frac{1}{2}, \frac{1}{10}, 1\frac{9}{10})$ is stable. We also observe that $c^{\varphi}(M) = d^{\varphi}(M)$ for all coalitions M other than $\{1,3\}$ and that $c^{\varphi}(\{1,3\}) > d^{\varphi}(\{1,3\})$. Indeed, zero servers is optimal for each coalition other than $\{1,3\}$. But coalition $\{1,3\}$ can improve by re-optimizing its number of servers: the costs of coalition $\{1,3\}$ are minimized by one server instead of zero. (Note that the optimal number of servers is not monotonic in this example: zero servers remains optimal for the grand coalition. So, by adding player 2 to coalition $\{1,3\}$, the optimal number of servers may decrease.) The game (N, d^{φ}) has a non-empty core because $d^{\varphi}(\{2\}) + d^{\varphi}(\{1,3\}) = 3\frac{11}{30} < 3\frac{1}{2} = d^{\varphi}(N)$. ◇

So, when comparing the core of a game with fixed numbers of servers to the core of the corresponding game with optimized numbers of servers, we cannot draw a general conclusion: Example 4.8 shows that the core of the former game may be smaller, whereas Example 4.9 shows that it may be larger. Thus, results for the former setting do not directly imply similar results for the latter setting, or vice versa.

The two effects of full pooling (introduced in Section 4.4.2) are also of no immediate help in characterizing the class of games with optimized numbers of servers with non-empty cores. As shown in the following example, the change in the weighted number of blocked customers (i.e., Effect 1, which is always beneficial for the setting with fixed numbers of servers, cf. Lemma 4.12) need not be beneficial for the setting with optimized numbers of servers. The intuition behind this is that, due to the risk pooling effect, fewer servers may suffice to jointly serve all arrival streams cost-effectively in the grand coalition.

Example 4.10. Reconsider Example 4.9. Suppose that the players had partitioned themselves into two separate pooling groups: one with player 2 and one with players 1 and 3. Then, under optimal numbers of servers, the expected number of blocked customers per unit

time for both groups together would be $\hat{B}(0,1)\cdot 1+\hat{B}(1,2)\cdot 2 = 2\frac{1}{3}$. Now suppose that the grand coalition would form and operate under zero servers, which is optimal. Then, resource costs are considerably reduced compared to the aforementioned arrangement in which players are partitioned in groups, but the expected number of blocked customers per unit time increases to $\hat{B}(0,3)\cdot 3 = 3$. As $3 > 2\frac{1}{3}$, full pooling under optimal numbers of servers may actually increase the number of blocked customers! This follows from the lack of monotonicity of the optimal number of servers. ◇

Because, for a given Erlang loss situation φ, there is no direct relation between balancedness of (N, c^φ) and (N, d^φ), and because the effects of full pooling cannot be used to establish intuitive conditions for balancedness of (N, d^φ) either, a separate analysis of (N, d^φ) is worthwhile. In the next section, we will derive several results for (N, d^φ) under the assumption that penalty costs are symmetric among players. The reason for that assumption lies in Example 4.9, which showed that (N, d^φ) need not be balanced when penalty costs are asymmetric.

4.5.3 The cost-based formulation

For the pure cost model, an easy way of allocating the total costs of the grand coalition—or any other coalition that may form—is by dividing these costs proportional to the arrival rate of each player. Formally, we define the resulting allocation scheme rule $\mathcal{P}$ for any Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ with unbounded H by

$$\mathcal{P}_{i,M}(\varphi) = d^\varphi(M)\cdot \lambda_i/\lambda_M. \tag{4.23}$$

for all $M \in 2^N_-$ and all $i \in M$. If costs are shared according to this rule, then a player with more frequent customer arrivals pays a greater share of the costs. The following example provides an illustration.

Example 4.11. Reconsider the Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ of Example 4.6. Recall that $\lambda_1 = \lambda_2 = \frac{1}{10}$ and $\lambda_3 = \frac{8}{10}$. The corresponding allocation scheme $\mathcal{P}(\varphi)$ is shown in Table 4.8. Notice that $\mathcal{P}_{1,\{1,2\}}(\varphi) = \frac{3}{4} > \frac{63}{160} = \mathcal{P}_{1,N}(\varphi)$ and similarly $\mathcal{P}_{2,\{1,2\}}(\varphi) > \mathcal{P}_{2,N}(\varphi)$, i.e., the amount that player 1 or 2 has to pay does not increase when player 3 joins them. This can be verified for the members of all such nested pairs of coalitions as well, implying that $\mathcal{P}(\varphi)$ is population monotonic.

We remark that the associated game (N, d^φ) is a single-attribute game. Indeed, recalling that $h = 1$ and $p_i = 15$ for all $i \in N$, we can construct a single-attribute situation $\gamma = (N, \lambda, \tilde{K})$ with $\tilde{K}(\ell) = \min_{s\in\mathbb{N}_0}\{s + \hat{B}(s,\ell)\cdot 15\ell\}$ for each $\ell > 0$ and $\tilde{K}(0) = 0$ such that the associated single-attribute game (N, c^γ) is equal to the Erlang loss game with optimized numbers of servers (N, d^φ). We remark that $\tilde{K}$, as illustrated in Figure 4.3 on page 90, is non-decreasing. Indeed, by Part (vi) of Theorem 4.1, $\tilde{K}$ is the minimum of a collection of non-decreasing functions. ◇

Coalition M	$d^{\varphi}(M)$	$\mathcal{P}_{1,M}(\varphi)$	$\mathcal{P}_{2,M}(\varphi)$	$\mathcal{P}_{3,M}(\varphi)$
$\{1\}$	$1\frac{3}{22}$	$1\frac{3}{22}$	*	*
$\{2\}$	$1\frac{3}{22}$	*	$1\frac{3}{22}$	*
$\{3\}$	$3\frac{384}{827}$	*	*	$3\frac{384}{827}$
$\{1,2\}$	$1\frac{1}{2}$	$\frac{3}{4}$	$\frac{3}{4}$	*
$\{1,3\}$	$3\frac{6561}{9706}$	$\frac{11893}{29118}$	*	$3\frac{3895}{14559}$
$\{2,3\}$	$3\frac{6561}{9706}$	*	$\frac{11893}{29118}$	$3\frac{3895}{14559}$
N	$3\frac{15}{16}$	$\frac{63}{160}$	$\frac{63}{160}$	$3\frac{24}{160}$

Table 4.8: *The values for the allocation scheme $\mathcal{P}_{i,M}(\varphi)$ in Example 4.11.*

This population monotonicity that we saw in this example is not a coincidence, as revealed by the following theorem.

Theorem 4.18. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation with unbounded H and $p_i = p_j$ for all $i, j \in N$. Then $\mathcal{P}(\varphi)$ is a PMAS of the associated Erlang loss game with optimized numbers of servers (N, d^{φ}).*

Proof. For any coalition $M \in 2^N_-$, we extend the domain of the cost function K_M, introduced in Equation (4.19), by a linear interpolation, i.e., we define, for all $s \in \mathbb{R}_+$,

$$K_M^{lin}(s) = (1 - (s - \lfloor s \rfloor)) \cdot K_M(\lfloor s \rfloor) + (s - \lfloor s \rfloor) \cdot K_M(\lceil s \rceil) = H^{lin}(s) + L(s, \lambda_M \tau) \cdot \lambda_M P.$$

We now fix two arbitrary coalitions $Q, R \in 2^N_-$ with $Q \subseteq R$, and we let $i \in Q$. Then,

$$\begin{aligned}
\mathcal{P}_{i,R}(\varphi) &= \frac{\lambda_i}{\lambda_R} \cdot K_R(S_R^*) \\
&\leq \frac{\lambda_i}{\lambda_R} \cdot K_R^{lin}\left(\frac{\lambda_R}{\lambda_Q} S_Q^*\right) \\
&= \frac{\lambda_i}{\lambda_R} \cdot H^{lin}\left(\frac{\lambda_R}{\lambda_Q} S_Q^*\right) + L\left(\frac{\lambda_R}{\lambda_Q} S_Q^*, \lambda_R \tau\right) \cdot \lambda_i p_i \\
&\leq \frac{\lambda_i}{\lambda_R} \cdot H^{lin}\left(\frac{\lambda_R}{\lambda_Q} S_Q^*\right) + L(S_Q^*, \lambda_Q \tau) \cdot \lambda_i p_i \\
&= \frac{\lambda_i}{\lambda_Q} \cdot \frac{\lambda_Q}{\lambda_R} \cdot H^{lin}\left(\frac{\lambda_R}{\lambda_Q} S_Q^*\right) + \hat{B}(S_Q^*, \lambda_Q \tau) \cdot \lambda_i p_i \\
&= \frac{\lambda_i}{\lambda_Q} \left[\frac{\lambda_Q}{\lambda_R} \cdot H^{lin}\left(\frac{\lambda_R}{\lambda_Q} S_Q^*\right) + \frac{\lambda_R - \lambda_Q}{\lambda_R} \cdot H^{lin}(0)\right] + \hat{B}(S_Q^*, \lambda_Q \tau) \cdot \lambda_i p_i \\
&\leq \frac{\lambda_i}{\lambda_Q} \cdot H^{lin}(S_Q^*) + \hat{B}(S_Q^*, \lambda_Q \tau) \cdot \lambda_i p_i \\
&= \frac{\lambda_i}{\lambda_Q} \cdot K_Q(S_Q^*) = \mathcal{P}_{i,Q}(\varphi).
\end{aligned}$$

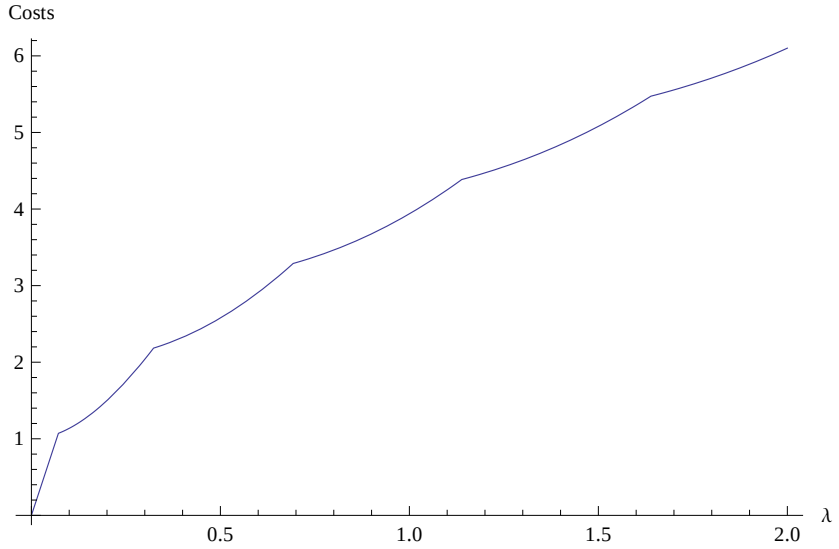

Figure 4.3: *A plot of the function* $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}$ *as defined in Example 4.11.*

The first inequality holds because S_R^* is a cost minimizing number of servers for coalition R, so both $K_R(\lfloor S_Q^* \lambda_R/\lambda_Q \rfloor)$ and $K_R(\lceil S_Q^* \lambda_R/\lambda_Q \rceil)$ are no smaller than $K_R(S_R^*)$, and the same holds for the associated linear interpolation. The second inequality holds because L is scalable (Theorem 4.11) while $S_Q^* \in \mathbb{N}_0$. The subsequent (third) equality holds because L is an extension. The consecutive (fourth) equality holds because, by assumption, $H^{lin}(0) = 0$. The third inequality holds because H^{lin} is concave function and $S_Q^* = [\lambda_Q/\lambda_R] \cdot (S_Q^* \lambda_R/\lambda_Q) + [(\lambda_R - \lambda_Q)/\lambda_R] \cdot 0$. We conclude that $\mathcal{P}(\varphi)$ is a PMAS. □

Özen et al. (2011) independently derived the same result as stated in Theorem 4.18, though for *linear* Erlang loss situations only. Their proof approach uses single-attribute games and elasticity of the function $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ defined by $\tilde{K}(0) = 0$ and $\tilde{K}(\ell) = \min_{s \in \mathbb{N}_0} \{hs + \hat{B}(s, \ell\tau)\ell P\}$ for all $\ell > 0$; see Figure 4.3. Our proof approach is different: we use structural properties of extensions of the Erlang loss function. In our proof, the linear interpolation L could not have been replaced by the continuous extension B because then, as shown in the following example, an inequality corresponding to the proof's first inequality would not hold anymore.

Example 4.12. Consider the 2-player linear Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ with $N = \{1, 2\}$, arbitrary S, $\lambda_1 = 0.6$, $\lambda_2 = 0.4$, $\tau = 1$, $h = 0.85$, and $p_1 = p_2 = 3$. For coalition $Q = \{1\}$, the optimal number of servers is $S_Q^* = 1$. For coalition $R = \{1, 2\}$, the optimal number of servers is $S_R^* = 2$ with associated optimal costs $K_R(S_R^*) = 2.3$. However, for $s = S_Q^* \lambda_R/\lambda_Q = 1\frac{2}{3}$, it holds that $0.85s + B(s, \lambda_R\tau) \cdot 3\lambda_R < 2.26 < 2.3 = K_R(S_R^*)$. So, the costs for coalition R under the optimal (integer) number of servers are actually *larger* than the costs for this coalition under $1\frac{2}{3}$ servers when the associated blocking probability

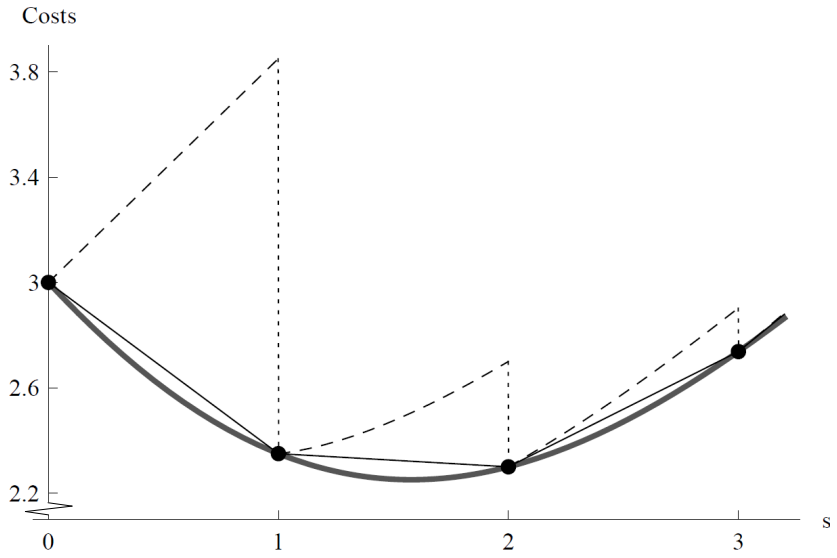

Figure 4.4: *For each of the three extensions of the Erlang loss function and for the Erlang loss function itself, the corresponding costs* $hs + B(s, \lambda_R\tau) \cdot \lambda_R P$ *(thick curve),* $hs + L(s, \lambda_R\tau) \cdot \lambda_R P$ *(straight line segments),* $hs + X(s, \lambda_R\tau) \cdot \lambda_R P$ *(dashed, discontinuous), and* $K_R(s)$ *(big dots) as functions of* s *for the coalition* R *as described in Example 4.12.*

is interpreted via the continuous extension B. This is graphically represented in Figure 4.4, which simultaneously illustrates validity of the first inequality in the proof of Theorem 4.18 due to the behavior of the extension L. ◇

By Sprumont (1990), population monotonicity of $\mathcal{P}(\varphi)$ not only signifies the advantageousness of any pooling group's growth but also implies the following corollary.

Corollary 4.19. *Let* $\varphi = (N, S, \lambda, \tau, H, p)$ *be an Erlang loss situation with unbounded* H *and* $p_i = p_j$ *for all* $i, j \in N$*. Then, the associated game* (N, d^φ) *is totally balanced and the allocation assigning* $\mathcal{P}_{i,N}(\varphi)$ *to all* $i \in N$ *is an element of its core.*

4.5.4 The service-based formulation

For an Erlang loss situation $\varphi = (N, S, \lambda, \tau, H, p)$ and a service constraint $\beta \in (0, 1)$, we propose a proportional allocation $\mathscr{P}(\varphi, \beta)$, defined by $\mathscr{P}_i(\varphi, \beta) = H^{lin}(\sigma_N) \cdot \lambda_i/\lambda_N$ for each player $i \in N$. The following theorem states that this allocation is stable, provided that there is an integer in the interval $[\sigma_N\lambda_M/\lambda_N, \sigma_N]$ for all subcoalitions M. This is a technical requirement that simultaneously illustrates the limitation and allure of scalability. This requirement is met if σ_N is a sufficiently large real number and each player is big enough. It is also met if σ_N is any integer number.

Theorem 4.20. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation and let $\beta \in (0, 1)$. If $\lfloor \sigma_N \rfloor \geq \sigma_N \cdot \lambda_M / \lambda_N$ for all $M \in 2^N_{--}$, then $\mathscr{P}(\varphi, \beta) \in \mathcal{C}(N, d^\varphi_\beta)$.*

Proof. Let $M \in 2^N_{--}$. Since $\beta \in (0, 1)$, we must have $\sigma_N > 0$. Assume that $\lfloor \sigma_N \rfloor \geq \sigma_N \cdot \lambda_M / \lambda_N$. As $\sigma_N > 0$ and $\lambda_M / \lambda_N > 0$, this assumption implies that $\lfloor \sigma_N \rfloor > 0$. Then,

$$L(\sigma_M, \lambda_M \tau) = \beta = L(\sigma_N, \lambda_N \tau) \leq L(\lfloor \sigma_N \rfloor, \lambda_N \tau \frac{\lfloor \sigma_N \rfloor}{\sigma_N}) \leq L(\sigma_N \frac{\lambda_M}{\lambda_N}, \lambda_M \tau). \tag{4.24}$$

The equalities hold by definition of σ_M and σ_N, respectively. The inequalities hold because L is scalable (Theorem 4.11), which we can employ because $\lfloor \sigma_N \rfloor \in \mathbb{N}$ and because $\sigma_N \geq \lfloor \sigma_N \rfloor \geq \sigma_N \lambda_M / \lambda_N$.

Moreover, $L(s, \lambda_M \tau)$ is decreasing in s on $\mathbb{R}_+$, by Part (iv) of Theorem 4.1. Hence, Inequality (4.24) implies that $\sigma_N \lambda_M / \lambda_N \leq \sigma_M$. This yields

$$\sum_{i \in M} \mathscr{P}_i(\varphi, \beta) = H^{lin}(\sigma_N) \cdot \lambda_M / \lambda_N \leq H^{lin}(\sigma_N \cdot \lambda_M / \lambda_N) \leq H^{lin}(\sigma_M) = d^\varphi_\beta(M).$$

In the first inequality, we use that H^{lin} is concave and non-negative. In the second inequality, we use that H^{lin} is non-decreasing. We conclude that $\mathscr{P}(\varphi, \beta)$ is stable. □

We remark that it remains an open question whether or not service constraint games admit a PMAS in general. If L could be proven to satisfy subhomogeneity of degree zero (as discussed in the concluding paragraph of Section 4.2.3), then this open question could be answered positively via a straightforward adjustment of the proof of Theorem 4.20.

Under the *natural* service constraint, σ_N would be an integer number by definition. Hence, we immediately obtain the following corollary to Theorem 4.20.

Corollary 4.21. *Let $\varphi = (N, S, \lambda, \tau, H, p)$ be an Erlang loss situation. Under the natural service constraint β^φ, it holds that $\mathscr{P}(\varphi, \beta^\varphi) \in \mathcal{C}(N, d^\varphi_{\beta^\varphi})$.*

This corollary shows an interesting result for the class of Erlang loss games with service constraints under the natural service constraint. In this class, each individual player is associated with a single attribute only: his demand rate. All other parameters are aggregate—the same for all players. At first glance, this would seem to imply that for a given resource cost function, given symmetric penalty cost, and given mean service times, the resulting class of Erlang loss games with service constraints under the natural service constraint would coincide with the class of all single-attribute games embedded in a specific cost function, so that Theorem 2.5 applies. This, however, is not the case. The reason is that this specific cost function would depend, through β^φ, on the value of λ_N, which is not allowed in the definition of a single-attribute game. This is illustrated in the following example.

Example 4.13. Reconsider the 3-player linear Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ from Example 4.7. Recall that $\lambda_1 = 0.1$. Taking the natural service constraint $\beta = \hat{B}(S^*_N, \lambda_N) =$

$\frac{1}{16}$, it holds that $d^{\varphi}_{\beta}(\{1\}) = 1\frac{221}{672}$; see Table 4.5 on page 84. So, if we were to construct a single-attribute cost function $\tilde{K}$ corresponding to h, p, and τ, then it would need to satisfy $\tilde{K}(0.1) = 1\frac{221}{672}$.

Now suppose that we remove player 3 to end up with the Erlang loss situation $\hat{\varphi} = (\hat{N}, S, \hat{\lambda}, \tau, h, p)$ where $\hat{N} = \{1, 2\}$ and $\hat{\lambda}_1 = \hat{\lambda}_2 = 0.1$. For this situation, Table 4.5 reveals that that the natural service constraint is $\hat{\beta} = \hat{B}(S^*_{\hat{N}}, \lambda_{\hat{N}}) = \frac{1}{6}$. Accordingly, $d^{\hat{\varphi}}_{\hat{\beta}}(\{1\}) = \frac{11}{12}$. So, if we were to construct a single-attribute cost function $\tilde{K}$ corresponding to h, p, and τ, then it would need to satisfy $\tilde{K}(0.1) = \frac{11}{12}$. As $1\frac{221}{672} \neq \frac{11}{12}$, we conclude that such a construction is impossible: a single attribute is not sufficient to capture all required information. $\diamond$

4.6 Optimal pooling, rationing, and transshipments

So far, we assumed full pooling: each arriving customer is always admitted if there is a free server available and only rejected when there is no free server. However, it is sometimes better to reject arrivals from customers with low penalty costs. This way, servers can be held back in anticipation of future arrivals from customers with high penalty costs. (This is related to the discussion that followed Example 4.4 on page 77.)

In this section, we will describe a model that allows for optimization of the pooling policy. Our objective is not to provide a comprehensive analysis, which is outside the scope of this chapter, but rather to offer preliminary insights into the impact of some of our assumptions made so far. In particular, we will show that games that were not balanced under full pooling can become balanced under an optimal pooling policy.

Although the Erlang loss model allowed for generally distributed service times, we will assume in this section that the service time for each player is exponentially distributed. This assumption is not too restrictive because inventory models for low-demand spare parts tend to be quite insensitive to variability in the lead time (Alfredsson and Verrijdt, 1999). The assumption is convenient because it facilitates analysis via Markov processes. A Markovian model formulation permits easy incorporation of positive reassignment or transshipment costs. Such costs are an interesting feature for, e.g., a network of spare parts stockpoints that reside at geographically dispersed locations. Because it requires almost no additional modeling effort, we will expand our model with this interesting feature.

Now, on to the analysis. Consider a linear Erlang loss situation $(N, S, \lambda, \tau, h, p)$ and a vector $r \in \mathbb{R}_+^{N \times N}$. The number $r_{i,j}$ describes the reassignment or transshipment costs that are incurred whenever a customer of player $i \in N$ is served by a server of player $j \in N$. Consider a coalition $M \in 2^N_-$ with server vector $s = (s_i)_{i \in M} \in \mathbb{N}_0^M$. The number s_i describes the amount of servers that player i will use. It may correspond to the exogenously given number of servers (i.e., $s_i = S_i$), or it may be the result of an optimization process. We write $\mathcal{S} = \prod_{i \in M} \{0, 1, \ldots, s_i\}$ for the state space, where state $(\sigma_i)_{i \in M} \in \mathcal{S}$ means that player i has σ_i available servers.

A policy, at least in this section, describes for every state what to do with any arriving customer. Formally, a *policy* for coalition M in state space $\mathcal{S}$ is a function $\psi : M \times \mathcal{S} \to M \cup \{R\}$ with $\psi(i, \mathbf{0}) = R$ for every $i \in M$.[35] The interpretation of such a policy is that an arrival for player $i \in M$ in state $\sigma \in \mathcal{S}$ is served by player j if $\psi(i, \sigma) = j$ and rejected if $\psi(i, \sigma) = R$. The requirement for state $\mathbf{0}$ means that a customer is always rejected if there are no free servers available. Note that these policies are non-preemptive: once a customer has started service, he cannot be rejected if another customer of a possibly more critical class arrives. We denote the set of policies for coalition M under server vector $s \in \mathbb{N}_0^M$ by $\Psi_M(s)$.

Given server vector $s \in \mathbb{N}_0^M$ and policy $\pi \in \Psi_M(s)$, we focus on the stochastic process $X = (X^t(s, \psi))_{t \in \mathbb{R}_+}$ that takes values in $\mathcal{S}$. State changes can only occur when a customer arrives or completes service. Since the interarrival and service times are assumed to be exponentially distributed, X can be represented as a Markov process. The corresponding transition rate vector q is then described for any pair $\sigma, \hat{\sigma} \in \mathcal{S}$ with $\sigma \neq \sigma'$ by

$$q_{\sigma,\hat{\sigma}} = \begin{cases} \sum_{i \in M : \psi(i,\sigma) = j} \lambda_i & \text{if } \hat{\sigma}_i = \sigma_i \text{ for all } i \in M \setminus \{j\} \text{ and } \hat{\sigma}_j = \sigma_j - 1; \\ (s_j - \sigma_j)/\tau & \text{if } \hat{\sigma}_i = \sigma_i \text{ for all } i \in M \setminus \{j\} \text{ and } \hat{\sigma}_j = \sigma_j + 1; \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

As the state space of the Markov process X is finite,[36] we can derive its steady-state probability distribution $(\pi_\sigma(s, \psi))_{\sigma \in \mathcal{S}}$ via standard linear algebra techniques, e.g., Gaussian elimination. Using this, the expected costs per time unit in steady state are described by

$$K_M(s, \psi) = \sum_{i \in M} \left(h_i s_i + \sum_{\sigma \in \mathcal{S} : \psi(i,\sigma) = R} p_i \lambda_i \pi_\sigma(s, \psi) + \sum_{j \in M} \sum_{\sigma \in \mathcal{S} : \psi(i,\sigma) = j} r_{i,j} \lambda_i \pi_\sigma(s, \psi) \right).$$

We are now ready to formulate two games.

Definition 4.6. Let $\varphi = (N, S, \lambda, \tau, h, p)$ be a linear Erlang loss situation and let $r \in \mathbb{R}_+^{N \times N}$.

We define the game $(N, \bar{c}^\varphi)$ with characteristic costs

$$\bar{c}^{\varphi,r}(M) = \min_{\psi \in \Psi_M((S_i)_{i \in M})} K_M((S_i)_{i \in M}, \psi) \tag{4.26}$$

for all $M \in 2_-^N$ to be the associated *optimal pooling game with fixed numbers of servers.*

If $h > 0$, then we define the game $(N, \bar{d}^\varphi)$ with characteristic costs

$$\bar{d}^{\varphi,r}(M) = \min_{s \in \mathbb{N}_0^M} \min_{\psi \in \Psi_M(s)} K_M(s, \psi) \tag{4.27}$$

for all $M \in 2_-^N$ to be the associated *optimal pooling game with optimized numbers of servers.*

[35] We tacitly assume that there is no player named R in the player set N.

[36] The "mandatory" blocking of customers when the state space is finite is responsible for this finiteness. In case of unlimited waiting space (in a queueing context) or backorders (in an inventory context), no customers would be rejected, so the state space would become infinite and the derivation of steady-state probabilties is less straightforward. Accordingly, we do not consider asymmetric penalty costs in the later chapters.

Optimal policies exist because $\Psi_M(s)$ is finite for every $M \in 2^N_-$ and every $s \in \mathbb{N}_0^M$. Furthermore, the first minimum in Equation (4.27) exists because $h > 0$ implies that costs grow unboundedly as the number of servers tends to infinity. We conclude that optimal pooling games are well-defined.

If $r = \mathbf{0}$ and $p_i = p_j$ for all $i, j \in N$, then full pooling, i.e., the policy with $\psi(i, \sigma) = R$ only if $\sigma = \mathbf{0}$, is optimal for every coalition (Miller, 1969; van Jaarsveld and Dekker, 2009) and, consequently, an optimal pooling game matches an Erlang loss game. Hence, by Examples 4.3 and 4.6, optimal pooling game with fixed and optimized numbers of servers, respectively, need not be concave. Nevertheless, optimal pooling games are easily shown to be subadditive because the policies used by two disjoint coalitions induce a feasible policy (operating two "separate" pools) for their union with the same costs. We omit the obvious proof.

Theorem 4.22. *Both optimal pooling games with fixed numbers of servers and optimal pooling games with optimized numbers of servers are subadditive.*

Since the stochastic process underlying the optimal pooling games is no longer an Erlang loss system, we can no longer draw upon the analytical properties of the Erlang loss function to prove, e.g., balancedness. What we *can* show is that a switch from full pooling to optimal pooling may turn a non-balanced game into a balanced one. We show this in the following two examples, respectively for game with fixed and optimized numbers of servers.

Example 4.14. Reconsider the Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ from Example 4.8. This situation featured server counts $S_1 = 0$, $S_2 = 0$, and $S_3 = 1$ and penalty costs $p_1 = 1\frac{3}{10}$, $p_2 = 1\frac{1}{10}$, and $p_3 = 2\frac{2}{5}$. Suppose that $r = \mathbf{0}$.

To determine the associated optimal pooling game with fixed numbers of servers, $(N, \bar{c}^{\varphi,r})$, we need to determine the optimal pooling policies. For the singleton coalitions $\{1\}$, $\{2\}$, and $\{3\}$, this is trivial because there is only one customer stream. For coalition $\{1, 2\}$, this is trivial because the coalition has 0 servers. For coalitions $\{1, 3\}$ and $\{2, 3\}$, we determine by enumeration that full pooling is optimal. So, for all $M \subset N$, it holds that $\bar{c}^{\varphi,r}(M) = c^{\varphi}(M)$, with $c^{\varphi}(M)$ as in Table 4.6.

For the grand coalition, we determine by enumeration that the optimal policy always accepts customers for players 1 and 3 and always rejects customers for player 2. The corresponding costs, by Markov chain analysis, are $\bar{c}^{\varphi,r}(N) = 3\frac{17}{30}$. The game $(N, \bar{c}^{\varphi,r})$ admits a stable allocation, e.g., $(1\frac{3}{10}, \frac{1}{10}, 2\frac{1}{6})$, and thus has a non-empty core. This is in contrast to the associated Erlang loss game with fixed numbers of servers, which had an empty core. $\diamond$

Example 4.15. Reconsider the Erlang loss situation $\varphi = (N, S, \lambda, \tau, h, p)$ from Example 4.9. This situation featured penalty costs $p_1 = 1\frac{1}{2}$, $p_2 = \frac{1}{10}$, and $p_3 = 1\frac{9}{10}$. Suppose that $r = \mathbf{0}$.

To determine the associated optimal pooling game with optimized numbers of servers, $(N, \bar{d}^{\varphi,r})$, we need to determine the optimal pooling policies. The singleton coalitions are again trivial. For coalitions $\{1, 2\}$ and $\{2, 3\}$, enumeration reveals that 0 servers remain

Conditions	Symmetrical p	Arbitrary p
Fixed numbers of servers; Symmetrical S/λ	Balanced (Theorem 4.14); PMAS (Theorem 4.16)	Balanced (Corollary 4.17); PMAS (Theorem 4.16)
Fixed numbers of servers; Arbitrary S/λ	Balanced (Theorem 4.14)	May be balanced or non-balanced (Example 4.4)
Optim. numbers of servers; Pure cost formulation	Balanced (Corollary 4.19); PMAS (Theorem 4.18)	May be balanced or non-balanced (Example 4.9)
Optim. numbers of servers; natural service constraint	Balanced (Corollary 4.21)	Not studied

Table 4.9: *Overview of the main results.*

optimal. For coalition $\{1,3\}$, enumeration shows that 1 server under full pooling is optimal. So, for all $M \subset N$, it holds that $\bar{c}^{\varphi,r}(M) = c^{\varphi}(M)$, with $c^{\varphi}(M)$ as in Table 4.7.

For the grand coalition, enumeration shows that 1 server is optimal under the policy that always accepts customers for players 1 and 3 and always rejects customers for player 2. The corresponding costs, by Markov chain analysis, are $\bar{d}^{\varphi,r}(N) = 3\frac{11}{30}$. The game $(N, \bar{d}^{\varphi,r})$ admits a stable allocation, e.g., $(1\frac{2}{5}, \frac{1}{10}, 1\frac{13}{15})$, and thus has a non-empty core. This is in contrast to the associated Erlang loss game with optimized numbers of servers, which had an empty core. ◇

It remains an open question whether optimal pooling games are balanced in general.

4.7 Conclusion

We presented cooperative games corresponding to a situation where several players pool their resources into a joint Erlang loss system to serve the union of their individual customer arrival streams. Although the corresponding Erlang loss games (with either fixed or optimized numbers of servers) are not concave, balanced, or even subadditive in general, they are balanced and sometimes even PMAS-admissible when penalty costs are symmetric. Table 4.9 provides an overview of our main results regarding balancedness and population monotonicity.

In the formulation of Erlang loss games, we made several assumptions. In particular, we assumed a full pooling policy and neglected transshipment/reassingment costs. We analyzed the impact of relaxing these assumptions in Section 4.6; we found that the corresponding optimal pooling games are always subadditive. Balancedness remains an interesting, though mathematically challenging, open question for future research. Alternatively, useful insights for optimal pooling games may be generated via a numerical experiment that investigates the stability of various allocation mechanisms in realistic parameter settings.

Another limitation is that we assumed symmetrical service time distributions. We made

this assumption because customer streams such as repair requests for the same type of failed component often come with the same repair lead times. However, if customers of different players would have vastly *different* mean service times, then full pooling can severely degrade performance and thus possibly lead to an increase in costs. The intuition behind this is that customers with long service times can adversely affect the blocking probability for customers with short service times. The issue may be circumvented by preferential treatment of more critical classes via partial pooling approaches, in line with Section 4.6. For discussions on the issue of asymmetric service times (without a game theoretical perspective), we refer to Smith and Whitt (1981) and Papier and Thonemann (2008). Tackling this asymmetry from a game theoretical perspective may make for an interesting avenue for future research.

A final direction for future research lies in the formulation and analysis of alternative Erlang loss games and scalability under heavy traffic formulas with square-root staffing (see, e.g., Jagerman, 1974; Janssen et al., 2008). Inspiration could be drawn from the queueing games in the asymptotic Halfin-Whitt regime that were studied by Yu et al. (2007) for the $M/M/s$ setting. It is important to note, however, that many of the motivating applications (e.g., spare parts pooling) often feature rather low traffic, such that a single resource is often optimal and a heavy traffic approximations are unsuitable and unnecessary. Moreover, the structural results that we proved via the exact Erlang loss formula will also remain valid for very large arrival rates. Nevertheless, a focus on asymptotic approximations may yield new insights and results, possibly regarding concavity of the games.

—Everything comes in time to those who can wait.

François Rabelais

5

Erlang delay games

5.1 Introduction

In this chapter, which is based on Karsten et al. (2011b), we formulate and analyze games, called Erlang delay games, to study the cost allocation problem in multi-server queueing systems with infinite waiting room. As in the previous chapter, our model features several service providers, who are associated with their own customer populations. The service providers can collaborate by sharing their servers and individual customer streams into a pooled resource system. In contrast to the previous chapter, we will allow waiting in a queue. Accordingly, we model the service system of any coalition as an Erlang delay system, i.e., an $M/M/s$ model with infinite waiting room. Resource pooling in this context results in reduced congestion, as measured by the expected time spent by customers waiting to be served (Smith and Whitt, 1981).

The $M/M/s$ queueing model has applications to a broad range of service systems in practice (see, e.g., the literature overview in Kolesar and Green, 1998). One may think of manufacturers of advanced technical equipment that employ a number of non-branded repairmen to maintain and repair machines at their customer's sites. One can also think of call centers sharing cross-trained telephone agents, airline companies pooling check-in counters, or clinical hospital departments sharing diagnostic equipment. The key is that customers can wait in a queue, and that both interarrival and service time distributions are exponential.

Costs consist of linear resource costs for servers and linear, symmetric delay costs for customers that have to wait before being served. Furthermore, all servers of the different

players have the same service rate for all incoming work, and all customers require the same (distribution of) service. We consider both the case of fixed numbers of servers and the case where the number of servers is optimized to minimize the sum of resource and delay costs.

As mentioned in Section 3.3.3, most of the previous work on cooperative queueing games (e.g., Anily and Haviv, 2010) has focused on the $M/M/1$ model, even though the $M/M/1$ model is not appropriate when service facilities consist of *multiple* servers whose service speeds are given. For such settings, the $M/M/s$ model is more accurate. And as we will show in this chapter, there are essential differences between $M/M/1$ games and Erlang delay games.

The first difference is due to the fact that if fewer than s customers are present, then the $M/M/1$ system can use its total service capacity, whereas the $M/M/s$ system has idle capacity. We note that as long as all servers are busy, the whole group of waiting customers in an $M/M/s$ system with service rate μ is served at the same rate as in an $M/M/1$ system with service rate $s\mu$, but this only holds when all servers are busy—it doesn't hold when fewer than s customers are present. Accordingly, the behavior of the two systems is different. As a result of this, Erlang delay games with fixed numbers of servers turn out not to admit a positive core allocation or a PMAS in general. This means that the simple proportional solution that worked for $M/M/1$ games with fixed service capacity does *not* work for Erlang delay games with fixed numbers of servers.

The second difference is due to the fact that the number of servers (i.e., total capacity) in an $M/M/s$ system can only be varied in discrete amounts—a limitation that is absent in $M/M/1$ systems where the service speed can be set at arbitrary levels. As it turns out, Erlang delay games with optimized numbers of servers are not balanced in general; we will show that the underlying reason is the integrality requirement. This means that the balancedness result that held true for $M/M/1$ games with optimized service capacity does *not* extend to Erlang delay games with optimized numbers of servers.

The integrality requirement leads to another complication, now from an analytical perspective. To prove that Erlang delay games with fixed numbers of servers are balanced, we would like to construct Erlang delay systems with fractional numbers of servers, similar to our approach in Chapter 4. However, the classic Erlang delay function, which describes the probability that an arrival must wait before beginning service, is not defined for non-integral numbers of servers, so we again have to dive into domain extensions. We will derive several new analytical properties of the (standard) continuous extension of the classic Erlang delay function. The properties that we will derive not only allow us to prove that certain Erlang delay games are balanced but also generalize and strengthen well-known characteristics of key performance measures in the $M/M/s$ model.

Altogether, the model, results, and analysis we present in this chapter make three primary contributions. First, we derive several new analytical properties of the (standard) continuous extension of the classic Erlang delay function. Second, we introduce Erlang delay games with fixed numbers of servers and prove that they are totally balanced, despite being markedly

different from $M/M/1$ games. Third, for Erlang delay games with optimized numbers of servers, we establish that the existence of a stable allocation and a PMAS is dependent on the domain over which optimization takes place. In particular, we show that if each coalition is required to choose an *integer* number of servers, this existence is not guaranteed.

The remainder of this chapter is organized as follows. We start in Section 5.2 with an analysis of the continuous extension of the classic Erlang delay function. Section 5.3 introduces Erlang delay situations. In Sections 5.4 and 5.5, we analyze the corresponding Erlang delay games with, respectively, fixed and optimized numbers of servers. We conclude in Section 5.6.

5.2 The continuous extension of the Erlang delay function

Consider an Erlang delay system, i.e., an $M/M/s$ queue. In such a system, as represented in Figure 2.4 on page 38, customers arrive according to a Poisson process with rate $\lambda > 0$. They are served by a group of $s \in \mathbb{N}$ homogeneous parallel servers. Service times are independent and exponentially distributed with rate $\mu > 0$. Customers who find all servers busy wait in an infinite capacity queue until served by the first available server. We let $a = \lambda/\mu$ denote the (offered) load.

The steady-state probability of delay (the probability that an arrival must wait before beginning service) in such a system is described by the classic *Erlang delay function*, first published by Erlang (1917). This function $\hat{C}$ is defined, for each $a > 0$ and $s \in \mathbb{N}$ with $s > a$, by

$$\hat{C}(s,a) = \left(1 + \sum_{y=0}^{s-1} \frac{s!(1-a/s)}{y!a^{s-y}}\right)^{-1}. \tag{5.1}$$

Another interesting performance measure, also derived by Erlang (1917), is the expected waiting time (delay before beginning service) experienced by an arbitrary customer in steady state. For any $\lambda > 0$, $\mu > 0$, and $s \in \mathbb{N}$ with $s > \lambda/\mu$, this waiting time equals

$$\hat{W}_q(s,\lambda,\mu) = \frac{\hat{C}(s,\lambda/\mu)}{s\mu - \lambda}. \tag{5.2}$$

Equations (5.1) and (5.2) are valid for any *non-biased* service discipline, i.e., a service discipline that selects the next customer to be served without taking the waiting customers' actual service lengths into account (cf. Cooper, 1981, pp. 95–98). Examples of non-biased service disciplines are service on a first-come first-serve basis, service in random order, or service on a last-come first-serve basis.

The following example illustrates the behavioral difference between $M/M/1$ and $M/M/s$ queues, which we alluded to earlier.

Example 5.1. Suppose that $\lambda = 0.5$, $s = 2$, and $\mu = 1$. Then, by Equation (5.2), the expected waiting time in an $M/M/1$ queue with arrival rate λ and service rate $s\mu$ is

$[\lambda/(s\mu)]/(s\mu - \lambda) = \frac{1}{6}$. In contrast, the expected waiting time in an $M/M/s$ queue with the same arrival rate and service rate μ per server is $\hat{C}(s, \lambda/\mu)/(s\mu - \lambda) = \frac{1}{15}$. Moreover, the service time differs between the two queueing models. So, approximating a multi-server system by a single-server system can be grossly inaccurate. ◇

For analytical purposes, it will be convenient to extend the domain of the Erlang delay function to non-integral values of s. Jagers and van Doorn (1991) have suggested a confluent hypergeometric function as a natural continuous extension. This function C is defined, for each $a > 0$ and $s \in \mathbb{R}$ with $s > a$, by

$$C(s, a) = \left(\int_0^\infty a e^{-ax} (1 + x)^{s-1} x dx \right)^{-1}. \tag{5.3}$$

For fixed $a > 0$, $C(s, a)$ is non-increasing and convex in s for $s \in (a, \infty)$ (Jagers and van Doorn, 1991). This analytic extension of the Erlang delay function enables a natural way to define the expected waiting time in an (artificial) queueing system with a non-integral number of servers: for any $\lambda > 0$, $\mu > 0$, and $s \in \mathbb{R}$ with $s > \lambda/\mu$, we define

$$W_q(s, \lambda, \mu) = \frac{C(s, \lambda/\mu)}{s\mu - \lambda}. \tag{5.4}$$

As observed by Jagers and van Doorn (1991), Equations (5.1) and (5.3) coincide for integer values of s, i.e., $\hat{C}(s, a) = C(s, a)$ for all $s \in \mathbb{N}$ and $a \in (0, s)$. Accordingly, Equations (5.2) and (5.4) coincide for those cases as well.

The performance measures described above satisfy various interesting structural properties. The literature dealing with these properties is rich (an excellent overview is provided in Whitt, 2002), but most research has focused on $\hat{C}$ and $\hat{W}_q$, thereby restricting the analysis to integer numbers of servers. In what follows, we will show that various well-known monotonicity, convexity, subadditivity, and subhomogeneity properties of $\hat{C}$ and $\hat{W}_q$ are also valid for C and W_q. Thus, we extend the analysis to non-integral numbers of servers by means of (5.3).

But given that all real-life queueing systems operate under an integral number of servers, why the fuss of this extended analysis? First there is the mathematical appeal of a generalization of known results; in fact, our analysis of the continuous extension (5.3) will provide simple alternative proofs of classic results in the $M/M/s$ model. But more importantly, the ensuing properties of the continuous extensions C and W_q will allow us to derive interesting results for Erlang delay games.

To obtain new structural results for the extensions C and W_q, we exploit a relation between the continuous extension of the Erlang delay function and the continuous extension B of the Erlang *loss* function, as defined in Equation (4.5), and we use a result that has already been established for the latter. The following lemma shows that the continuous extension of the Erlang delay function can be expressed in terms of the continuous extension of the Erlang

loss function, and vice versa. For integer s, this relation is well known (see, e.g., Cooper, 1981, p. 92). For non-integer s, this relation is retained by C and B, as observed by Janssen et al. (2011). We provide a proof for completeness.

Lemma 5.1. *Let $a > 0$ and $s \in \mathbb{R}$ with $s > a$. Then,*

$$C(s,a) = \frac{B(s,a)}{1-(a/s)(1-B(s,a))}.$$

Proof. For notational ease, let $C = C(s,a)$, $B = B(s,a)$, $C^{-1} = 1/C$, $B^{-1} = 1/B$, and $\rho = a/s$. (Note that $C > 0$ and $B > 0$.) Then,

$$\begin{aligned} C^{-1} - (1-\rho)B^{-1} &= \int_0^\infty ae^{-ax}(1+x)^{s-1}x\,dx - \int_0^\infty ae^{-ax}(1+x)^s(1-\rho)\,dx \\ &= \int_0^\infty ae^{-ax}(1+x)^{s-1}[x-(1+x)(1-\rho)]\,dx \\ &= \int_0^\infty ae^{-ax}(1+x)^{s-1}[\rho(1+x)-1]\,dx \\ &= -ae^{-ax}(1+x)^s/s\big|_{x=0}^{x=\infty} = \rho. \end{aligned}$$

This implies $BC^{-1} - 1 + \rho = B\rho$, which in turn implies $BC^{-1} - 1 = -\rho(1-B)$, and thus $C = B/[1-\rho(1-B)]$. This completes the proof. □

Next, we show that when the load per server is held constant, the probability of delay is decreased by adding servers. In other words, C is subhomogeneous of degree zero. When the domain is restricted to integer s, this property has already been proven by Calabrese (1992, Proposition 1).

Lemma 5.2. *Fix $a > 0$ and $s \in \mathbb{R}$ with $s > a$. Then, $C(ts,ta)$ is decreasing in t for $t > 0$.*

Proof. Note that, by assumption, $s > 0$. Hence, for $t > 0$, it holds that $B(ts,ta) > 0$ and, by Lemma 5.1,

$$C(ts,ta) = \frac{1}{\dfrac{1-a/s}{B(ts,ta)} + a/s}. \tag{5.5}$$

Now, as shown in Theorem 4.5, $B(ts,ta)$ is decreasing in t for $t > 0$, given that $s > 0$. This result, in combination with Equation (5.5), implies that $C(ts,ta)$ is also decreasing in t for $t > 0$. □

The following theorem states that when the load per server is held constant again, the expected waiting time is decreased by adding servers. Benjaafar (1995, p. 377) provides a proof of this result for integer s.

Theorem 5.3. *Fix $\lambda, \mu > 0$ and $s \in \mathbb{R}$ with $s > \lambda/\mu$. Then, $W_q(ts, t\lambda, \mu)$ is decreasing in t for $t > 0$.*

Proof. Let $t_1, t_2 > 0$ with $t_1 < t_2$. Then,

$$\begin{aligned} W_q(t_1 s, t_1\lambda, \mu) &= C(t_1 s, t_1\lambda/\mu)/[t_1(s\mu - \lambda)] \\ &> C(t_2 s, t_2\lambda/\mu)/[t_1(s\mu - \lambda)] \\ &> C(t_2 s, t_2\lambda/\mu)/[t_2(s\mu - \lambda)] = W_q(t_2 s, t_2\lambda, \mu), \end{aligned}$$

where the first inequality holds by Lemma 5.2. The second inequality is strict as $C(t_2 s, t_2\lambda/\mu) > 0$. We conclude that $W_q(ts, t\lambda, \mu)$ is decreasing in t for $t > 0$. □

In the process of proving the next property of the continuous extension of the Erlang delay function, we will use the following lemma. Although the result is straightforward, we were unable to find a proof in the literature, and therefore we provide a proof for completeness.

Lemma 5.4. *Let $f : D \to \mathbb{R}_{++}$ be a positive, non-increasing, convex function on an interval $D \subseteq \mathbb{R}$. Let $g : D \to \mathbb{R}_{++}$ be a positive, decreasing, strictly convex function on the same domain as f. Then the function $h : D \to \mathbb{R}_{++}$, defined by $h(s) = f(s)g(s)$ for all $s \in D$, is decreasing and strictly convex.*

Proof. The case $|D| \leq 1$ is trivial. In the remainder of the proof, we will assume $|D| > 1$. Let $s_1, s_2 \in D$ with $s_1 < s_2$. We first show that h is decreasing. We have

$$h(s_1) = f(s_1)g(s_1) \geq f(s_2)g(s_1) > f(s_2)g(s_2) = h(s_2),$$

where the first inequality holds because f is non-increasing while $g(s_1) > 0$, and the second inequality holds because $f(s_2) > 0$ and g is decreasing. We conclude that h is decreasing.

Next, we show that h is strictly convex. Let $x \in (0, 1)$. Then,

$$\begin{aligned} h(xs_1 + (1-x)s_2) &= f(xs_1 + (1-x)s_2) \cdot g(xs_1 + (1-x)s_2) \\ &\leq \big[xf(s_1) + (1-x)f(s_2)\big] \cdot g(xs_1 + (1-x)s_2) \\ &< \big[xf(s_1) + (1-x)f(s_2)\big] \cdot \big[xg(s_1) + (1-x)g(s_2)\big] \\ &= x^2 h(s_1) + (1-x)^2 h(s_2) + x(1-x)f(s_1)\big[g(s_2) - g(s_1) + g(s_1)\big] \\ &\quad + (1-x)xf(s_2)\big[g(s_1) - g(s_2) + g(s_2)\big] \\ &= xh(s_1) + (1-x)h(s_2) \\ &\quad + x(1-x)\big[f(s_1)[g(s_2) - g(s_1)] + f(s_2)[g(s_1) - g(s_2)]\big] \\ &\leq xh(s_1) + (1-x)h(s_2), \end{aligned}$$

The first inequality holds because f is convex, while g is a positive function. The second inequality holds because g is strictly convex and f is a positive function. The third inequality holds because $x(1-x) > 0$, $g(s_2) - g(s_1) < 0$, and $f(s_1) - f(s_2) \geq 0$. We conclude that h is indeed strictly convex. □

The following theorem says that the expected waiting time is decreasing and strictly convex in the number of servers. For integer s, these properties have already been proven by Dyer and Proll (1977).

Theorem 5.5. *Let $\lambda, \mu > 0$. Then, $W_q(s, \lambda, \mu)$ is a decreasing and strictly convex function of s for $s \in \mathbb{R}$ with $s > \lambda/\mu$.*

Proof. For the fixed choice of λ and μ, we denote $a = \lambda/\mu$. Define the set $D = \{s \in \mathbb{R} \mid s > a\}$. We next define three functions with domain D. First, the function $f : D \to \mathbb{R}_{++}$ with $f(s) = C(s, a)$ for all $s \in D$. Second, the function $g : D \to \mathbb{R}_{++}$ with $g(s) = 1/(s\mu - \lambda)$ for all $s \in D$. Third, $h : D \to \mathbb{R}_{++}$, defined by $h(s) = f(s)g(s) =$ for all $s \in D$. Note that $h(s) = W_q(s, \lambda, \mu)$ for all $s \in D$. Since, as shown by Jagers and van Doorn (1991), the continuous extension of the Erlang delay function is positive, non-increasing and convex in the number of servers, f has the same properties. Moreover, g is positive, decreasing, and strictly convex. Hence, by Lemma 5.4, h is decreasing and strictly convex. □

Next, we consider a subadditivity property that describes the economy-of-scale effect associated with larger service systems. Specifically, the following theorem states that combining two separate $M/M/s$ queues with common service rates into a joint system will lead to a reduction in the average (per-arrival) delay. Smith and Whitt (1981) provide a proof of this result for integer s, although not with strict inequality.

Theorem 5.6. *Let $\lambda_1, \lambda_2, \mu > 0$. Then, for all $s_1 \in \mathbb{R}$ with $s_1 > \lambda_1/\mu$ and for all $s_2 \in \mathbb{R}$ with $s_2 > \lambda_2/\mu$, it holds that*

$$W_q(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \cdot (\lambda_1 + \lambda_2) < W_q(s_1, \lambda_1, \mu) \cdot \lambda_1 + W_q(s_2, \lambda_2, \mu) \cdot \lambda_2.$$

Proof. Using the convexity property of Theorem 5.5 and subsequently using Theorem 5.3, we obtain

$$\begin{aligned} W_q(s_1 + s_2, \lambda_1 + \lambda_2, \mu) &\leq \frac{\lambda_1}{\lambda_1 + \lambda_2} W_q\left(\frac{\lambda_1 + \lambda_2}{\lambda_1} s_1, \lambda_1 + \lambda_2, \mu\right) \\ &\quad + \frac{\lambda_2}{\lambda_1 + \lambda_2} W_q\left(\frac{\lambda_1 + \lambda_2}{\lambda_2} s_2, \lambda_1 + \lambda_2, \mu\right) \\ &< \frac{\lambda_1}{\lambda_1 + \lambda_2} W_q(s_1, \lambda_1, \mu) + \frac{\lambda_2}{\lambda_1 + \lambda_2} W_q(s_2, \lambda_2, \mu), \end{aligned}$$

Multiplying both sides with $\lambda_1 + \lambda_2 > 0$ completes the proof. □

We conclude this section with a comment on linear interpolations. As shown in Chapter 4, the linear interpolation of the Erlang loss function is scalable. Surprisingly, as shown in the following example, the same does not hold for the linear interpolation of the Erlang delay function.

Example 5.2. Consider the linear interpolation C^{lin} of the Erlang delay function, defined for any $a > 0$ and $s \in \mathbb{N}$ with $\lfloor s \rfloor > a$ by

$$C^{lin}(s,a) = (1 - (s - \lfloor s \rfloor)) \cdot C(\lfloor s \rfloor, a) + (s - \lfloor s \rfloor) \cdot C(\lceil s \rceil, a).$$

Let $s = 1$, $a = 0.5$, and $t = 1.2$. We obtain that $C^{lin}(s,a) = C(1, 0.5) = 0.5$, whereas $C^{lin}(ts, ta) = 0.8 \cdot C(1, 0.6) + 0.2 \cdot C(2, 0.6) = 0.8 \cdot \frac{3}{5} + 0.2 \cdot \frac{9}{65} = \frac{33}{65}$. So, if we start with an Erlang delay system with 1 server and offered load 0.5, then scaling up by a factor of 1.2 results in an *increased* delay probability. This is in contrast to Erlang loss systems, where by Theorem 4.11 the same type of scaling leads to a *decreased* blocking probability: $L(s,a) = B(1, 0.5) = \frac{1}{3}$ and $L(ts, ta) = 0.8 \cdot B(1, 0.6) + 0.2 \cdot B(2, 0.6) = 0.8 \cdot \frac{3}{8} + 0.2 \cdot \frac{9}{89} = \frac{57}{178}$. $\diamond$

5.3 Model description

In this section, we introduce Erlang delay situations. Consider several service organizations, which we will simply refer to as players. Each player witnesses a Poisson arrival process of customers. The arrival processes of the players are mutually independent. Service times (for an arbitrary customer of any player) are exponential and i.i.d. Each player has an exogenously given number of servers to provide service to their customer streams; this number of servers may be either fixed or adjustable. Either way, servers are costly. An arriving customer who finds all servers busy upon arrival waits in a queue, incurring delay costs that are proportional to their waiting time. These delay costs, which are symmetrical across players, represent customer dissatisfaction, lost goodwill, and/or contractual penalties; they are borne by the player to whom the customer belongs.

Players are interested in their long-term average costs per unit time, which they may be able to reduce by collaborating, i.e., pooling their resources to serve their customer streams together. Our aim is to determine (the existence of) fair allocations of costs to support such collaboration. To analyze this, we have the following definition.

Definition 5.1. An *Erlang delay situation* is a tuple $(N, \lambda, \mu, S, h, w)$, where

- N is the nonempty, finite set of players;
- $\lambda \in \mathbb{R}^N_{++}$ is the vector of arrival rates, where λ_i describes the arrival rate of customers that belong to player $i \in N$;
- $\mu > 0$ is the rate of the exponential service time distribution;
- $S \in \mathbb{R}^N_{++}$ is the vector of numbers of servers with $S_i > \lambda_i/\mu$ for all $i \in N$, where S_i describes the number of servers that player $i \in N$ brings to any coalition;
- $h = (h_M)_{M \in 2^N_-} \in \mathbb{R}_+^{2^N_-}$ is the vector of resource cost rates that satisfies $h_M \geq h_L$ for all $M, L \in 2^N_-$ with $M \subseteq L$, where h_M describes the resource cost incurred per unit time for each server operated by coalition M;

- $w > 0$ is the delay or waiting cost incurred by any customer for waiting one unit of time in the queue.

For each coalition $M \in 2^N_-$, we denote $\lambda_M = \sum_{i \in M} \lambda_i$ and $S_M = \sum_{i \in M} S_i$.

In the formulation of a Erlang delay situation, we imposed several requirements. We discuss four of them.

First, an Erlang delay situation only has a natural interpretation when each player has an integer number of servers, but our formulation does allow a player to possess a non-integral number of servers. This simplifies proofs and allows a good fit between Sections 5.4 and 5.5.

Second, the requirement that $S_i > \lambda_i/\mu$ for all $i \in N$ means that each player possesses enough servers to ensure that the expected waiting time in his own service facility is finite. This assumption is not essential, but allows a clear exposition.

Third, the requirement that $h_M \geq h_L$ for all $M, L \in 2^N_-$ with $M \subseteq L$ means that the resource cost rate does not increase as a coalition grows. Our formulation captures not only the natural situation in which each coalition has the same resource cost rate but also more general settings wherein larger coalitions can acquire and maintain servers at a reduced cost rate because more players gives stronger negotiation or buying power.[37]

Fourth, we assumed that delay costs are symmetric, for two reasons. First, if delay costs would *not* be symmetric across players, then complete pooling of servers need not be superior anymore; we provide a counterexample in Subsection 5.5.2. This adverse effect is in keeping with the observations made in Section 4.4.2 for the Erlang loss model. In addition, Anily and Haviv (2010) analyzed their $M/M/1$ games under symmetric delay costs. To allow a more crisp comparison to the $M/M/1$ games, we assume symmetric delay costs in this chapter.

This model is sufficiently general to cover a wide variety of situations in which a resource pooling arrangement can arise between independent service providers that operate service facilities with infinite waiting room. Our model is simple, yet it has all ingredients to capture a concrete setting. The following example illustrates this and simultaneously highlights the difference with the Erlang loss model.

Example 5.3. Reconsider the departments of the manufacturer of advanced technical equipment from Example 4.2, who could pool service technicians to repair failed machines at their customers' sites. In Example 4.2, if no repairman was immediately available upon a failure, then the repair was taken over by a separate repair organization. Instead, we will now assume that there is no separate repair organization, and that the customer is entitled to a monetary compensation of 100 euro for every business hour he has to wait for a regular repairman.

For brevity, consider Departments 1 and 2 only. As before, both departments face repair requests at a rate of 0.1 per day while employing 1 repairman. The salary of a repairman

[37] This represents an additional factor that leads to synergy due to pooling. We could also allow a concave resource cost function, in line with Chapter 4. See the discussion in Section 4.3.2.

is unaffected by coalition formation and remains 100 euro per day. If the service time is exponentially distributed with mean 1 day, then the $M/M/s$ model applies. We can represent this situation as an Erlang delay situation $(N, \lambda, \mu, S, h, w)$ by letting $N = \{1, 2\}$, $S_1 = 1$, $S_2 = 1$, $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\mu = 1$, $h_M = 100$ for all coalitions $M \in 2^N_-$, and $w = 100 \cdot 8 = 800$, where we measure time in business hours and money in euros. $\diamond$

5.4 Fixed numbers of servers

In this section, we consider a setting in which each player brings a predetermined number of servers to any coalition. This model captures situations where adjusting the number of servers is practically or legally impossible. We first define the associated game. Subsequently, we analyze structural properties of this game, identify stable cost allocations, and study the existence of a PMAS.

5.4.1 Game

Consider any Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$ and an arbitrary coalition $M \in 2^N_-$. We assume that the players in this coalition collaborate by complete pooling of their respective arrival streams and servers into a joint system. Since the superposition of independent Poisson processes is also a Poisson process, this coalition now faces a combined Poisson arrival process with aggregate rate λ_M. The coalition has S_M servers at its disposal. We assume that each server can handle all types of customers with equal ease and that all customers can effortlessly access the joint service facility. A non-biased service discipline, such as service in order of arrival, is used.

Based on these assumptions, the pooled system behaves as an Erlang delay system. The expected waiting time that an arbitrary customer spends in the queue before starting service is equal to $W_q(s_M, \lambda_M, \mu)$. Multiplying this by $\lambda_M w$, we obtain the delay costs per unit time in steady state. We can now formulate a game corresponding to Erlang delay situation φ.

Definition 5.2. For any Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$, we call the game (N, c^φ) with

$$c^\varphi(M) = h_M S_M + W_q(S_M, \lambda_M, \mu) \cdot \lambda_M w. \tag{5.6}$$

for all $M \in 2^N_-$ the *associated Erlang delay game with fixed numbers of servers.*

By Theorem 5.6, cooperation in the context of pooling Erlang delay systems with symmetric service times and delay costs always leads to a reduction in costs. However, this does not yet imply the existence of a stable cost allocation or a PMAS. If Erlang delay games with fixed numbers of servers would be concave, existence of both would be guaranteed. The following example, however, shows that these games need not be concave. The example also shows, even stronger, that the Shapley value is not necessarily in the core.

Coalition M	S_M	λ_M	$C(S_M, \lambda_M/\mu)$	$W_q(S_M, \lambda_M, \mu)$	$c^\varphi(M)$
$\{1\}$	1	$\frac{1}{2}$	$\frac{1}{2}$	1	5
$\{2\}$	1	$\frac{3}{4}$	$\frac{3}{4}$	3	$22\frac{1}{2}$
$\{3\}$	3	$\frac{1}{4}$	$\frac{1}{452}$	$\frac{1}{1243}$	$\frac{5}{2486}$
$\{1,2\}$	2	$1\frac{1}{4}$	$\frac{25}{52}$	$\frac{25}{39}$	$8\frac{1}{78}$
$\{1,3\}$	4	$\frac{3}{4}$	$\frac{27}{3524}$	$\frac{27}{11453}$	$\frac{405}{22906}$
$\{2,3\}$	4	1	$\frac{1}{49}$	$\frac{1}{147}$	$\frac{10}{147}$
$\{1,2,3\}$	5	$1\frac{1}{2}$	$\frac{81}{4022}$	$\frac{81}{14077}$	$\frac{1215}{14077}$

Table 5.1: *The Erlang delay game and expected delay times of Example 5.4.*

Example 5.4. Consider the Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$ with player set $N = \{1, 2, 3\}$, service rate $\mu = 1$, resource cost rate $h_M = 0$ for all coalitions M, delay cost rate $w = 10$, and

$$\lambda_1 = 1/2; \quad \lambda_2 = 3/4; \quad \lambda_3 = 1/4;$$
$$S_1 = 1; \qquad S_2 = 1; \qquad S_3 = 3.$$

The characteristic cost function c^φ of the associated Erlang delay game with fixed numbers of servers (N, c^φ) is represented in Table 5.1, along with the expected waiting time of an arbitrary customer in any coalition's service system. This game has a nonempty core: for example, the allocation x given by $x_1 = 2$, $x_2 = 3$, and $x_3 = -4\frac{12862}{14077}$ is stable. However, this game is not concave since $c^\varphi(\{1,2\}) - c^\varphi(\{2\}) = -14\frac{19}{39} < \frac{5405}{295617} = c^\phi(\{1,2,3\}) - c^\phi(\{2,3\})$. In other words, player 1's marginal cost contribution may increase if he joins a larger coalition. Moreover, the game's Shapley value $\Phi(N, c^\varphi)$, which is approximately equal to $\Phi_1(N, c^\varphi) \approx -0.74$, $\Phi_2(N, c^\varphi) \approx 8.04$, and $\Phi_3(N, c^\varphi) \approx -7.21$ (rounded to 2 decimals), is not in the core of this game since $\Phi_2(N, c^\varphi) + \Phi_3(N, c^\varphi) > c^\varphi(\{2,3\})$. ◇

We remark that the characteristic cost function in the preceding example is not monotonically decreasing. In fact, expected waiting times may increase when a new player joins. For instance, when player 2 joins player 3, the expected delay experienced by a customer of player 3 increases from 1/1243 to 1/147. This is because player 3 possesses a relatively large number of servers and, as a result, observes few delays when acting independently.

5.4.2 Cost allocation: stability

In this section, we present general results on the existence and multiplicity of stable cost allocations for Erlang delay games with fixed numbers of servers. First, we show that any such game, as well as each of its sub-games, has a strictly stable allocation. The proof is based on the notion of strict balancedness, cf. Part (i) of Theorem 2.4.

Theorem 5.7. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation with $|N| > 1$. The associated game (N, c^φ) is strictly totally balanced.*

Proof. Let $\kappa \in \mathcal{W}^N$ be an arbitrary minimally balanced map. Then, we have

$$\begin{aligned}
c^\varphi(N) &= h_N S_N + W_q\left(S_N, \lambda_N, \mu\right) \cdot \lambda_N w \\
&= h_N S_N + W_q\left(\sum_{M \in \mathbb{B}(\kappa)} \kappa(M) S_M \cdot \frac{\lambda_N}{\lambda_M} \cdot \frac{\lambda_M}{\lambda_N}, \lambda_N, \mu\right) \cdot \lambda_N w \\
&\leq h_N S_N + \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \frac{\lambda_M}{\lambda_N} \cdot W_q\left(S_M \cdot \frac{\lambda_N}{\lambda_M}, \lambda_N, \mu\right) \cdot \lambda_N w \\
&= h_N S_N + \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) W_q\left(S_M \cdot \frac{\lambda_N}{\lambda_M}, \lambda_N, \mu\right) \cdot \lambda_M w \\
&< h_N S_N + \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) W_q\left(S_M, \lambda_M, \mu\right) \cdot \lambda_M w \\
&= \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \left[h_N S_M + W_q\left(S_M, \lambda_M, \mu\right) \cdot \lambda_M w\right] \\
&\leq \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \left[h_M S_M + W_q\left(S_M, \lambda_M, \mu\right) \cdot \lambda_M w\right] = \sum_{M \in \mathbb{B}(\kappa)} \kappa(M) c^\varphi(M).
\end{aligned}$$

The second and second-to-last equalities are valid because $\sum_{M \in \mathbb{B}(\kappa)} \kappa(M) S_M = S_N$. The first inequality holds by the combination of the convexity property of Theorem 5.5 and the fact that $\sum_{M \in \mathbb{B}(\kappa)} \kappa(M) \lambda_M / \lambda_N = 1$; that is, we employ this convexity by using that the nonnegative convex weights $\kappa(M)\lambda_M/\lambda_N$ add up to 1. The second inequality holds by Theorem 5.3. This inequality is strict because κ is minimally balanced and thus, by definition, every $M \in \mathbb{B}(\kappa)$ is a proper subset of N. The third and final equality holds because $h_N \leq h_M$ by assumption.

We conclude that the game (N, c^φ) is strictly balanced. Noting that every sub-game of (N, c^φ) is a game associated with an Erlang delay situation itself completes the proof. □

By Part (*ii*) of Theorem 2.4, we immediately obtain the following corollary to Theorem 5.7.

Corollary 5.8. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation with $|N| > 1$. The associated game (N, c^φ) has infinitely many (strict) core allocations.*

By Part (*vii*) of Theorem 2.4, the nucleolus always accomplishes an allocation in the strict core if nonempty. Therefore, it accomplishes strictly stable cost allocations for our Erlang delay games. However, computation of the nucleolus may be difficult (see, e.g., Leng and Parlar, 2010). In light of this downside and because there are infinitely many core allocations, one may well ask whether the core contains a simple cost allocation, e.g., proportional with respect to arrival rates or to numbers of servers. The following example shows that such (positive) proportional allocations will not necessarily be in the core, as there are instances in which a player is assigned a negative cost (i.e., a reward) in every core allocation.

Example 5.5. Consider the Erlang delay game with fixed numbers of servers (N, c^{φ}) of Example 5.4 again. For any allocation x in the core of (N, c^{φ}), it holds that $x_1 + x_3 \leq c^{\varphi}(\{1,3\})$, $x_2 + x_3 \leq c^{\varphi}(\{2,3\})$, and $x_1 + x_2 + x_3 = c^{\varphi}(\{1,2,3\})$. Hence,

$$x_3 = x_3 + x_1 + x_2 + x_3 - c^{\varphi}(N) \leq c^{\varphi}(\{1,3\}) + c^{\varphi}(\{2,3\}) - c^{\varphi}(N) = \frac{405}{22906} + \frac{10}{147} - \frac{1215}{14077} < 0.$$

Thus, player 3 is assigned a negative cost in every core allocation. The intuition behind this is that player 3 should be compensated for the relatively large number of servers that he adds to any coalition. ◇

In the corresponding $M/M/1$ queueing game (cf. Anily and Haviv, 2010), non-negative core allocations always existed; thus, in this respect, the multi-server models exhibit different behavior than their single-server counterparts. So, the simple proportional allocation that "worked" for $M/M/1$ games, as discussed in Section 3.3.1, does *not* work for Erlang delay games with fixed numbers of servers.

5.4.3 Cost allocation: population monotonicity

As $M/M/1$ games always admitted a PMAS, we next pose the question whether or not Erlang delay games with fixed numbers of servers possess a PMAS. The following example answers this question negatively: Erlang delay games with fixed numbers of servers and four or more players need not admit a PMAS.[38] This non-existence contrasts with results that we will obtain in Section 5 for Erlang delay games with optimized (real) numbers of servers. Absence of a PMAS may complicate coalition formation: it implies for any fixed allocation scheme that under some sequence of adding players one-by-one to a pooling group, there is at least one player who, at a certain point, becomes worse off when another player is added.

Example 5.6. Consider the Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$ with $N = \{1,2,3,4\}$, $\mu = 1$, $w = 10$, $h_M = 0$ for all coalitions M, and

$$\begin{aligned} &\lambda_1 = \lambda_2 = 0.1; \quad \lambda_3 = \lambda_4 = 0.9; \\ &S_1 = S_2 = 2; \qquad S_3 = S_4 = 1. \end{aligned}$$

The associated Erlang delay game with fixed numbers of servers (N, c^{φ}) is represented in Table 5.2. To show that this game does not admit a PMAS, we use the dual description of the class of games with a PMAS, introduced in Norde and Reijnierse (2002). They provide a set of necessary conditions (their Theorem 8) to determine whether a game has a PMAS or not. For our 4-player game, one of these conditions — which follows from $y_{2,\{1,2,3\}} \leq y_{2,\{2,3\}}$ and five similar monotonicity inequalities, combined with efficiency — is given by (cf. p. 331 of Norde and Reijnierse, 2002):

$$c^{\varphi}(\{1,2,3\}) + c^{\varphi}(\{2,3,4\}) \leq c^{\varphi}(\{1,3\}) + c^{\varphi}(\{2,3\}) + c^{\varphi}(\{2,4\}). \tag{5.7}$$

[38] Every balanced game with 3 or fewer players always admits a PMAS; see Sprumont (1990).

Coalition M	S_M	λ_M	$C(S_M, \lambda_M/\mu)$	$W_q(S_M, \lambda_M, \mu)$	$c^\varphi(M)$
$\{1\},\{2\}$	2	0.1	$\frac{1}{210}$	$\frac{1}{399}$	$\frac{1}{399}$
$\{3\},\{4\}$	1	0.9	$\frac{9}{10}$	9	81
$\{1,2\}$	4	0.2	$\frac{1}{17405}$	$\frac{1}{66139}$	$\frac{2}{66139}$
$\{3,4\}$	2	1.8	$\frac{81}{95}$	$4\frac{5}{19}$	$76\frac{14}{19}$
$\{1,3\},\{1,4\},\{2,3\},\{2,4\}$	3	1	$\frac{1}{11}$	$\frac{1}{22}$	$\frac{5}{11}$
$\{1,2,3\},\{1,2,4\}$	5	1.1	$\frac{161051}{28127210}$	$\frac{161051}{109696119}$	$\frac{1771561}{109696119}$
$\{1,3,4\},\{2,3,4\}$	4	1.9	$\frac{130321}{867190}$	$\frac{130321}{1821099}$	$1\frac{655000}{1821099}$
N	6	2	$\frac{2}{111}$	$\frac{1}{222}$	$\frac{10}{111}$

Table 5.2: *The Erlang delay game and expected delay times of Example 5.6.*

Inequality (5.7) may be interpreted as stating that an arrangement in which players 1, 2, and 3 work one unit of time together, incurring costs $c^\varphi(1,2,3)$ per time unit, and in which player 2, 3, and 4 work one unit of time together, incurring costs $c^\varphi(2,3,4)$ per time unit, generates lower costs than an alternative schedule in which players work the same amount of time as before, but in smaller coalitions. As in our game it holds that

$$c^\varphi(\{1,2,3\})+c^\varphi(\{2,3,4\}) = \frac{1771561}{109696119}+1\frac{655000}{1821099} > \frac{5}{11}\cdot 3 = c^\varphi(\{1,3\})+c^\varphi(\{2,3\})+c^\varphi(\{2,4\}),$$

Inequality (5.7) is not satisfied. We conclude that this game lacks a PMAS.

To make this nonexistence intuitively plausible, notice that there are two types of players: on the one hand there are players 1 and 2, both with low arrival rates and many servers, and on the other hand there are players 3 and 4, both with high arrival rates and few servers. Any two-player coalition containing one player of each type can already attain most of the benefits of pooling. Combining this fact for three of those two-player coalitions leads to an incompatibility with the costs that should be paid in two related three-player coalitions. $\diamond$

Conditions for existence of a PMAS in our games follow from the characterization result of Norde and Reijnierse (2002). However, these conditions, when applied to our game, are technical and do not convey direct insight for our setting. The same problem is encountered when trying to provide general conditions under which all core allocations are non-negative. Instead, we will provide a more intuitive sufficient condition for our games to admit a PMAS and a positive core allocation. To this end, we introduce an allocation scheme under which the expected waiting cost of any coalition is allocated proportional to the arrival rates of its members. Under an assumption on the ratios between players' servers and arrival rates, this allocation scheme will turn out to be population monotonic.

Allocation scheme $\mathcal{P}^{FIX}$ for Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$ is defined, for all $M \in 2^N_-$ and all $i \in M$, by

$$\mathcal{P}^{FIX}_{i,M}(\varphi) = h_M S_i + W_q(S_M, \lambda_M, \mu) \cdot \lambda_i w. \tag{5.8}$$

Now, suppose that the ratio of the number of servers to arrival rates is identical among players. This symmetry condition implies that players with larger arrival rates possess more servers. This symmetry is also in place when players represent equally sized service providers, all with the same number of servers and arrival rates. The following theorem states that, under this symmetry condition, the average delay experienced by an arbitrary customer decreases as a coalition grows larger and, as a result, the amount a player has to pay under $\mathcal{P}(\varphi)$ decreases when the coalition to which he belongs grows.

Theorem 5.9. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be a Erlang delay situation with $S_i/\lambda_i = S_j/\lambda_j$ for all $i, j \in N$.*

(i) For any two coalitions $M, L \in 2^N_-$ with $M \subset L$, $W_q(S_M, \lambda_M, \mu) > W_q(S_L, \lambda_L, \mu)$.

(ii) The proportional scheme $\mathcal{P}^{FIX}(\varphi)$ is an SPMAS for the associated game (N, c^φ).

Proof. Part (i). Let $M, L \in 2^N_-$ with $M \subset L$. Then,

$$W_q(S_M, \lambda_M, \mu) > W_q\left(S_M \cdot \frac{S_L}{S_M}, \lambda_M \cdot \frac{S_L}{S_M}, \mu\right) = W_q(S_L, \lambda_L, \mu),$$

where the inequality follows by Theorem 5.3 and the equality holds because $S_M/\lambda_M = S_L/\lambda_L$.

Part (ii). Let $M, L \in 2^N_-$ with $M \subset L$, and let $i \in M$. Then,

$$\begin{aligned}\mathcal{P}^{FIX}_{i,M}(\varphi) &= h_M S_i + W_q(S_M, \lambda_M, \mu) \cdot \lambda_i w \\ &\geq h_L S_i + W_q(S_M, \lambda_M, \mu) \cdot \lambda_i w \\ &> h_L S_i + W_q(S_L, \lambda_L, \mu) \cdot \lambda_i w = \mathcal{P}^{FIX}_{i,L}(\varphi),\end{aligned}$$

where the first inequality follows by assumption on h and the second inequality follows by Part (i). We conclude that $\mathcal{P}^{FIX}(\varphi)$ is indeed an SPMAS.

□

Recall that, by Part (iii) of Theorem 2.4, the fact that $\mathcal{P}^{FIX}(\varphi)$ is an SPMAS immediately implies the properties stated in following corollary.

Corollary 5.10. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be a Erlang delay situation with $S_i/\lambda_i = S_j/\lambda_j$ for all $i, j \in N$. Then, the associated game (N, c^φ) is strictly totally balanced and the allocation assigning $\mathcal{P}_{i,N}(\varphi)$ to all $i \in N$ is an element of its strict core.*

5.5 Optimized numbers of servers

In this section, we consider a setting in which the number of servers can be jointly optimized by each coalition. This model captures situations where the number of servers can be easily adjusted against negligible costs.[39] By allowing re-optimization of the number of servers, the

[39] If there are costs involved (e.g., a cost to shed an existing server or to purchase a new server) then these will likely be negligible in comparison to resource and delay costs in the long run.

grand coalition will be no worse off than in the case with fixed numbers of servers. After all, due to the resource pooling effect, fewer servers may suffice to jointly serve all customer streams in a cost-effective way. This reduction in the costs of the grand coalition would make it easier to find a stable allocation. Yet, sub-coalitions will also choose a cost-minimizing number of servers, reducing their costs and possibly shrinking the core. Due to these two opposite effects, which are explained in more detail in the context of Erlang loss games in Section 4.5.2, there is no direct relation between balancedness of Erlang delay games with fixed and optimized numbers of servers.

In this section, we will investigate whether the balancedness of Erlang delay games with fixed numbers of servers remain valid when optimization is introduced. To analyze this, we will define and analyze two corresponding games, which differ in the domain on which this optimization takes place. Throughout this entire section, we will assume for every Erlang delay situation $(N, \lambda, \mu, S, h, w)$ that the resource cost rates are positive, i.e., $h_M > 0$ for all $M \in 2^N_-$. This assumption will ensure the existence of optimal numbers of servers.

5.5.1 Games

Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation, and consider a coalition $M \in 2^N_-$. As before, this coalition will jointly serve customers that arrive according to a Poisson process with combined rate $\lambda_M = \sum_{i \in M} \lambda_i$. Suppose that this coalition would pick $s > \lambda_M/\mu$ common servers. Then, this coalition's joint service facility behaves as an Erlang delay system, and the expected costs per unit of time in steady state incurred by coalition M are equal to

$$K^{\varphi}_M(s) = h_M s + W_q(s, \lambda_M, \mu) \cdot \lambda_M w. \tag{5.9}$$

Figure 5.1 illustrates this cost function. Next, we formulate two different games corresponding to Erlang delay situation φ. In the first game, any coalition M optimizes $K^{\varphi}_M(s)$ over integer numbers of servers, i.e., over domain $\mathscr{N}^{\varphi}_M = \{s \in \mathbb{N} \mid s > \lambda_M/\mu\}$. In the second game, the optimization is taken over real numbers of servers, i.e., over domain $\mathscr{R}^{\varphi}_M = \{s \in \mathbb{R} \mid s > \lambda_M/\mu\}$.

We emphasize that our main interest lies in the first game because it represents the exact discrete optimization problem. The second game will help in understanding the discrete game: structural properties of the second game will imply structural properties of the first game.

Although a system with a non-integral number of servers as described by the continuous extension W_q does not lend itself to a natural interpretation, we remark that one might view it as, e.g., an approximation of a system with a part-time worker. We also point out that Borst et al. (2004), in dealing with the staffing problem of large call centers, have approximated costs based on the continuous extension (5.3) to find an approximately optimal number of servers.

We first define the game with optimization over the integers and subsequently define the game with optimization over the reals.

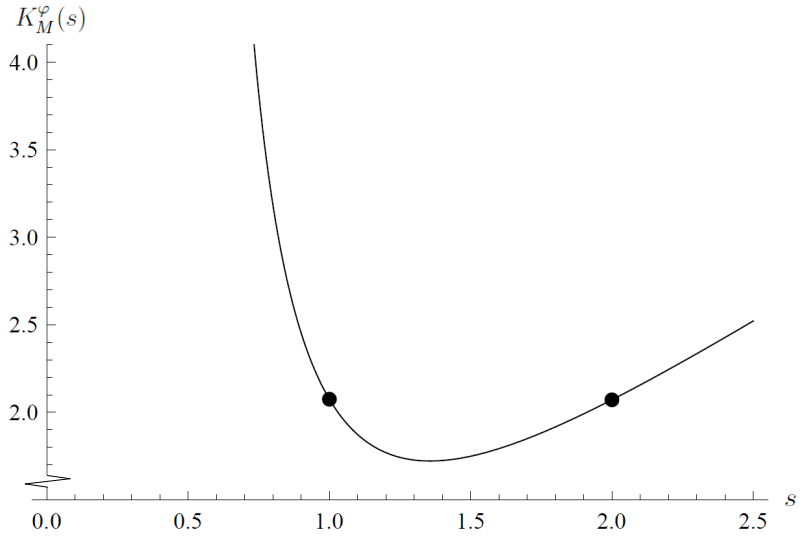

Figure 5.1: *The costs $K_M^\varphi(s)$ as a function of the number of servers s, for $\lambda_M = 0.5$, $\mu = 1$, $h_M = 1$, and $w = 2.15$. This function is defined on domain $\{s \in \mathbb{R} \mid s > \lambda_M/\mu\}$. The values $K_M^\varphi(1) = 2\frac{3}{40}$ and $K_M^\varphi(2) = 2\frac{43}{600}$ are identified by black dots.*

Definition 5.3. For any Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$, we call the game $(N, \hat{d}^\varphi)$ with

$$\hat{d}^\varphi(M) = \min_{s \in \mathscr{N}_M^\varphi} K_M^\varphi(s) \tag{5.10}$$

for all $M \in 2_-^N$ the associated *Erlang delay game with $\mathbb{N}$-optimization.*

Consider any coalition $M \in 2_-^N$. When restricted to domain $\mathscr{N}_M^\varphi$, the cost function K_M^φ is strictly convex (due to Theorem 5.5) and it achieves a minimum (since the costs grow unboundedly as the number of servers tends to infinity). Hence, an optimal integer number of servers is given by the smallest $s \in \mathscr{N}_M^\varphi$ satisfying $K_M^\varphi(s+1) \geq K_M^\varphi(s)$, and we denote this optimizer by $\hat{s}_M^*$.

Definition 5.4. For any Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$, we call the game (N, d^φ) with

$$d^\varphi(M) = \min_{s \in \mathscr{R}_M^\varphi} K_M^\varphi(s) \tag{5.11}$$

for all $M \in 2_-^N$ the associated *Erlang delay game with $\mathbb{R}$-optimization.*

Consider any coalition $M \in 2_-^N$. By strict convexity of K_M^φ and because $\lim_{s\to\infty} K_M^\varphi(s) = \lim_{s\downarrow\lambda_M/\mu} K_M^\varphi(s) = \infty$, the cost function $K_M^\varphi(s)$ achieves a minimum on domain $\mathscr{R}_M^\varphi$, implying that the game is well-defined. Now, the optimal real number of servers is unique (due to the strict convexity), and we denote it by s_M^*.

The following theorem states that cooperation is beneficial in both games and establishes a link between the two games.

Theorem 5.11. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation.*

(i) Both associated games $(N, \hat{d}^\varphi)$ and (N, d^φ) are strictly subadditive.

(ii) Let $M \in 2^N_-$ be a coalition. Then, either $\hat{s}^*_M = \lfloor s^*_M \rfloor$ or $\hat{s}^*_M = \lceil s^*_M \rceil$. Furthermore, $\hat{c}^\varphi(M) \geq c^\varphi(M)$, with equality if and only if $s^*_M \in \mathbb{N}$.

Proof. Part (i). Let $M, L \in 2^N_-$ with $M \cap L = \emptyset$. To show strict subadditivity Erlang delay game with $\mathbb{R}$-optimization (N, d^φ), we have

$$\begin{aligned} d^\varphi(M \cup L) &= h_{M\cup L} s^*_{M\cup L} + W_q(s^*_{M\cup L}, \lambda_M + \lambda_L, \mu) \cdot (\lambda_M + \lambda_L) w \\ &\leq h_{M\cup L}(s^*_M + s^*_L) + W_q(s^*_M + s^*_L, \lambda_M + \lambda_L, \mu) \cdot (\lambda_M + \lambda_L) w \\ &\leq h_M s^*_M + h_L s^*_L + W_q(s^*_M + s^*_L, \lambda_M + \lambda_L, \mu) \cdot (\lambda_M + \lambda_L) w \\ &< h_M s^*_M + W_q(s^*_M, \lambda_M, \mu) \cdot \lambda_M d + h_L s^*_L + W_q(s^*_L, \lambda_L, \mu) \cdot \lambda_L w \\ &= d^\varphi(M) + d^\varphi(L), \end{aligned}$$

where the first inequality holds because $s^*_{M\cup L}$ is an optimal number of servers for coalition $M \cup L$, the second inequality is valid because $h_{M\cup L} \leq h_M$ and $h_{M\cup L} \leq h_L$ by assumption on φ, and the third inequality holds by Theorem 5.6. We conclude that (N, d^φ) is strictly subadditive. The proof of strict subadditivity of the Erlang delay game with $\mathbb{N}$-optimization $(N, \hat{d}^\varphi)$ goes analogously.

Part (ii). First, the claim that $\hat{s}^*_M = \lfloor s^*_M \rfloor$ or $\hat{s}^*_M = \lceil s^*_M \rceil$ follows immediately from strict convexity of $K^\varphi_M(s)$. The relation between $\hat{d}^\varphi(M)$ and $d^\varphi(M)$ follows from uniqueness of s^*_M and from the observation that in the $OPT^{\mathbb{R}}$ game any coalition $M \in 2^N_-$ optimizes over a larger domain than in the $OPT^{\mathbb{N}}$ game, which implies that $\hat{d}^\varphi(M) = d^\varphi(M)$ if and only if $s^*_M = \hat{s}^*_M$, but that occurs if and only if $s^*_M \in \mathbb{N}$. □

5.5.2 Cost allocation: stability and population monotonicity

In this section, we investigate whether or not cost allocation can be carried out in a stable and population monotonic way. We start by introducing two simple rules that allocate costs proportional to arrival rates. The first rule, $\hat{\mathscr{P}}$, divides the costs of the grand coalition in Erlang delay games with $\mathbb{N}$-optimization. The second rule, $\mathscr{P}$, does this for Erlang delay games with $\mathbb{R}$-optimization. Formally, they are defined, for any Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$ and $i \in N$, by $\hat{\mathscr{P}}_i(\varphi) = \hat{d}^\varphi(N)\lambda_i/\lambda_N$ and by $\mathscr{P}_i(\varphi) = d^\varphi(N)\lambda_i/\lambda_N$, respectively.

Extending this idea to every coalition, we define the proportional allocation scheme rules $\hat{\mathcal{P}}$ and $\mathcal{P}$, for any Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$, coalition $M \in 2^N_-$, and player $i \in M$, by $\hat{\mathcal{P}}_{i,M}(\varphi) = \hat{d}^\varphi(M)\lambda_i/\lambda_M$ and by $\mathcal{P}_{i,M}(\varphi) = d^\varphi(M)\lambda_i/\lambda_M$, respectively. The following example illustrates the rules and schemes for Erlang delay games with $\mathbb{R}$-optimization and

Coalition M	s^*_M	$d^\varphi(M)$	$\mathcal{P}_{1,M}(\varphi)$	$\mathcal{P}_{2,M}(\varphi)$	$\mathcal{P}_{3,M}(\varphi)$
$\{1\}$	0.75878	1.01675	1.01675	*	*
$\{2\}$	0.75878	1.01675	*	1.01675	*
$\{3\}$	0.50511	0.70489	*	*	0.70489
$\{1,2\}$	1.17171	1.50706	0.75353	0.75353	*
$\{1,3\}$	0.97478	1.27524	0.85016	*	0.42508
$\{2,3\}$	0.97478	1.27524	*	0.85016	0.42508
N	1.35662	1.72219	0.68887	0.68887	0.34444

Table 5.3: *The Erlang delay game with $\mathbb{R}$-optimization, optimal numbers of servers, and proportional allocation scheme of Example 5.7. (All values are rounded to 5 decimals.)*

Coalition M	$\hat{s}^*_M$	$\hat{d}^\varphi(M)$	$\hat{\mathcal{P}}_{1,M}(\varphi)$	$\hat{\mathcal{P}}_{2,M}(\varphi)$	$\hat{\mathcal{P}}_{3,M}(\varphi)$
$\{1\}$	1	1.10750	1.10750	*	*
$\{2\}$	1	1.10750	*	1.10750	*
$\{3\}$	1	1.02389	*	*	1.02389
$\{1,2\}$	1	1.57333	0.78667	0.78667	*
$\{1,3\}$	1	1.27643	0.85095	*	0.42548
$\{2,3\}$	1	1.27643	*	0.85095	0.42548
N	2	2.07167	0.82867	0.82867	0.41433

Table 5.4: *The Erlang delay game with $\mathbb{N}$-optimization, optimal numbers of servers, and proportional allocation scheme of Example 5.7. (All values are rounded to 5 decimals.)*

simultaneously shows that Erlang delay games with $\mathbb{N}$-optimization need not admit a stable cost allocation.

Example 5.7. Consider the Erlang delay situation $\varphi = (N, \lambda, \mu, S, h, w)$ with player set $N = \{1,2,3\}$, arrival rates $\lambda_1 = \lambda_2 = 0.2$ and $\lambda_3 = 0.1$, service rate $\mu = 1$, resource cost rate $h_M = 1$ for all $M \in 2^N_-$, and delay cost rate $w = 2.15$. The cost function for the grand coalition corresponds to the one displayed in Figure 5.1 on page 115. The characteristic cost functions of the associated games (N, d^φ) and $(N, \hat{d}^\varphi)$ are represented in Tables 5.3 and 5.4, respectively, along with the optimal numbers of servers for each coalition and the proportional allocation schemes.

For the Erlang delay game with $\mathbb{R}$-optimization, notice that $\mathcal{P}_{1,\{1,2\}}(\varphi) > \mathcal{P}_{1,N}(\varphi)$ and similarly $\mathcal{P}_{2,\{1,2\}}(\varphi) > \mathcal{P}_{2,N}(\varphi)$, i.e., the amount that player 1 or 2 has to pay does not increase when player 3 joins them. This population monotonicity can be verified for the members of all other nested pairs of coalitions as well, implying that $\mathcal{P}(\varphi)$ is population monotonic. Accordingly, the game (N, c^φ) has a nonempty core containing $\mathscr{P}(\varphi)$.

In contrast, the Erlang delay game with $\mathbb{N}$-optimization $(N, \hat{d}^\varphi)$ has an empty core! Indeed, $2\hat{d}^\varphi(N) > 4.1433 > 4.1263 > \hat{d}^\varphi(\{1,2\}) + \hat{d}^\varphi(\{1,3\}) + \hat{d}^\varphi(\{2,3\})$, and thus a necessary balancedness condition is not satisfied. This means that no stable allocation exists. As an illustration, notice that $\hat{\mathcal{P}}(\varphi)$ is not a PMAS because for $i \in \{1,2\}$ it holds that $\hat{\mathcal{P}}_{i,\{1,2\}}(\varphi) < \hat{\mathcal{P}}_{i,N}(\varphi)$. Accordingly, $\hat{\mathscr{P}}(\varphi)$ is not a stable allocation because $\sum_{i\in\{1,2\}} \hat{\mathscr{P}}_i(\varphi) > \hat{d}^\varphi(\{1,2\})$. ◇

It is worth pointing out that Yu et al. (2007) gave a (2-player) counterexample indicating that their games corresponding to server optimization in Erlang delay systems may have an empty core. However, their counterexample included players with asymmetrical delay costs, and as a result their game was not subadditive, i.e., complete pooling was detrimental. In contrast, in our subadditive game $(N, \hat{d}^\varphi)$, all customers are homogeneous in delay costs and full pooling is beneficial. Despite the subadditivity, a stable allocation is lacking.

In Example 5.7, we observed that the proportional allocation scheme rule $\mathcal{P}$ accomplished a population monotonic allocation scheme for the Erlang delay game with $\mathbb{R}$-optimization. The following theorem shows that this is not a coincidence.

Theorem 5.12. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation. Then, $\mathcal{P}(\varphi)$ is a PMAS for the associated Erlang delay game with $\mathbb{R}$-optimization (N, d^φ).*

Proof. Let $M, L \in 2^N_-$ with $M \subseteq L$, and let $i \in M$. Then,

$$\begin{aligned}
\mathcal{P}_{i,L}(\varphi) &= \frac{\lambda_i}{\lambda_L} K^\varphi_L(s^*_L) \\
&\leq \frac{\lambda_i}{\lambda_L} K^\varphi_L\left(s^*_M \frac{\lambda_L}{\lambda_M}\right) \\
&= h_L s^*_M \frac{\lambda_i}{\lambda_M} + W_q\left(s^*_M \frac{\lambda_L}{\lambda_M}, \lambda_L, \mu\right) \cdot \lambda_i w \\
&\leq h_L s^*_M \frac{\lambda_i}{\lambda_M} + W_q(s^*_M, \lambda_M, \mu) \cdot \lambda_i w \\
&\leq h_M s^*_M \frac{\lambda_i}{\lambda_M} + W_q(s^*_M, \lambda_M, \mu) \cdot \lambda_i w \\
&= \frac{\lambda_i}{\lambda_M} K^\varphi_M(s^*_M) = \mathcal{P}_{i,M}(\varphi),
\end{aligned}$$

where the first inequality holds because s^*_L is the optimal number of servers for coalition L, the second inequality is valid by Theorem 5.3, and the third inequality holds because $h_L \leq h_M$ by assumption on φ. We conclude that $\mathcal{P}(\varphi)$ is indeed a PMAS for game (N, c^φ). □

By Sprumont (1990), population monotonicity of $\mathcal{P}(\varphi)$ not only signifies the advantageousness of any pooling group's growth but also, as described in Section 2.3.7, the following corollary.

Corollary 5.13. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation. Then, the associated Erlang delay game with $\mathbb{R}$-optimization (N, d^φ) is totally balanced and $\mathscr{P}(\varphi) \in \mathcal{C}(N, d^\varphi)$.*

The following theorem states sufficient conditions for Erlang delay game with $\mathbb{N}$-optimization to possess a core allocation and to admit a PMAS.

Theorem 5.14. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation.*

(i) If $s_N^* \in \mathbb{N}$, then the game $(N, \hat{d}^\varphi)$ has a non-empty core that contains $\hat{\mathscr{P}}(\varphi)$.

(ii) If $s_M^* \in \mathbb{N}$ for all $M \in 2_-^N$, then $\hat{\mathcal{P}}(\varphi)$ is a PMAS for game $(N, \hat{d}^\varphi)$.

Proof. Part (i). Assuming $s_N^* \in \mathbb{N}$, it holds that $\hat{d}^\varphi(N) = d^\varphi(N)$ and $\hat{d}^\varphi(M) \geq d^\varphi(M)$ for all $M \in 2_-^N$, by Part (ii) of Theorem 5.11. Using this in combination with Theorem 5.12, we obtain $\hat{\mathscr{P}}(\varphi) = \mathscr{P}(\varphi) \in \mathcal{C}(N, d^\varphi) \subseteq \mathcal{C}(N, \hat{d}^\varphi)$. Thus, $\hat{\mathscr{P}}(\varphi)$ is a stable allocation for game $(N, \hat{d}^\varphi)$.

Part (ii). Let $M, L \in 2_-^N$ with $M \subseteq L$, and let $i \in M$. Note that s_M^* and s_L^* are, by assumption, integers. By Part (ii) of Theorem 5.11, d and $\hat{d}$ coincide. Accordingly, $\mathcal{P}$ and $\hat{\mathcal{P}}$ coincide. So, by Theorem 5.12, we can conclude that $\hat{\mathcal{P}}(\varphi)$ is indeed a PMAS for game $(N, \hat{d}^\varphi)$. □

Several insights emerge from our analysis thus far. First, Erlang delay game with $\mathbb{R}$-optimization show nice properties: they have nonempty cores and admit a population monotonic allocation scheme. These properties are not satisfied by Erlang delay game with $\mathbb{N}$-optimization in general; the integrality requirement is the sole culprit, as clearly illustrated by the preceding theorem. Interestingly, the $M/M/1$ games where coalitions optimize service capacity to reduce their customers' system sojourn times (reviewed in Section 3.3.2) as well as the Erlang loss games with optimized numbers of servers (analyzed in Chapter 4) all had non-empty cores. As such, Erlang delay game with $\mathbb{N}$-optimization exhibit fundamentally different behavior.

5.5.3 Approximate stability and population monotonicity

To expand our understanding of potential instability in Erlang delay game with $\mathbb{N}$-optimization, we will introduce approximate core and PMAS concepts, and we will use these concepts to derive insights regarding the impact of integrality. The first general concept for cooperative games that we introduce in this section can be seen as a generalization of the core. For any vector $\delta = (\delta_i)_{i \in N} \in \mathbb{R}^N$, we define the *$\delta$-core* of game (N, c) as

$$\mathcal{C}_\delta(N, c) = \left\{ x \in \mathbb{R}^N \,\middle|\, \sum_{i \in N} x_i = c(N) \text{ and } \sum_{i \in M} x_i \leq c(M) + \sum_{i \in M} \delta_i \text{ for all } M \in 2_-^N \right\}.$$

This δ-core is the set of all cost allocations where no coalition M can obtain lower costs by leaving the grand coalition, if upon leaving it must pay a penalty of δ_i for member i. Naturally, the core of a game coincides with its $\mathbf{0}$-core. For any game (N, c), an allocation in $\mathcal{C}_\delta(N, c)$ for some vector δ is called a *δ-stable* allocation.

Our *δ-core* is reminiscent of the weak ϵ-core (with $\epsilon \in \mathbb{R}$) introduced by Shapley and Shubik (1966), but differs from it by associating a number with each player rather than a single number with all players. This change allows us to describe a stronger upper bound on the impact of integrality than is possible with Shapley and Shubik's concept. In fact, our δ-core concept encompasses their weak ϵ-core as a special case by setting $\delta_i = \epsilon$ for all $i \in N$.

We next introduce another new concept analogous to the (vector) δ-core. For any vector $\delta = (\delta_i)_{i\in N} \in \mathbb{R}^N$, we say that an allocation scheme y for a game (N, c) is a *δ-PMAS* if $y_{i,M} + \delta_i \geq y_{i,L}$ for all coalitions $M, L \in 2^N_-$ with $M \subset L$ and $i \in M$. Here, δ_i may be interpreted as an exogenous bonus received by player i if the coalition to which he belongs grows.

The following theorem uses these newly defined notions to capture the influence of the integrality requirement.

Theorem 5.15. *Let $\varphi = (N, \lambda, \mu, S, h, w)$ be an Erlang delay situation.*

(i) Fix δ by $\delta_i = h_N\lambda_i/\lambda_N$ for all $i \in N$. Then, the game $(N, \hat{d}^\varphi)$ has a non-empty (vector) δ-core, and $\hat{\mathscr{P}}(\varphi)$ is a δ-stable allocation.

(ii) Fix δ by $\delta_i = h_{\{i\}}$ for all $i \in N$. Then, $\hat{\mathcal{P}}(\varphi)$ is a δ-PMAS for the game $(N, \hat{d}^\varphi)$.

Proof. Part (i). Let $M \in 2^N_-$. Clearly, $\sum_{i\in N} \hat{\mathscr{P}}_i(\varphi) = \hat{d}^\varphi(N)$. Moreover,

$$\begin{aligned}
\sum_{i\in M} \hat{\mathscr{P}}_i(\varphi) &= K_N^\varphi(\hat{s}_N^*) \cdot \frac{\lambda_M}{\lambda_N} \\
&\leq K_N^\varphi\Big(\Big\lceil \hat{s}_M^* \frac{\lambda_N}{\lambda_M} \Big\rceil\Big) \cdot \frac{\lambda_M}{\lambda_N} \\
&= \left(h_N \Big\lceil \hat{s}_M^* \frac{\lambda_N}{\lambda_M} \Big\rceil + W_q\Big(\Big\lceil \hat{s}_M^* \frac{\lambda_N}{\lambda_M} \Big\rceil, \lambda_N, \mu\Big) \cdot \lambda_N w\right) \cdot \frac{\lambda_M}{\lambda_N} \\
&\leq \left(h_N \Big\lceil \hat{s}_M^* \frac{\lambda_N}{\lambda_M} \Big\rceil + W_q\Big(\hat{s}_M^* \frac{\lambda_N}{\lambda_M}, \lambda_N, \mu\Big) \cdot \lambda_N w\right) \cdot \frac{\lambda_M}{\lambda_N} \\
&\leq \left(h_N \Big\lceil \hat{s}_M^* \frac{\lambda_N}{\lambda_M} \Big\rceil + W_q\big(\hat{s}_M^*, \lambda_M, \mu\big) \cdot \lambda_N w\right) \cdot \frac{\lambda_M}{\lambda_N} \\
&\leq \left(h_N \Big(\hat{s}_M^* \frac{\lambda_N}{\lambda_M} + 1\Big) + W_q\big(\hat{s}_M^*, \lambda_M, \mu\big) \cdot \lambda_N w\right) \cdot \frac{\lambda_M}{\lambda_N} \\
&= h_N \hat{s}_M^* + \sum_{i\in M} \delta_i + W_q\big(\hat{s}_M^*, \lambda_M, \mu\big) \cdot \lambda_M w \\
&\leq h_M \hat{s}_M^* + \sum_{i\in M} \delta_i + W_q\big(\hat{s}_M^*, \lambda_M, \mu\big) \cdot \lambda_M w = \hat{d}^\varphi(M) + \sum_{i\in M} \delta_i.
\end{aligned}$$

The first inequality holds because $\hat{s}_N^*$ is an optimal number of servers for the grand coalition. The second inequality holds by the monotonicity property of Theorem 5.5. The third inequality is due to Theorem 5.3. The final inequality holds since $h_N \leq h_M$ by assumption on φ. We conclude that $\hat{\mathscr{P}}(\varphi)$ is a δ-stable allocation for game $(N, \hat{d}^\varphi)$.

Part (ii). Let $M, L \in 2^N_-$ with $M \subset L$, and let $i \in M$. Clearly, $\sum_{i\in M} \hat{\mathscr{P}}_{i,M}(\varphi) = \hat{d}^{\varphi}(M)$. Moreover, in line with the proof of Part (i) of this theorem,

$$\begin{aligned}
\hat{\mathcal{P}}_{i,L}(\varphi) &= K^{\varphi}_L(\hat{s}^*_L) \cdot \frac{\lambda_i}{\lambda_L} \\
&\leq K^{\varphi}_L\left(\left\lceil \hat{s}^*_M \frac{\lambda_L}{\lambda_M} \right\rceil\right) \cdot \frac{\lambda_i}{\lambda_L} \\
&= \left(h_L \left\lceil \hat{s}^*_M \frac{\lambda_L}{\lambda_M} \right\rceil + W_q\left(\left\lceil \hat{s}^*_M \frac{\lambda_L}{\lambda_M} \right\rceil, \lambda_M, \mu\right) \cdot \lambda_L d\right) \cdot \frac{\lambda_i}{\lambda_L} \\
&\leq \left(h_L \left\lceil \hat{s}^*_M \frac{\lambda_L}{\lambda_M} \right\rceil + W_q(\hat{s}^*_M, \lambda_M, \mu) \cdot \lambda_L d\right) \cdot \frac{\lambda_i}{\lambda_L} \\
&\leq \left(h_L \left(\hat{s}^*_M \frac{\lambda_L}{\lambda_M} + 1\right) + W_q(\hat{s}^*_M, \lambda_M, \mu) \cdot \lambda_L d\right) \cdot \frac{\lambda_i}{\lambda_L} \\
&\leq h_M \hat{s}^*_M \frac{\lambda_i}{\lambda_M} + h_{\{i\}} + W_q(\hat{s}^*_M, \lambda_M, \mu) \cdot \lambda_i d = \hat{\mathcal{P}}_{i,M}(\varphi) + \delta_i,
\end{aligned}$$

where the last inequality holds since $h_L \leq h_M \leq h_{\{i\}}$ by assumption on φ and since $\lambda_i/\lambda_L \leq 1$. We conclude that $\hat{\mathcal{P}}(\varphi)$ is a δ-PMAS for game $(N, \hat{d}^{\varphi})$. □

Part (i) of Theorem 5.15 constructively shows that any possible instability disappears if coalitions would have to pay an amount, no greater than h_N, to leave the grand coalition. Part (ii) of this theorem states a similar conclusion regarding the effect of discrete service capacity on population monotonicity.[40]

Altogether, our analysis suggests that Erlang delay game with $\mathbb{N}$-optimization associated with large service facilities have nonempty δ-cores and a δ-PMAS for relatively small δ, as the resource cost incurred for a single server is small relative to the total costs faced by any coalition if the optimal number of servers is large. This is in line with the observation of Yu et al. (2007) that a variant of Erlang delay game with $\mathbb{N}$-optimization where the true characteristic costs are approximated via the Halfin-Whitt heavy traffic regime—an approximation that is asymptotically exact as the arrival rate tends to infinity—always have nonempty cores.

5.6 Conclusion

In this chapter, we considered multi-server queueing systems with infinite waiting room. Sharing servers between providers is beneficial from the whole system point of view, and we analyzed the corresponding games. In both cases considered—fixed and optimized numbers of servers—we studied (existence of) stable allocations of the resource costs for common servers and delay costs for waiting customers. Our analysis reveals that collaboration is

[40] If $|N| > 1$, our proof approach immediately reveals that a stronger, but less crisp, bound is possible: $\hat{\mathcal{P}}(\varphi)$ is also an δ'-PMAS for game $(N, \hat{c}^{\varphi})$ if we set $\delta'_i = \max\{h_M \lambda_i/\lambda_M : M \in 2^N_-,\ |M| = 2, \text{ and } i \in M\}$ for all $i \in N$. Note that $\delta'_i \leq h_{\{i\}}$.

	Fixed service capacity	Optimized service capacity
$M/M/1$ games	Proportional allocation stable and reachable via PMAS	Stable allocation exists
Erlang delay games	Stable allocation exists but PMAS need not exist	Stable allocation need not exist

Table 5.5: *Comparison between the main results of $M/M/1$ games and Erlang delay games.*

always supportable by a stable allocation if players' numbers of servers are exogenously given (in line with the corresponding $M/M/1$ game of Anily and Haviv, 2010), whereas a stable allocation need not exist in general if each coalition optimizes over *integer* numbers of servers (in contrast to $M/M/1$ games considered in, e.g., Yu et al., 2009). Table 5.5 provides an overview.

There are various directions in which our work may be extended. One interesting avenue is to relax the assumption that delay costs and service times are symmetric across players. Opposed to the setting considered in this paper, complete pooling of servers need not be superior anymore if such asymmetries are allowed (see, e.g., Smith and Whitt, 1981). This issue may be circumvented by preferential treatment of more critical classes via priority disciplines, although incorporating this is far from trivial in the analysis.

Another direction is on general service time distributions and/or positive switchover times. To generate preliminary insights for settings with non-exponential service times, one could start by studying games derived from pooling of $M/G/1$ queues. To understand resource pooling in settings where the server incurs a certain setup or traveling time when switching from one customer class to the next, one could study games derived from polling systems (see, e.g., the survey of Boon et al., 2011, for an introduction to polling systems).

Finally, we concede that completely pooling service systems is not always feasible, especially if service facilities of the players are operated at geographically dispersed locations. Nevertheless, some degree of partial pooling may still be feasible. To study such a setting, challenging as it may be, one may consider a model variation in which players partition themselves into separate service groups, in line with Whitt (1999).

—The whole is greater than the sum of its parts.

Aristotle

6

Spare parts games with backordering

6.1 Introduction

In this chapter, which is based on Karsten and Basten (2013), we study the cost sharing problem that arises when several players, each facing a Poisson demand process for an expensive, low-usage item, share a stock point that is controlled by an $(S-1, S)$ base stock policy under full backordering. The players incur symmetrical holding costs for items on hand and symmetrical shortage costs for items that are backlogged. Ordering costs are negligible in comparison, and all players face the same replenishment lead times. The key benefit of pooling in this setting is that one player's on-hand part can cancel out against another player's backorder, thereby saving both holding and shortage costs. The games that we will formulate to study the resulting cost sharing problem may have various applications, e.g., pooled inventories of expensive, low-demand luxury cars or time-shared aircraft. However, our main motivating application is the pooling of spare parts between companies in the capital goods industry. Consequently, the games will be referred to as spare parts games.

We already described examples of spare parts pools and their monetary savings potential in Section 1.2. We now add a few more. Tram operators in the Netherlands are a good example. In the Netherlands, the local public transport in the three largest cities (Amsterdam, Rotterdam, and The Hague, all of which are within an hour's driving distance of each other) is operated by a separate company per city. Even though the operators use trams of different models, there is still a lot of commonality on the component level. Consequently, the tram operators could gain a lot by pooling their inventories. Another example is that

of independently managed plants of a large energy company, as described in Guajardo et al. (2012): the plants currently hold their inventory separately, but annual savings of 44% may be obtainable if pooling is taken into account. While promising, Guajardo et al. mention that it does raise the question of how the plants should share these cost savings. This chapter will answer that question.

The cost and behavior of an inventory system greatly depend on what happens to demands when the system is out of stock. In practice, there are two common ways to deal with stock-outs: using emergency procedures or backlogging. An emergency procedure typically refers to the instantaneous delivery of a part from an alternative supplier. This results in lost sales for the inventory system under consideration, and it is exactly what we studied in Chapter 4. Under backlogging, there is no such alternative supplier with a negligible lead time—we simply have to wait for a part to arrive as a regular replenishment. The setting with backlogging is exactly what we will tackle in the present chapter.

It is well-known in the inventory literature (see, e.g. Feeney and Sherbrooke, 1966) that results obtained for a model with lost sales need not carry over to a model with backordering, or vice versa, and that the two models require different analysis. Indeed, in Chapter 4, we showed the existence of a core allocation by exploiting extensions of the Erlang loss function. The Erlang loss function has no direct relation to the $(S-1, S)$ inventory model with backordering that will be considered in the present chapter, and thus our analysis will be different from Chapter 4.

The inventory model that our analysis is based on has been studied extensively in the literature, due to its practical relevance. As a result, the steady-state distributions of the number of items on hand and on backorder are well-known (see, e.g. Feeney and Sherbrooke, 1966); the same holds for the average long-term costs and the behavior of these costs as a function of the base stock level (see, e.g. Zipkin, 2000). These results, however, do not directly help in identifying a suitable cost sharing mechanism for the problem at hand. For that, we need to understand how the number of items in resupply and the average long-term costs behave when the demand rate varies (as a result of new players joining the pool). We do this by deriving several new properties, including new convexity and elasticity properties of the total costs as a function of the demand rate. These properties allow us to constructively prove that spare parts games, both under fixed and optimized base stock levels, have a non-empty (strict) core. For the latter game, the allocation proportional to players' demand rates is again stable. The reason why we propose a similar proportional rule in three different settings (Erlang loss games, Erlang delay games, and spare parts games, all under optimized numbers of servers) is twofold. First, all come down to a single-attribute game and thus Theorem 2.5 on the proportional rule applies. Second, all feature Poisson demand processes, which turns out to ensure that a proportional allocation can be easily implemented.

This question of implementation, which is relevant for both fixed and optimized numbers of resources, is this: Having found core allocations for the games in expected terms, how to

allocate the costs that are actually realized in, say, a given year? After all, realized costs may differ greatly from expected costs. Although the problem of realization vs. expectation is also present in the models considered in Chapters 4 and 5, we postponed it to the present chapter. There are two reasons for this postponement. First, in the inventory model of this chapter, holding costs are only paid for items on-hand, not for items in the replenishment pipeline. Hence, holding cost payments fluctuate over time depending on the demand realizations; this yields an enriched setting in which holding costs realizations are also important. Second, the process that we propose for assigning cost realizations under *fixed* stocking levels heavily depends on the structure of this chapter's inventory model; it does not extend to $M/G/s/s$ or $M/M/s$ queueing models. That's why the discussion of cost realizations is contained in the present chapter. Nevertheless, the process that we propose for assigning cost realizations under *optimized* stocking levels exploits properties of Poisson processes only. Consequently, the idea behind it also works for Erlang loss and Erlang delay games with optimized numbers of servers.

This chapter makes three main contributions. First, we derive novel structural properties of the cost function in our $(S-1,S)$ inventory model. Second, we introduce spare parts games with fixed base stock levels and identify a (strictly) stable allocation that makes all players better off in every possible sample path. Third, we introduce spare parts games with optimized base stock levels and show that the allocation scheme proportional to players' demand rates is (strictly) population monotonic with appealing properties that enable easy implementation in practice.

The structure of this chapter is similar to the preceding ones. We first describe and analyze the inventory model in Section 6.2. Next, in Section 6.3, we formulate the spare parts situations that induce our games. Sections 6.4 and 6.5 treat the associated games with fixed and optimized base stock levels, respectively. Finally, we draw conclusions and suggest directions for future research in Section 6.6.

6.2 Analysis of the $(S-1,S)$ model with backlogging

In this section we formally describe the inventory model and provide an expression for the expected costs per unit time. Subsequently, we analyze the behavior of these costs as a function of the demand rate.

6.2.1 The inventory model

We consider a single location that stocks one item. Initially, there are $S \in \mathbb{N}_0$ parts on stock. The demand process is a Poisson process with rate $\lambda > 0$. A demand is immediately fulfilled from stock if a part is available. Otherwise, it is backordered and fulfilled first come first serve. In either case, an order for a new part is instigated immediately. This means that the

stock point operates under a continuous-review base stock policy with base stock level S and one-for-one replenishments.

The stock point orders parts from an external supplier. The time that elapses between demand occurrence and receipt of the new part is called the lead time. Lead times are assumed to be i.i.d. according to some general distribution function, and we assume without loss of generality (by rescaling time) that its mean is 1 time period. Independence of successive lead times typically means that the supplier has no capacity constraints.

In the remainder, we will analyze the resulting inventory system in steady-state at an arbitrary point in time. First, we consider the number of parts on order, the so-called pipeline stock, denoted by $X(\lambda)$. By Palm (1938), $X(\lambda)$ is Poisson distributed with mean λ, i.e., for all $x \in \mathbb{N}_0$ it holds that

$$\mathbb{P}[X(\lambda) = x] = \frac{\lambda^x}{x!} e^{-\lambda}. \tag{6.1}$$

We are mainly interested in two related random variables: the number of backorders, $B(S, \lambda)$, and the stock on hand, $I(S, \lambda)$, both as functions of the base stock level S and the demand rate λ. Backorders exist if the pipeline stock is larger than the base stock level, so $B(S, \lambda) = \max\{X(\lambda) - S, 0\}$. Similarly, $I(S, \lambda) = \max\{S - X(\lambda), 0\}$. Thus, using (6.1), we can obtain the distributions and expectations of the number of backorders and the total stock on hand. For instance, the distribution of the number of backorders, $B(S, \lambda)$, is given by[41]

$$\mathbb{P}[B(S, \lambda) = x] = \begin{cases} \sum_{y=0}^{S} \mathbb{P}[X(\lambda) = y] & \text{if } x = 0; \\ \mathbb{P}[X(\lambda) = x + S] & \text{if } x \in \mathbb{N}. \end{cases} \tag{6.2}$$

Accordingly, the expected number of backorders is

$$\begin{aligned} \mathbb{E}B(S, \lambda) &= \sum_{x=S+1}^{\infty} (x - S)\mathbb{P}[X(\lambda) = x] \\ &= \sum_{x=0}^{\infty} x \cdot \mathbb{P}[X(\lambda) = x] - \sum_{x=0}^{\infty} S \cdot \mathbb{P}[X(\lambda) = x] + \sum_{x=0}^{S} (S - x)\mathbb{P}[X(\lambda) = x] \\ &= \lambda - S + \sum_{x=0}^{S} (S - x)\mathbb{P}[X(\lambda) = x]. \end{aligned} \tag{6.3}$$

Similarly, the expected on-hand stock is

$$\mathbb{E}I(S, \lambda) = \sum_{x=0}^{S} (S - x)\mathbb{P}[X(\lambda) = x]. \tag{6.4}$$

We consider holding costs $h > 0$ per unit time per spare part in the on-hand stock. These costs encompass warehousing, insurance, and interest costs on the capital tied up by the

[41] In Chapters 4 and 5, we used B to denote the continuous extension of the Erlang loss function. In this chapter, we use B to denote the number of backorders cf. Equation (6.2).

inventory. Furthermore, we consider penalty costs $b > 0$ per unit time per backordered demand. We disregard the part procurement price or holding costs for pipeline stock because these cost factors would represent constant terms, unaffected by decisions on base stock level or collaboration. The long-term average costs per unit time are given by

$$K(S,\lambda) = h \cdot \mathbb{E}I(S,\lambda) + b \cdot \mathbb{E}B(S,\lambda). \tag{6.5}$$

Although the above-described model is general and might also apply for, e.g., inventories of luxury cars, its formulation and underlying assumptions are driven by inventory systems of expensive, low-demand spare parts meant for technologically advanced capital goods. Indeed, the critical assumptions (the stationary Poisson demand process, the continuous-review one-for-one replenishment strategy, and the independence of successive lead times) are justifiable for this spare parts setting: See Example 4.1 on page 69.

In a typical spare parts setting, demands are triggered by failures of either consumable or repairable machine components. Although we formulated our model for consumable parts, it is also applicable for repairable parts if—instead of placing orders for new parts—any failed component is immediately sent to an uncapacitated repair facility that returns the component to the stock point as a ready-for-use spare part after an i.i.d. repair lead time whose mean is scaled to 1 time unit. Again, direct repair costs and holding costs for parts in repair would represent constant terms and can be disregarded without loss of generality.

6.2.2 Behavior of the costs as a function of the demand rate

In this section we first provide a characterization of the optimal base stock levels and subsequently derive partial derivatives of the cost function K with respect to the demand rate. These intermediate results ultimately enable us to analyze how the costs under optimal base stock levels behave as the demand rate varies on $\mathbb{R}_{++}$. Later on, these properties will prove helpful for analyzing a spare parts game under optimal base stock levels: we use them to prove stability and population monotonicity of proportional allocations in Section 6.5.2 and to show which player benefits the most from pooling in Section 6.5.3. The holding and backorder cost rates, h and b, will remain fixed in the ensuing analysis.

We start by stating several properties of the steady-state probability of having no backorders, $\mathbb{P}[B(S,\lambda) = 0]$, in Lemmas 6.1 and 6.2. These are useful because, as we will show in Lemma 6.3, the optimal base stock levels are intricately related to $\mathbb{P}[B(S,\lambda) = 0]$. Although the properties stated in the following two lemmas are rather straightforward, we were unable to find a proof in the literature, and therefore we provide a proof for completeness.

Lemma 6.1. *Let the demand rate $\lambda > 0$ be fixed.*

(*i*) $\mathbb{P}[B(S,\lambda) = 0]$ *is increasing as a function of* S (*for* S *on* $\mathbb{N}_0$).

(*ii*) $\lim_{S\to\infty} \mathbb{P}[B(S,\lambda) = 0] = 1$.

Proof. Part (i) follows immediately from (6.2) as $\mathbb{P}[X(\lambda) = y] > 0$ for all $y \in \mathbb{N}_0$. Part (ii) holds because $\lim_{S\to\infty} \mathbb{P}[B(S,\lambda) = 0] = \sum_{y=0}^{\infty} \mathbb{P}[X(\lambda) = y] = 1$ as the infinite sum covers the entire support of the Poisson distribution. □

Lemma 6.2. *Let the base stock level $S \in \mathbb{N}_0$ be fixed.*

(i) $\mathbb{P}[B(S,\lambda) = 0]$ *is continuous and differentiable as a function of* λ *(for* λ *on* $\mathbb{R}_{++}$*).*

(ii) $\mathbb{P}[B(S,\lambda) = 0]$ *is decreasing as a function of* λ *(for* λ *on* $\mathbb{R}_{++}$*).*

(iii) $\lim_{\lambda\downarrow 0} \mathbb{P}[B(S,\lambda) = 0] = 1$ *and* $\lim_{\lambda\to\infty} \mathbb{P}[B(S,\lambda) = 0] = 0$*.*

Proof. Part (i) follows immediately from the fact that, for each $y \in \mathbb{N}_0$, $\mathbb{P}[X(\lambda) = y]$ is continuous and differentiable in λ. For Part (ii), we distinguish between two cases. First, if $S = 0$, it holds that

$$\frac{\partial}{\partial\lambda}\mathbb{P}[B(S,\lambda) = 0] = -e^{-\lambda} < 0.$$

Second, if $S \in \mathbb{N}$, we obtain

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\mathbb{P}[B(S,\lambda) = 0] &= \frac{\partial}{\partial\lambda}\sum_{y=0}^{S}\frac{\lambda^y}{y!}e^{-\lambda} \\
&= -e^{-\lambda} + \sum_{y=1}^{S}\frac{\lambda^{y-1}}{(y-1)!}e^{-\lambda} - \sum_{y=1}^{S}\frac{\lambda^y}{y!}e^{-\lambda} \\
&= -e^{-\lambda} + \sum_{y=0}^{S-1}\frac{\lambda^y}{y!}e^{-\lambda} - \sum_{y=1}^{S}\frac{\lambda^y}{y!}e^{-\lambda} \\
&= \sum_{y=0}^{S-1}\left[\left(\frac{\lambda^y}{y!} - \frac{\lambda^y}{y!}\right)e^{-\lambda}\right] - \frac{\lambda^S}{S!}e^{-\lambda} = -\frac{\lambda^S}{S!}e^{-\lambda} < 0.
\end{aligned}$$

We conclude that $\mathbb{P}[B(S,\lambda) = 0]$ is decreasing in λ. Part (iii) follows from

$$\lim_{\lambda\downarrow 0}\mathbb{P}[B(S,\lambda) = 0] = \sum_{y=0}^{S}\lim_{\lambda\downarrow 0}\mathbb{P}[X(\lambda) = y] = \lim_{\lambda\downarrow 0}\mathbb{P}[X(\lambda) = 0] + \sum_{y=1}^{S} 0 = 1,$$

since $\lambda^y e^{-\lambda} \to 0$ as $\lambda \downarrow 0$ for any $y \in \mathbb{N}$, and from

$$\lim_{\lambda\to\infty}\mathbb{P}[B(S,\lambda) = 0] = \sum_{y=0}^{S}\lim_{\lambda\to\infty}\mathbb{P}[X(\lambda) = y] = \sum_{y=0}^{S} 0 = 0,$$

since $\lambda^y e^{-\lambda} \to 0$ as $\lambda \to \infty$ for any $y \in \mathbb{N}_0$. This completes the proof. □

The following lemma states that the cost function in our model is strictly convex in the base stock level and provides a standard characterization of the cost-minimizing base stock level(s) in terms of a newsvendor fractile. This characterization is illustrated in Figure 6.1. Although convexity is relatively well-known for the inventory model under consideration (see, e.g., Zipkin, 2000, p. 215) we show *strict* convexity and address the uniqueness and multiplicity of optimal base stock levels more formally, which will facilitate our analysis.

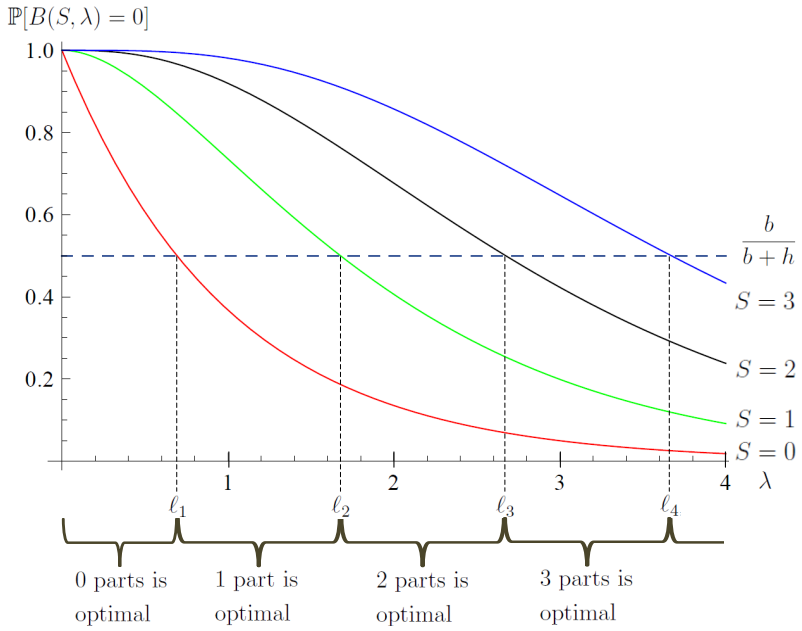

Figure 6.1: *The probability of having no backorders, $\mathbb{P}[B(S,\lambda)=0]$, as a function of the demand rate λ for various base stock levels S. ($h=b=1$.)*

Lemma 6.3. *Let the demand rate $\lambda>0$ be fixed.*

(*i*) *$K(S,\lambda)$ is strictly convex as a function of S (for S on $\mathbb{N}_0$).*

(*ii*) *There is at least one $S\in\mathbb{N}_0$ for which $\mathbb{P}[B(S,\lambda)=0]\geq b/(b+h)$; let S^* denote the smallest such S. Then, S^* is the unique optimal base stock level unless $\mathbb{P}[B(S^*,\lambda)=0]=b/(b+h)$; in that case, both S^* and S^*+1 (and no other) are optimal.*

Proof. Part (*i*). We start by defining the function $\Delta K:\mathbb{N}_0\to\mathbb{R}$ by $\Delta K(S)=K(S+1,\lambda)-K(S,\lambda)$ for all $S\in\mathbb{N}_0$. Consider (6.3) and (6.4), and notice that an equivalent way of writing them is by letting their summations run to $S-1$ instead of S. Using this in combination with (6.2) and (6.5), we obtain for all $S\in\mathbb{N}_0$ that

$$\begin{aligned}\Delta K(S,\lambda)&=h\cdot[\mathbb{E}I(S+1,\lambda)-\mathbb{E}I(S,\lambda)]+b\cdot[\mathbb{E}B(S+1,\lambda)-\mathbb{E}B(S,\lambda)]\\&=h\left[\sum_{x=0}^{S}(S+1-x)\mathbb{P}[X(\lambda)=x]-\sum_{x=0}^{S}(S-x)\mathbb{P}[X(\lambda)=x]\right]\\&\quad+b\left[-(S+1)+\sum_{x=0}^{S}(S+1-x)\mathbb{P}[X(\lambda)=x]+S-\sum_{x=0}^{S}(S-x)\mathbb{P}[X(\lambda)=x]\right]\\&=h\sum_{x=0}^{S}\mathbb{P}[X(\lambda)=x]+b\sum_{x=0}^{S}\mathbb{P}[X(\lambda)=x]-b\\&=(b+h)\mathbb{P}[B(S,\lambda)=0]-b.\end{aligned}\tag{6.6}$$

From (6.6), from Part (i) of Lemma 6.1, and from the fact that $b > 0$ and $h > 0$, we conclude that ΔK is increasing. Therefore, $K(S, \lambda)$ is strictly convex in S.

Part (ii). By Lemma 6.1, Part (ii), there always exists an $S \in \mathbb{N}_0$ for which $\mathbb{P}[B(S, \lambda) = 0] \geq b/(b + h)$. By the convexity result established in Part (i) above, it immediately follows from (6.6) that the cost $K(S, \lambda)$ as a function of S achieves a minimum at the smallest $S \in \mathbb{N}_0$ for which $(b+h)\mathbb{P}[B(S, \lambda) = 0]-b \geq 0$, which corresponds to the definition of S^* in the lemma.

We now distinguish two cases. First, if $\mathbb{P}[B(S^*, \lambda) = 0] > b/(b + h)$, then $\Delta K(S^*, \lambda) > 0$, which implies that S^* is the unique optimal base stock level. Second, if $\mathbb{P}[B(S^*, \lambda) = 0] = b/(b + h)$, then $\Delta K(S^*, \lambda) = 0$, and thus $K(S^*, \lambda) = K(S^* + 1, \lambda)$, which implies that both S^* and $S^* + 1$ are optimal base stock levels, and these are the only two optimal base stock levels due to the strict convexity established in Part (i). □

We now introduce some additional notation: for any $n \in \mathbb{N}$ (note that $0 \notin \mathbb{N}$), we define ℓ_n to be the unique positive real number that satisfies

$$\mathbb{P}[B(n - 1, \ell_n) = 0] = b/(b + h),$$

which is well-defined due to Lemma 6.2. Notice that, by Part (ii) of Lemma 6.3, for any $n \in \mathbb{N}$ it holds that ℓ_n is the demand rate for which both base stock levels $n - 1$ and n are optimal. The following lemma formally establishes several additional properties, which are illustrated in Figure 6.1.

Lemma 6.4. *Let $n \in \mathbb{N}$. Then, the following statements hold.*

- (i) $\ell_n < \ell_{n+1}$.
- (ii) *For any λ in the interval (ℓ_n, ℓ_{n+1}), the unique optimal base stock level for the inventory system with demand rate λ is n.*
- (iii) $\lim_{n\to\infty} \ell_n = \infty$.

Proof. Part (i). By Lemma 6.1, Part (i), and by definition of ℓ_n, it holds that

$$\mathbb{P}[B(n, \ell_n) = 0] > \mathbb{P}[B(n - 1, \ell_n) = 0] = b/(b + h). \tag{6.7}$$

Because $\mathbb{P}[B(n, \lambda) = 0]$ is decreasing as a function of λ (by Part (ii) of Lemma 6.2) while, by definition, $\mathbb{P}[B(n, \ell_{n+1}) = 0] = b/(b + h)$, (6.7) implies that $\ell_n < \ell_{n+1}$.

Part (ii). Let $\lambda \in (\ell_n, \ell_{n+1})$. Thus, $\lambda < \ell_{n+1}$; this, in combination with Part (ii) of Lemma 6.2, implies that

$$\mathbb{P}[B(n, \lambda) = 0] > \mathbb{P}[B(n, \ell_{n+1}) = 0] = b/(b + h).$$

Yet, for each $S \in \mathbb{N}_0$ with $S < n$, we observe, again by Part (ii) of Lemma 6.2, that

$$\mathbb{P}[B(S, \lambda) = 0] < \mathbb{P}[B(S, \ell_{S+1}) = 0] = b/(b + h),$$

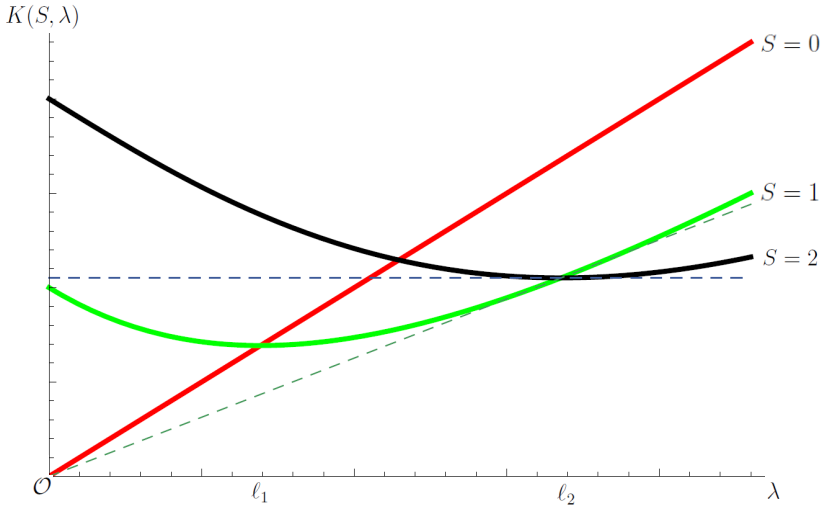

Figure 6.2: *The costs, $K(S,\lambda)$, as a function of the demand rate λ for various base stock levels S. Also shown (dashed) are the tangent lines to $K(1,\lambda)$ and $K(2,\lambda)$ at $\lambda = \ell_2$.*

where the inequality holds because $\lambda > \ell_n \geq \ell_{S+1}$. (To see that $\ell_n \geq \ell_{S+1}$, note that either $S+1 = n$, in which case $\ell_n = \ell_{S+1}$, or $S+1 < n$, in which case $\ell_n > \ell_{S+1}$ by Part (i) of this lemma.) Hence, by Part (ii) of Lemma 6.3, the unique optimal base stock level is n.

Part (iii). By the monotonicity result established in Part (i) above, there are two possibilities: $\lim_{n\to\infty} \ell_n = \infty$ or $\lim_{n\to\infty} \ell_n = A$ for some $A \in \mathbb{R}$. Suppose that the latter is true, aiming for a contradiction, and consider any $\lambda > A$. Then, by Lemma 6.1, there must exist an $S \in \mathbb{N}_0$ such that $\mathbb{P}[B(S,\lambda) = 0] > b/(b+h)$. Given this S, there must exist a $\Lambda > \lambda$ such that $\mathbb{P}[B(S,\Lambda) = 0] = b/(b+h)$ by virtue of Lemma 6.2. Then, by definition, $\ell_{S+1} = \Lambda$. But because $\ell_{S+1} = \Lambda > \lambda > A$ and because of the monotonicity result from Part (i), we obtain a contradiction with the assumption that $\lim_{n\to\infty} \ell_n = A$. We conclude that $\lim_{n\to\infty} \ell_n = \infty$. □

The following lemma considers, for a fixed base stock level greater than zero, the expected steady-state costs per unit time as a function of the demand rate: the lemma states a simple expression for its derivative and shows that this cost function is strictly convex. This convexity is illustrated in Figure 6.2.

Lemma 6.5. *Let the base stock level $S \in \mathbb{N}$ be fixed.*

(i) $\dfrac{\partial}{\partial\lambda} K(S,\lambda) = b - (b+h)\cdot\mathbb{P}[B(S-1,\lambda) = 0]$.

(ii) *$K(S,\lambda)$ is twice differentiable and strictly convex as a function of λ (for λ on $\mathbb{R}_{++}$).*

Proof. Part (i). First, observe that

$$K(S,\lambda) = b\lambda - bS + (b+h)\cdot\left[Se^{-\lambda} + \sum_{x=1}^{S-1}(S-x)e^{-\lambda}\frac{\lambda^x}{x!}\right].$$

Taking the partial derivative with respect to the demand rate,

$$\begin{aligned}
\frac{\partial}{\partial\lambda}K(S,\lambda) &= b + (b+h)\cdot e^{-\lambda}\left[-S + \sum_{x=1}^{S-1}(S-x)\frac{\lambda^{x-1}}{(x-1)!} - \sum_{x=1}^{S-1}(S-x)\frac{\lambda^{x}}{x!}\right] \\
&= b + (b+h)\cdot e^{-\lambda}\left[-S + \sum_{x=0}^{S-2}(S-x-1)\frac{\lambda^{x}}{x!} - \sum_{x=0}^{S-2}(S-x)\frac{\lambda^{x}}{x!} - \frac{\lambda^{S-1}}{(S-1)!} + S\right] \\
&= b + (b+h)\cdot e^{-\lambda}\left[-\sum_{x=0}^{S-2}\frac{\lambda^{x}}{x!} - \frac{\lambda^{S-1}}{(S-1)!}\right] \\
&= b - (b+h)\cdot\sum_{x=0}^{S-1}\mathbb{P}[X(\lambda)=x] \\
&= b - (b+h)\cdot\mathbb{P}[B(S-1,\lambda)=0]. \qquad (6.8)
\end{aligned}$$

Part (ii). This follows upon combining (6.8) with the facts that $\mathbb{P}[B(S-1,\lambda)=0]$ is differentiable and strictly decreasing in λ (cf. Parts (i) and (ii) of Lemma 6.2). □

The following lemma provides insightful expressions for the partial derivatives of the cost function with respect to the demand rate, evaluated at any ℓ_n (the demand rate at which both base stock levels n and $n-1$ are optimal). In particular, when this cost function is considered as a function of λ, the tangent line to $K(n,\lambda)$ at $\lambda=\ell_n$ is flat and the tangent line to $K(n-1,\lambda)$ at $\lambda=\ell_n$ goes through the origin, as illustrated in Figure 6.2.

Lemma 6.6. *Let $n\in\mathbb{N}$ be an arbitrary positive integer.*

(i) $\dfrac{\partial}{\partial\lambda}K(n,\lambda)\big|_{\lambda=\ell_n} = 0.$

(ii) $\dfrac{\partial}{\partial\lambda}K(n-1,\lambda)\big|_{\lambda=\ell_n} = \dfrac{K(n-1,\ell_n)}{\ell_n}.$

Proof. Part (i). By Lemma 6.5, Part (i), and by definition of ℓ_n,

$$\frac{\partial}{\partial\lambda}K(n,\lambda)\big|_{\lambda=\ell_n} = b - (b+h)\cdot\mathbb{P}[B(n-1,\ell_n)=0] = b - (b+h)\cdot\frac{b}{b+h} = 0. \qquad (6.9)$$

Part (ii). We distinguish between two cases. First, if $n=1$, then clearly $K(n-1,\lambda)=b\lambda$, and the result follows trivially. Second, if $n>1$, then we define $\Delta K(\lambda)=K(n,\lambda)-K(n-1,\lambda)$ for all $\lambda>0$. (This ΔK should not be confused with the different ΔK that we considered in the proof of Lemma 6.3.) Differentiating and evaluating at ℓ_n, we obtain

$$\begin{aligned}
\frac{d}{d\lambda}\Delta K(\lambda)\big|_{\lambda=\ell_n} &= \frac{\partial}{\partial\lambda}K(n,\lambda)\big|_{\lambda=\ell_n} - \frac{\partial}{\partial\lambda}K(n-1,\lambda)\big|_{\lambda=\ell_n} \\
&= \Big(b - (b+h)\cdot\mathbb{P}[B(n-1,\ell_n)=0]\Big) - \Big(b - (b+h)\cdot\mathbb{P}[B(n-2,\ell_n)=0]\Big) \\
&= -(b+h)\cdot\left(\sum_{x=0}^{n-1}\mathbb{P}[X(\ell_n)=x] - \sum_{x=0}^{n-2}\mathbb{P}[X(\ell_n)=x]\right) \\
&= -(b+h)\mathbb{P}[X(\ell_n)=n-1], \qquad (6.10)
\end{aligned}$$

where the second equality follows from Part (i) of Lemma 6.5. Combining (6.9) and (6.10), we obtain

$$\begin{aligned}\frac{\partial}{\partial\lambda}K(n-1,\lambda)\big|_{\lambda=\ell_n} &= \frac{\partial}{\partial\lambda}K(n,\lambda)\big|_{\lambda=\ell_n} - \frac{d}{d\lambda}\Delta K(\lambda)\big|_{\lambda=\ell_n}\\ &= (b+h)\mathbb{P}[X(\ell_n)=n-1]. \end{aligned} \tag{6.11}$$

We will now rewrite Expression (6.11), multiplied by ℓ_n. For this, it is convenient to denote $S^* = n-1$ (as $n-1$ is an optimal base stock level) and $Z = (b+h)\left(\sum_{x=0}^{S^*}(S^*-x)\mathbb{P}[X(\ell_n)=x]\right)$. Then, we obtain

$$\begin{aligned}\ell_n(b+h)\mathbb{P}[X(\ell_n)=S^*] &= (b+h)\left(\ell_n\cdot\mathbb{P}[X(\ell_n)=S^*] - \sum_{x=0}^{S^*}(S^*-x)\mathbb{P}[X(\ell_n)=x]\right)+Z\\ &= (b+h)e^{-\ell_n}\left(\ell_n\cdot\frac{(\ell_n)^{S^*}}{S^*!} - S^* - \sum_{x=1}^{S^*}(S^*-x)\frac{(\ell_n)^x}{x!}\right)+Z\\ &= (b+h)e^{-\ell_n}\left(\frac{(\ell_n)^{S^*+1}}{S^*!} + \sum_{x=1}^{S^*}\frac{(\ell_n)^x\cdot x}{x!} - S^* - \sum_{x=1}^{S^*}S^*\frac{(\ell_n)^x}{x!}\right)+Z\\ &= (b+h)e^{-\ell_n}\left(\sum_{x=1}^{S^*+1}\frac{(\ell_n)^x}{(x-1)!} - \sum_{x=0}^{S^*}\frac{(\ell_n)^x}{x!}\cdot S^*\right)+Z\\ &= (b+h)e^{-\ell_n}\left(\sum_{x=0}^{S^*}\frac{(\ell_n)^x}{x!}\cdot\ell_n - \sum_{x=0}^{S^*}\frac{(\ell_n)^x}{x!}\cdot S^*\right)+Z\\ &= (b+h)\sum_{x=0}^{S^*}\mathbb{P}[X(\ell_n)=x](\ell_n-S^*)+Z\\ &= (b+h)\mathbb{P}[B(S^*,\ell_n)=0](\ell_n-S^*)+Z\\ &= b(\ell_n-S^*)+Z = K(S^*,\ell_n). \end{aligned} \tag{6.12}$$

The penultimate equality holds because $\mathbb{P}[B(S^*,\ell_n)=0] = b/(b+h)$ by Part (ii) of Lemma 6.3. By (6.12), we obtain

$$(b+h)\mathbb{P}[X(\ell_n)=n-1] = \frac{K(n-1,\ell_n)}{\ell_n}. \tag{6.13}$$

Combining (6.11) and (6.13) completes the proof. □

We finally consider how the cost of an inventory system with optimal base stock levels behaves as the demand rate varies. To this end, we define the optimal cost function $\tilde{K}:\mathbb{R}_+\to\mathbb{R}$ by

$$\tilde{K}(\lambda) = \begin{cases} 0 & \text{if } \lambda = 0;\\ \min_{S\in\mathbb{N}_0} K(S,\lambda) & \text{if } \lambda > 0. \end{cases} \tag{6.14}$$

This function is well-defined due to Part (ii) of Lemma 6.3. The following theorem states that this function is elastic, as illustrated in Figure 6.3.

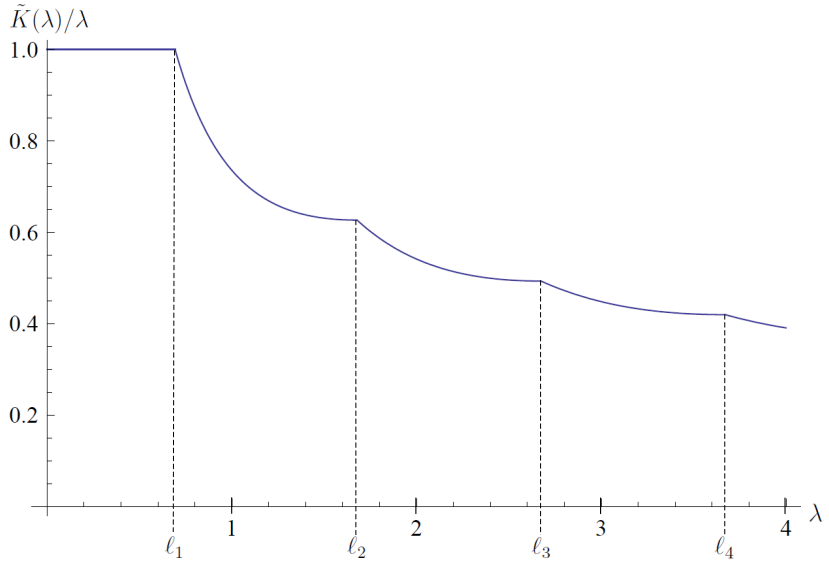

Figure 6.3: *The optimal per-demand costs,* $\tilde{K}(\lambda)/\lambda$. *(*$h = b = 1$*.)*

Theorem 6.7. *The function* $\tilde{K}$ *is elastic. In particular,* $\tilde{K}(\lambda)/\lambda$ *is constant for* λ *on* $(0, \ell_1]$ *and strictly decreasing for* λ *on* $[\ell_1, \infty)$.

Proof. First, for any $\lambda \in (0, \ell_1]$, a base stock level of zero is optimal, by the combination of Part (ii) of Lemma 6.2, Part (ii) of Lemma 6.3, and the definition of ℓ_1. Hence, for λ on $(0, \ell_1]$, $\tilde{K}(\lambda) = K(0, \lambda) = b\lambda$, and thus $\tilde{K}(\lambda)/\lambda$ is constant.

Next, let $n \in \mathbb{N}$. By Part (ii) of Lemma 6.4, for any $\lambda \in [\ell_n, \ell_{n+1})$, it holds that $\tilde{K}(\lambda) = K(n, \lambda)$. We now fix $\tilde{\lambda} \in [\ell_n, \ell_{n+1})$ arbitrarily. Note that, by Part (i) of Lemma 6.4, $\tilde{\lambda} < \ell_{n+1}$. Using this, we find that

$$\begin{aligned} K(n, \tilde{\lambda}) &> K(n, \ell_{n+1}) + (\tilde{\lambda} - \ell_{n+1}) \cdot \frac{\partial}{\partial \lambda} K(n, \lambda)\big|_{\lambda = \ell_{n+1}} \\ &= K(n, \ell_{n+1}) + (\tilde{\lambda} - \ell_{n+1}) \cdot \frac{K(n, \ell_{n+1})}{\ell_{n+1}} \\ &= K(n, \ell_{n+1}) \cdot \frac{\tilde{\lambda}}{\ell_{n+1}}. \end{aligned} \tag{6.15}$$

The inequality holds because, cf. Part (ii) of Lemma 6.5, $K(n, \lambda)$ as a function of λ (for λ on $\mathbb{R}_{++}$) is strictly convex and twice differentiable; consequentially, it lies strictly above its tangent line through ℓ_{n+1} at $\tilde{\lambda} < \ell_{n+1}$. The first equality holds by Part (ii) of Lemma 6.6.

Using this, we obtain

$$\begin{aligned} \frac{\partial}{\partial \lambda} \left(\frac{K(n, \lambda)}{\lambda} \right) \Big|_{\lambda = \tilde{\lambda}} &= \left(\tilde{\lambda} \cdot \frac{\partial}{\partial \lambda} K(n, \lambda)\big|_{\lambda = \tilde{\lambda}} - K(n, \tilde{\lambda}) \right) / \tilde{\lambda}^2 \\ &< \left(\tilde{\lambda} \cdot \frac{\partial}{\partial \lambda} K(n, \lambda)\big|_{\lambda = \ell_{n+1}} - K(n, \tilde{\lambda}) \right) / \tilde{\lambda}^2 \end{aligned}$$

$$= \left(\tilde{\lambda} \cdot \frac{K(n, \ell_{n+1})}{\ell_{n+1}} - K(n, \tilde{\lambda}) \right) / \tilde{\lambda}^2$$
$$< \left(\tilde{\lambda} \cdot \frac{K(n, \tilde{\lambda})}{\tilde{\lambda}} - K(n, \tilde{\lambda}) \right) / \tilde{\lambda}^2 = 0.$$

The first inequality holds by Part (*ii*) of Lemma 6.5. The subsequent equality holds by Part (*ii*) of Lemma 6.6. The second inequality holds by Inequality (6.15). We conclude that $\tilde{K}(\lambda)/\lambda$ is strictly decreasing for λ on $[\ell_n, \ell_{n+1})$.

As, by Part (*ii*) of Lemma 6.3, both n and $n+1$ are optimal base stock levels for demand rate ℓ_{n+1}, it holds that $\tilde{K}(\ell_{n+1}) = K(n, \ell_{n+1}) = K(n+1, \ell_{n+1})$. Furthermore, it follows from Part (*ii*) of Lemma 6.5 that both $K(n, \lambda)$ and $K(n+1, \lambda)$ as functions of λ are continuous at $\lambda = \ell_{n+1}$. We conclude that $\tilde{K}$ is continuous at ℓ_{n+1}.

To summarize, we have established that $\tilde{K}(\lambda)/\lambda$ is non-increasing for λ on $(0, \ell_1]$ and that, for arbitrary positive integer n, this function is strictly decreasing on $[\ell_n, \ell_{n+1})$ and continuous at ℓ_{n+1}. Now, since by Part (*iii*) of Lemma 6.4 it holds that $\bigcup_{n \in \mathbb{N}} [\ell_n, \ell_{n+1}) = [\ell_1, \infty)$, this implies that $\tilde{K}(\lambda)/\lambda$ is strictly decreasing for λ on $[\ell_1, \infty)$. Elasticity of $\tilde{K}$ follows, and the proof is complete. □

This concludes our analysis of the inventory model without a game theoretical perspective. In the next section, we introduce spare parts situations, which feature multiple players.

6.3 Model description

Consider several players who may pool inventories of a common item. Each player witnesses a stationary Poisson demand process, and the demand processes of the players are mutually independent. Each player possesses an exogenously given, unadjustable number of (repairable) spare parts a priori. The players have the same replenishment lead time distribution, possibly because they use the same supplier or repair facility, and without loss of generality we rescale its mean to 1 time unit. The replenishment lead times of the players are mutually independent. The players face symmetric holding costs and backorder costs, possibly because they operate in the same industry under similar operating conditions.

Initially, each player has an exogenously given stock of (repairable) spare parts; this stocking level may be either fixed or adjustable. The players may collaborate by pooling their spare parts. The following definition captures all relevant parameters.

Definition 6.1. A *spare parts situation* is a tuple (N, λ, S, h, b), where

- N is the non-empty, finite set of players;
- $\lambda \in \mathbb{R}^N_{++}$ is the vector of demand rates, where λ_i describes the demand rate of customers that belong to player $i \in N$;

- $S \in \mathbb{N}_0^N$ is the vector of stocking levels, where S_i describes the number of (repairable) spare parts that player $i \in N$ brings to any coalition;
- $h > 0$ is the holding cost rate;
- $b > 0$ is the backorder cost rate.

In line with notation of the preceding chapters, we write $\lambda_M = \sum_{i \in M} \lambda_i$ and $S_M = \sum_{i \in M} S_i$ for any coalition $M \in 2_-^N$.

Example 6.1. Reconsider the air carrier companies from Example 4.1, who could pool spare gearboxes. In Example 4.1, an emergency procedure (a lease from an exogenous infinite supplier or an expedited repair process) was instigated upon a stock-out. In contrast, we now suppose that such emergency procedures are not possible: if a part is not available, then we simply have to wait for a regular repair to complete. One possible reason for this is that the parts are no longer in production.

For brevity, consider Airline A only. This airline owned 1 spare gearbox and expected 4 demands per year. Recall that downtime costs were incurred at a rate of 500,000 euro per grounded aircraft per day, which comes down to 182,500,000 per year. Holding costs were incurred at a rate of 10,000 euro per on-hand gearbox per year. We can represent this situation as spare parts situation (N, λ, S, h, b) by letting $N = \{A\}$, $\lambda_A = 4/4 = 1$, $S_A = 1$, $h = 10,000/4 = 2,500$, and $b = 182,500,000/4 = 45,625,000$, where we measure time in quarter years (i.e., the mean repair lead time, which is scaled to 1) and money in euros. $\diamond$

6.4 Fixed base stock levels

In this section we consider the game that arises when each coalition maintains the aggregate of the individual players' base stock levels. Pooling allows a reduction in backorders and on-hand stocks. After describing this setting in more detail, we identify a strictly stable cost allocation.

6.4.1 Game

Consider any spare parts situation $\varphi = (N, \lambda, S, h, b)$ and any coalition $M \in 2_-^N$. Coalition M can set up a single stock point from which the combined demand streams of the coalition members are fulfilled first-come first-served. Since the superposition of independent Poisson processes is also a Poisson process, this single stock point would face a Poisson arrival process with merged rate $\lambda_M = \sum_{i \in M} \lambda_i$. The stock point takes over the stocks of (repairable) spare parts of the coalition members in full.

Based on these assumptions, the pooled system behaves as an $(S-1, S)$ inventory model as described in Section 6.2.1. We assume throughout that players are interested in reducing their long-term average holding and backorder costs, and that other, smaller effects of setting

up the pool (e.g., additional transportation costs or economies of scale in warehousing) are insignificant in comparison. Recalling the expected steady-state costs per unit time K as defined in (6.5), we can now formulate a game corresponding to spare parts situation φ.

Definition 6.2. For any spare parts situation $\varphi = (N, \lambda, S, h, b)$, we call the game (N, c^φ) with

$$c^\varphi(M) = K(S_M, \lambda_M) \tag{6.16}$$

for all $M \in 2^N_-$ the *associated spare parts game with fixed base stock levels*.

6.4.2 A strict core allocation

In this subsection, we propose payment rates that depend on the current pipeline stocks. As we show, these rates make every coalition better off under every realization of the pipeline stocks and result in a strict core allocation in expectation. First, for a given spare parts situation (N, λ, S, h, b), we introduce some additional terminology and notation.

A *pipeline stock realization* is a vector $x = (x_i)_{i \in N} \in \mathbb{N}_0^N$. The number x_i should be interpreted as the total number of replenishment parts in the pipeline, at an arbitrary point in time, that were triggered by demands for items at player i. If we were to observe the inventory system of any coalition $M \in 2^N_-$ in steady-state at an arbitrary point in time, then the number x_i for any $i \in M$ would be drawn from a Poisson distribution with mean λ_i.[42] We will, however, not focus on the underlying stochastic process for now. Rather, we consider the inventory system at an arbitrary point in time, and observe the pipeline stock realization x. Taking x as a given, we denote $x_M = \sum_{i \in M} x_i$ for any coalition M. The total rate at which costs are incurred per unit time in coalition M as long as the pipeline stock realization remains x is equal to

$$K_M^{\text{rate}}(x_M) = h \cdot \max\{S_M - x_M, 0\} + b \cdot \max\{x_M - S_M, 0\}.$$

We are particularly interested in x_N, the total number of parts in the pipeline for the grand coalition. If $S_N > x_N$, then we say that *backorders are scarce*. Note that backorders being scarce need not imply that $S_i > x_i$ for all $i \in N$. If $S_N < x_N$, then we say that *on-hand parts are scarce*. And if $S_N = x_N$, then we say that *nothing is scarce*.

Suppose that the grand coalition has formed. For any pipeline stock realization x, we will propose a way to fully assign $K_N^{\text{rate}}(x)$ over the players. Specifically, we propose that any

[42] This number is thus unaffected by any pooling arrangement. This is a crucial observation; it differentiates the model in this chapter from the preceding ones! The observation only holds because the supply or repair process is uncapacitated and corresponds to an $M/G/\infty$ queue. In the Erlang loss or Erlang delay system, the number of servers is finite, which has an important implication: even if for any sample path the same demands arrive no matter the coalition structure, the service can differ. For instance, if player i has 5 servers and 0 arrivals over a certain period, while player j has 1 server and 4 arrivals over a certain period, then the number of servers that are busy with customers of player i (the analogue of our pipeline stock realization x_i) would depend on whether or not the players formed a coalition.

player $i \in N$ should incur net cost at a rate of

$$A_i(x) = \begin{cases} h(S_i - x_i) & \text{if backorders are scarce;} \\ b(x_i - S_i) & \text{if on-hand parts are scarce;} \\ 0 & \text{if nothing is scarce.} \end{cases} \tag{6.17}$$

per unit time. Note that these rates may be negative at some time for some players, and that $\sum_{i \in N} A_i(x) = K_N^{\text{rate}}(x)$. The following example illustrates these rates, indicates that they never make anyone worse off, and describes how they may be implemented via concrete transfer payments.

Example 6.2. Gyro Gearloose has invented a new machine that produces chocolate bars at a constant rate. He has sold it to several people in Duckburg: besides Gyro himself, Scrooge McDuck and Launchpad McQuack are also operating a number of these machines.[43] The machine has one critical component that is prone to failures. This component can be bought for 3000 euros from a supplier, due upon receipt of the component, and there is a fixed lead time of 1 month. We represent the set of players by $N = \{G, S, L\}$. Initially, each player $i \in N$ has $S_i = 2$ spare parts on hand, with nothing in the pipeline. The monthly interest rate is 1%, i.e., monthly holding costs are $h = 30$ per part. Downtime costs due to lost chocolate bar production are $b = 10000$ per month per broken machine.

The three players store their parts in a joint warehouse, though each puts their own parts on a separate shelf. Consider the system on a certain day, say January 1. Suppose that up till today, all players had always been able to satisfy their demands from their own shelves. But today will be different. At the beginning of the day, Gyro has $x_G = 2$ parts in the pipeline, Scrooge has $x_S = 1$ part in the pipeline, and Launchpad has $x_L = 1$ part in the pipeline. At some point during the day, one of Gyro's machines fails. As always, Gyro immediately orders a new part. This increases the size of the pipeline triggered by Gyro's demands to $x_G = 3$. However, since $S_G = 2 < 3 = x_G$, Gyro is unable to fulfill his demand with an on-hand part. Yet, since $S_N = 6 > 5 = x_N$, backorders are scarce and Gyro can obtain the required part from another player.

Suppose that the players contractually agreed to fully pool their parts and to incur costs as described by the payment rates A. For convenience of determining transfer payments, the players agreed that everyone first always incurs the holding and backlogging costs for their *own* inventory and backlogs. Since no shortages had arisen in the past, no transfer payments had been necessary up till now. However, given the current pipeline stock realization $x = (3, 1, 1)$, the payment rates should be $A_G(x) = -h$, $A_S(x) = h$, and $A_L(x) = h$, and thus transfer payments are necessary. Gyro calls Scrooge, and the following discussion ensues.

Gyro: Mr. McDuck, I would like to buy your spare part for the regular price of 3000 euros.

[43] The Dutch reader may recognize these characters more readily as Willie Wortel, Dagobert Duck, and Turbo McKwek, respectively.

You will receive and eventually pay for the new part that I just ordered, while rewarding me at a rate of $h = 30$ per month as long as that order is in the pipeline.

Scrooge: Curse me kilts... I have to give you my spare part and pay you for the privilege as well? How is that what we agreed to?

Gyro: Gee, Mr. McDuck, lighten up. My proposal indeed lets us implement the payment rates that we all contractually agreed to. Besides, what's more important: a few measly euros, or helping out a friend?

Scrooge: Is this a multiple choice question? Kindness is of little use in this world. Anyway, this allocation make no sense to me. Why should I pay you?

Gyro: Okay, let me explain why this allocation is reasonable. By giving me your part, I certainly save downtime costs at a rate of b per month. But you also get money upfront that I'm sure you can put to good use, i.e., you save on holding costs. However, Launchpad and I could achieve exactly the same thing. I recently read a paper by Owen (1975) on linear production games and a paper by Sánchez-Soriano et al. (2001) on a refinement called transshipment games, which perfectly describe our current predicament. Indeed, each coalition of two or more players that includes me can save $h + b$, while no other coalition can achieve anything. Such a game only has one core allocation: the one where I get all the benefits.

Scrooge: I guess that makes sense. But what if both Launchpad's and my own machine fail tomorrow? Launchpad tends to crash everything he touches, and I have so many machines that one is bound to fail soon. If that happens, then there will be no spare part left to fix my machine, and I would regret giving away my part.

Gyro: In that case, on-hand parts become scarce, and the payments are switched. Rather than you paying me, I will start paying you at a monthly rate of $b = 10000$ to compensate you for your downtime. One way or another, you won't be worse off!

Scrooge: Fair enough... although I am still not reaping any positive benefits from this. I had heard that these pooling arrangements make everyone *strictly* better off. How about that? As you know, my favorite hobby is actually increasing my wealth.

Gyro: Well, in the future, if you face a stock-out while I have a part available, I will be happy to help you out. At that point in time, you will be able to obtain positive benefits from this collaboration.

Scrooge: Sure, you convinced me. Let's do this!

In the end, Gyro got his part, and everyone collaborated happily ever after. ◇

The following lemma shows that the above-described rates indeed always never make any coalition worse off.

Lemma 6.8. *Consider a spare parts situation (N, λ, S, h, b). Let x be a pipeline stock realization. Let $M \in 2^N_{--}$ be a subcoalition. The following inequalities hold:*

(i) $\sum_{i\in M} A_i(x) = K^{rate}_M(x_M)$ *if backorders are scarce and* $S_M \geq x_M$.
(ii) $\sum_{i\in M} A_i(x) < K^{rate}_M(x_M)$ *if backorders are scarce and* $S_M < x_M$.
(iii) $\sum_{i\in M} A_i(x) = K^{rate}_M(x_M)$ *if on-hand parts are scarce and* $S_M \leq x_M$.
(iv) $\sum_{i\in M} A_i(x) < K^{rate}_M(x_M)$ *if on-hand parts are scarce and* $S_M > x_M$.
(v) $\sum_{i\in M} A_i(x) = K^{rate}_M(x_M)$ *if nothing is scarce and* $S_M = x_M$.
(vi) $\sum_{i\in M} A_i(x) < K^{rate}_M(x_M)$ *if nothing is scarce and* $S_M \neq x_M$.

Proof. Part (i). Suppose that backorders are scarce and that $S_M \geq x_M$. Then, $\sum_{i\in M} A_i(x) = h\,(S_M - x_M) = K^{\text{rate}}_M(x_M)$.

Part (ii). Suppose that backorders are scarce and that $S_M < x_M$. Then, $\sum_{i\in M} A_i(x) = h\,(S_M - x_M) < 0 < b\,(x_M - S_M) = K^{\text{rate}}_M(x_M)$, where the inequalities hold because, by assumption, $S_M - x_M < 0$.

Part (iii). Suppose that on-hand parts are scarce and that $S_M \leq x_M$. Then, $\sum_{i\in M} A_i(x) = b\,(x_M - S_M) = K^{\text{rate}}_M(x_M)$.

Part (iv). Suppose that on-hand parts are scarce and that $S_M > x_M$. Then, $\sum_{i\in M} A_i(x) = b\,(S_M - x_M) < 0 < h\,(S_M - x_M) = K^{\text{rate}}_M(x_M)$, where the inequalities hold because, by assumption, $S_M - x_M > 0$.

Part (v). Suppose that nothing is scarce and that $S_M = x_M$. Then, $\sum_{i\in M} A_i(x) = 0 = K^{\text{rate}}_M(x_M)$.

Part (vi). Suppose that nothing is scarce and that $S_M \neq x_M$. Then, $\sum_{i\in M} A_i(x) = 0 < K^{\text{rate}}_M(x_M)$, where the inequality holds because due to the mismatch in S_M and x_M, coalition M either pays holding or backlogging costs at a positive rate. □

For any spare parts situation $\varphi = (N, \lambda, S, h, d)$, we define the vector $\mathscr{A}(\varphi) \in \mathbb{R}^N$ by

$$\mathscr{A}_i(\varphi) = \sum_{x\in\mathbb{N}_0^N} \mathbb{P}\big[(X(\lambda_i))_{i\in N} = x\big]\, A_i(x)$$

for all $i \in N$, where

$$\mathbb{P}\big[(X(\lambda_i))_{i\in N} = x\big] = \prod_{i\in N} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}$$

represents the probability of observing pipeline stock realization x at an arbitrary point in time. Using these definitions and Lemma 6.8, we can show that these rates induce a strict core allocation in expectation, as stated in the following theorem. This theorem focuses on settings with more than one player who have at least one spare part between them to ensure that positive pooling benefits are obtainable.

Theorem 6.9. *Consider a spare parts situation $\varphi = (N, \lambda, S, h, b)$ with $|N| \geq 2$ and $S_N > 0$. Then, $\mathscr{A}(\varphi)$ is a strictly stable allocation for the associated spare parts game with fixed base stock levels (N, c^φ).*

Proof. Let $M \in 2^N_{--}$ be a subcoalition. Then,

$$\begin{aligned}
\sum_{i\in M} \mathscr{A}_i(\varphi) &= \sum_{x\in\mathbb{N}_0^N} \mathbb{P}\big[(X(\lambda_i))_{i\in N} = x\big] \sum_{i\in M} A_i(x) \\
&< \sum_{x\in\mathbb{N}_0^N} \mathbb{P}\big[(X(\lambda_i))_{i\in N} = x\big] K_M^{\text{rate}}(x_M) \\
&= \sum_{y\in\mathbb{N}_0} \sum_{x\in\mathbb{N}_0^N : x_M = y} \mathbb{P}\big[(X(\lambda_i))_{i\in N} = x\big] K_M^{\text{rate}}(y) \\
&= \sum_{y\in\mathbb{N}_0} \mathbb{P}\left[\sum_{i\in M} X(\lambda_i) = y\right] K_M^{\text{rate}}(y) \\
&= \sum_{y\in\mathbb{N}_0} \mathbb{P}[X(\lambda_M) = y] K_M^{\text{rate}}(y) = c^\varphi(M),
\end{aligned}$$

where the inequality holds by Lemma 6.8. The inequality is strict because the case as described in Part (*vi*) of Lemma 6.8 occurs with positive probability. Indeed, since $S_N > 0$ by assumption, it holds that $S_M < x_M = x_N = S_N$ has a positive occurrence probability if $S_M < S_N$ and that, alternatively, $x_M < S_M = x_N = S_N$ has a positive occurrence probability if $S_M = S_N$. The penultimate inequality, i.e., $\mathbb{P}\left[\sum_{i\in M} X(\lambda_i) = y\right] = \mathbb{P}[X(\lambda_M) = y]$ holds because the sum of several Poisson distributed random variables is again Poisson distributed according to the sum of the parameters. We conclude that $\mathscr{A}(\varphi)$, if it were an (efficient) allocation, would be stable.

It remains to check that $\mathscr{A}(\varphi)$ is indeed an (efficient) allocation, i.e., $\sum_{i\in N} \mathscr{A}_i(x) = c^\varphi(N)$. This, however, can be proven analogous to the above derivation, except that the inequality becomes an equality. We conclude that $\mathscr{A}(\varphi)$ is a strictly stable allocation for (N, c^φ). □

From this theorem, we obtain the next one, which states that pooling the inventory and demand streams of several spare parts inventory systems (with common lead time and cost structure) leads to a strict reduction in expected backorders (6.3), expected on-hand inventory (6.4), and expected costs (6.5). The intuition behind this is that pooling allows one player's backorder to cancel against another player's on-hand part.

Theorem 6.10. *Consider a spare parts situation $\varphi = (N, \lambda, S, h, b)$ with $N = \{1, 2\}$ and $S_N > 0$. Then the following strict subadditivity properties hold:*

(*i*) $K(S_1, \lambda_1) + K(S_2, \lambda_2) > K(S_1 + S_2, \lambda_1 + \lambda_2)$.

(*ii*) $\mathbb{E}B(S_1, \lambda_1) + \mathbb{E}B(S_2, \lambda_2) > \mathbb{E}B(S_1 + S_2, \lambda_1 + \lambda_2)$.

(*iii*) $\mathbb{E}I(S_1, \lambda_1) + \mathbb{E}I(S_2, \lambda_2) > \mathbb{E}I(S_1 + S_2, \lambda_1 + \lambda_2)$.

Proof. By Theorem 6.9, it holds that

$$K(S_1, \lambda_1) + K(S_2, \lambda_2) = c^{\varphi}(\{1\}) + c^{\gamma}(\{2\}) > c^{\varphi}(N) = K(S_1, S_2, \lambda_1 + \lambda_2), \tag{6.18}$$

which shows validity of Part (i). Recall that, for any $\hat{S} \in \mathbb{N}_0$ and $\hat{\lambda} > 0$, it holds that $K(\hat{S}, \hat{\lambda}) = b(\hat{\lambda} - \hat{S}) + (b+h)\sum_{x=0}^{\hat{S}}(\hat{S} - x)\mathbb{P}[X(\hat{\lambda}) = x]$. So, subtracting $b(\lambda_1 + \lambda_2 - S_1 - S_2)$ from both sides of (6.18) and dividing by $h + b > 0$, we obtain

$$\sum_{x=0}^{S_1}(S_1 - x)\mathbb{P}[X(\lambda_1) = x] + \sum_{x=0}^{S_2}(S_2 - x)\mathbb{P}[X(\lambda_2) = x] > \sum_{x=0}^{S_1+S_2}(S_1 + S_2 - x)\mathbb{P}[X(\lambda_1 + \lambda_2) = x].$$

Comparison with Equations (6.3) and (6.4) completes the proof of Parts (ii) and (iii), respectively. □

6.5 Optimized base stock levels

The previous section showed that it is beneficial to share inventories while maintaining the aggregate of the base stock levels, but costs can of course be reduced further by re-optimizing the joint base stock level. After all, due to the risk pooling effect, lower base stock levels may suffice to jointly serve all demand streams in a cost-effective way. Although such a lower aggregate base stock level might result in an *increase* in the expected backorders compared to the no-pooling situation, this may be counterweighted by the reduction in the expected on-hand inventory. However, due to the opposite effects explained in Section 4.5.2, there is no direct relation between games with fixed and optimized based stock levels. In this section we consider the games corresponding to the setting where base stock levels are optimized.

6.5.1 Game

Let (N, λ, S, h, b) be a spare parts situation, and consider a coalition $M \in 2^N_-$. We remark that since base stock levels are adjustable in this section, S has becomes superfluous in the tuple. For any particular choice of the base stock level $s \in \mathbb{N}_0$, the behavior of the stock point would correspond to the model described in Section 6.2, and thus the expected costs per time unit faced by coalition M would be equal to $K(s, \lambda_M)$. Assuming that any coalition picks an optimal base stock level, which exists by Lemma 6.3, we formulate the following game.

Definition 6.3. For any spare parts situation $\varphi = (N, \lambda, S, h, b)$, we call the game (N, d^{φ}) with

$$d^{\varphi}(M) = \min_{s \in \mathbb{N}_0} K(s, \lambda_M) \tag{6.19}$$

for all $M \in 2^N_-$ the associated *spare parts game with optimized base stock levels.*

Before proceeding with our analysis, we discuss our choice to concentrate on a pure cost model. The reason is that the pure cost model permits adequate comparisons between

coalitions. To illustrate, suppose that we would instead have chosen to study a service constraint model where each coalition would minimize the *integer* base stock level subject to, say, a 95% fill rate constraint. Then, improvements in service beyond that 95% would unjustly appear worthless: In such a service model, a player would prefer a coalition in which he would face 1000 euro/month in holding costs under a 95.0% fill rate over a coalition costing 1001 euro/month for 99.9%. Such unnatural outcomes are avoided by considering backlogging costs explicitly. We avoided this issue in Section 4.5 by using the linear interpolation of the Erlang loss function to formulate a service constraint game. For the $(S-1,S)$ inventory model with backlogging, however, our analysis in Section 6.2.2 focused on optimal costs directly, not on linear interpolations. Studying a service constraint model would therefore require considerable additional analysis of a linear interpolation, which we will not go into here.

We next propose an allocation rule that divides costs proportional to the demand rate of each player: the rule $\mathscr{P}$ is defined by $\mathscr{P}_i(\varphi) = d^\varphi(N) \cdot \lambda_i/\lambda_N$ for each $i \in N$ in spare parts situation (N, λ, S, h, b). Extending this idea to every coalition, we define the proportional allocation scheme rule $\mathcal{P}$ by

$$\mathcal{P}_{i,M}(\varphi) = d^\varphi(M) \cdot \lambda_i/\lambda_M \tag{6.20}$$

for each coalition M and $i \in M$ in spare parts situation (N, λ, S, h, b). Note that in contrast to the state-dependent cost rate that we had in the previous section, we now moved to a constant cost rate. The following example illustrates this proportional allocation scheme rule numerically and simultaneously shows that the Shapley value is not necessarily in the core.

Example 6.3. Consider three companies in the capital goods industry that wish to pool common parts. One company expects to face 0.1 demands per month on average. The monthly demand rates are 0.8005 and $\ln 2$ for the other two companies. The part in question is very expensive; a single part on hand costs 10,000 euros per month, and a machine that is down will cost 10,000 euros per month as well.[44] This can be modeled as a spare parts situation $\varphi = (N,\, \lambda,\, S,\, h,\, b)$ with player set $N = \{1,2,3\}$, $\lambda_1 = 0.1$, $\lambda_2 = 0.8005$, $\lambda_3 = \ln 2 (\approx 0.6931)$, arbitrary S, and $h = b = 10000 = 10^4$.

To illustrate the determination of an optimal base stock level and associated costs, consider the singleton coalition $\{3\}$. By Equation (6.1), $\mathbb{P}[X(\lambda_3) = 0] = 0.5$. Combining this with Equations (6.3), (6.4), and (6.5), we obtain for the case were player 3 would decide to stock zero parts that $\mathbb{E}B(0,\lambda_3) = \ln 2$, $\mathbb{E}I(0,\lambda_3) = 0$, and $K(0,\lambda_3) = 10^4 \cdot \ln 2$. If player 3 would decide to use a base stock level of one instead, then $\mathbb{E}B(1,\lambda_3) = \ln 2 - 1 + 0.5$, $\mathbb{E}I(1,\lambda_3) = 0.5$, and $K(1,\lambda_3) = 10^4 \cdot \ln 2$. As $K(0,\lambda_3) = K(1,\lambda_3)$ and this cost function is strictly convex in the base stock level (by Lemma 6.3), minimal costs are achieved with a base stock level

[44] We are aware that these downtime costs are rather low relative to the holding costs. We chose these parameter values because they simultaneously yield a computationally convenient illustration of the costs and allocations involved, in addition to a game whose Shapley value lies outside the core.

Coalition M	Optimal base stock levels	$d^{\varphi}(M)$	$\mathcal{P}_{1,M}(\varphi)$	$\mathcal{P}_{2,M}(\varphi)$	$\mathcal{P}_{3,M}(\varphi)$
$\{1\}$	0	1000	1000	*	*
$\{2\}$	1	6987	*	6987	*
$\{3\}$	0 and 1	6931	*	*	6931
$\{1,2\}$	1	7132	792	6340	*
$\{1,3\}$	1	6980	880	*	6100
$\{2,3\}$	1	9428	*	5053	4375
N	1	10000	628	5023	4349

Table 6.1: *The spare parts game with optimized base stock levels and the proportional allocation scheme of Example 6.3.*

of either 0 or 1, and $d^{\varphi}(\{3\}) = 10^4 \cdot \ln 2$. In the remainder of this example, we will round monthly cost values to whole euros for notational convenience.

If player 1 would join to form coalition $\{1,3\}$, then it can be verified that a base stock level of one for their combined stock point is optimal; thus, $d^{\varphi}(\{1,3\}) = K(1, \lambda_1 + \lambda_3) \approx 6980$. Under the proportional allocation scheme rule $\mathcal{P}$, player 3 would have to pay $\mathcal{P}_{3,\{1,3\}}(\varphi) \approx 6100$ in coalition $\{1,3\}$, which is lower than $\mathcal{P}_{3,\{3\}}(\varphi) \approx 6931$ (see Table 6.1). This strict population monotonicity can be verified for the members of all other nested pairs of coalitions as well, implying that $\mathcal{P}(\varphi)$ is strictly population monotonic. Accordingly, the spare parts game with optimized base stock levels (N, d^{φ}) has a non-empty strict core containing $\mathscr{P}(\varphi)$.

Although this allocation is stable, it need not make every coalition better off in *every* sample path, as opposed to the setting with fixed base stock levels. Indeed, the optimal base stock level for the grand coalition is less than the sum of the optimal base stock levels in the partitioning $\{\{1,2\},\{3\}\}$. This implies that in case of unexpectedly high demand in a certain year, the grand coalition may face higher overall costs. Players have to content themselves with the knowledge that they get better off in expectation, i.e., in the long run.

The game's Shapley value $\Phi(N, c^{\varphi})$, which assigns $\Phi_1(N, c^{\varphi}) \approx 556$, $\Phi_2(N, c^{\varphi}) \approx 4774$, and $\Phi_3(N, c^{\varphi}) \approx 4670$, is not in the core of this game because $\Phi_2(N, c^{\varphi}) + \Phi_3(N, c^{\varphi}) > d^{\varphi}(\{2,3\})$. Accordingly, the game is not concave; indeed, $d^{\varphi}(\{1,3\}) - d^{\psi}(\{3\}) < 100 < 500 < d^{\phi}(\{1,2,3\}) - d^{\phi}(\{2,3\})$. ◇

6.5.2 A strictly population monotonic allocation scheme

In Example 6.3, cost allocation could be carried out in a stable and population monotonic way via the proportional rules. In this section we prove this by using that each player is associated with a single attribute. We discern two different proof approaches. The first approach is based on the elasticity stated in Theorem 6.7 and allows us to show *strict* stability and population

monotonicity. The second approach is based on a connection with newsvendor games.

We start off with the first approach. We let

$$S^*(\lambda) = \min\{S \in \mathbb{N}_0 : K(S, \lambda) = \tilde{K}(\lambda)\}$$

denote the (smallest) optimal base stock level for any demand rate $\lambda > 0$. This number is well-defined by Lemma 6.3. The following theorem shows that the proportional allocation scheme rule $\mathcal{P}(\varphi)$ always results in a PMAS. Moreover, it accomplishes an SPMAS if demand rates are sufficiently high for each coalition with two or more players to have an optimal base stock level greater than zero.

Theorem 6.11. *Let $\varphi = (N, \lambda, S, h, b)$ be a spare parts situation. For the associated game (N, d^φ), the following holds.*

(i) $\mathcal{P}(\varphi)$ is a PMAS.

(ii) If $S^*(\lambda_L) > 0$ for each $L \in 2^N_-$ with $|L| \geq 2$, then $\mathcal{P}(\varphi)$ is an SPMAS.

Proof. Part *(i)*. Let $M, L \in 2^N_-$ with $M \subset L$, and let $i \in M$. By Theorem 6.7,[45]

$$\mathcal{P}_{i,L}(\varphi) = d^\varphi(L)\frac{\lambda_i}{\lambda_L} = \tilde{K}(\lambda_L)\frac{\lambda_i}{\lambda_L} \leq \tilde{K}(\lambda_M)\frac{\lambda_i}{\lambda_M} = d^\varphi(M)\frac{\lambda_i}{\lambda_M} = \mathcal{P}_{i,M}(\varphi). \tag{6.21}$$

Part *(ii)*. For arbitrary $M, L \in 2^N_-$ with $M \subset L$, assume that $S^*(\lambda_L) > 0$. This implies that λ_L, the collective demand rate of coalition L, is strictly larger than ℓ_1, the demand rate for which both base stock levels 0 and 1 are optimal. Therefore, the inequality in (6.21) is strict by Theorem 6.7. Accordingly, $\mathcal{P}(\varphi)$ is an SPMAS. □

From this theorem, we immediately obtain the following corollary.

Corollary 6.12. *Let $\varphi = (N, \lambda, S, h, b)$ be a spare parts situation.*

(i) The associated game (N, d^φ) is totally balanced, and its core contains $\mathscr{P}(\varphi)$.

(ii) If $S^*(\lambda_L) > 0$ for each $L \in 2^N_-$ with $|L| \geq 2$, then (N, d^φ) has a non-empty strict core containing $\mathscr{P}(\varphi)$.

We now turn to the alternative, second proof approach, which is based on two lemmas. The first lemma is based on the observation that, mathematically, the problem of finding optimal base stock levels in our (infinite-horizon) inventory model corresponds to a (single-period) newsvendor problem. Recall that newsvendor games were defined in Section 3.2.3.

Lemma 6.13. *Let $\varphi = (N, \lambda, S, h, b)$ be a spare parts situation. Then, the associated spare parts game with optimized base stock levels is a newsvendor game.*

[45] This result would also follow upon combining Theorem 6.7 and Theorem 2.5. However, Equation (6.21) is short and direct, and it requires no invocation of Theorem 2.5.

Proof. We set oversupply cost $c_o = h$, undersupply cost $c_u = b$, and stochastic demand D_i for any player $i \in N$ distributed according to a Poisson distribution with mean λ_i. Then, for any coalition M, we obtain

$$\begin{aligned} d^{\varphi}(M) &= \min_{S \in \mathbb{N}_0} \left(h \cdot \mathbb{E}I(S, \lambda_M) + b \cdot \mathbb{E}B(S, \lambda_M) \right) \\ &= \min_{S \in \mathbb{N}_0} \left(h \cdot \mathbb{E}\max\{S - X(\lambda_M), 0\} + b \cdot \mathbb{E}\max\{X(\lambda_M) - S, 0\} \right) \\ &= \min_{S \in \mathbb{N}_0} \left(c_o \cdot \mathbb{E}\max\left\{S - \sum_{i \in M} D_i, 0\right\} + c_u \cdot \mathbb{E}\max\left\{\sum_{i \in M} D_i - S, 0\right\} \right) \\ &= \min_{s \geq 0} \left(c_o \cdot \mathbb{E}\max\left\{s - \sum_{i \in M} D_i, 0\right\} + c_u \cdot \mathbb{E}\max\left\{\sum_{i \in M} D_i - s, 0\right\} \right). \end{aligned}$$

The first equality holds by Equations (6.5) and (6.19). The third equality holds because both $X(\lambda_M)$ and $\sum_{i \in M} D_i$ are both Poisson distributed with mean λ_M. The final equality holds because, even if we allow optimization over the real numbers, there will always be an optimal integer-valued order size due to the discreteness of the demand distribution. We conclude that the game (N, d^{φ}) is a newsvendor game. □

The second lemma uses the notion of a single-attribute game, as introduced in Section 2.3.10.

Lemma 6.14. *Consider a non-decreasing function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(0) = 0$, and suppose that all games embedded in f are newsvendor games. Then, for any player set N and attribute vector $a \in \mathbb{R}^N_{++}$, the allocation scheme assigning $c(M) \cdot a_i / a_M$ to any $i \in M$ for each $M \in 2^N_-$ is a PMAS for the corresponding single-attribute game.*

Proof. Newsvendor games are known to have a non-empty core in general—a result that was derived independently by Müller et al. (2002) and Slikker et al. (2001). The result then follows upon combination of Parts (i) and (iii) of Theorem 2.5. □

These lemmas provide an alternative way to prove Part (i) of Theorem 6.11. We briefly restate this result for convenience.

Theorem 6.15. *Let $\varphi = (N, \lambda, S, h, b)$ be a spare parts situation. Then, $\mathcal{P}(\varphi)$ is a PMAS for the associated game (N, c^{φ}).*

Proof. Consider the class of all spare parts situations with holding and backorder costs h and b. All associated spare parts games with optimized base stock levels comprise all single-attribute games embedded in the optimal cost function $\tilde{K}$, with attributes corresponding to the players' demand rates. By Lemma 6.13, all these games are newsvendor games. This implies, by Lemma 6.14, that the allocation scheme assigning costs proportional to players' attributes, which is exactly what $\mathcal{P}$ prescribes, is a PMAS. □

We conclude this subsection with several remarks. Firstly, even though the class of spare parts games with optimized base stock levels turned out to coincide with the class of newsvendor games where all players have Poisson distributed demand, our focus on this specific *subclass* of newsvendor games—a subclass that had never been explicitly studied before—enabled us to find novel results that do not extend to newsvendor games in general. Secondly, the combination of the connection with newsvendor games and Theorem 2.5 allowed a short proof for population monotonicity, but it does not provide insights into the structure of the problem as our analysis in Section 6.2.2 did. Moreover, our structural analysis in Section 6.2.2 allowed us to identify a *strictly* stable allocation (Part (*ii*) of Theorem 6.11) and it will allow us to identify the player who benefits most from the collaboration (in Section 6.5.3); these additional results do not follow from the alternative proof approach.

The following example illustrates an application of Lemma 6.14 to newsvendor games beyond the spare parts context. Key in the example, which deals with normally distributed demands, is that the demand information of any player can be fully represented as a single real number (i.e., a single attribute) and that the demand distribution of each coalition belong to the same class.

Example 6.4. For given oversupply cost $c_o \geq 0$, undersupply cost $c_u \geq 0$, and variance-to-mean ratio C, consider the function $f : \mathbb{R}_{++} \to \mathbb{R}$ defined by

$$f(\mu) = \min_{s \geq 0} \left(c_o \cdot \mathbb{E} \max \left\{ s - D(\mu), 0 \right\} + c_u \cdot \mathbb{E} \max \left\{ D(\mu) - s, 0 \right\} \right),$$

where $D(\mu) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = \mu C$. Then, the set of single-attribute games embedded in f is comprised of all newsvendor games with oversupply cost c_o, undersupply cost c_u, and stochastic demand $D_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ with $\sigma_i^2 = \mu_i C$ and $\mu_i > 0$ for all players i. It is known that these games are convex (Özen et al., 2011, Theorem 2). What's new is that, by Lemma 6.14, any such newsvendor game (N, c) admits a *proportional* PMAS that assigns $c(M) \cdot \mu_i / \sum_{j \in M} \mu_j$ to any $M \in 2^N_-$ and $i \in M$. ◇

6.5.3 Who reaps the benefits?

One might wonder who actually reaps most of the benefits of the collaboration. By benefits we mean the difference between the costs incurred by a player when acting alone and the cost assigned to this player under rule $\mathscr{P}$. Thus, the benefits for a player with demand rate $\lambda > 0$ when participating in a spare parts pool with aggregate demand rate $\Lambda \geq \lambda$ are given by

$$\mathcal{B}(\lambda, \Lambda) = \tilde{K}(\lambda) - \lambda \cdot \tilde{K}(\Lambda)/\Lambda.$$

Now, will the smallest player (i.e., the player with the lowest demand rate) always benefit most, or will the largest player always take the lion's share? The following example shows that it could actually be neither of them. Thus, the proportional rule does not favor smaller or larger players categorically.

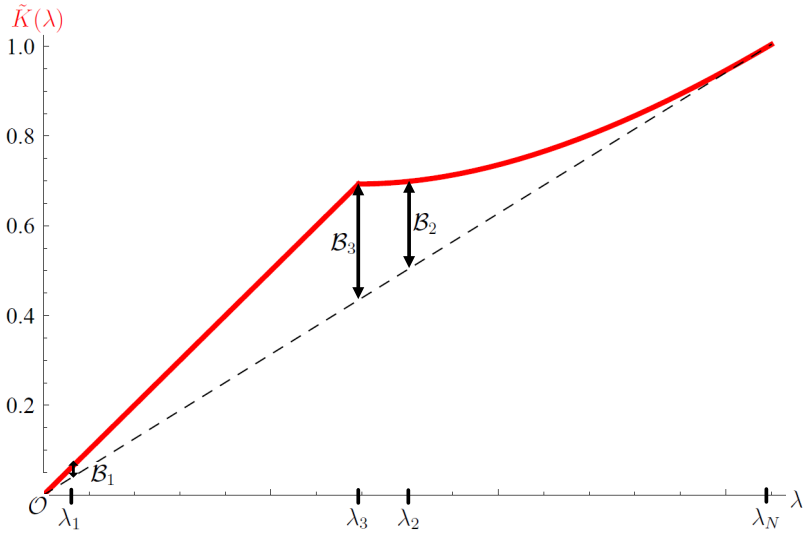

Figure 6.4: *The optimal costs $\tilde{K}(\lambda)$ as a solid line, a dotted line through the origin with slope $\tilde{K}(\lambda_N)/\lambda_N$, and the benefits $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$ for the players in Example 6.5.*

Example 6.5. Reconsider the spare parts situation φ as described in Example 6.3. For this situation, the players' benefits (rounded to four decimals) are $\mathcal{B}_1 = \mathcal{B}(\lambda_1, \lambda_N) = 0.0372$, $\mathcal{B}_2 = \mathcal{B}(\lambda_2, \lambda_N) = 0.1964$, $\mathcal{B}_3 = \mathcal{B}(\lambda_3, \lambda_N) = 0.2582$. So, $\mathcal{B}_3 > \mathcal{B}_2 > \mathcal{B}_1$, even though $\lambda_2 > \lambda_3 > \lambda_1$. Hence, a large player (e.g., 2) might reap less benefit than a smaller player (e.g., 3). Yet, a large player (e.g., 3) might reap more benefit than a smaller player (e.g., 1) as well. This is represented graphically in Figure 6.4. ◇

This example suggests that the largest benefits are typically reaped by a player with demand rate equal to ℓ_n for some $n \in \mathbb{N}$, i.e., a demand rate for which two base stock levels are optimal. To prove that this indeed holds in general, we will use the following lemma.

Lemma 6.16. *Let a pooling group's total demand rate $\Lambda > \ell_1$ be fixed.*

(i) For all $\lambda \in (0, \ell_1)$, $\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_1, \Lambda)$.

(ii) If $S^*(\Lambda) \geq 2$, then for any $n \in \{1, \ldots, S^*(\Lambda) - 1\}$ and $\lambda \in (\ell_n, \ell_{n+1})$, at least one of the following two inequalities hold: $\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_{n+1}, \Lambda)$ and/or $\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_n, \Lambda)$.

(iii) For all $\lambda \in (\ell_{S^*(\Lambda)}, \Lambda]$, $\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_{S^*(\Lambda)}, \Lambda)$.

Proof. Part *(i)*. For any $\lambda \in (0, \ell_1)$, it holds that $S^*(\lambda) = S^*(\ell_1) = 0$, which implies that $\tilde{K}(\lambda)/\lambda = \tilde{K}(\ell_1)/\ell_1 = b$. Using this, we find for any $\lambda \in (0, \ell_1)$ that

$$\mathcal{B}(\lambda, \Lambda) = \lambda \left(\frac{\tilde{K}(\lambda)}{\lambda} - \frac{\tilde{K}(\Lambda)}{\Lambda} \right) = \lambda \left(\frac{\tilde{K}(\ell_1)}{\ell_1} - \frac{\tilde{K}(\Lambda)}{\Lambda} \right) < \ell_1 \left(\frac{\tilde{K}(\ell_1)}{\ell_1} - \frac{\tilde{K}(\Lambda)}{\Lambda} \right) = \mathcal{B}(\ell_1, \Lambda)$$

where the inequality holds because $\tilde{K}(\ell_1)/\ell_1 > \tilde{K}(\Lambda)/\Lambda$ by Theorem 6.7.

Part (*ii*). Assume that $S^*(\Lambda) \geq 2$, and let $n \in \{1, \ldots, S^*(\Lambda) - 1\}$. For λ on $[\ell_n, \ell_{n+1}]$, it holds that $\mathcal{B}(\lambda, \Lambda) = K(n, \lambda) - \lambda \cdot \tilde{K}(\Lambda)/\Lambda$, i.e., the sum of a term that is strictly convex in λ by Lemma 6.5 and a term that is linear in λ. This means that $\mathcal{B}(\lambda, \Lambda)$ is strictly convex in λ for $[\ell_n, \ell_{n+1}]$. By convexity, it reaches a maximum at one of the two endpoints, i.e., either at ℓ_n or ℓ_{n+1}. Because the convexity is strict, any λ that is not equal to ℓ_n or ℓ_{n+1} cannot lead to a maximum.

Part (*iii*). For λ on $[\ell_{S^*(\Lambda)}, \Lambda]$, it holds that $\mathcal{B}(\lambda, \Lambda) = K(S^*(\Lambda), \lambda) - \lambda \cdot \tilde{K}(\Lambda)/\Lambda$ is strictly convex in λ, for the same reason as in Part (*ii*). Combining this strict convexity with $\mathcal{B}(\lambda, \Lambda) > 0$ for all $\lambda \in [\ell_{S^*(\Lambda)}, \Lambda)$, which follows from Part (*ii*) of Corollary 6.12, and $\mathcal{B}(\lambda, \Lambda) = 0$, we conclude that a unique maximum on $[\ell_{S^*(\Lambda)}, \Lambda]$ is achieved at $\ell_{S^*(\Lambda)}$. Hence, $\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_{S^*(\Lambda)}, \Lambda)$ for all $\lambda \in (\ell_{S^*(\Lambda)}, \Lambda]$. □

The following theorem states that a demand rate for which two base stock levels are optimal yields maximal benefits, provided that it is not optimal for the grand coalition to stock zero parts (in which case there would clearly be no pooling benefits at all).

Theorem 6.17. *Consider a spare parts pool with total demand rate $\Lambda > \ell_1$. Then, there exists an $n \in \{1, \ldots, S^*(\Lambda)\}$ such that $\mathcal{B}(\ell_n, \Lambda) > \mathcal{B}(\lambda, \Lambda)$ for all $\lambda \in (0, \Lambda]$ with $\lambda \neq \ell_{n'}$ for any $n' \in \mathbb{N}$.*

Proof. Let $n^* \in \operatorname{argmax}_{n \in \{1, \ldots, S^*(\Lambda)\}} \mathcal{B}(\ell_n, \Lambda)$. Let $\lambda \in (0, \Lambda]$ with $\lambda \neq \ell_{n'}$ for any $n' \in \mathbb{N}$. If $\lambda < \ell_1$, then

$$\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_1, \Lambda) \leq \mathcal{B}(\ell_{n^*}, \Lambda),$$

where the first inequality is due to Lemma 6.16, Part (*i*), and the second inequality holds by choice of n^* as a maximizer. Similarly, if $\lambda > \ell_1$, then $\mathcal{B}(\lambda, \Lambda) < \mathcal{B}(\ell_{n^*}, \Lambda)$ by Part (*ii*) or (*iii*) of Lemma 6.16, and by choice of n^* as a maximizer. □

From this theorem, we immediately obtain the following corollary, which concerns situations where the grand coalition optimally stocks a single part. This is quite common for low-demand, expensive spare parts.

Corollary 6.18. *Let $\varphi = (N, \lambda, S, h, b)$ be a spare parts situation with optimized base stock levels such that $S^*(\lambda_N) = 1$ and $\lambda_i = \ell_1$ for some $i \in N$. Then, $\mathcal{B}(\lambda_i, \lambda_N) > \mathcal{B}(\lambda_j, \lambda_N)$ for all $j \in N$ with $\lambda_j \neq \lambda_i$.*

6.5.4 A truth-inducing implementation

In this subsection, we show that the proportional allocation rule can be easily implemented via a simple cost division per realization. Although our games have been formulated in expected terms to investigate a priori attractiveness of pooling, fair assignments of *realized* costs in any finite time period will be required to sustain cooperation in practice. We will propose a process to do that, and subsequently discuss the implications of this process for

truthful information disclosure in the context of a non-cooperative game. These issues have been previously considered in the context of collaborating newsvendors with no inventory carryover: fair divisions of realized costs are studied by Dror et al. (2008), Chen and Zhang (2009, Remark 3), and Kemahlıoğlu-Ziya and Bartholdi III (2011, Section 6) while schemes that induce truthful revelation of private demand information are studied by Norde et al. (2011). We, in contrast, tackle these issues in an infinite-horizon continuous-review inventory model. Our approach differs from existing approaches in the newsvendor context because we exploit the PASTA property of Poisson processes.

To propose a method for assigning realized costs, we make the natural assumption of a first-in first-out (FIFO) stock discipline: whenever more than one part is available in the on-hand stock when a demand is placed, the demand is fulfilled by the oldest part in the on-hand stock. We now propose the following method to allocate costs as they materialize in an inventory system with any base stock level $S \in \mathbb{N}_0$ operated by the grand coalition in any spare parts situation $(N, (\lambda_i)_{i\in N}, S, h, b)$.

Process 6.1. Realized costs for the grand coalition are assigned as follows:

- Each player, upon placing a demand when the on-hand stock is positive, pays all holding costs incurred for the part taken (according to FIFO). That is, if the taken part was delivered at the stock point at a time τ and the player's demand occurs at time t, then this player incur $h(t-\tau)$ upon placing his demand.

- Each player, upon placing a demand that is backordered, pays all backorder costs incurred for this backorder. That is, if the demand occurs at time t and the associated backorder is later fulfilled via delivery of a new part at time τ, then this player incurs $b(\tau - t)$ over the duration of his backorder.

This process assigns the holding costs incurred for some part in an intuitively fair way to the player who directly benefits from the part. No holding costs are assigned to a player who does not benefit from on-hand inventory (as a result of not facing any demands over a certain time period or due to unfortunate stock-outs at his demand epochs). Additionally, since players fully incur all costs for their own backorders, the process eliminates the need for transfer payments of backorder costs, thereby avoiding disputes about their exact magnitude — this is an important property for the capital goods context wherein backorder costs typically comprise the downtime costs of a player's machine due to unavailability of a spare part. Moreover, as stated in the following lemma, the process can indeed be used to *implement*[46] the proportional allocation.

[46]Here, we assume that any player takes a part from the pool if *and only if* he faces a component failure. Opportunistic behavior (not demanding a part upon a failure, or demanding a part before a failure occurs) may be prohibited by requiring a failed part in exchange for any part taken from the inventory.

Lemma 6.19. *Under Process 6.1, the share of long-term average costs (of an inventory system operated by the grand coalition with base stock level $S \in \mathbb{N}_0$) borne by player $i \in N$ is λ_i/λ_N. In particular, if the grand coalition optimally stocks $S^*(\lambda_N)$ parts, the long-term average costs assigned to each player under Process 6.1 coincide with the assignment of expected costs under the proportional allocation rule $\mathscr{P}$.*

Proof. Follows directly from PASTA. □

A final appealing property of Process 6.1 is that it removes any incentive for players to lie about their demand rates a priori. Although thus far we have adhered to the assumption of full and open information (a standard assumption in cooperative game theory), in reality a player's demand rate may be private information, and true demand rates may not be verifiable as, e.g., there may only be a handful demands over an entire 30-year collaboration.

During initial negotiations, when all cooperating players have to state their demand rate for the purpose of joint base stock level optimization, a player might lie if the collaboration would be implemented via an inappropriate cost realization assignment method. For example, under a method that charges each player a yearly fee based on his stated demand rates independent of that player's realized demand volume in that year, a player might have a reason to understate his actual demand rate a priori. However, under Process 6.1, truth telling is a Nash equilibrium in the corresponding noncooperative information disclosure game. In this game, the player set is N; the strategy set of each player is $\mathbb{R}_{++}$, where strategy λ_i for player i should be interpreted as stating demand rate λ_i; and the payoff to each player for any strategy profile $(\hat{\lambda}_i)_{i\in N}$ is equal to the long-term average costs assigned under Process 6.1 in the inventory system with base stock level $S^*(\sum_{i\in N} \hat{\lambda}_i)$.

Theorem 6.20. *The strategy profile $(\lambda_i)_{i\in N}$, in which each player $i \in N$ states his true demand rate λ_i, is a Nash equilibrium in this noncooperative game.*

Proof. Consider a player $i \in N$ and suppose that all other players $j \in N \setminus \{i\}$ announce their true demand rate λ_j. By lying, i.e., stating any demand rate $\mathcal{L} \in \mathbb{R}_{++}$ other than λ_i, player i can only accomplish a possibly suboptimal base stock level since $K(S^*(\lambda_N), \lambda_N) \leq K(S^*(\sum_{j\in N\setminus\{i\}} \lambda_j + \mathcal{L}), \lambda_N)$ by definition of S^*. Yet, the fraction of long-term average costs assigned to player i under Process 6.1 is equal to λ_i/λ_N by Lemma 6.19, i.e., is independent of the demand rate that he states. Thus, player i minimizes his costs by stating λ_i. This completes the proof. □

6.6 Conclusion

In this chapter we studied the cost allocation problem in a spare parts inventory model with backordering. We derived new structural properties of the resulting cost function, in particular concerning its behavior for varying demand rates, which may be relevant beyond

	Fixed base stock levels	Optimized base stock levels
Game in expectation	Pipeline allocation stable	Proportional allocation stable
Implementation	Players get better off in every realization	Players cannot get better off in every realization

Table 6.2: *Overview and contrast of the main results.*

the context of our games. Using these properties, we were able to show that the associated cooperative games, both under fixed and optimized base stock levels, have non-empty (strict) cores. In particular, for the game with fixed based stock levels, an allocation based on pipeline stock realizations is stable and makes all players better off in every realization. For the game with optimized base stock levels, the allocation scheme that assigns costs proportional to each player's demand rate was shown to be population monotonic whilst being easy to implement via a truth-inducing process. Table 6.2 provides an overview and contrast.

Possible limitations to the practical implications of the findings are that we assumed symmetric shortage costs and negligible transshipment costs. Full pooling under a FIFO discipline may be detrimental if players have different shortage costs, in line with observations made in Chapter 4, or if there are huge costs for transshipping a part from one player to the other. That said, Slikker et al. (2005) did show that "generalized" newsvendor games featuring asymmetric shortage costs and non-negligible transshipment costs are balanced, assuming that all coalitions operate under *optimal* ordering and pooling policies. Since we showed that "basic" spare parts games with optimized base stock levels are "basic" newsvendor games, it would be interesting to investigate whether or not there is an analogous connection for their "generalized" variants under optimal policies. A related idea would be to study whether a newsvendor connection might allow us to extend the results for spare parts games with Poisson demand processes to spare parts games with general demand processes.

Another possible direction for further research is to extend the model to two echelon levels: a one warehouse, multiple retailers setting. The spare parts at the central warehouse may be owned by a coalition of retailers, or by a third party. If this third party is the original equipment manufacturer, then it may also be interesting to allow this party to exert additional design effort to improve component reliability. This may be beneficial from the whole system's point of view, but it also raises the question of what share of the benefits the manufacturer is entitled to. Cooperative game theory may provide the tools to determine (existence of) fair allocations of collective costs.

—Winter is coming.

Eddard Stark

7

Inventory pooling games with non-stationary, dynamic demands

7.1 Introduction

The previous chapters dealt with stationary infinite-horizon models. That is, we assumed that the demand rate is constant over time, and we were concerned with steady-state behavior. Moreover, we assumed that the demand rate is known with certainty, that there are no correlations over time or between players, and that the cost parameters are stationary as well. These assumptions enabled us to formulate tractable models. However, the real word is often not stationary; in reality, things tend to change over time (Silver, 1981, p. 639).

To deal with this the present chapter, which is based on Karsten et al. (2013a), will tackle inventory pooling games in an environment that is not only stochastic, but possibly non-stationary and dynamic as well. With non-stationary, we mean that demand distributions and cost parameters *vary* over time, though in a way that is known beforehand. With dynamic, we mean that probabilities and beliefs are *related* over time—related due to correlations and/or uncertainty about the true process governing the demands. Before moving towards the modeling details and the results, let us first try to understand for which kind of settings the incorporation of non-stationarity and dynamics into an inventory model are important. For concreteness, we discuss two illustrative (single-player) examples: spare parts and style goods.

In the spare parts context, demand is never really stationary over the entire life cycle.

The life cycle of an installed base of machines can be divided into three stages: increment, steady-state, and retirement (Jin and Tian, 2012, p.157). In the increment stage, which typically lasts several years to a decade (Jin and Tian, 2012, p.157), a new generation of machines is released to the market. In this stage, the demand process for spare parts is non-stationary due to the gradual increase of the installed base (Jin and Tian, 2012). In the steady-state stage, the installed base remains more or less constant over a period of multiple decades. Nevertheless, spare parts demands can sometimes be non-stationary due to, e.g., changing maintenance policies, changing operating conditions, or reliability growth programs (van Jaarsveld and Dekker, 2011, p.425). Finally, in the retirement stage, when new models start to replace the old machines, obsolescence of spare parts is an important issue that leads to non-stationarity of demand (Pinçe and Dekker, 2011). Accordingly, the number of failures per year changes over time. Moreover, even if the component failure rate λ remains constant over time, the value of λ is actually unknown! Especially at the time of initial provisioning, estimates on the reliability of a component may be quite tentative due to the absence of operational data (Azoury, 1985; Aronis et al., 2004). The best we can often do is to provide a prior distribution for λ, based on past experience with similar components or on initial estimates made by engineers. However, as more failure data is obtained and the component's reliability becomes better known, the distribution that we assigned to λ can be updated dynamically. As shown in the above-mentioned papers, the incorporation of non-stationarity and dynamics can result in substantially improved stocking policies in settings where the above-described effects play an important role.

For the next example that illustrates non-stationarity and dynamics, we turn to style goods such as toys, electronics, sporting goods, or apparel. These goods are usually on the market for a short amount of time due to rapid shifts in fashion trends or high rates of technological innovation. Accordingly, they are often sold in a concentrated, short selling season with a limited number of replenishment opportunities. During this season, demand first ramps up, then peaks, and eventually dies out. Moreover, there are sales spikes before Christmas and other seasonal effects (Neale and Willems, 2009, p.388). On top of all that, the degree of uncertainty regarding an item's sales appeal varies over time as well. Indeed, when a new item is introduced, it is difficult to tell whether it will be hot or not; the best one can do is to make a priori estimates based on history, experience, or intuition (Murray and Silver, 1966; Popović, 1987). These estimates can then be updated once actual sales roll in. An excellent account of this is given in Fisher and Raman (1996), who studied inventory control at Sport Obermeyer, a major fashion skiwear designer and manufacturer. Taking into account high correlations of demand over time, Fisher and Raman propose the policy of ordering only a modest amount at first, improving forecasts based on observing the first 20% of the demand, and subsequently placing a second order based on the improved forecasts. They show that their policy has the potential of doubling profits compared to a policy without demand updating.

To summarize, we argued that it is highly likely for demand not to come from a stationary i.i.d. process. So does this mean that all the models we analyzed thus far are ineffective and worthless? Of course not. In fact, most of the literature on inventory management assumes that demand is a stationary process, and for good reasons. We mention three. First of all, the stationarity assumption allows for easy derivations of the steady-state performance, e.g., via queueing theory. This simplifies the computations considerably and leads to easily understandable policies. The second reason, which is a consequence of the first, is that the relative simplicity of the stationarity models allows us to derive nice subhomogeneity, convexity, and elasticity properties that improve our theoretical understanding of inventory systems. The more bells and whistles we add to a model, the less tractable the analysis becomes. Finally, even though the demand process is rarely truly stationary, there are many settings where it is sufficiently *close* to stationary, at least within the planning horizon for which the model is intended to aid. Especially for, e.g., the steady-state stage in the lifecycle of a capital good, the demand process for spare parts will likely fluctuate only very little.

Relative to inventory models with stationary i.i.d. demand processes, much less work exists on models with non-stationary, unknown, and/or correlated demand processes. Still, the need to address these factors has been recognized by a number of authors, including many of the founding fathers and early pioneers of stochastic inventory theory. Early works that incorporate non-stationarity in inventory models, though assuming known demand distributions, are Karlin (1960) and Hadley and Whitin (1962). They propose stochastic dynamic programming as the method for deriving an optimal ordering policy. Early works on inventory models where the demand parameters are unknown, though stationary over time, are Dvoretzky et al. (1952) and Scarf (1959). They propose Bayesian updating as the method for learning about future demand from past observations. Inventory models featuring the *combination* of non-stationarity, unknown demand processes, and correlations over time—which require both stochastic dynamic programming and Bayesian updating—are studied by, e.g., Fisher and Raman (1996), Eppen and Iyer (1997), Petruzzi and Dada (2001), and Treharne and Sox (2002). All consider a periodic-review, finite-horizon model with known linear ordering, holding, and penalty costs.

All of the above-mentioned papers consider models with a *single* decision maker who is responsible for all inventories. However, as argued before, inventories (and information) may often be shared between multiple self-interested players. One can think of tram operators that pool spare parts, independent retailers that pool style goods, or different municipalities that share snow salt. This chapter deals with the corresponding inventory pooling games in a non-stationary, dynamic environment. In line with most of the above-mentioned papers, we will consider a periodic-review, finite-horizon model with known linear ordering, holding, and backlogging costs. Our focus will be on the resulting cost sharing problem.

While tackling this problem, we run into two interesting questions that were not relevant for the papers above, because they only come to the fore when collaborating over time in a

dynamic, uncertain environment. First, how does the ability of a player to observe demand realizations of other players—even if those players do not participate in an inventory pool—influence the game? Second, would it be beneficial for an opportunistic player to stay solitary at first and only try to join a pooling group after observing a sudden demand spike? The answers we will provide to these questions may be interesting not only in our inventory model but also in other contexts with dynamic uncertainties.

Altogether, the model, results, and analysis we present in this chapter make three primary contributions. First, we introduce stochastic dynamic inventory pooling games, a rich modeling framework that extends basic newsvendor games and that incorporates the impact of being able to observe other players' demand realizations. Second, we prove that these games are always balanced and, despite their inherent complexity, have a simple proportional allocation—proportional to the players' mean demands—in their cores under certain assumptions. Third, we show for two players that if demand realizations of one player are always observable by the other, then it is never beneficial to strategically join late.

The remainder of this chapter is organized as follows. We start by defining stochastic dynamic inventory pooling situations in Section 7.2 and the associated stochastic dynamic inventory pooling games in Section 7.3. In Sections 7.4 and 7.5, we study the balancedness of these games and the stability of proportional cost allocations, respectively. The combination of multiple players, multiple periods, stochastic demands, correlations, dynamics, and strategic behavior culminates in the notational nightmare that is Section 7.6, in which we analyze settings where players may strategically join late. We conclude in Section 7.7.

7.2 Model description

Consider a number of players that are all facing stochastic demands for a particular item over a finite time horizon. They may collaborate by coordinating orders and pooling inventories.

Definition 7.1. We define a *stochastic dynamic inventory pooling situation* γ as a tuple of input parameters $(N, \Omega, \mathbb{P}, I^0, T, D, p, b, h)$, where

- N is the nonempty, finite set of players;
- Ω is the nonempty sample space (i.e., the set of scenarios or possible futures), which is assumed to be either finite or countably infinite;
- $\mathbb{P}$ is the probability measure on Ω, i.e., a function $\mathbb{P} : \Omega \to [0, 1]$ with $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$;
- $I^0 = (I_i^0)_{i \in N} \in \mathbb{R}^N$ is the initial inventory vector specifying that player i owns I_i^0 products at the beginning of the first period;

- $T = \{T_{\text{start}}, T_{\text{start}} + 1, \ldots, T_{\text{end}}\}$ is a non-empty, finite set[47] of consecutive integers that represent the periods in the planning horizon;
- $D = (D_i^t)_{i \in N, t \in T}$ is the sequence of demand vectors, where $D_i^t : \Omega \to \mathbb{R}$ is the stochastic demand faced by player i in period t with $\mathbb{E}[D_i^t] < \infty$;
- $p = (p^t)_{t \in T} > \mathbf{0}$ is the vector of production cost vectors, where p^t is the cost to produce or order one item in period t;
- $b = (b^t)_{t \in T} \geq \mathbf{0}$ is the vector of backlogging cost vectors, where b^t is the cost incurred for every unit of backlogged demand at the end of period t;
- $h = (h^t)_{t \in T} \geq \mathbf{0}$ is the vector of holding cost vectors, where h^t is the cost incurred for every unit of on-hand inventory at the end of period t.

We remark that his chapter's focus on a finite-horizon model with a fixed number of periods differs from the infinite-horizon models of the preceding chapters: In preceding chapters, a coalition aimed to minimize the expected costs per unit time in steady-state; in the present chapter, a coalition will aim to minimize the total expected costs over the horizon. A period in our model can represent any granularity: a day, a month, or a decade, depending on the application.

We next describe the sequence of events. Within each period, events for any coalition occur in the following order. First, a replenishment order is placed, which arrives immediately. This order is used to satisfy any outstanding backlog and to increase the coalition's inventory level. Next, demand realizes. (As costs among players are identical, there is no question of which set of demands to satisfy.) Unsatisfied demand is backlogged. At the end of the period, backlogs and inventories are observed, and penalty and holding costs, respectively, are incurred. The inventory is carried, as usual, to the next period.

Example 7.1. Imagine two farmers, Alice and Bob, each possessing a field of grain at different geographical locations. Each farmer owns a Harvester 2000X, a sophisticated machine that is used to harvest crops in the summer. Both farmers desire to harvest in July and August. Weather conditions are not in their favor, however: meteorologists have determined that a huge storm is going to occur, although the meteorologists are not sure when or where. Due to peculiar meteorological conditions, this storm will occur either on July 1 over the field of Alice, or on August 1 over the field of Bob. The former scenario is twice as likely as the latter. A storm destroys all harvesting equipment in its path.

Destroyed harvesting equipment leads to a monthly production loss of 250K euros in both July and August. From September on, no crops can be harvested anymore, and any Harvester 2000X machine, destroyed or not, is depreciated and worthless.

[47] Customarily, $T = \{1, 2, \ldots, \mathcal{T}\}$ for some $\mathcal{T} \in \mathbb{N}$. However, we also allow more general sets, such as $T = \{3, 4, 5\}$, to allow comparisons to collaborations that start at a later time period.

To guard against the impact of this upcoming storm, the farmers can buy a set of service parts, which can be used to repair any destroyed Harvester 2000X. This set of service parts costs 60K euros and can only be bought at a store that is open only on the last day of June and the last day of July. Neither of the farmers currently owns any service parts. Warehouse space where the heavy service parts can be stored can be rented at a monthly fee of 3K euros. Unused service parts are worthless and can be freely disposed of on August 2. Parts cannot be discarded before that date due to environmental regulations.

The farmers could prepare for the storm separately. Alternatively, they could collaborate by jointly purchasing and pooling service parts. This situation may be modeled as a stochastic dynamic inventory pooling situation $\gamma = (N, \Omega, \mathbb{P}, I^0, T, D, p, b, h)$, where

- $N = \{A, B\}$;
- $\Omega = \{\omega_1, \omega_2\}$;
- $\mathbb{P}(\omega_1) = \frac{2}{3}$; $\mathbb{P}(\omega_2) = \frac{1}{3}$;
- $I_1^0 = I_2^0 = 0$;
- $T = \{1, 2\}$;
- $D_A^1(\omega_1) = 1$; $D_A^2(\omega_1) = 0$; $D_A^1(\omega_2) = 0$; $D_A^2(\omega_2) = 0$;
- $D_B^1(\omega_1) = 0$; $D_B^2(\omega_1) = 0$; $D_B^1(\omega_2) = 0$; $D_B^2(\omega_2) = 1$;
- $p^1 = 60$; $p^2 = 60$;
- $b^1 = 250$; $b^2 = 250$;
- $h^1 = 3$; $h^2 = 0$.

This farmer story was devised to illustrate concepts throughout in a setting that is as simple as possible; we will treat a more realistic setting later on. ◇

7.3 Observation possibilities and games

In this section, we describe how the cooperation between players may take place. Given any stochastic dynamic inventory pooling situation, we will formulate various associated games that differ in the degree to which coalitions can observe the demand realizations of players outside the coalition. Before defining our games, we need to introduce additional concepts and notation.

Consider any stochastic dynamic inventory pooling situation γ. For any coalition $M \in 2_-^N$, period $t \in \{T_{\text{start}} + 1, \ldots, T_{\text{end}}\}$, and scenario $\omega \in \Omega$, we systematically denote that coalition's previous demand realizations $(D_i^\tau(\omega))_{i \in M, \tau \in \{T_{\text{start}}, \ldots, t-1\}}$ more succinctly as $\mathcal{D}_M^t(\omega)$.

Additionally, we denote $\mathcal{D}_M^{T_{\text{start}}}(\omega) = \emptyset$. Using this notation, we define for any coalition $M \in 2_-^N$ and period $t \in T$ the set $\mathscr{D}_M^t = \{\mathcal{D}_M^t(\omega) \mid \omega \in \Omega\}$. This set contains all possible demand realizations for coalition M in periods $T_{\text{start}}, \ldots, t-1$.[48] Note that $\mathscr{D}_M^{T_{\text{start}}} = \{\emptyset\}$.

We call any function $\mathcal{O} : 2_-^N \to 2_-^N$ with $M \subseteq \mathcal{O}(M)$ for all $M \in 2_-^N$ an *observation possibility*. Such a function $\mathcal{O}$ describes the extent to which coalitions can observe the demand realizations of other players; $\mathcal{O}(M) = L$ should be interpreted as saying that coalition M can always observe the demand realizations of all players in the set L, where by assumption $M \subseteq L$. The observation possibilities that we are most interested in are the one where everything can be observed, i.e., $\mathcal{O}(M) = N$ for all $M \in 2_-^N$, and the one where nothing can be observed outside any coalition, i.e., $\mathcal{O}(M) = M$ for all $M \in 2_-^N$. Monotonicity in $\mathcal{O}$, i.e., the property that $\mathcal{O}(M) \subseteq \mathcal{O}(L)$ for all $M, L \in 2_-^N$ with $M \subseteq L$, is a seemingly natural property, but we do not impose it a priori. Irrespective of $\mathcal{O}$, the sample space Ω and probability measure $\mathbb{P}$ is common knowledge among players.

A *policy* $\pi = (\pi^t)_{t \in T}$ for a coalition $M \in 2_-^N$ under observation possibility $\mathcal{O}$ is a collection of functions $\pi^t : \mathscr{D}_{\mathcal{O}(M)}^t \to \mathbb{R}_+$ for all periods $t \in T$. The value $\pi^t(\mathcal{D})$ for any $\mathcal{D} \in \mathscr{D}_{\mathcal{O}(M)}^t$ signifies the number of products that coalition M will order in period t, given the previously observed demands $\mathcal{D}$ for all players in the set $\mathcal{O}(M)$.[49] Let $\Pi_M^{\gamma,\mathcal{O}}$ denote the set of all such policies for a coalition $M \in 2_-^N$ under observation possibility $\mathcal{O}$ in situation γ.

Example 7.2. Recall the farmers Alice and Bob. As an illustration of the demand observation sets, we have $\mathscr{D}_{\{A\}}^2 = \{(1), (0)\}$, $\mathscr{D}_{\{B\}}^2 = \{(0)\}$, and $\mathscr{D}_{\{A,B\}}^2 = \{(1,0), (0,0)\}$.

Suppose that the farmers decide not to form a coalition. Then, a question arises: when Bob makes his independent ordering decision on July 31, does he know whether the storm has hit Alice on July 1 or not? Knowledge of the "demand" realization of Alice gives full information because the storm will hit Bob if and only if it did not hit Alice. However, if Bob was unable to observe what happened to Alice, then all he knows is that there is a probability of $\mathbb{P}(\omega_2) = \frac{1}{3}$ that the storm will hit him on the next day.

At first glance, it seems reasonable to assume that the farmers would be able to observe each other's "demand" realization, simply by tuning in to the Weather Channel or by driving by the other farmer's field. Under this assumption, captured by the observation possibility $\mathcal{O}$ with $\mathcal{O}(M) = \{A, B\}$ for all coalitions M, the (unique) optimal policy $\pi_{\{B\}}$ for Bob alone is described by $\pi_{\{B\}}^1(\emptyset) = 0$, $\pi_{\{B\}}^2((1,0)) = 0$, and $\pi_{\{B\}}^2((0,0)) = 1$.

If, however, Bob would live exactly as people did in 1900, without a car or electricity—Ward (2008) chronicles his interesting experience of doing just that in today's world—then he

[48] This set may contain demand histories under scenarios that occurs with zero probability. We allow this because it will simplify notation later on.

[49] Coalition M's ordering decision in period t is also influenced by its previous ordering decisions and by the current inventory position, of course, but that dependence is captured implicitly by including the previously observed demands in the domain of π^t and by letting coalitions coordinate $(\pi^t)_{t \in T}$, i.e., all decisions in all periods simultaneously.

would have no way of knowing what had happened to Alice. If Alice in the meantime would enjoy the luxuries of the present century, then this would correspond to $\mathcal{O}(\{B\}) = \{B\}$, $\mathcal{O}(\{A\}) = \{A, B\}$, and $\mathcal{O}(\{A, B\}) = \{A, B\}$. In this case, the (unique) optimal policy $\pi_{\{B\}}$ for Bob alone is described by $\pi^1_{\{B\}}(\emptyset) = 0$ and $\pi^2_{\{B\}}((0)) = 1$. Indeed, the expected total costs when ordering the service parts on August 1, p^2, is lower than the expected total costs when ordering nothing, $\mathbb{P}(\omega_2)b^2$, and ordering anything on July 1 is not optimal because it leads to unnecessary holding costs.

The two other ordering possibility functions can be analyzed in similar fashion, but we disregard them for brevity. ◇

We will refer to any pair $(\gamma, \mathcal{O})$, where γ is a stochastic dynamic inventory pooling situation and $\mathcal{O}$ a corresponding observation possibility, as an *environment*. We next describe what happens under a specific policy.

Consider environment $(\gamma, \mathcal{O})$, coalition $M \in 2^N_-$, scenario $\omega \in \Omega$, and policy $\pi \in \Pi^{\gamma,\mathcal{O}}_M$. For any period $t \in T$, the net inventory at the end of period t is given by

$$I^t_M(\pi, \omega) = \sum_{i \in M} I^0_i + \sum_{\tau \in \{T_{\text{start}}, \ldots, t\}} \left(\pi^\tau(\mathcal{D}^\tau_{\mathcal{O}(M)}(\omega)) - \sum_{i \in M} D^\tau_i(\omega) \right),$$

with a negative value of $I^t_M(\pi, \omega)$ signifying backorders. The total *realized* costs in period $t \in T$ for coalition M under scenario ω and policy π is given by

$$K^{\gamma,t}_M(\pi, \omega) = p^t \pi^t(\mathcal{D}^t_{\mathcal{O}(M)}(\omega)) + h^t(I^t_M(\pi, \omega))^+ + b^t(I^t_M(\pi, \omega))^-.$$

where $x^+ = \max\{0, x\}$ and $x^- = \max\{0, -x\}$ for any $x \in \mathbb{R}$. The total *realized* costs over the time horizon for coalition M under scenario ω and policy π is given by

$$K^\gamma_M(\pi, \omega) = \sum_{t \in T} K^{\gamma,t}_M(\pi, \omega).$$

Finally, the total *expected* costs over the time horizon for coalition M under policy π is given by

$$K^\gamma_M(\pi) = \sum_{\omega \in \Omega} \mathbb{P}(\omega) K^\gamma_M(\pi, \omega).$$

This number is precisely what coalition M aims to minimize by picking an optimal policy. Using the notation introduced above, we are now ready to define our games, one per environment.

Definition 7.2. Consider any environment $(\gamma, \mathcal{O})$. The game (N, c^γ) with characteristic costs

$$c^{\gamma,\mathcal{O}}(M) = \min_{\pi \in \Pi^{\gamma,\mathcal{O}}_M} K^\gamma_M(\pi) \tag{7.1}$$

for all $M \in 2^N_-$ is called the associated *stochastic dynamic inventory pooling game*.

M	$c^{\gamma,\mathcal{O}}(M)$ when $\mathcal{O}(\{B\}) = \{A, B\}$	$c^{\gamma,\mathcal{O}}(M)$ when $\mathcal{O}(\{B\}) = \{B\}$
{A}	61	61
{B}	20	60
{A,B}	61	61

Table 7.1: *The collection of games corresponding to the situation of Example 7.3.*

An optimal policy indeed exists by the combination of two facts. First, for every scenario, the realized cost is continuous in any period's order quantity; hence, the same holds for the expected costs. Second, because $p^t > 0$ for any $t \in T$, we can restrict attention to orders in a compact, non-empty set; indeed, above a certain order quantity in any period, the production cost alone already becomes larger than the total remaining costs-to-go under an order of, say, zero. This implies that the game is well-defined.

We illustrate (collections of) stochastic dynamic inventory pooling games in the following example.

Example 7.3. Recall the farmers Alice and Bob. First, consider coalition $\{B\}$, for whom the optimal policies in different observation possibilities were already described in Example 7.2. When $\mathcal{O}(\{B\}) = \{A, B\}$, Bob faces expected costs $\mathbb{P}(\omega_2)p^2 = 20$. When $\mathcal{O}(\{B\}) = \{B\}$, Bob faces expected costs $p^2 = 60$.

For coalition $\{A\}$, the expected costs are unaffected by the observation possibility, as Alice already learns the true scenario on July 1. It is easy to verify that Alice should order 1 item on June 30, with corresponding expected costs $p^1 + \mathbb{P}(\omega_2)h^1 = 61$.

For the grand coalition $\{A, B\}$, which always fully observes all demand realizations, it is easy to verify that ordering 1 item on June 30 is optimal, with corresponding expected costs $p^1 + \mathbb{P}(\omega_2)h^1 = 61$.

The games $(N, c^{\gamma,\mathcal{O}})$ for each observation possibility $\mathcal{O}$ are described in Table 7.1. Note that all of these games are balanced. Also, note that additional information is beneficial: "enlarging" the observation possibility for Bob reduces his expected costs, as it should. $\diamond$

This example dealt with demands that were correlated across periods and across players. If demands of a player are independent of earlier demands of other players, then preceding demand observations of other players cannot inform ordering decisions. If that independence holds for all players and all periods, then each observation possibility will obviously lead to the same game.

7.4 Balancedness

First and foremost, we are interested in the existence of a stable cost allocation for stochastic dynamic inventory pooling games. If these games would always be concave, then this existence

would be immediate. However, newsvendor games need not be concave.[50] Since any stochastic dynamic inventory pooling game with a single period is a newsvendor game, concavity is not guaranteed. Instead, we follow another approach: we make use of the Bondareva-Shapley theorem. The gist of the idea behind the proof is as follows. First, for every balanced map, we construct a *feasible* policy for the grand coalition out of the corresponding balanced combination of *optimal* policies of subcoalitions. Then, we show that the costs of this constructed policy for the grand coalition does not exceed the corresponding balanced combination of the optimal costs for the subcoalitions. Consequently, it is possible to allocate the expected costs of an *optimal* policy for the grand coalition in a stable way.

The construction is straightforward. Consider an arbitrary environment $(\gamma, \mathcal{O})$, an arbitrarily balanced map $\kappa : 2^N_- \to [0,1]$, and an arbitrary optimal policy $\pi_M \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_M} K^\gamma_M(\pi)$ for each coalition $M \in 2^N_-$. Using these policies, we construct a feasible policy π_κ for the grand coalition, which attempts to mirror the orders that subcoalitions would place. It is defined by $\pi^t_\kappa(\mathcal{D}^t_N(\omega)) = \sum_{M \in 2^N_-} \kappa(M) \pi^t_M(\mathcal{D}^t_{\mathcal{O}(M)}(\omega))$ for all periods $t \in T$ and all scenarios $\omega \in \Omega$. Note that this indeed represents a policy, for two reasons. First, for all $t \in T$ and all $\omega \in \Omega$, we stay in the domain of π^t_κ, i.e., $\mathcal{D}^t_N(\omega) \in \mathscr{D}^t_{\mathcal{O}(N)} = \mathscr{D}^t_N$. Second, $\mathcal{D}^t_M(\omega)$ for each $M \in 2^N_-$ is fully determined by $\mathcal{D}^t_N(\omega)$. We illustrate the construction of this policy in the following example.

Example 7.4. Recall the farmers Alice and Bob. Suppose that $\mathcal{O}(\{A\}) = \{A, B\}$, $\mathcal{O}(\{B\}) = \{B\}$, and $\mathcal{O}(\{A, B\}) = \{A, B\}$. In this case, let $\pi_{\{A\}}$ and $\pi_{\{B\}}$ be the optimal policy for Alice and Bob, respectively, when acting separately. Recall that $\pi^1_{\{A\}}(\emptyset) = 1$, $\pi^2_{\{A\}}((1,0)) = 0$, and $\pi^2_{\{A\}}((0,0)) = 0$ for Alice, while $\pi^1_{\{B\}}(\emptyset) = 0$ and $\pi^2_{\{B\}}((0)) = 1$ for Bob. For the balanced map $\kappa : 2^N_- \to [0,1]$ described by $\kappa(\{A\}) = \kappa(\{B\}) = 1$ and $\kappa(\{A, B\}) = 0$, the policy π_κ is constructed by

- $\pi^1_\kappa(\emptyset) = \pi^1_{\{A\}}(\emptyset) + \pi^1_{\{B\}}(\emptyset) = 1$;
- $\pi^2_\kappa((1,0)) = \pi^2_{\{A\}}((1,0)) + \pi^2_{\{B\}}((0)) = 1$;
- $\pi^2_\kappa((0,0)) = \pi^2_{\{A\}}((0,0)) + \pi^2_{\{B\}}((0)) = 1$.

The expected costs of this policy would be 121. This matches the balanced combination of the optimal costs for the subcoalitions, i.e., $\sum_{M \in 2^N_-} \kappa(M) c^{\gamma,\mathcal{O}}(M) = 1 \cdot 61 + 1 \cdot 60 = 121$. $\diamond$

In the process of proving that stochastic dynamic inventory pooling games are balanced, we will use the following lemma. This lemma states that, for every demand realization, the balanced combinations of the on-hand inventories (resp. backlogs) associated with the optimal policies for subcoalitions are at least as large as the number of on-hand inventories (resp. backlogs) under the constructed policy π_κ.

[50] This follows from Özen et al. (2011). They consider newsvendor games in profit rather than cost terms, but their results carry over straightforwardly.

Lemma 7.1. *Consider an environment $(\gamma, \mathcal{O})$, a balanced map $\kappa : 2^N_- \to [0,1]$, and an optimal policy $\pi_M \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_M} K^\gamma_M(\pi)$ for each $M \in 2^N_-$. Let $t \in T$ and let $\omega \in \Omega$.*
(i) $\sum_{M \in 2^N_-} \kappa(M)(I^t_M(\pi_M, \omega))^+ \geq (I^t_N(\pi_\kappa, \omega))^+$.
(ii) $\sum_{M \in 2^N_-} \kappa(M)(I^t_M(\pi_M, \omega))^- \geq (I^t_N(\pi_\kappa, \omega))^-$.

Proof. We will split up each of the two parts of the lemma into two mutually exclusive cases that are exhaustive of all possibilities, and subsequently show that each desired inequality holds in both cases.

Part (i) - Case $I^t_N(\pi_\kappa, \omega) > 0$. Then,

$$\begin{aligned}\sum_{M \in 2^N_-} \kappa(M)(I^t_M(\pi_M, \omega))^+ &\geq \sum_{M \in 2^N_-} \kappa(M) I^t_M(\pi_M, \omega) \\ &= \sum_{M \in 2^N_-} \kappa(M) \left(\sum_{\tau \in \{T_{\text{start}}, \ldots, t\}} \left(\pi^\tau_M(\mathcal{D}^\tau_M(\omega)) - \sum_{i \in M} D^\tau_i(\omega) \right) \right) \\ &= \sum_{\tau \in \{T_{\text{start}}, \ldots, t\}} \left(\pi^\tau_\kappa(\mathcal{D}^\tau_N(\omega)) - \sum_{i \in N} D^\tau_i(\omega) \right) \\ &= I^t_N(\pi_\kappa) = (I^t_N(\pi_\kappa))^+.\end{aligned}$$

The second equality holds by definition of π_κ and because κ, as any balanced map, satisfies $\sum_{M \in 2^N_-} \kappa(M) \sum_{i \in M} f(i) = \sum_{i \in N} f(i)$ for all functions $f : N \to \mathbb{R}$.

Part (i) - Case $I^t_N(\pi_\kappa, \omega) \leq 0$. Then,

$$\sum_{M \in 2^N_-} \kappa(M)(I^t_M(\pi_M, \omega))^+ \geq 0 = (I^t_N(\pi_\kappa, \omega))^+.$$

Part (ii) - Case $I^t_N(\pi_\kappa, \omega) < 0$. Then,

$$\begin{aligned}\sum_{M \in 2^N_-} \kappa(M)(I^t_M(\pi_M, \omega))^- &\geq - \sum_{M \in 2^N_-} \kappa(M) I^t_M(\pi_M, \omega) \\ &= -I^t_N(\pi_\kappa) = (I^t_N(\pi_\kappa))^-,\end{aligned}$$

where the first equality holds as shown in the first case of Part (i) above.

Part (ii) - Case $I^t_N(\pi_\kappa, \omega) \geq 0$. Then,

$$\sum_{M \in 2^N_-} \kappa(M)(I^t_M(\pi_M, \omega))^- \geq 0 = (I^t_N(\pi_\kappa, \omega))^-.$$

This completes the proof. □

The following theorem implies that every stochastic dynamic inventory pooling game has a nonempty core.

Theorem 7.2. *Stochastic dynamic inventory pooling games are balanced.*

Proof. Consider an environment $(\gamma, \mathcal{O})$, a balanced map $\kappa : 2^N_- \to [0,1]$, and an optimal policy $\pi_M \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_M} K^\gamma_M(\pi)$ for each $M \in 2^N_-$. Then, we find

$$\begin{aligned}
&\sum_{M \in 2^N_-} \kappa(M) c^{\gamma,\mathcal{O}}(M) \\
&\quad = \sum_{M \in 2^N_-} \kappa(M) K^\gamma_M(\pi_M) \\
&\quad = \sum_{M \in 2^N_-} \kappa(M) \sum_{\omega \in \Omega} \mathbb{P}(\omega) \sum_{t \in T} \left(p^t \pi^t_M(\mathcal{D}^t_{\mathcal{O}(M)}(\omega)) + h^t(I^t_M(\pi_M, \omega))^+ + b^t(I^t_M(\pi_M, \omega))^- \right) \\
&\quad = \sum_{\omega \in \Omega} \mathbb{P}(\omega) \sum_{t \in T} \left(p^t \pi^t_\kappa(\mathcal{D}^t_N(\omega)) + \sum_{M \in 2^N_-} \kappa(M) \big(h^t(I^t_M(\pi_M, \omega))^+ + b^t(I^t_M(\pi_M, \omega))^- \big) \right) \\
&\quad \geq \sum_{\omega \in \Omega} \mathbb{P}(\omega) \sum_{t \in T} \big(p^t \pi^t_\kappa(\mathcal{D}^t_N(\omega)) + h^t(I^t_N(\pi_M, \omega))^+ + b^t(I^t_N(\pi_M, \omega))^- \big) \\
&\quad = K^\gamma_N(\pi_\kappa) \\
&\quad \geq K^\gamma_N(\pi_N) = c^{\gamma,\mathcal{O}}(N).
\end{aligned}$$

The second equality holds by definition of K^γ_M. The third equality holds by interchanging terms, where we use that by construction of π_κ, its production costs always match the balanced combination of production costs for the subcoalitions. The first inequality holds by Lemma 7.1. The subsequent equality holds by definition of K^γ_N. The final inequality holds because π_κ is merely a feasible policy for the grand coalition, whereas π_N is an optimal (cost-minimizing) policy for the grand coalition. We conclude that the game $(N, c^{\gamma,\mathcal{O}})$ is balanced. □

An observation possibility is called monotonic if $\mathcal{O}(M) \subseteq \mathcal{O}(L)$ for all $M, L \in 2^N_-$ with $M \subseteq L$. For an environment $(\gamma, \mathcal{O})$ with a monotonic observation possibility, it holds that every sub-game of $(N, c^{\gamma,\mathcal{O}})$ is a game associated with an environment with a monotonic observation possibility as well. Hence, we immediately obtain the following corollary to Theorem 7.2.

Corollary 7.3. *Let $(\gamma, \mathcal{O})$ be an environment with a monotonic observation possibility. The associated stochastic dynamic inventory pooling game $(N, c^{\gamma,\mathcal{O}})$ is totally balanced.*

If the observation possibility is *not* monotonic, then the game is still balanced (by Theorem 7.2), but not necessarily totally. This is illustrated in the following example.

Example 7.5. Recall the farmers Alice and Bob. We now add a third farmer, Claire, who faces a stochastic demand in June that completely reveals the true state of the world: if Claire has demand of exactly 1 in June, which occurs with probability $\frac{2}{3}$, then the storm will hit Alice; otherwise, it will hit Bob. Suppose that $\mathcal{O}(\{A\}) = \{A, C\}$, $\mathcal{O}(\{B\}) = \{B, C\}$, and $\mathcal{O}(\{A, B\}) = \{A, B\}$. So, if Alice and Bob act independently, then each can observe Claire's

demand realization. However, once Alice and Bob form a coalition, Claire will close her fence and won't reveal anything.

Since coalitions $\{A\}$ and $\{B\}$ can now use perfect information in their ordering decisions, their characteristic costs become $\frac{2}{3}\cdot 60 = 40$ for $\{A\}$ and $\frac{1}{3}\cdot 60 = 20$ for $\{B\}$. At the same time, the characteristic cost for $\{A, B\}$ remains 61—the same as in Example 7.3. Since $20+40 < 61$, the sub-game for $\{A, B\}$ is not balanced. ◇

7.5 Cost allocation

The previous section taught us that stochastic dynamic inventory pooling games are balanced. Hence, we can propose the nucleolus, which is guaranteed to be in the core of any balanced game, as a solution to the cost allocation problem. However, calculating the nucleolus is not easy, especially when the number of players is large. In previous chapters, we found that a simpler proportional rule was frequently stable, too. In this section, we investigate whether or not such a rule would "work" for stochastic dynamic inventory pooling games as well. We start with the definition: the *demand-proportional rule* $\mathscr{P}$ is defined by allocating

$$\mathscr{P}_i(\gamma, \mathcal{O}) = \frac{\sum_{t\in T} \mathbb{E}[D_i^t]}{\sum_{j\in N}\sum_{t\in T} \mathbb{E}[D_j^t]} c^{\gamma,\mathcal{O}}(N)$$

to player $i \in N$ in environment $(\gamma, \mathcal{O})$. This rule assigns more costs to players with higher expected total demands over the horizon, but disregards important information such as the variance or specific timing of demands. The following example shows that the demand-proportional rule does not accomplish stable allocations in general.

Example 7.6. Recall the farmers Alice and Bob and the games described in Table 7.1 on page 161. The nucleolus is given by $(51, 10)$ if $\mathcal{O}(\{B\}) = \{A, B\}$ and by $(31, 30)$ if $\mathcal{O}(\{B\}) = \{B\}$. So, if expected costs are always split according to the nucleolus, then Alice has an economic incentive to convince Bob to live exactly as people did in 1900. Also, note that even though the allocation $(51, 10)$ was derived under the assumption that $\mathcal{O}(\{B\}) = \{A, B\}$, this allocation is still in the core of the game corresponding to the observation possibility with $\mathcal{O}(\{B\}) = \{B\}$.

The demand-proportional rule assigns $(40\frac{2}{3}, 20\frac{1}{3})$, irrespective of the observation possibility. This allocation is not in the core when $\mathcal{O}(\{B\}) = \{A, B\}$. ◇

In this example, a stable allocation remained stable when the observation possibility was "narrowed." The intuition is that subcoalitions' costs do not decrease upon such a "narrowing," as we observed in Example 7.3. The following theorem, which relates allocations for different environments to each other, formalizes this in general.

Theorem 7.4. *Consider two environments $(\gamma, \mathcal{O})$ and $(\gamma, \tilde{\mathcal{O}})$ with $\mathcal{O}(M) \subseteq \tilde{\mathcal{O}}(M)$ for all $M \in 2^N_-$. Then, $\mathcal{C}(N, c^{\gamma,\mathcal{O}}) \supseteq \mathcal{C}(N, c^{\gamma,\tilde{\mathcal{O}}})$.*

Proof. Let $x \in \mathcal{C}(N, c^{\gamma,\tilde{\mathcal{O}}})$. This represents an efficient allocation for the game $(N, c^{\gamma,\mathcal{O}})$ because $\sum_{i\in N} x_i = c^{\gamma,\tilde{\mathcal{O}}}(N) = c^{\gamma,\mathcal{O}}(N)$, where the first equality holds because $x \in \mathcal{C}(N, c^{\gamma,\tilde{\mathcal{O}}})$ and the second equality holds because the grand coalition's cost is the same in both environments as $\mathcal{O}(N) = \tilde{\mathcal{O}}(N) = N$.

The allocation x is stable for the game $(N, c^{\gamma,\mathcal{O}})$ because $\sum_{i\in M} x_i \leq c^{\gamma,\tilde{\mathcal{O}}}(M) \leq c^{\gamma,\mathcal{O}}(M)$, where the first inequality holds because $x \in \mathcal{C}(N, c^{\gamma,\tilde{\mathcal{O}}})$ and the second inequality holds because an optimal policy $\pi_M \in \operatorname{argmin}_{\pi\in\Pi_M^{\gamma,\mathcal{O}}} K_M^{\gamma}(\pi)$ merely induces a *feasible* policy $\tilde{\pi}_M \in \Pi_M^{\gamma,\tilde{\mathcal{O}}}$ which does the same as π_M by "disregarding" demand information from players in $\tilde{\mathcal{O}}(M) \setminus \mathcal{O}(M)$. This completes the proof. □

The following theorem presents a sufficient condition for stability of the demand-proportional allocations. The proof is based on single-attribute games, similar to the approach described in Section 6.5.2 for spare parts games with optimized base stock levels.

Theorem 7.5. *Let $(\gamma, \mathcal{O})$ be an environment, let $a \in \mathbb{R}_+^N$, and let $G \in \mathbb{R}_+^T$. If $I_i^0 = 0$ for all $i \in N$ and $D_i^t \sim Poi(a_i G_t)$ for all $i \in N$ and all $t \in T$, with all D_i^t independent, then $\mathscr{P}_i(\gamma, \mathcal{O}) \in \mathcal{C}(N, c^{\gamma,\mathcal{O}})$.*

Proof. Fix the time horizon T, the vector G, and the cost parameters p, b, and h. Define the function $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ by setting $\tilde{K}(\lambda)$, for all $\lambda \in \mathbb{R}_+$, equal to the expected optimal costs faced by a stock point that starts with zero inventory, faces independent Poisson distributed demands with parameter λG_t in any period $t \in T$, while incurring costs cf. the inventory model as described in Section 7.2 according to the cost vectors p, b, and h.

Our game $(N, c^{\gamma,\mathcal{O}})$ is a single-attribute game embedded in $\tilde{K}$. Indeed, $c^{\gamma,\mathcal{O}}(M) = \tilde{K}(\sum_{i\in M} a_i)$ because the sum of independent Poisson distributed random variables is itself Poisson distributed according to the sum of the parameters and because D_i^t is independent of D_j^τ for all $i, j \in N$ and $t, \tau \in T$. For analogous reasons, all single-attribute games embedded in $\tilde{K}$ are stochastic dynamic inventory pooling games and, hence, have a non-empty core by Theorem 7.2. This implies that $\tilde{K}$ is elastic by Part (*iii*) of Theorem 2.5. Consequently, $\mathscr{P}_i(\gamma, \mathcal{O}) \in \mathcal{C}(N, c^{\gamma,\mathcal{O}})$ by Part (*i*) of Theorem 2.5. □

The following example provides an illustration of this theorem.

Example 7.7. Imagine three independent wind turbine farms, named player 1, player 2, and player 3. On January 1, 2014, player 1 owns 5 wind turbines, player 2 owns 10, and player 3 owns 15. Due to a governmental renewable energy decree, each player will have to double their installed base on January 1, 2015. Each wind turbine consists of exactly one critical component, whose lifetime is exponentially distributed with mean 20 years.

A failure must be resolved by either replacing the failed component with an available spare part or by sending a team of specialized repairmen who fix the component on-site. The latter option, which comes at a cost of 1000 euro, can be considered as an emergency procedure because it doesn't require a spare part.

The manufacturer of the spare part offers only two ordering possibilities: an initial production run on January 1, 2014, and an alternative run on January 1, 2015. The unit purchase price is 200 euro in the initial production run and 300 euro in the alternative run. Holding and warehousing costs are negligible. All wind turbines will be replaced by newer models on January 1, 2016; accordingly, the spare parts are only useful in 2014 and 2015. All cost and demand parameters are common knowledge.

This situation may be modeled as an environment $(\gamma, \mathcal{O})$, with

- $N = \{1, 2\}$; $T = \{1, 2\}$;
- $D_1^1 \sim Poi(\frac{1}{4})$; $D_1^2 \sim Poi(\frac{1}{2})$;
- $D_2^1 \sim Poi(\frac{1}{2})$; $D_2^2 \sim Poi(1)$;
- $D_3^1 \sim Poi(\frac{3}{4})$; $D_3^2 \sim Poi(\frac{3}{2})$;
- $I_1^0 = I_2^0 = 0$;
- $p^1 = 200$; $p^2 = 300$;
- $b^1 = 700$; $b^2 = 1000$;
- $h^1 = 0$; $h^2 = 0$.

The observation possibility $\mathcal{O}$ is arbitrary because demands are independent between periods and between players. Note that we transformed the cost for an emergency procedures into a cost for a backlog. This modeling trick is only possible because the lead time is zero and the purchase cost is lower than the backlogging cost; hence, in any optimal policy, a backlog is always immediately resolved and only leads to a one-time cost: $b^1 + p^2$ in case of a stock-out in period 1 and b^2 in case of a stock-out in period 2.

Any coalition has to pick an ordering policy. The standard approach for finding the best decisions in a sequential decision problem such as this one is known as stochastic dynamic programming. We illustrate the approach for coalition $\{3\}$. As holding costs are zero, we focus on backlogging and production costs only. We will round all values to whole euros.

We start the analysis in period 2. Let $C^2(S)$ denote the expected backlogging costs for $\{3\}$ in period 2 under a post-order inventory of $S \in \mathbb{N}_0$ parts, i.e., $C^2(S) = b^2 \cdot \mathbb{E}[(S - D_3^2)^-]$. Then, we find that $C^2(0) = 1500$, $C^2(1) = 723$, and $C^2(2) = 281$. Combining this with the ordering cost of $p^2 = 300$, we conclude that ordering up to 2 parts is optimal. We denote this optimal order-up-to level by $S^*_{\{3\}}$.

We use this optimal policy for period 2 to determine the optimal order in period 1. We let $C^1(q)$ denote the expected ordering and backlogging costs for $\{3\}$ over the horizon when starting with an order of $q \in \mathbb{N}_0$ parts in period 1, i.e., $C^1(q) = p^1 q + \mathbb{E}\left[b^1(q - D_3^1)^- + p^2\big(S^*_{\{3\}} - (q - D_3^1)\big)^+ + C^2(S^*_{\{3\}})\right]$. Enumerating, we find that $C^1(0) =$

M	q_M^*	S_M^*	$c^{\gamma,\mathcal{O}}(M)$	$\sum_{i\in M}\mathscr{P}_i(\gamma,\mathcal{O})$
{1}	1	1	402	247
{2}	1	2	651	494
{3}	3	2	863	741
{1,2}	3	2	863	741
{1,3}	4	3	1079	988
{2,3}	5	3	1282	1235
{1,2,3}	6	4	1482	1482

Table 7.2: *For every coalition M, The optimal order quantities q_M^* in period 1, the optimal order-up-to-levels S_M^* in period 2, the characteristic costs $c^{\gamma,\mathcal{O}}(M)$, and the total costs to be paid under the allocation $\mathscr{P}(\gamma,\mathcal{O})$ (rounded to whole euros) in Example 7.7.*

1631, $C^1(1) = 1161$, $C^1(2) = 940$, $C^1(3) = 863$, $C^1(4) = 907$, and $C^1(q) > 1000$ for all $q \geq 5$. Hence, an order of 3 parts is optimal.

In similar fashion, we find that the optimal policy for any coalition $M \in 2^N_-$ is to order a (unique) amount q_M^* in period 1 and then order up to S_M^* in period 2. These optimal numbers[51] and each coalition's costs in the corresponding stochastic dynamic inventory pooling game $(N, c^{\gamma,\mathcal{O}})$ are described in Table 7.2.

The demand-proportional rule $\mathscr{P}$ assigns $(247, 494, 741)$, rounding to whole euros. Table 7.2 shows that this allocation is stable. Although this allocation is in expected terms, it can be implemented for any cost realization by assigning a fraction $\mathscr{P}_i(\gamma,\mathcal{O})/\sum_{j\in N}\mathscr{P}_j(\gamma,\mathcal{O})$ of the *realized* costs in any period to player $i \in N$. ◇

7.6 Strategically joining late

In this section, we study what might happen if a player doesn't show up immediately, but arrives belatedly after some periods of demand have already been realized. In particular, would it be beneficial for an opportunistic player to stay solitary for a while and only try to join a pooling group after facing a demand spike? Such a strategic late arrival is analogous to buying medical insurance only after getting sick. Note that this is about strategic, not cooperative, behavior: We can contract the future, but not the past.

So, strategic late arrival is a risk that could happen, but it requires extensive, comprehensive analysis if considered in full generality. In this section, we take a preliminary, explorative first step to analyze this risk in a restricted setting. We make three main assumptions throughout this section. First, there are only two players, i.e., $|N| = 2$. Second,

[51] More formally, the optimal policy π_M for any coalition $M \in 2^N_-$ is described by $\pi_M^1(\emptyset) = q_M^*$ and $\pi_M^2((d_i^1)_{i\in\mathcal{O}(M)}) = \left(S_M^* - (q_M^* - \sum_{i\in M} d_i^1)\right)^+$ for all $(d_i^1)_{i\in\mathcal{O}(M)} \in \mathscr{D}^2_{\mathcal{O}}(M)$.

the benefits of collaboration will always be split equally, i.e., the Shapley value or nucleolus is used. Third, all demand realizations are fully observable, i.e., $\mathcal{O}(M) = N$ for all coalitions M. We discuss the impact of these assumptions at the end of this section.

To determine how costs will be shared when a player shows up late, say in period t, we aim to construct a stochastic dynamic inventory pooling game corresponding to the situation that captures the relevant setting from periods t onwards.

Definition 7.3. Consider any stochastic dynamic inventory pooling situation $\gamma = (N, \Omega, \mathbb{P}, I^0, T, D, p, b, h)$, period $t \in T$, demand observations $\mathcal{D} = (d_i^\tau)_{i \in N, \tau \in \{T_{\text{start}}, \ldots, t-1\}} \in \mathscr{D}_N^t$, and a net inventory vector $\hat{I}^0 \in \mathbb{R}^N$. (The origination of $\hat{I}^0$ is arbitrary.) The corresponding *sub-situation* $\hat{\gamma} = (N, \Omega, \hat{\mathbb{P}}, \hat{I}^0, \hat{T}, \hat{D}, \hat{p}, \hat{b}, \hat{h})$ is the stochastic dynamic inventory pooling situation for which:

- $\hat{\mathbb{P}}$ is obtained via straightforward Bayesian updating: using the set of possible scenarios $\Omega^{\mathcal{D}} = \{\omega \in \Omega \mid D_i^\tau(\omega) = d_i^\tau \text{ for all } \tau \in \{T_{\text{start}}, \ldots, t-1\} \text{ and all } i \in N\}$, we get

$$\hat{\mathbb{P}}(\omega) = \begin{cases} \dfrac{\mathbb{P}(\omega)}{\sum_{\omega' \in \Omega^{\mathcal{D}}} \mathbb{P}(\omega')} & \text{if } \omega \in \Omega^{\mathcal{D}}; \\ 0 & \text{otherwise.} \end{cases} \tag{7.2}$$

- $\hat{I}^0$ is the net inventory vector that was given.
- $\hat{T} = \{\hat{T}_{\text{start}}, \hat{T}_{\text{start}} + 1, \ldots, \hat{T}_{\text{end}}\}$ with $\hat{T}_{\text{start}} = t$ and $\hat{T}_{\text{end}} = T_{\text{end}}$ is the planning horizon that is restricted to start at period t.
- $\hat{D} = (D^\tau)_{\tau \in \hat{T}}$, $\hat{p} = (p^\tau)_{\tau \in \hat{T}}$, $\hat{b} = (b^\tau)_{\tau \in \hat{T}}$, and $\hat{h} = (h^\tau)_{\tau \in \hat{T}}$ are obtained by merely restricting the original parameters to start at period t.

The following example provides an illustration.

Example 7.8. Recall the farmers Alice and Bob. Suppose that $\mathcal{O}(M) = N$ for all $M \in 2^N_-$. Bob cunningly ponders the following strategy: say no to the coalition at the start, order $\pi^1_{\{B\}}(\emptyset) = 0$ items, but always ask to form the coalition at the beginning of period 2. Is this worthwhile? Suppose that Alice, in the meantime, orders $\pi^1_{\{A\}}(\emptyset) = 1$ for herself at the start of period 1. (These numbers $\pi^1_{\{B\}}$ and $\pi^1_{\{A\}}$, which represent our assumption on the policies followed by the farmers as long as they stay solitary, are chosen for illustrative purposes.)

If the demand observation in period 1 would be $(0, 0)$, then this would lead to the sub-situation described by

- $N = \{A, B\}$;
- $\Omega = \{\omega_1, \omega_2\}$;
- $\hat{\mathbb{P}}(\omega_1) = 0$; $\mathbb{P}(\omega_2) = 1$;

M	$c^{\hat{\gamma}(0,0),\mathcal{O}}(M)$	$c^{\hat{\gamma}(1,0),\mathcal{O}}(M)$
{A}	0	0
{B}	60	0
{A,B}	0	0

Table 7.3: *The games corresponding to the two sub-situations in Example 7.8.*

- $\hat{I}_1^0 = 1$; $\hat{I}_2^0 = 0$;
- $T = \{2\}$;
- $D_A^2(\omega_1) = 0$; $D_A^2(\omega_2) = 0$;
- $D_B^2(\omega_1) = 0$; $D_B^2(\omega_2) = 1$;
- $p^2 = 60$;
- $b^2 = 250$;
- $h^2 = 0$.

We will denote this sub-situation by $\hat{\gamma}(0,0)$. If the demand observation in period 1 would be $(1,0)$ instead, then this would lead to a similar sub-situation, denoted by $\hat{\gamma}(1,0)$, which has $\hat{I}_1^0 = 0$ and $\hat{\mathbb{P}}(\omega_1) = 1$.

The games corresponding to both sub-situations are listed in Table 7.3. Recall that we assume that the Shapley value Φ is always used for cost allocation. For the "original" game, described in Table 7.1 on page 161, $\Phi(N, c^{\gamma,\mathcal{O}}) = (51, 10)$. For the games associated with the sub-situations, $\Phi(N, c^{\hat{\gamma}(0,0),\mathcal{O}}) = (-30, 30)$ and $\Phi(N, c^{\hat{\gamma}(1,0),\mathcal{O}}) = (0, 0)$.

Although Alice might feel uneasy about Bob getting cold feet after rejecting the coalition at the start, it is still always in her interest to form the coalition at the beginning of period 2 because $\Phi_A(N, c^{\hat{\gamma}(0,0),\mathcal{O}}) \leq c^{\hat{\gamma}(0,0),\mathcal{O}}(\{A\})$ and, likewise, $\Phi_A(N, c^{\hat{\gamma}(1,0),\mathcal{O}}) \leq c^{\hat{\gamma}(1,0),\mathcal{O}}(\{A\})$.

Bob's opportunistic strategy leads to the same expected costs that he would have to pay if coalition $\{A, B\}$ had formed right away:

$$\mathbb{P}(\omega_1)[p^1\pi^1_{\{B\}}(\emptyset) + \Phi_B(N, c^{\hat{\gamma}(1,0),\mathcal{O}})] + \mathbb{P}(\omega_2)[p^1\pi^1_{\{B\}}(\emptyset) + \Phi_B(N, c^{\hat{\gamma}(0,0),\mathcal{O}})] = 10 = \Phi_B(N, c^{\gamma,\mathcal{O}}).$$

So, it is in Bob's best interests to join right from the start. ◇

In the preceding example, opportunistically joining late did not pay off. To analyze whether that holds in general, we formulate a non-cooperative game. In this game, a strategy of a player is the combination of a policy to follow while alone and, for every period, a collection of demand histories that entice the player to say "yes" to coalition formation. Once both players say "yes" at the beginning of a period, they join (i.e., form the coalition) and stay together for the remainder of the horizon.

Definition 7.4. Consider an environment $(\gamma, \mathcal{O})$ with $N = \{1,2\}$ and $\mathcal{O}(M) = N$ for all $M \in 2^N_-$. The corresponding *strategic join game* (N, S, f) is defined by

- $S_i = \left\{ \left(\pi_{\{i\}}, \left(\mathscr{D}^{t,\text{join}}_{\{i\}} \right)_{t \in T} \right) \middle| \pi_{\{i\}} \in \Pi^{\gamma,\mathcal{O}}_{\{i\}} \text{ and } \mathscr{D}^{t,\text{join}}_{\{i\}} \subseteq \mathscr{D}^t_{\mathcal{O}(\{i\})} \text{ for all } t \in T \right\}$ for all $i \in N$;
- $f_i(s) = \sum_{\omega \in \Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} K^{\gamma,t}_{\{i\}} \left(\pi_{\{i\}}, \omega \right) + \Phi_i \left(N, c^{\hat{\gamma}(\omega),\mathcal{O}} \right) \right]$ for all strategy profiles $s = \left(\pi_{\{i\}}, \mathscr{D}^{\text{join}}_{\{i\}} \right)_{i \in N} \in S$, where
  - For each $\omega \in \Omega$, $T_{\text{join}}(\omega)$ represents the first period $t \in T$ at which for both $i \in N$ it holds that $\mathcal{D}^t_N(\omega) \in \mathscr{D}^{t,\text{join}}_{\{i\}}$, if such a period exists; otherwise, $T_{\text{join}}(\omega) = T_{\text{end}} + 1$ and $\Phi(N, c^{\hat{\gamma}(\omega),\mathcal{O}}) = \mathbf{0}$.
  - For each $\omega \in \Omega$, $\hat{\gamma}(\omega)$ represents the sub-situation derived from γ by restricting the time horizon and the cost parameters to start at period $T_{\text{join}}(\omega)$, by updating the probability distributions given observations $\mathcal{D}^{T_{\text{join}}(\omega)-1}_N(\omega)$, and by setting the starting inventory of player $i \in N$ equal to $\hat{I}^0_i = I^{T_{\text{join}}(\omega)-1}_{\{i\}} \left(\pi_{\{i\}}, \omega \right)$.

The following lemma states that if one player goes for an opportunistic strategy while the other adheres to an optimal policy for as long as he stays solitary and is always ready to join in each period, then the expected costs incurred by the opportunistic player are no smaller than the expected costs that he would pay if the grand coalition had formed at the beginning of the time horizon.

Lemma 7.6. *Consider environment $(\gamma, \mathcal{O})$ with $N = \{1,2\}$ and $\mathcal{O}(M) = N$ for all $M \in 2^N_-$. Suppose that player 2 adheres to an arbitrary strategy $\left(\pi_{\{2\}}, \mathscr{D}^{join}_{\{2\}} \right)$ while player 1 adheres to a strategy $\left(\pi^*_{\{1\}}, \mathscr{D}^{join}_{\{1\}} \right)$ with $\pi^*_{\{1\}} \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_{\{1\}}} K^{\gamma}_{\{1\}}(\pi)$ and $\mathscr{D}^{t,join}_{\{1\}} = \mathscr{D}^t_N$ for all $t \in T$. Then, $f_2 \left(\pi^*_{\{1\}}, \mathscr{D}^{join}_{\{1\}}, \pi_{\{2\}}, \mathscr{D}^{join}_{\{2\}} \right) \geq \Phi_2(N, c^{\gamma,\mathcal{O}})$.*

Proof. We start by fixing a number of policies. For situation γ, let $\pi^*_{\{2\}} \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_{\{2\}}} K^{\gamma}_{\{2\}}(\pi)$ and let $\pi^*_N \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_N} K^{\gamma}_N(\pi)$. For sub-situation $\hat{\gamma}(\omega)$ under any $\omega \in \Omega$, let $\pi^{\omega,*}_{\{2\}} \in \operatorname{argmin}_{\pi \in \Pi^{\hat{\gamma}(\omega),\mathcal{O}}_{\{2\}}} K^{\hat{\gamma}(\omega)}_{\{2\}}(\pi)$ and let $\pi^{\omega,*}_N \in \operatorname{argmin}_{\pi \in \Pi^{\hat{\gamma}(\omega),\mathcal{O}}_N} K^{\hat{\gamma}(\omega)}_N(\pi)$.

Further, for sub-situation $\hat{\gamma}(\omega)$ under any $\omega \in \Omega$, let $\pi^{\omega,*}_{\{1\}}$ be the policy in $\Pi^{\hat{\gamma}(\omega),\mathcal{O}}_{\{1\}}$ that coincides with $\pi^*_{\{1\}}$ from periods $T_{\text{join}}(\omega)$ on. That is, given that we arrived in sub-situation $\hat{\gamma}(\omega)$ via demands $\mathcal{D}_N = (D^t_i(\omega))_{i \in N, t \in \{T_{\text{start}}, \ldots, T_{\text{join}}(\omega)-1\}}$ and given any possible subsequent demand observations $\mathcal{D}^t = (d^\tau_i)_{i \in N, \tau \in \{T_{\text{join}}(\omega), \ldots, t\}}$, we determine $\pi^{t,\omega,*}_{\{1\}}(\mathcal{D}^t) = \pi^{t,*}_{\{1\}}(\mathcal{D}, \mathcal{D}^t)$ in every period $t \in \hat{T}$. The policy $\pi^{\omega,*}_{\{1\}}$ is an element of $\operatorname{argmin}_{\pi \in \Pi^{\hat{\gamma}(\omega),\mathcal{O}}_{\{1\}}} K^{\hat{\gamma}(\omega)}_{\{1\}}(\pi)$ by the combination of $\mathcal{O}(\{1\}) = \{1,2\}$, optimality of $\pi^*_{\{1\}}$ within $\Pi^{\gamma,\mathcal{O}}_{\{1\}}$, and $\hat{I}^0_i = I^{T_{\text{join}}(\omega)-1}_{\{1\}}(\pi^*_{\{1\}}, \omega)$.

Next, for situation γ, let $\pi^{\text{feas}}_{\{2\}}$ be the policy in $\Pi^{\gamma,\mathcal{O}}_{\{2\}}$ that coincides with $\pi_{\{2\}}$ before joining, i.e., such that $\pi^{t,\text{feas}}_{\{2\}} = \pi^t_{\{2\}}$ for all $t \in \{1, \ldots, T_{\text{join}}(\omega) - 1\}$, and that always orders

the same amount as $\pi^{\omega,*}_{\{2\}}$ afterwards, i.e., such that $\pi^{t,\text{feas}}_{\{2\}}(\mathcal{D},\mathcal{D}^t) = \pi^{t,\omega,*}_{\{2\}}(\mathcal{D}^t)$ in every period $t \in \hat{T}$, using the same notation as before. Note that $\pi^{\text{feas}}_{\{2\}}$ indeed represents a policy because $\mathcal{O}(\{2\}) = \{1,2\}$.

Finally, for situation γ, let π^{feas}_N be the policy in $\Pi^{\gamma,\mathcal{O}}_N$ that always orders the same total amount as $\pi^*_{\{1\}}$ and $\pi_{\{2\}}$ before joining, i.e., such that $\pi^{t,\text{feas}}_N = \pi^{t,*}_{\{1\}} + \pi^t_{\{2\}}$ for all $t \in \{1,\dots,T_{\text{join}}(\omega)-1\}$, and that coincides with $\pi^{\omega,*}_N$ afterwards, i.e., such that $\pi^{t,\text{feas}}_N(\mathcal{D},\mathcal{D}^t) = \pi^{t,\omega,*}_N(\mathcal{D}^t)$ in every period $t \in \hat{T}$, using the same notation as before. We then obtain

$$
\begin{aligned}
&f_2\left(\pi^*_{\{1\}}, \mathscr{D}^{\text{join}}_{\{1\}}, \pi_{\{2\}}, \mathscr{D}^{\text{join}}_{\{2\}}\right) \\
&= \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) + \Phi_2(N, c^{\gamma(\omega),\mathcal{O}})\right] \\
&= \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) + \frac{c^{\hat{\gamma}(\omega),\mathcal{O}}(\{2\}) - c^{\hat{\gamma}(\omega),\mathcal{O}}(\{1\}) + c^{\hat{\gamma}(\omega),\mathcal{O}}(N)}{2}\right] \\
&= \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) + \sum_{t=T_{\text{join}}(\omega)}^{T_{\text{end}}} \frac{K^{\hat{\gamma}(\omega),t}_{\{2\}}(\pi^{\omega,*}_{\{2\}},\omega) - K^{\hat{\gamma}(\omega),t}_{\{1\}}(\pi^{\omega,*}_{\{1\}},\omega) + K^{\hat{\gamma}(\omega),t}_{N}(\pi^{\omega,*}_{N},\omega)}{2}\right] \\
&= \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} \left(\frac{K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) - K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2} + \frac{K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) + K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2}\right)\right. \\
&\qquad \left. + \sum_{t=T_{\text{join}}(\omega)}^{T_{\text{end}}} \left(\frac{K^{\hat{\gamma}(\omega),t}_{\{2\}}(\pi^{\omega,*}_{\{2\}},\omega) - K^{\hat{\gamma}(\omega),t}_{\{1\}}(\pi^{\omega,*}_{\{1\}},\omega)}{2} + \frac{K^{\hat{\gamma}(\omega),t}_{N}(\pi^{\omega,*}_{N},\omega)}{2}\right)\right] \\
&= \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{end}}} \frac{K^{\gamma,t}_{\{2\}}(\pi^{\text{feas}}_{\{2\}},\omega) - K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2}\right. \\
&\qquad \left. + \sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} \frac{K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) + K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2} + \sum_{t=T_{\text{join}}(\omega)}^{T_{\text{end}}} \frac{K^{\hat{\gamma}(\omega),t}_{N}(\pi^{\omega,*}_{N},\omega)}{2}\right] \\
&\geq \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{end}}} \frac{K^{\gamma,t}_{\{2\}}(\pi^*_{\{2\}},\omega) - K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2}\right. \\
&\qquad \left. + \sum_{t=T_{\text{start}}}^{T_{\text{join}}(\omega)-1} \frac{K^{\gamma,t}_{\{2\}}(\pi_{\{2\}},\omega) + K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2} + \sum_{t=T_{\text{join}}(\omega)}^{T_{\text{end}}} \frac{K^{\hat{\gamma}(\omega),t}_{N}(\pi^{\omega,*}_{N},\omega)}{2}\right] \\
&\geq \sum_{\omega\in\Omega} \mathbb{P}(\omega) \left[\sum_{t=T_{\text{start}}}^{T_{\text{end}}} \frac{K^{\gamma,t}_{\{2\}}(\pi^*_{\{2\}},\omega) - K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega)}{2} + \sum_{t=T_{\text{start}}}^{T_{\text{end}}} \frac{K^{\gamma,t}_{N}(\pi^{\text{feas}}_{N},\omega)}{2}\right] \\
&\geq \sum_{\omega\in\Omega} \mathbb{P}(\omega) \sum_{t=T_{\text{start}}}^{T_{\text{end}}} \frac{K^{\gamma,t}_{\{2\}}(\pi^*_{\{2\}},\omega) - K^{\gamma,t}_{\{1\}}(\pi^*_{\{1\}},\omega) + K^{\gamma,t}_{N}(\pi^*_{N},\omega)}{2} \\
&= \Phi_2(N, c^{\gamma,\mathcal{O}})
\end{aligned}
$$

The fifth equality holds by construction of the policies $\pi^{\text{feas}}_{\{2\}}$ and $\pi^{\omega,*}_{\{1\}}$. The first inequality holds because the expected costs of a feasible policy are never lower than the expected costs of an optimal policy. The second inequality holds for two reasons. First, in periods in periods $T_{\text{start}}, \dots, T_{\text{join}}(\omega) - 1$, the ordering cost under π^{feas}_N always matches the sum of the ordering costs of $\pi^*_{\{1\}}$ and $\pi_{\{2\}}$, whereas the holding and backlogging cost under π^{feas}_N never exceeds those of $\pi^*_{\{1\}}$ and $\pi_{\{2\}}$ together, as demonstrated by Lemma 7.1. Second, since by construction $I^{T_{\text{join}}(\omega)-1}_{\{1\}}(\pi^*_{\{1\}}, \omega) + I^{T_{\text{join}}(\omega)-1}_{\{2\}}(\pi_{\{2\}}, \omega) = I^{T_{\text{join}}(\omega)-1}_N(\pi^{\text{feas}}_N, \omega)$ while $\pi^{t,\text{feas}}_N$ and $\pi^{t,\omega,*}_N$ coincide for $t \in \{1, \dots, T_{\text{join}}(\omega) - 1\}$, the total post-joining costs of the policies π^{feas}_N and $\pi^{\omega,*}_N$ match. The third inequality holds because the expected costs of a feasible policy are never lower than the expected costs of an optimal policy. This completes the proof. □

The following theorem states that the strategy profile where both players are always ready to join, while "threatening" to operate under an optimal policy otherwise, is a Nash equilibrium. Hence, formation of coalition $\{1, 2\}$ at the beginning of the time horizon is a natural outcome.

Theorem 7.7. *Consider environment $(\gamma, \mathcal{O})$ with $N = \{1, 2\}$ and $\mathcal{O}(M) = N$ for all $M \in 2^N_-$. Any strategy profile in which each player $i \in N$ plays $(\pi^*_{\{i\}}, \mathscr{D}^{join}_{\{i\}})$ with $\pi^*_{\{i\}} \in \operatorname{argmin}_{\pi \in \Pi^{\gamma,\mathcal{O}}_{\{i\}}} K^{\gamma}_{\{i\}}(\pi)$ and $\mathscr{D}^{t,join}_{\{i\}} = \mathscr{D}^t_N$ for all $t \in T$ is a Nash equilibrium for the corresponding strategic join game.*

Proof. By Lemma 7.6, the strategy of player 2 is a best reply to the one of player 1 and, by symmetry, vice versa. Hence, the strategy profile is a Nash equilibrium. □

We conclude this section with some comments on the three key assumptions: equal split of benefits, two players, and full observability. First off, the assumption of an equal split of benefits can be justified on fairness grounds because it corresponds to both the Shapley value and the nucleolus in a two-player game.

Next, the two-player assumption. It is restrictive, but extending our analysis to more than two players would result in additional notational burden and modeling complexities. To illustrate some of these issues, suppose that player 1 would show up belatedly to a situation in which players 2 and 3 had already formed a coalition. Is player 2 then allowed to break away and form a coalition with player 1 only? If so, how to split up the joint inventory that is currently shared by players 2 and 3? And which cost allocation rule to use? Answering these questions to study dynamic coalition formation for more than two players is outside the scope of this chapter.

Our final assumption was full observability. This is not restrictive if both players' demands are independent of previous demands of the other. However, extending our analysis to settings with dependent demands and incomplete observability would again result in additional notational burden and modeling complexities. In particular, upon (belated) coalition formation, do players fully share previous demand histories *before* signing cost

sharing agreements, or *after*? The resulting cost allocations are different. To illustrate, suppose that Bob can only observe his own demands and that Alice was hit by the storm. If Bob shows up late at the start of period 2, he would not want to pay more than 0 if he learns that Alice was hit by the storm *before* signing a cost sharing agreement, but would accept paying 60 if he only learns Alice's history *after* signing a cost sharing agreement. It is easy to check that in the former case, strategically joining late *can* be beneficial.

7.7 Conclusion

In this chapter, we presented cooperative games corresponding to a situation where several players face stochastic demands over a finite time horizon. We allowed demands to be non-stationary and correlated. We analyzed the cost sharing problem, which arises when players coordinate orders and pool inventories, by applying concepts from cooperative game theory. An important insight is that information matters: different observation possibilities for the historical demands lead to different games. Accordingly, if you don't keep track of what people know, it is worth extending effort to do so.

Our focus was not on solution algorithms. Earlier, we described stochastic dynamic programming as the standard method for finding optimal policies in sequential decision problems. However, we should mention that this method runs into computational intractability and the curse of dimensionality when dealing with large-scale problems. Approximate dynamic programming (see, e.g., Powell, 2011) is capable of handling larger problems, though often at the cost of optimality.

Our model's inherent complexity did not deter us from proving that a stable cost allocation exists. We expect this result to extend to more general settings with, e.g., infinite sample spaces, asymmetric costs, nonzero lead times, costly transshipments, market signals that help explain future demand, and/or capacity restrictions. The idea behind the proof approach described in Section 7.4, i.e., constructing a feasible policy for the grand coalition out of a balanced combination of optimal policies of subcoalitions, can act as a guidance to prove balancedness for those more general models. However, our proof approach does not directly extend to non-linear costs, such as concave production costs, because the resulting policy that would be constructed for the grand coalition might lead to higher costs than for the subcoalitions.

In Section 7.6, we showed that, under certain assumptions, we need not worry about players gaming the allocation by strategically arriving late. Future research might extend our explorative analysis to a more general dynamic coalition formation problem with three or more players. Some of the resulting challenges and hurdles were discussed at the end of that section.

—Equals should be treated equally, and unequals unequally in proportion to relevant similarities and differences.

Artistotle

8

Allocation rules on elastic single-attribute situations

8.1 Introduction

In the previous chapters, we frequently found that a proportional rule, which simply divides the collective cost of the grand coalition proportional to players' arrival or demand rates, was guaranteed to result in stable allocations. This was shown to be true for, e.g., spare parts games with optimized base stock levels in Section 6.5.2. Indeed, for given holding and backlogging costs, the class of spare parts games with optimized numbers of servers coincides with all single-attribute games embedded in a specific elastic cost function. Due to this elasticity, stability of the allocation prescribed by the proportional rule was guaranteed by Theorem 2.5.

A proportional rule is easy to understand and is computationally attractive. Additionally, as we showed in Section 6.5.4, when the arrival or demand process is a Poisson process, then the proportional allocation of expected infinite-horizon costs can be easily implemented in practice via a simple process for cost realizations. Moreover, as we will show in Section 8.2, the proportional rule can be axiomatically characterized (on the domain of single-attribute situations with elastic cost functions) as the *unique* continuous rule that is immune to manipulations of the players via artificial splitting or merging. This means that if a small change in players' attributes can only cause a small change in the allocation and if no group of players can have an incentive to artificially represent themselves together as a single player, or vice versa, then we cannot escape the proportional rule.

The proportional rule has a downside as well: a player with a lower attribute may

reap more benefits (defined as the costs a player would incur when acting alone minus the cost assigned to this player when collaborating with everyone) than a player with a higher attribute. Although, as we showed in Chapter 6, a larger player *might* receive more benefits than a smaller player, this is by no means guaranteed. For instance, as we will show in more detail in Example 8.2, a "small" player with attribute 9 may reap a benefit of 1.2 under the proportional rule, while a "large" player with attribute 16 only reaps a benefit of 0.8 as a result of collaboration. Then, the large player may feel like he is treated unfairly. Even if the allocation may be stable, the small player is almost free-riding on the large attribute that the large player brought in.

In this chapter, which is based on Karsten et al. (2013b), we study various alternative allocation rules to see if we can do "better" than the proportional rule, in particular with regard to this benefit ordering issue. It would be nice to have an allocation rule that always accomplishes core allocations *and* always gives larger benefits to players with larger attributes. This, however, turns out to be impossible: there is no rule on the class of elastic single-attribute situations that satisfies both requirements simultaneously!

After relaxing our requirements to compatible ones, we introduce new allocation rules on elastic single-attribute situations and evaluate their performance with respect to core inclusion and benefit ordering of their allocations. Here, we are inspired by existing cost allocation rules from various strands of literature, including such seminal works as Shapley (1953) on the Shapley value, Shapley (1971) on marginal rules, and Moulin and Shenker (1992) on the serial cost sharing rule. The new rules we study explicitly take into account the specific shape of the cost function, as opposed to the proportional rule.

The main contribution of this chapter is that it constitutes the first comparison of cost allocation rules on the domain of elastic single-attribute situations. Our main motivation for considering this domain is that it includes the inventory situations with optimized numbers of servers from Chapters 4 and 6 in which each player was associated with his demand rate only. Hence, the advantages and disadvantages of the proportional rule and alternative allocation rules that we will analyze in the present chapter teach us more about cost allocation in those resource pooling games. The domain of elastic single-attribute situations also includes symmetric $M/M/1$ optimized situations, as argued in Section 3.3.2. Moreover, it contains EOQ situations. Indeed, in any EOQ situation $(N, (m_i^2)_{i\in N}, a)$, as defined in Section 3.2, each player has an attribute $m_i^2 \in \mathbb{R}_{++}$, and the costs for any coalition depend on its members' attributes only through its sum: under total attribute $\ell > 0$, the cost is $2a\ell^{1/2}$. This cost function is concave, hence elastic, in ℓ.

The remainder of this chapter is organized as follows. We start in Section 8.2 with the axiomatic characterization of the proportional rule based on non-manipulability. Section 8.3 presents our impossibility result saying that benefit ordering and coalitional rationality are not compatible in general. Subsequently, we introduce and analyze several new allocation rules: Section 8.4 proposes the serial cost sharing rule, Section 8.5 discusses the rule

that assigns benefits proportionally, and Section 8.6 forms the main part of our work by introducing variants of marginal allocations based on concavicated single-attribute situations. We conclude in Section 8.7.

8.2 An axiomatic characterization of the proportional rule

In this section, we show that there is only one continuous allocation rule on elastic single-attribute situations that is immune to manipulations of the players via splitting or merging: the proportional rule. We start with two definitions, which are based on elasticity of a function (introduced in Section 2.2.2) and on single-attribute situations (introduced in Section 2.3.10).

Definition 8.1. A single-attribute situation $(N, \tilde{K}, \lambda)$ is called *elastic* if $\tilde{K}$ is elastic.

The set of all elastic single-attribute situations with finite but variable N is denoted by $\mathcal{E}$.

Definition 8.2. An *allocation rule on elastic single-attribute situations* (or *rule* for short) is a mapping $\mathscr{F}$ on $\mathcal{E}$ such that for every $(N, \tilde{K}, \lambda) \in \mathcal{E}$ it holds that $\mathscr{F}(N, \tilde{K}, \lambda) \in \mathbb{R}^N$ and that $\sum_{i \in N} \mathscr{F}_i(N, \tilde{K}, \lambda) = \tilde{K}(\lambda_N)$.

So, a rule $\mathscr{F}$ assigns to any elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ an allocation for the associated single-attribute game (N, c^φ).

The following definition formally describes the proportional rule. This rule is often referred to as "average pricing" in the cost sharing literature (e.g., Moulin and Shenker, 1992).

Definition 8.3. The *proportional rule* $\mathscr{P}$ on the class of elastic single-attribute situations $\mathcal{E}$ is defined by allocating $\mathscr{P}_i(N, \tilde{K}, \lambda) = \tilde{K}(\lambda_N) \cdot \lambda_i / \lambda_N$ to player $i \in N$ in situation $(N, \tilde{K}, \lambda) \in \mathcal{E}$.

Non-manipulability means that no group of players has an incentive to artificially represent themselves together as a single player, or vice versa. The following two definitions describe this concept more formally.

Definition 8.4. Two elastic single-attribute situations $(N, \tilde{K}, \lambda)$ and $(\bar{N}, \tilde{K}, \bar{\lambda})$ are said to be *manipulations* if

- $|\bar{N} \setminus N| = 1$;
- $\sum_{i \in N \setminus \bar{N}} \lambda_i = \sum_{i \in \bar{N} \setminus N} \lambda_i$; and
- $\bar{\lambda}_i = \lambda_i$ for all $i \in \bar{N} \cap N$.

It may be helpful to think of $N \setminus \bar{N}$ as the set of players that is merged into the single element of $\bar{N} \setminus N$, while $\bar{N} \cap N$ is the set of unaffected players that N and $\bar{N}$ have in common.

Definition 8.5. A rule $\mathscr{F}$ satisfies *non-manipulability* (NM) on $\mathcal{E}$ if, for any two elastic single-attribute situations that are manipulations, say $\varphi = (N, \tilde{K}, \lambda)$ and $\bar{\varphi} = (\bar{N}, \tilde{K}, \bar{\lambda})$ with $\bar{N} \setminus N = \{j\}$, it holds that $\mathscr{F}_j(\bar{\varphi}) = \sum_{i \in N \setminus \bar{N}} \mathscr{F}_i(\varphi)$.

The following example illustrates these definitions.

Example 8.1. Consider the 3-player single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ with $N = \{1, 2, 3\}$, $\lambda_1 = 1$, $\lambda_2 = 9$, $\lambda_3 = 16$, and $\tilde{K}(\ell) = 12$ for all $\ell \in \mathbb{R}_+$. If players 2 and 3 would artificially represent themselves together as a single player, say 4, then this is described by the manipulation $\bar{\varphi} = (\bar{N}, \tilde{K}, \bar{\lambda})$ with $\bar{N} = \{1, 4\}$, $\bar{\lambda}_1 = 1$, and $\bar{\lambda}_4 = 25$.

To illustrate manipulability, consider the rule $\mathscr{U}$ on $\mathcal{E}$ that assigns costs proportional to the square root of the players' attributes. This rule assigns $\mathscr{U}_2(\varphi) = 4.5$, $\mathscr{U}_3(\varphi) = 6$, and $\mathscr{U}_4(\bar{\varphi}) = 10$. So, under $\mathscr{U}$, players 2 and 3 have an incentive to merge. Hence, $\mathscr{U}$ does not satisfy non-manipulability. ◇

The following property says that a small change in the attributes can only cause a small change in the allocation.

Definition 8.6. A rule $\mathscr{F}$ is called *continuous in attributes* (CA) on $\mathcal{E}$ if $\mathscr{F}(N, \tilde{K}, \lambda)$ is continuous in λ for every $(N, \tilde{K}, \lambda) \in \mathcal{E}$.

The following lemma states that the proportional rule satisfies NM and CA.

Lemma 8.1. *The rule $\mathscr{P}$ on $\mathcal{E}$ is non-manipulable and continuous in attributes.*

Proof. Let $\varphi = (N, \tilde{K}, \lambda)$ and $\bar{\varphi} = (\bar{N}, \tilde{K}, \bar{\lambda})$ be two elastic single-attribute situations that are manipulations with $\bar{N} \setminus N = \{j\}$. Because they are manipulations,

$$\mathscr{P}_j(\bar{\varphi}) = K(\lambda_{\bar{N}}) \cdot \bar{\lambda}_j / \bar{\lambda}_{\bar{N}} = K(\lambda_{\bar{N}}) \cdot \sum_{i \in N \setminus \bar{N}} \lambda_i / \bar{\lambda}_{\bar{N}} = K(\lambda_N) \cdot \sum_{i \in N \setminus \bar{N}} \lambda_i / \lambda_N = \sum_{i \in N \setminus \bar{N}} \mathscr{P}_i(\varphi).$$

Hence, $\mathscr{P}$ satisfies non-manipulability.

To establish continuity, observe that $\lambda_i > 0$ for all $i \in N$, and thus $\lambda_N > 0$. Accordingly, by Theorem 2.2, $\tilde{K}(\lambda_N)$ is continuous in the vector λ. This implies that $\mathscr{P}$ is continuous in attributes as well. □

The following lemma states that the costs assigned to any player by a non-manipulable rule can only depend on the vector of attributes through that player's own attribute and an aggregate parameter. The lemma also states that these costs are independent of the total number and names of participating players. We let $\mathbb{E}$ denote the set of all elastic, non-decreasing functions mapping $\mathbb{R}_+$ to $\mathbb{R}_+$.

Lemma 8.2. *Let $\mathscr{F}$ be a non-manipulable rule on $\mathcal{E}$. Then there exists an associated function $g : \mathbb{R}_{++} \times \mathbb{R}_{++} \times \mathbb{E} \to \mathbb{R}$ such that $\mathscr{F}_j(N, \tilde{K}, \lambda) = g(\lambda_j, \lambda_N, \tilde{K})$ for all $(N, \tilde{K}, \lambda) \in \mathcal{E}$ and all $j \in N$.*

Proof. Let j be a player name. Our proof will proceed as follows. First, we show that $\mathscr{F}_j(\varphi)$ for $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$ with $j \in N$ can only depend on the vector λ through the numbers λ_j and λ_N and can only depend on the player set N through the name of player j. Second, we show that $\mathscr{F}_j(\varphi)$ for $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$ with $j \in N$ does not depend on the name of player j either.

For the first part, let $\hat{\varphi} = (\hat{N}, \tilde{K}, \hat{\lambda}) \in \mathcal{E}$ such that $j \in \hat{N}$, $\lambda_j = \hat{\lambda}_j$, and $\lambda_N = \hat{\lambda}_{\hat{N}}$. We will show that $\mathscr{F}_j(\varphi) = \mathscr{F}_j(\hat{\varphi})$ and distinguish two cases.

Case 1: $N = \{j\}$. Then, $\hat{N} = \{j\}$ as well, so clearly $\varphi = \hat{\varphi}$, and thus $\mathscr{F}_j(\varphi) = \mathscr{F}_j(\hat{\varphi})$.

Case 2: $|N| \geq 2$. Merge the set $N \setminus \{j\}$ into a player $k \neq j$ to obtain the manipulation $\bar{\varphi} = (\bar{N}, \tilde{K}, \bar{\lambda})$ with $\bar{N} = \{j, k\}$. By the combination of assumptions on φ and $\hat{\varphi}$, situation $\bar{\varphi}$ is a manipulation of $\hat{\varphi}$ as well. Hence, $\lambda_N = \hat{\lambda}_{\hat{N}} = \bar{\lambda}_{\bar{N}}$. We then obtain

$$\mathscr{F}_j(\varphi) = \tilde{K}(\lambda_N) - \textstyle\sum_{i \in N \setminus \bar{N}} \mathscr{F}_i(\varphi) = \tilde{K}(\bar{\lambda}_{\bar{N}}) - \sum_{i \in N \setminus \bar{N}} \mathscr{F}_i(\varphi) = \tilde{K}(\bar{\lambda}_{\bar{N}}) - \mathscr{F}_k(\bar{\varphi}) = \mathscr{F}_j(\bar{\varphi}),$$

where the third equality holds because $\mathscr{F}$ satisfies non-manipulability. Analogously, we can also show $\mathscr{F}_j(\hat{\varphi}) = \mathscr{F}_j(\bar{\varphi})$, and thus $\mathscr{F}_j(\varphi) = \mathscr{F}_j(\hat{\varphi})$.

Combining both cases, we conclude that $\mathscr{F}_j(\varphi)$ can only depend on the vector λ through λ_j and λ_N and does not depend on the size of N or on the names of players other than j.

It remains to prove that $\mathscr{F}_j(\varphi)$ does not depend on the name of player j either. Again, we distinguish two cases.

Case 1: $N = \{j\}$. Then, by efficiency, $\mathscr{F}_j(\varphi) = \tilde{K}(\lambda_N)$, independent of player j's name.

Case 2: $|N| \geq 2$. Consider the elastic single-attribute situation φ' that is obtained from φ by merely relabeling player j to j' in the entire tuple. Because φ' is a manipulation of φ, we obtain $\mathscr{F}_j(\varphi) = \mathscr{F}_{j'}(\varphi')$.

Combining both cases, we conclude that $\mathscr{F}_j(\varphi)$ does not depend on the name of player j. This completes the proof. □

We are now ready to state this section's main result.

Theorem 8.3. *The proportional rule $\mathscr{P}$ is the unique allocation rule on the class of elastic single-attribute situations $\mathcal{E}$ satisfying both non-manipulability and continuity in attributes.*

Proof. Let $\mathscr{F}$ be a non-manipulable rule that is continuous in attributes, which exists by Lemma 8.1. Take an associated function g as in Lemma 8.2. We will show that $\mathscr{F} = \mathscr{P}$. Let $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$. Keeping λ_N and $\tilde{K}$ fixed, we define the function $\hat{g} : (0, \lambda_N] \to \mathbb{R}$ by $\hat{g}(\ell) = g(\ell, \lambda_N, \tilde{K})$.

We first show that $\hat{g}$ is linear. To this end, let $\ell_1, \ell_2 \in (0, \lambda_N]$ be such that $\ell_1 + \ell_2 \leq \lambda_N$. Consider an elastic single-attribute situation $\varphi' = (N', \lambda', \tilde{K})$ such that $N' = \{1, 2, \ldots, n\}$ for some integer $n \geq 2$, $\lambda'_{N'} = \lambda_N$, $\lambda'_1 = \ell_1$, and $\lambda'_2 = \ell_2$. Merging players 1 and 2 into a player j, we define the manipulation $\bar{\varphi}' = (\bar{N}', \bar{\lambda}', \tilde{K})$ with $\bar{N}' = \{j\} \cup \{3, \ldots, n\}$. Then,

$$\hat{g}(\ell_1 + \ell_2) = \hat{g}(\lambda'_1 + \lambda'_2) = \hat{g}(\bar{\lambda}'_j) = \mathscr{F}_j(\bar{\varphi}') = \mathscr{F}_1(\varphi') + \mathscr{F}_2(\varphi') = \hat{g}(\lambda'_1) + \hat{g}(\lambda'_2) = \hat{g}(\ell_1) + \hat{g}(\ell_2).$$

The second equality holds since $\bar{\varphi}'$ is a manipulation of φ', and hence $\lambda'_1 + \lambda'_2 = \bar{\lambda}_j$. The third and fifth equalities hold by Lemma 8.2, and the fourth equality holds because $\mathscr{F}$ is a non-manipulable allocation rule. Hence, $\hat{g}$ is an additive function.

Moreover, $\hat{g}$ is a continuous function because, for any $\ell \in (0, \lambda_N]$, there exists an elastic single-attribute situation $\hat{\varphi} = (\hat{N}, \hat{\lambda}, \tilde{K})$ such that $j \in N$, $\hat{\lambda}_j = \ell$, and $\hat{\lambda}_{\hat{N}} = \lambda_N$, while by Lemma 8.2 it holds that $\hat{g}(\lambda_j) = \mathscr{F}_j(\varphi)$. Thus, continuity in attributes of $\mathscr{F}$ implies continuity of $\hat{g}$. As $\hat{g}$ is both continuous and additive, we conclude in line with Cauchy (1821, Chapter 5) that $\hat{g}$ is linear.

To finish the proof, we return to our original situation φ. As $\mathscr{F}_i(\varphi) = \hat{g}(\lambda_i)$ for all $i \in N$ and $\hat{g}$ is linear, it holds that $\sum_{i \in N} \mathscr{F}_i(\varphi) = \lambda_N \cdot C$ for some $C \in \mathbb{R}$. Since $\sum_{i \in N} \mathscr{F}_i(\varphi) = \tilde{K}(\lambda_N)$, we find that $C = \tilde{K}(\lambda_N)/\lambda_N$. Hence, $\mathscr{F}_i(\varphi) = \tilde{K}(\lambda_N) \cdot \lambda_i/\lambda_N = \mathscr{P}_i(\varphi)$ for all $i \in N$. We conclude that $\mathscr{P}$ is the unique rule satisfying both non-manipulability and continuity in attributes. □

Similar axiomatic characterizations have appeared in the literature. For example, Mosquera et al. (2008) and García-Sanz et al. (2008) uniquely characterize a proportional allocation rule for, respectively, EOQ situations and (symmetric) $M/M/1$ queueing situations with optimized numbers of servers. Likewise, Banker (1981), O'Neill (1982), Chun (1988), and de Frutos (1999) uniquely characterize a proportional allocation rule for bankruptcy situations. (In a bankruptcy situation, each player has a non-negative claim over an amount of money, which is similar to an attribute.) These authors use axioms that are in the same spirit of non-manipulability; however, they consider situation classes that differ from our class of elastic single-attribute situations.

All employ a secondary axiom to accomplish a unique characterization. Banker (1981) assumed that attributes are elements of $\mathbb{Q}$; de Frutos (1999), Mosquera et al. (2008), and García-Sanz et al. (2008) assumed that cost allocations are non-negative; and O'Neill (1982) assumed a continuity property. We chose continuity for the characterization on our specific domain because we view exclusion of irrational-valued attributes or negative cost allocations a priori as unnatural. For example, the allocation $(-1, \pi, 13 - \pi)$ for the single-attribute game (N, c^{φ}) of Example 8.1 is individually rational[52] and should not be excluded a priori. We view continuity as more natural. In fact, continuity in at least one point would already have enough bite, cf. O'Neill (1982), but it would have a less natural interpretation.

Although we gave our axiomatic characterization for the entire class of single-attribute situations with *elastic* cost functions, our axiomatization could just as well have been carried out on a class induced by a *specific* elastic cost function (e.g., the optimal cost function in an Erlang loss model for given resource and penalty costs) or on a more general class of

[52]This allocation is not stable, as it must be. For any single-attribute situation $(N, \lambda, \tilde{K})$, every allocation x for the single-attribute game (N, c^{φ}) with $x_i < 0$ for some $i \in N$ is not stable because $\sum_{j \in N \setminus \{i\}} x_j = c(N) - x_i > c(N) = \tilde{K}(\lambda_N) \geq \tilde{K}(\lambda_{N \setminus \{i\}}) = c^{\varphi}(N \setminus \{i\})$, where the first inequality holds because $x_i < 0$ and the second inequality holds because $\tilde{K}$ is non-decreasing.

single-attribute situations with *continuous* cost functions. Indeed, the only time we used any property of the cost function $\tilde{K}$ at all is in Lemma 8.1, where we merely required continuity of $\tilde{K}$ on $\mathbb{R}_{++}$.

8.3 An impossibility result

In the previous section, we showed that the proportional rule $\mathscr{P}$ is not manipulable on $\mathcal{E}$. So in a collaboration between, e.g., company A with a single business unit and company B with two business units, the costs assigned to company A by $\mathscr{P}$ are unaffected by whether the business units comprising company B claim they should be treated as one player together or two players separately. For such situations, $\mathscr{P}$ may be compelling. In other situations, non-manipulability may be of less importance. Then, we have to consider other fairness criteria. One such criterion is described in the following definition.

Definition 8.7. A rule $\mathscr{F}$ on $\mathcal{E}$ is said to have the *coalitional rationality property* (CR) if $\mathscr{F}(\varphi) \in \mathcal{C}(N, c^\varphi)$ for any elastic single-attribute situation $\varphi = (N, \lambda, \tilde{K})$.

Coalitional rationality says that a rule should always generate stable allocations, i.e., a core element for the associated single-attribute game. Stability of allocations is important in our resource pooling context since a coalition could credibly threaten to split off and set up a separate pooling group. Hence, in this context, any reasonable rule should satisfy the coalitional rationality property.

The following example describes a disadvantage of $\mathscr{P}$.

Example 8.2. Suppose that player 1 (with demand rate 9 per month) and player 2 (with demand rate 16 per month) aim to set up a joint service system, whose total monthly costs increase concavely according to the square root of the total demand rate served. This may be modeled via the elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ with $N = \{1, 2\}$, $\lambda_1 = 9$, $\lambda_2 = 16$, and $\tilde{K}(\ell) = \sqrt{\ell}$ for all $\ell \in \mathbb{R}_+$. The single-attribute game associated with this situation, (N, c^φ), is given by $c^\varphi(\{1\}) = 3$, $c^\varphi(\{2\}) = 4$, and $c^\varphi(N) = 5$.

Clearly, $\mathscr{P}_1(\varphi) = 5 \cdot 9/25 = 9/5$ and $\mathscr{P}_2(\varphi) = 5 \cdot 16/25 = 16/5$. This means that the cost savings allocated to player 1 by the proportional rule, $c^\varphi(\{1\}) - \mathscr{P}_1(\varphi) = 6/5$, are larger than the cost savings allocated to player 2, $c^\varphi(\{2\}) - \mathscr{P}_2(\varphi) = 4/5$. So, even though collaboration under $\mathscr{P}$ does in fact produce a small saving for player 2, he gets less out of the collaboration than player 1. One could argue that this is unfair because the total savings were only made possible because player 2 allowed player 1 to piggyback on his large attribute. ◇

The following definition formalizes the idea that an allocation rule should avoid the issue described in Example 8.2.

Definition 8.8. A rule $\mathscr{F}$ on $\mathcal{E}$ is said to have the *benefit ordering property* (BO) if, for every elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ and every $i, j \in N$ with $\lambda_i \leq \lambda_j$, we have that $c^\varphi(\{i\}) - \mathscr{F}_i(\varphi) \leq c^\varphi(\{j\}) - \mathscr{F}_j(\varphi)$.

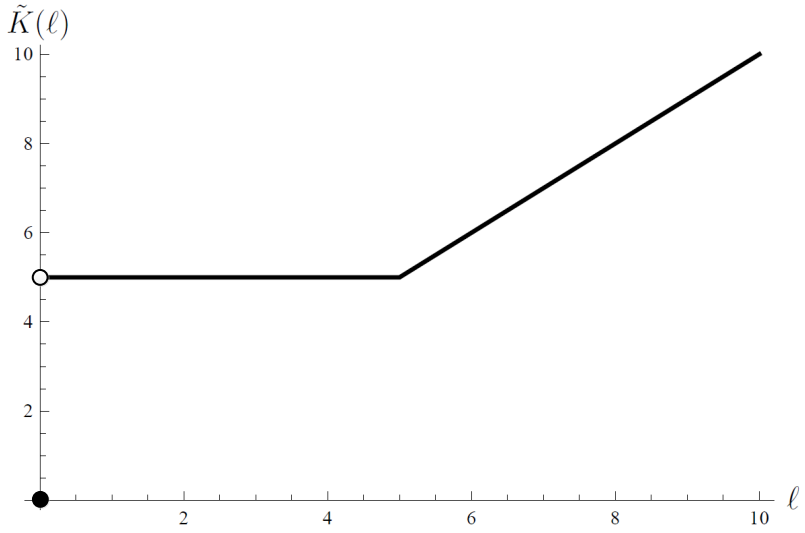

Figure 8.1: *A plot of the function $\tilde{K}$ in Example 8.3.*

Benefit ordering means that a player with a larger attribute should always reap at least as much benefit (allocated cost savings) from the collaboration as a player with a smaller attribute. As shown in Example 8.2, the proportional rule does not satisfy the benefit ordering property. However, by Theorem 2.5, the proportional rule does satisfy the coalitional rationality property. A natural follow-up question is whether there is a rule on $\mathcal{E}$ that satisfies both the benefit ordering property and the coalitional rationality property. The following example, however, proves that such a rule does not exist: *no* rule on $\mathcal{E}$ can satisfy both CR and BO simultaneously. This is a remarkable impossibility result.

Example 8.3. Consider the single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ with $N = \{1, 2, 3\}$, $\lambda_1 = \lambda_2 = 2.5$, $\lambda_3 = 5$, and cost function $\tilde{K}$ described by

$$\tilde{K}(x) = \begin{cases} 0 & \text{if } x = 0; \\ 5 & \text{if } x \in (0, 5]; \\ x & \text{if } x > 5. \end{cases}$$

We showed that this function is elastic in Example 2.2 on page 20. For convenience, this function is graphically represented in Figure 8.1. The associated single-attribute game (N, c^φ) is given by

$$c^\varphi(M) = \begin{cases} 5 & \text{if } |M| = 1 \text{ or } M = \{1, 2\}; \\ 7.5 & \text{if } M = \{1, 3\} \text{ or } M = \{2, 3\}; \\ 10 & \text{if } M = N. \end{cases}$$

This game's core has only one element: $\mathcal{C}(N, c^\varphi) = \{(2.5, 2.5, 5)\}$.[53] The benefits for player 3 under the core allocation are zero, while the benefits to players 1 and 2 are (strictly) positive. Since the attributes of players 1 and 2 are smaller than the attribute of player 3, we conclude that any rule satisfying the coalitional rationality property cannot satisfy the benefit ordering property.[54] ◇

Example 8.3 shows that the properties CR and BO are incompatible on $\mathcal{E}$. If we want to arrive at properties that *are* compatible, then we have to relax some of our requirements. The following two definitions propose relaxations of CR and BO, respectively.

Definition 8.9. A rule $\mathscr{F}$ is said to have the *individual rationality property* (IR) on $\mathcal{E}$ if $\mathscr{F}(\varphi) \in \mathcal{I}(N, c^\varphi)$ for any elastic single-attribute situation φ.

Definition 8.10. A rule $\mathscr{F}$ on $\mathcal{E}$ is said to have the *benefit ordering property under concavity* (BOC) if, for every elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ with concave $\tilde{K}$ and every $i, j \in N$ with $\lambda_i \leq \lambda_j$, we have that $c^\varphi(\{i\}) - \mathscr{F}_i(\varphi) \leq c^\varphi(\{j\}) - \mathscr{F}_j(\varphi)$.

It is easy to see that CR implies IR and that BO implies BOC. Indeed, CR extends the rationality requirement of IR from single players to coalitions, and BO extends the ordering requirement of BOC from concave cost functions to arbitrary cost functions. These relaxations help to illustrate the extent to which CR and BO are incompatible. In Section 8.5, we construct a rule on $\mathcal{E}$ satisfying both IR and BO. In Section 8.6, we construct a rule on $\mathcal{E}$ satisfying both CR and BOC.

We remark that in many situations, the cost function is concave, e.g., EOQ situations and symmetric $M/M/1$ queueing situations with optimized numbers of servers. However, even in such concave situations, the proportional rule $\mathscr{P}$ is not guaranteed to dish out larger benefits to larger players, as shown in Example 8.2 on page 181. This implies that $\mathscr{P}$ does not satisfy BOC. Accordingly, a rule that *does* satisfy BOC can be rightfully said to do "better" than $\mathscr{P}$ with regard to the ordering of players' benefits. In the next three sections, we introduce several new rules and analyze their properties.

8.4 The serial rule

Single-attribute situations have the same mathematical structure as so-called cost sharing situations, which are well-studied in the literature. Indeed, a cost sharing situation is a tuple (N, q, C), with a given set $N = \{1, \ldots, n\}$ of users who have demands $q \in \mathbb{R}_+^N$ and who share

[53] To see this, first note that $(2.5, 2.5, 5)$ is a stable allocation, so the core is non-empty. Let x be a core allocation. This implies that $x_3 \leq c(\{3\}) = 5$, $x_1 + x_2 \leq c(\{1, 2\}) = 5$, and $x_1 + x_2 + x_3 = 10$, which together yield $x_3 = x_1 + x_2 = 5$. Stability of x in combination with $x_3 = 5$ yields $x_1 \leq c^\varphi(\{1, 3\}) - x_3 = 2.5$ and $x_2 \leq c^\varphi(\{1, 3\}) - x_3 = 2.5$. But since we established that $x_1 + x_2 = 5$, we must have $x_1 = x_2 = 2.5$.

[54] The same incompatibility would occur if, e.g., $\tilde{K} = \sqrt{5\ell}$ on $[0, 5]$. The discontinuity of the cost function at 0 in the example does not drive the incompatibility; it is merely for expositional ease.

a joint production process described by a cost function $C : \mathbb{R}_+ \to \mathbb{R}_+$. The argument of C is interpreted as the sum of demands to be served.

The typical *story* behind cost sharing situations differs from our single-attribute situations, however. In a cost sharing situation, there is only one production process that is jointly owned by all the players, and they are basically forced to collaborate with each other. Applications range from distributing the total production costs for output produced on a single advanced machine, setting fees for the use of a common parking facility in a shopping mall, to sharing the joint costs of a shared water supply transportation network. See, e.g., Moulin and Shenker (1992). In particular, no group of users can get their desired output if they split off and act independently. This different interpretation means that certain considerations which are natural for resource pooling situations, such as individual rationality or stability, do not apply to cost sharing situations.

The literature on cost sharing situations, however, contains interesting cost sharing mechanisms. We translate the serial rule (Moulin and Shenker, 1992), one of the most-studied mechanisms, to a rule on $\mathcal{E}$. This rule takes into account some of the intermediate behavior of the cost function, as opposed to the proportional rule which entirely ignores the behavior of the cost function between zero and the total demand of the grand coalition.

Definition 8.11. The *serial rule* $\mathscr{S}$ on $\mathcal{E}$ is defined by allocating

$$\mathscr{S}_i(N, \tilde{K}, \lambda) = \sum_{j=1}^{\sigma(i)} \frac{\tilde{K}(\Lambda_j) - \tilde{K}(\Lambda_{j-1})}{|N| + 1 - j}$$

to player $i \in N$ in the elastic single-attribute situation $(N, \tilde{K}, \lambda)$, where σ is an ordering on N such that $\lambda_{\sigma^{-1}(1)} \leq \lambda_{\sigma^{-1}(2)} \leq \ldots \leq \lambda_{\sigma^{-1}(|N|)}$, which orders the players from small to large attributes[55], where $\Lambda_0 = 0$, and where

$$\Lambda_j = (n + 1 - j)\lambda_{\sigma^{-1}(j)} + \sum_{k=1}^{\sigma(j)-1} \lambda_{\sigma^{-1}(k)}$$

for any $j \in \{1, \ldots, |N|\}$. Note that $\Lambda_1 \leq \Lambda_2 \leq \ldots \leq \Lambda_{|N|}$.

To illustrate, suppose for notational ease that $N = \{1, 2, \ldots, n\}$ and that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Then, serial cost sharing says that player 1, with the lowest attribute λ_1, pays $(1/n)$th of the cost of $\Lambda_1 = n\lambda_1$, i.e., the total costs if everyone would have player 1's attribute. Player 2, with the next lowest attribute λ_2, pays player 1's cost share plus $1/(n-1)$th of the incremental cost from $\Lambda_1 = n\lambda_1$ to $\Lambda_2 = (n-1)(\lambda_2 - \lambda_1) + n\lambda_1$, i.e., to the total costs if everyone except for player 1 would have player 2's attribute. Player 3, with the next lowest attribute λ_3 pays player 2's cost share, plus $1/(n-2)$th of the incremental cost from $\Lambda_2 = (n-1)(\lambda_2 - \lambda_1) + n\lambda_1$ to $\Lambda_3 = (n-2)(\lambda_3 - \lambda_2) + (n-1)(\lambda_2 - \lambda_1) + n\lambda_1$. And so on.

[55] If several players have the same attribute, then all such orderings lead to the same allocation.

The serial rule $\mathscr{S}$ is clearly continuous in attributes and thus, by Theorem 8.3, manipulable. The following theorem deals with two more properties.

Theorem 8.4. *The serial rule $\mathscr{S}$ on $\mathcal{E}$ satisfies the individual rationality property and the benefit ordering property under concavity.*

Proof. Let $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$. Without loss of generality, assume for notational convenience that $N = \{1, 2, \ldots, n\}$ and that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Let $i \in N$. Then,

$$\begin{aligned}
\mathscr{S}_i(\varphi) &= \sum_{j=1}^{i} \frac{\tilde{K}(\Lambda_j) - \tilde{K}(\Lambda_{j-1})}{n+1-j} \\
&= \frac{\tilde{K}(\Lambda_i)}{n-i+1} - \sum_{j=1}^{i-1} \frac{\tilde{K}(\Lambda_j)}{(n-j+1)(n-j)} \\
&= \frac{\Lambda_i}{n-i+1} \cdot \frac{\tilde{K}(\Lambda_i)}{\Lambda_i} - \sum_{j=1}^{i-1} \frac{\Lambda_j}{(n-j+1)(n-j)} \cdot \frac{\tilde{K}(\Lambda_j)}{\Lambda_j} \\
&\leq \frac{\Lambda_i}{n-i+1} \cdot \frac{\tilde{K}(\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\})}{\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}} \\
&\quad - \sum_{j=1}^{i-1} \frac{\Lambda_j}{(n-j+1)(n-j)} \cdot \frac{\tilde{K}(\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\})}{\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}} \\
&= \frac{\tilde{K}(\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\})}{\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}} \cdot \left(\frac{\Lambda_1}{n} + \sum_{j=2}^{i} \frac{\Lambda_j - \Lambda_{j-1}}{n+1-j} \right) \\
&= \frac{\tilde{K}(\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\})}{\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}} \cdot \left(\lambda_1 + \sum_{j=2}^{i} (\lambda_j - \lambda_{j-1}) \right) \\
&= \frac{\tilde{K}(\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\})}{\max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}} \cdot \lambda_i \\
&\leq \frac{\tilde{K}(\lambda_i)}{\lambda_i} \cdot \lambda_i = c^{\varphi}(\{i\}),
\end{aligned}$$

where both inequalities hold by elasticity of $\tilde{K}$. In the first inequality, we use that $\Lambda_i \geq \max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}$, while $\Lambda_j \leq \max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}$ for any $j \in \{1, \ldots, i-1\}$. In the second inequality, we use that $\lambda_i \leq \max\{\Lambda_1, \ldots, \Lambda_{i-1}, \lambda_i\}$. We conclude that $\mathscr{S}(\varphi) \in \mathcal{I}(N, c^{\varphi})$.

To study benefit ordering under concavity, assume that $\tilde{K}$ is concave. Suppose that $i \in \{1, 2, \ldots, n-1\}$. By assumption, $\lambda_i \leq \lambda_{i+1}$. It suffices to prove that $c^{\varphi}(\{i\}) - \mathscr{S}_i(\varphi) \leq c^{\varphi}(\{i+1\}) - \mathscr{S}_{i+1}(\varphi)$. To this end, we first derive that

$$\begin{aligned}
\tilde{K}(\lambda_{i+1}) - \tilde{K}(\lambda_i) &= \frac{\tilde{K}(\lambda_i + \lambda_{i+1} - \lambda_i) - \tilde{K}(\lambda_i)}{\lambda_{i+1} - \lambda_i} \cdot (\lambda_{i+1} - \lambda_i) \\
&\geq \frac{\tilde{K}\big(\Lambda_i + (n-i)(\lambda_{i+1} - \lambda_i)\big) - \tilde{K}(\Lambda_i)}{(n-i)(\lambda_{i+1} - \lambda_i)} \cdot (\lambda_{i+1} - \lambda_i)
\end{aligned}$$

$$= \frac{\tilde{K}\big(\Lambda_i + (n-i)(\lambda_{i+1} - \lambda_i)\big) - \tilde{K}(\Lambda_i)}{n-i}$$
$$= \frac{\tilde{K}(\Lambda_{i+1}) - \tilde{K}(\Lambda_i)}{n-i},$$

where the inequality holds because $\tilde{K}$ is concave and thus the difference quotient obtained when adding an amount (in particular, $\lambda_{i+1} - \lambda_i$) to λ_i is at least as large as the difference quotient obtained when adding an amount (in particular, $(n-i)(\lambda_{i+1} - \lambda_i)$) to Λ_i, since $\lambda_i \leq \Lambda_i$.

Subtracting $\tilde{K}(\lambda_{i+1})$ from both sides and multiplying by -1 yields

$$\tilde{K}(\lambda_i) \leq \tilde{K}(\lambda_{i+1}) - \frac{\tilde{K}(\Lambda_{i+1}) - \tilde{K}(\Lambda_i)}{n-i}.$$

Using this inequality, we obtain

$$c^\varphi(\{i\}) - \mathscr{S}_i(\varphi) = \tilde{K}(\lambda_i) - \sum_{j=1}^{i} \frac{\tilde{K}(\Lambda_j) - \tilde{K}(\Lambda_{j-1})}{n+1-j}$$
$$\leq \tilde{K}(\lambda_{i+1}) - \sum_{j=1}^{i+1} \frac{\tilde{K}(\Lambda_j) - \tilde{K}(\Lambda_{j-1})}{n+1-j} = c^\varphi(\{i+1\}) - \mathscr{S}_{i+1}(\varphi).$$

We conclude that $\mathscr{S}$ satisfies the benefit ordering property under concavity. □

The following example shows that $\mathscr{S}$ neither satisfies the coalitional rationality property nor the benefit ordering property.

Example 8.4. Consider the single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ with $N = \{1,2,3\}$, $\lambda_1 = 1$, $\lambda_2 = 4$, $\lambda_3 = 5$, and the elastic, but not concave, cost function $\tilde{K}$ as in Example 8.3. The associated single-attribute game (N, c^φ) is given by

$$c^\varphi(M) = \begin{cases} 5 & \text{if } |M| = 1; \\ \sum_{i \in M} \lambda_i & \text{otherwise.} \end{cases}$$

The serial rule allocates $\mathscr{S}_1(\varphi) = \tilde{K}(3)/3 = 5/3$, $\mathscr{S}_2(\varphi) = \tilde{K}(3)/3 + (\tilde{K}(9) - \tilde{K}(3))/2 = 11/3$, and $\mathscr{S}_3(\varphi) = 14/3$. Since $\mathscr{S}_1(\varphi) + \mathscr{S}_2(\varphi) = 16/3 > 5 = c^\varphi(\{1,2\})$, we conclude that $\mathscr{S}(\varphi) \notin \mathcal{C}(N, c^\varphi)$. Hence, $\mathscr{S}$ does not satisfy the coalitional rationality property.

Furthermore, since $c^\varphi(\{1\}) - \mathscr{S}_1(\varphi) = 10/3$ and $c^\varphi(\{2\}) - \mathscr{S}_2(\varphi) = 4/3$, player 1 obtains a larger cost saving than player 2, which implies that $\mathscr{S}$ does not satisfy the benefit ordering property either. ◇

We conclude that $\mathscr{S}$ satisfies CA, IR, and BOC, but lacks NM, CR, and BO.

8.5 The benefit-proportional rule

Our next alternative allocation rule is a variation on the proportional rule. Rather than allocating the total *costs* proportional to the attribute of each player, we allocate the *benefits* proportionally instead.

Definition 8.12. The *benefit-proportional rule* $\mathscr{B}$ on $\mathcal{E}$ is defined by allocating $\mathscr{B}_i(N, \tilde{K}, \lambda) = \tilde{K}(\lambda_i) - \big[\sum_{k\in N} \tilde{K}(\lambda_k) - \tilde{K}(\lambda_N)\big] \cdot \lambda_i/\lambda_N$ to player $i \in N$ in situation $(N, \tilde{K}, \lambda) \in \mathcal{E}$.

The following example illustrates this rule and shows that it does not satisfy the coalitional rationality property.

Example 8.5. Reconsider the single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ of Example 8.3. The allocation of the benefit-proportional rule $\mathscr{B}(\varphi)$, which is given by $(3.75, 3.75, 2.5)$, differs from the unique core allocation, $(2.5, 2.5, 5)$; hence $\mathscr{B}(\varphi)$ is not in the core of the associated single-attribute game. This implies that it does not satisfy the coalitional rationality property.

We remark that for this situation, $\mathscr{B}(\varphi)$ assigns less costs to player 3 than to player 2, even though player 3 has a larger attribute. ◇

The benefit-proportional rule is clearly continuous in attributes and thus, by Theorem 8.3, manipulable. The following theorem deals with two more properties.

Theorem 8.5. *The benefit-proportional rule $\mathscr{B}$ satisfies the benefit ordering property and the individual rationality property.*

Proof. Let $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$. Let $i, j \in N$ with $\lambda_i \le \lambda_j$. Then,

$$\begin{aligned} c^{\varphi}(\{i\}) - \mathscr{B}_i(\varphi) &= \tilde{K}(\lambda_i) - \tilde{K}(\lambda_i) + \big[\textstyle\sum_{k\in N} \tilde{K}(\lambda_k) - \tilde{K}(\lambda_N)\big] \cdot \lambda_i/\lambda_N \\ &= \big[\textstyle\sum_{k\in N} \tilde{K}(\lambda_k) - \tilde{K}(\lambda_N)\big] \cdot \lambda_i/\lambda_N \\ &\le \big[\textstyle\sum_{k\in N} \tilde{K}(\lambda_k) - \tilde{K}(\lambda_N)\big] \cdot \lambda_j/\lambda_N \\ &= \tilde{K}(\lambda_j) - \tilde{K}(\lambda_j) + \big[\textstyle\sum_{k\in N} \tilde{K}(\lambda_k) - \tilde{K}(\lambda_N)\big] \cdot \lambda_j/\lambda_N \\ &= c^{\varphi}(\{j\}) - \mathscr{B}_j(\varphi), \end{aligned}$$

where the inequality holds because $\lambda_i \le \lambda_j$. Hence, $\mathscr{B}$ satisfies BO. Furthermore, $\mathscr{B}$ satisfies IR because $\sum_{j\in N} \tilde{K}(\lambda_j) \ge \tilde{K}(\lambda_N)$ by Theorem 2.5; hence, by definition, $\mathscr{B}_i(\varphi) \le c^{\varphi}(\{i\})$ for each $i \in N$. □

We conclude that $\mathscr{B}$ shows that the properties IR and BO are compatible.

8.6 Concavicated marginal rules

This section introduces and analyzes two new rules on elastic single-attribute situations: the concavicated increasing marginal rule and the concavicated average marginal rule. Section

8.6.1 focuses on marginal allocations and the Shapley value in single-attribute situations with concave cost functions. Section 8.6.2 describes how to construct an order-specific concave function under an elastic function. Section 8.6.3 defines the two concavicated rules and analyzes their properties.

8.6.1 Concave single-attribute situations

This subsection considers single-attribute situations with concave cost functions. For the corresponding single-attribute games, we will study marginal allocations and the Shapley value (introduced in Section 2.3.6). We remark that these are rules on games, not on elastic single-attribute situations. We start with a simple preliminary result.

Lemma 8.6. *Let $\varphi = (N, \tilde{K}, \lambda)$ be a single-attribute situation with concave $\tilde{K}$. Then,*

(i) *(N, c^φ) is concave.*

(ii) *For all orderings σ on N, $m^\sigma(N, c^\varphi) \in \mathcal{C}(N, c^\varphi)$.*

(iii) *$\Phi(N, c^\varphi) \in \mathcal{C}(N, c^\varphi)$.*

Proof. *(i)*. Let $i \in N$ and let $M, L \subseteq N \setminus \{i\}$ with $M \subseteq L$. Then,

$$\begin{aligned} c^\varphi(M \cup \{i\}) - c^\varphi(M) &= \tilde{K}(\lambda_M + \lambda_i) - \tilde{K}(\lambda_M) \\ &\geq \tilde{K}(\lambda_M + \lambda_{L\setminus M} + \lambda_i) - \tilde{K}(\lambda_M + \lambda_{L\setminus M}) \\ &= c^\varphi(L \cup \{i\}) - c^\varphi(L), \end{aligned}$$

where the inequality holds because $\tilde{K}$ is concave. This means that (N, c^φ) is concave.

(ii). Follows from Part *(i)* since, by Shapley (1971), any marginal vector is in the core of a concave game.

(iii). Follows from Part *(ii)* since the Shapley value is the average of the marginal vectors and the core is a convex set. □

We next describe a restriction on orderings.

Definition 8.13. Given an elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$, the set $\Theta(\varphi)$ is defined as the set of all orderings on N for which players are ordered in non-decreasing order of their attributes, i.e., $\Theta(\varphi) = \{\sigma \in \Pi(N) \mid \lambda_{\sigma^{-1}(1)} \leq \lambda_{\sigma^{-1}(2)} \leq \ldots \leq \lambda_{\sigma^{-1}(|N|)}\}$.

The following theorem considers such orderings and states that any corresponding marginal allocation yields a proper ordering of the players' benefits.

Theorem 8.7. *Let $\varphi = (N, \tilde{K}, \lambda)$ be a single-attribute situation with concave $\tilde{K}$. Let $\sigma \in \Theta(\varphi)$. Then, $c^\varphi(\{i\}) - m_i^\sigma(N, c^\varphi) \leq c^\varphi(\{j\}) - m_i^\sigma(N, c^\varphi)$ for each $i, j \in N$ with $\sigma(i) \leq \sigma(j)$.*

σ^{-1}	$m_1^\sigma(N,c^\varphi)$	$m_2^\sigma(N,c^\varphi)$	$m_3^\sigma(N,c^\varphi)$
(1,2,3)	1	$\sqrt{2}-1$	$2-\sqrt{2}$
(1,3,2)	1	$2-\sqrt{3}$	$\sqrt{3}-1$
(2,1,3)	$\sqrt{2}-1$	1	$2-\sqrt{2}$
(2,3,1)	$2-\sqrt{3}$	1	$\sqrt{3}-1$
(3,2,1)	$2-\sqrt{3}$	$\sqrt{3}-\sqrt{2}$	$\sqrt{2}$
(3,1,2)	$\sqrt{3}-\sqrt{2}$	$2-\sqrt{3}$	$\sqrt{2}$
Sum	$5-\sqrt{3}$	$5-\sqrt{3}$	$2+2\sqrt{3}$
Average	$\frac{5}{6}-\frac{1}{6}\sqrt{3}$	$\frac{5}{6}-\frac{1}{6}\sqrt{3}$	$\frac{1}{3}+\frac{1}{3}\sqrt{3}$

Table 8.1: *All marginal allocations and their average for the game in Example 8.6.*

Proof. Let $i,j\in N$ with $\sigma(i)\leq\sigma(j)$, which implies that $\lambda_i\leq\lambda_j$. Then,

$$\begin{aligned}
c^\varphi(\{i\})-m_i^\sigma(N,c^\varphi) &= c^\varphi(\{i\})-c(P_i^\sigma\cup\{i\})+c(P_i^\sigma)\\
&= \tilde{K}(\lambda_i)-\tilde{K}(\lambda_{P_i^\sigma}+\lambda_i)+\tilde{K}(\lambda_{P_i^\sigma})\\
&\leq \tilde{K}(\lambda_i+(\lambda_j-\lambda_i))-\tilde{K}(\lambda_{P_i^\sigma}+\lambda_i+(\lambda_j-\lambda_i))+\tilde{K}(\lambda_{P_i^\sigma})\\
&\leq \tilde{K}(\lambda_i+(\lambda_j-\lambda_i))-\tilde{K}(\lambda_{P_j^\sigma}+\lambda_i+(\lambda_j-\lambda_i))+\tilde{K}(\lambda_{P_j^\sigma})\\
&= c^\varphi(\{j\})-m_j^\sigma(N,c^\varphi),
\end{aligned}$$

where the first inequality holds because $\tilde{K}$ is concave, the second inequality because $\sigma(i)\leq\sigma(j)$ and consequently $P_i^\sigma\subseteq P_j^\sigma$. □

The following example provides an illustration and shows that a marginal allocation need not assign the same costs to players with identical attributes.

Example 8.6. Consider the concave single-attribute situation $\varphi=(N,\tilde{K},\lambda)$ with $N=\{1,2,3\}$, $\lambda_1=1$, $\lambda_2=1$, $\lambda_3=2$, and $\tilde{K}(\ell)=\sqrt{\ell}$ for all $\ell\in\mathbb{R}_+$. The associated single-attribute game (N,c^φ) is described by

$$c^\varphi(M)=\begin{cases}1 & \text{if } M=\{1\} \text{ or } M=\{2\};\\ \sqrt{2} & \text{if } M=\{1,2\} \text{ or } M=\{3\};\\ \sqrt{3} & \text{if } M=\{1,3\} \text{ or } M=\{2,3\};\\ 2 & \text{if } M=N.\end{cases}$$

Note that this game is concave. All marginal allocations and their average (i.e., the Shapley value) for this game are described in Table 8.1. Note that they are in the core of (N,c^φ).

The set $\Theta(\varphi)$ contains two orderings: one whose inverse is $(1,2,3)$ and one whose inverse is $(2,1,3)$. Let us consider the first-mentioned one and denote it by σ. So, player 1 is in first position, player 2 is in second position, and player 3 is in last position. The marginal

allocation $m^\sigma(N, c^\varphi)$ according to this ordering is given by $(1, \sqrt{2}-1, 2-\sqrt{2})$. Note that even though players 1 and 2 have identical attributes, they get a different cost assignment. Due to the ordering, player 1's benefit of $c^\varphi(\{1\}) - m_1^\sigma(N, c^\varphi) = 0$ is smaller than player 2's benefit of $c^\varphi(\{2\}) - m_2^\sigma(N, c^\varphi) = 2 - \sqrt{2}$, which in turn is smaller than player 3's benefit of $c^\varphi(\{3\}) - m_3^\sigma(N, c^\varphi) = 2\sqrt{2} - 2$. ◇

This example raised the issue that an arbitrary marginal allocation corresponding to an ordering in $\Theta(\varphi)$ treats identical players differently. This can be avoided by averaging over $\Theta(\varphi)$, as shown in the following theorem.

Theorem 8.8. *Let $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$ with concave $\tilde{K}$. Then,*

$$c^\varphi(\{i\}) - \frac{1}{|\Theta(\varphi)|} \sum_{\sigma \in \Theta(\varphi)} m_i^\sigma(N, c^\varphi) \leq c^\varphi(\{j\}) - \frac{1}{|\Theta(\varphi)|} \sum_{\sigma \in \Theta(\varphi)} m_j^\sigma(N, c^\varphi)$$

for each $i, j \in N$ with $\lambda_i \leq \lambda_j$.

Proof. Let $i, j \in N$ with $\lambda_i \leq \lambda_j$. We distinguish between two cases.

Case 1: $\lambda_i < \lambda_j$. Then, $\sigma(i) < \sigma(j)$ for each $\sigma \in \Theta(\varphi)$. The desired inequality then holds by Theorem 8.7.

Case 2: $\lambda_i = \lambda_j$. Then, obviously, $c^\varphi(\{i\}) = c^\varphi(\{j\})$. Moreover, players i and j are symmetric, and thus for every ordering $\check{\sigma} \in \Theta(\varphi)$ with $\check{\sigma}(i) = a$ and $\check{\sigma}(j) = b$ there exists another ordering $\hat{\sigma} \in \Theta(\varphi)$ with $\hat{\sigma}(i) = b$, $\hat{\sigma}(j) = a$, and $\hat{\sigma}(k) = \sigma(k)$ for all $k \in N \setminus \{i, j\}$. For these orderings, $m_i^{\check{\sigma}}(N, c^\varphi) = m_j^{\hat{\sigma}}(N, c^\varphi)$, so $\sum_{\sigma \in \Theta(\varphi)} m_i^\sigma(N, c^\varphi) = \sum_{\sigma \in \Theta(\varphi)} m_j^\sigma(N, c^\varphi)$. Hence, the desired inequality holds with equality. □

Remarkably, the average of all marginal allocations (i.e., the Shapley value) also has the players' benefits ordered in the same way as their attributes.

Theorem 8.9. *Let $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$ with concave $\tilde{K}$. Then,*

$$c^\varphi(\{i\}) - \Phi_i(N, c^\varphi) \leq c^\varphi(\{j\}) - \Phi_j(N, c^\varphi)$$

for each $i, j \in N$ with $\lambda_i \leq \lambda_j$.

Proof. Define $\alpha(M) = (|M|)! \cdot (|N| - |M| - 1)!$ for all $M \subset N$. (This number $\alpha(M)$ may be interpreted as follows: given a fixed player $k \in N \setminus M$, $\alpha(M)$ is the number of different orderings of N where positions 1 through $|M|$ are are taken by players in M, position $|M|+1$ is taken by player k, and any remaining positions are taken by players in $N \setminus (M \cup \{k\})$. So, $\sum_{M \subseteq N \setminus \{k\}} \alpha(M) = |N|!)$

Let $i, j \in N$ with $\lambda_i \geq \lambda_j$. Recalling the alternative definition of the Shapley value from Section 2.3.6, we obtain

$$
\begin{aligned}
&c^{\varphi}(\{i\}) - \Phi_i(N, c^{\varphi}) \\
&= c^{\varphi}(\{i\}) - \sum_{M \subseteq N \setminus \{i\}} \frac{\alpha(M)}{|N|!} \Big[c^{\varphi}(M \cup \{i\}) - c^{\varphi}(M) \Big] \\
&= \frac{1}{|N|!} \cdot \sum_{M \subseteq N \setminus \{i\}} \alpha(M) \Big[c^{\varphi}(\{i\}) - c^{\varphi}(M \cup \{i\}) + c^{\varphi}(M) \Big] \\
&= \frac{1}{|N|!} \cdot \Big[\sum_{M \subseteq N \setminus \{i\} : j \notin M} \alpha(M) \left(\tilde{K}(\lambda_i) - \tilde{K}(\lambda_M + \lambda_i) + \tilde{K}(\lambda_M) \right) \\
&\qquad + \sum_{M \subseteq N \setminus \{i\} : j \in M} \alpha(M) \left(\tilde{K}(\lambda_i) - \tilde{K}(\lambda_M + \lambda_i) + \tilde{K}(\lambda_M) \right) \Big] \\
&\leq \frac{1}{|N|!} \cdot \Big[\sum_{M \subseteq N \setminus \{i\} : j \notin M} \alpha(M) \left(\tilde{K}(\lambda_i + (\lambda_j - \lambda_i)) - \tilde{K}(\lambda_M + \lambda_i + (\lambda_j - \lambda_i)) + \tilde{K}(\lambda_M) \right) \\
&\qquad + \sum_{M \subseteq N \setminus \{i\} : j \in M} \alpha(M) \left(\tilde{K}(\lambda_i) - \tilde{K}(\lambda_M + \lambda_i) + \tilde{K}(\lambda_M) \right) \Big] \\
&\leq \frac{1}{|N|!} \cdot \Big[\sum_{M \subseteq N \setminus \{i\} : j \notin M} \alpha(M) \left(\tilde{K}(\lambda_j) - \tilde{K}(\lambda_M + \lambda_j) + \tilde{K}(\lambda_M) \right) \\
&\qquad + \sum_{M \subseteq N \setminus \{i\} : j \in M} \alpha(M) \left(\tilde{K}(\lambda_j) - \tilde{K}(\lambda_M + \lambda_i) + \tilde{K}(\lambda_M - \lambda_j + \lambda_i) \right) \Big] \\
&= c^{\varphi}(\{j\}) - \frac{1}{|N|!} \cdot \Big[\sum_{M \subseteq N \setminus \{i\} : j \notin M} \alpha(M) \Big(c^{\varphi}(M \cup \{j\}) - c^{\varphi}(M) \Big) \\
&\qquad + \sum_{M \subseteq N \setminus \{i\} : j \in M} \alpha(M) \Big(c^{\varphi}(M \cup \{i\}) - c^{\varphi}((M \setminus \{j\}) \cup \{i\}) \Big) \Big] \\
&= c^{\varphi}(\{j\}) - \frac{1}{|N|!} \cdot \Big[\sum_{M \subseteq N \setminus \{j\} : i \notin M} \alpha(M) \Big(c^{\varphi}(M \cup \{j\}) - c^{\varphi}(M) \Big) \\
&\quad + \sum_{M \subseteq N \setminus \{j\} : i \in M} \alpha(M) \Big(c^{\varphi}(M \cup \{j\}) - c^{\varphi}(M) \Big) \Big] \\
&= c^{\varphi}(\{j\}) - \frac{1}{|N|!} \cdot \sum_{M \subseteq N \setminus \{j\}} \alpha(M) \Big(c^{\varphi}(M \cup \{j\}) - c^{\varphi}(M) \Big) \\
&= c^{\varphi}(\{j\}) - \Phi_j(N, c^{\varphi}),
\end{aligned}
$$

where both inequalities hold by concavity of $\tilde{K}$. In the first inequality, we use that $\lambda_i \leq \lambda_j$. In the second inequality, we also use $\tilde{K}(\lambda_M - \lambda_j + \lambda_j) - \tilde{K}(\lambda_j) \leq \tilde{K}(\lambda_M - \lambda_j + \lambda_i)) - \tilde{K}(\lambda_i)$, which implies that $\tilde{K}(\lambda_i) + \tilde{K}(\lambda_M) \leq \tilde{K}(\lambda_j) + \tilde{K}(\lambda_M - \lambda_j + \lambda_i)$. □

The following example provides an illustration of Theorems 8.8 and 8.9.

Example 8.7. Reconsider the elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ of Example 8.6. The allocation $\sum_{\sigma \in \Theta(\varphi)} m^\sigma(N, c^\varphi)/|\Theta(\varphi)|$ is given by $(\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2}, 2 - \sqrt{2})$. Under this allocation, the benefit to player 1, $1 - \frac{1}{2}\sqrt{2} \approx 0.29$, is the same as the benefit to player 2, which in turn is smaller than the benefit to player 3, $2\sqrt{2} - 2 \approx 0.83$.

The Shapley value of the game (N, c^φ) is given by $(\frac{5}{6} - \frac{1}{6}\sqrt{3}, \frac{5}{6} - \frac{1}{6}\sqrt{3}, \frac{1}{3} + \frac{1}{3}\sqrt{3})$; see Table 8.1. Under the Shapley value, the benefit to player 1, $c^\varphi(\{1\}) - \Phi_1(N, c^\varphi) = \frac{1}{6}(1 + \sqrt{3}) \approx 0.46$, is the same as the benefit to player 2, which in turn is smaller than the benefit to player 3, $c^\varphi(\{1\}) - \Phi_1(N, c^\varphi) = \sqrt{2} - \frac{1}{3}(1 + \sqrt{3}) \approx 0.50$. ◇

We now know that when the cost function is concave, players with larger attributes get larger benefits under the Shapley value. Since the Shapley value is the average of all marginal vectors, a natural question is whether or not this result extends to all marginal vectors. The following example shows that this is not the case.

Example 8.8. Reconsider the elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ of Example 8.6. Consider the ordering σ on N described by $\sigma^{-1} = (3, 2, 1)$. So, player 3 is in first position, player 2 is in second position, and player 1 is in last position. The marginal allocation $m^\sigma(N, c^\varphi)$ according to this ordering is given by $(2 - \sqrt{3}, \sqrt{3} - \sqrt{2}, \sqrt{2})$. We see that $c^\varphi(\{1\}) - m_1^\sigma(N, c^\varphi) = -1 + \sqrt{3} > 0 = c^\varphi(\{3\}) - m_3^\sigma(N, c^\varphi)$ even though $\lambda_1 < \lambda_3$. ◇

Marginal allocations for single-attribute games are not guaranteed to be stable if the cost function is merely elastic (as opposed to concave). Indeed, in Example 6.3, we saw that the Shapley value can lie outside the core in such a case. In the remainder, we aim to remedy this.

8.6.2 Concave functions under elastic functions

The preceding subsection focused on concave cost functions. We now return to elastic cost functions. Given the positive results of Theorems 8.8 and 8.9 for concave cost functions, we next propose a number of ways—one per ordering of the players—to approximate an elastic function with a concave function.[56] Our ultimate goal is to use marginal allocations on the single-attribute games induced by these concave functions. Any marginal allocation depends on the cost function only through the value of that function at $|N|$ distinct arguments, and we will construct a concave function that approximates the original elastic function as closely as possible at these $|N|$ arguments. These arguments may differ across marginal allocations. However, no marginal allocation depends on the cost function beyond the maximum argument λ_N, which allows us to restrict ourselves to constructing a function with domain $[0, \lambda_N]$. Hence, we will construct a concave function on $[0, \lambda_N]$ for every possible ordering of the players. This function will be made up of straight, consecutive line segments.

[56] Sometimes there is only one way. Take spare parts games, for example. As shown in Figure 6.1 on page 129, if multiple base stock levels are optimal for the grand coalition, then there is no flexibility: the relevant tangent line goes through the origin, as shown in Lemma 6.6.

Definition 8.14. For any $\lambda, \mu, q, r \in \mathbb{R}_+$ with $\lambda < \mu$ and $q \leq r$, we define the function $L^{(\mu,r)}_{(\lambda,q)} : \mathbb{R} \to \mathbb{R}$ by

$$L^{(\mu,r)}_{(\lambda,q)}(x) = \frac{r-q}{\mu-\lambda}(x-\mu) + r \qquad \text{for all } x \in \mathbb{R} .$$

This function should be interpreted as the straight line through the points (λ, q) and (μ, r).

We next describe how to draw an order-specific concave function under an elastic function. Figure 8.2 may prove helpful as an illustration.

Procedure 8.1. We are given an elastic single-attribute situation $(N, \tilde{K}, \lambda)$ and an ordering σ on N. For notational ease, write $n = |N|$. Define $\Lambda^\sigma_0 = 0$ and $\Lambda^\sigma_i = \lambda_{\sigma^{-1}(1)} + \ldots + \lambda_{\sigma^{-1}(i)}$ for all $i \in \{1, 2, \ldots, n\}$. Note that $\Lambda_n = \lambda_N$. We now aim to construct a continuous function that is made up of line segments between points $(\Lambda^\sigma_0, Q^\sigma_0)$, $(\Lambda^\sigma_1, Q^\sigma_1)$, $\ldots$, $(\Lambda^\sigma_n, Q^\sigma_n)$ such that the resulting function is non-negative, concave, non-decreasing, and not above $\tilde{K}$ on $[0, \Lambda^\sigma_n]$. We now present the procedure for determining the numbers $Q^\sigma_0, Q^\sigma_1, \ldots, Q^\sigma_n$.

We start from the right and set $Q^\sigma_n = \tilde{K}(\Lambda^\sigma_n)$. We then fix Q^σ_{n-1} by drawing the highest possible line from $(\Lambda^\sigma_n, Q^\sigma_n)$ to $(\Lambda^\sigma_{n-1}, Q^\sigma_{n-1})$ that is non-negative and not above $\tilde{K}$ on the interval $[\Lambda^\sigma_{n-1}, \Lambda^\sigma_n]$. That is, we take

$$Q^\sigma_{n-1} = \max\left\{ q \in \left[L^{(\Lambda^\sigma_n, Q^\sigma_n)}_{(0,0)}(\Lambda^\sigma_{n-1}), \tilde{K}(\Lambda^\sigma_{n-1}) \right] \,\middle|\, L^{(\Lambda^\sigma_n, Q^\sigma_n)}_{(\Lambda^\sigma_{n-1}, q)}(\ell) \leq \tilde{K}(\ell) \;\forall\; \ell \in [\Lambda^\sigma_{n-1}, \Lambda^\sigma_n] \right\}.$$

The number Q^σ_{n-1} is well-defined because $\tilde{K}$ is elastic.[57]

For $j \in \{0, \ldots, n-2\}$, recursively, we then fix Q^σ_j by drawing the highest possible line from $(\Lambda^\sigma_{j+1}, Q^\sigma_{j+1})$ to $(\Lambda^\sigma_j, Q^\sigma_j)$ that is non-negative and not above $\tilde{K}$ on the interval $[\Lambda^\sigma_j, \Lambda^\sigma_{j+1}]$ and, moreover, that is at least as steep as the previous line. That is, we take

$$Q^\sigma_j = \max\left\{ q \in \left[L^{(\Lambda^\sigma_{j+1}, Q^\sigma_{j+1})}_{(0,0)}(\Lambda^\sigma_j), L^{(\Lambda^\sigma_{j+2}, Q^\sigma_{j+2})}_{(\Lambda^\sigma_{j+1}, Q^\sigma_{j+1})}(\Lambda^\sigma_j) \right] \,\middle|\, L^{(\Lambda^\sigma_{j+1}, Q^\sigma_{j+1})}_{(\Lambda^\sigma_j, q)}(\ell) \leq \tilde{K}(\ell) \;\forall\; \ell \in [\Lambda^\sigma_j, \Lambda^\sigma_{j+1}] \right\},$$

The number Q^σ_j is well-defined because, again, $\tilde{K}$ is elastic.[58]

Given the numbers $Q^\sigma_0, Q^\sigma_1, \ldots, Q^\sigma_n$ as defined above, the σ-concavication $\tilde{K}^{\text{conc}}_\sigma : [0, \lambda_N] \to \mathbb{R}_+$ is given by

$$\tilde{K}^{\text{conc}}_\sigma(\ell) = L^{(\Lambda^\sigma_j, Q^\sigma_j)}_{(\Lambda^\sigma_{j-1}, Q^\sigma_{j-1})}(\ell) \text{ for } j \in \{1, \ldots, n\} \text{ with } \ell \in [\Lambda^\sigma_{j-1}, \Lambda^\sigma_j].$$

for all $\ell \in (0, \lambda_N]$ and $\tilde{K}^{\text{conc}}_\sigma(0) = 0$.

[57] The set over which we take the maximum is non-empty because if we would take $q = L^{(\Lambda^\sigma_n, Q^\sigma_n)}_{(0,0)}(\Lambda_{n-1})$, then the resulting line is not above $\tilde{K}$ on $[\Lambda^\sigma_{n-1}, \Lambda^\sigma_n]$ by Theorem 2.2. The maximum actually exists because if $n > 1$ then $\tilde{K}$ is continuous on $[\Lambda^\sigma_{n-1}, \Lambda^\sigma_n]$ by Theorem 2.2; otherwise, if $n = 1$ then $Q_0 = 0$ because the interval from which we are to pick q only includes 0.

[58] The set over which we take the maximum is non-empty, for two reasons. First, the interval from which we are to pick q is non-empty because, by construction, $L^{(\Lambda^\sigma_{j+2}, Q^\sigma_{j+2})}_{(0,0)}(\Lambda_{j+1}) \leq L^{(\Lambda_{j+2}, Q_{j+2})}_{(\Lambda_{j+1}, Q_{j+1})}(\Lambda_{j+1})$; this implies that $L^{(\Lambda_{j+1}, Q_{j+1})}_{(0,0)}$ is steeper than $L^{(\Lambda_{j+2}, Q_{j+2})}_{(\Lambda_{j+1}, Q_{j+1})}$. Second, by construction, $Q^\sigma_{j+1} \leq \tilde{K}(\Lambda^\sigma_{j+1})$; hence, if we would take $q = L^{(\Lambda^\sigma_{j+1}, Q^\sigma_{j+1})}_{(0,0)}(\Lambda_j)$, then on $[0, \Lambda_{j+1}]$ the resulting line is not above $L^{(\Lambda^\sigma_{j+1}, \tilde{K}(\Lambda^\sigma_{j+1}))}_{(0,0)}$, which in turn is not above $\tilde{K}$ by Theorem 2.2. The maximum actually exists because if $j \geq 1$ then $\tilde{K}$ is continuous on $[\Lambda^\sigma_{n-1}, \Lambda^\sigma_n]$ by Theorem 2.2; otherwise, if $j = 0$ then $\tilde{K}(0) = 0$ implies $Q_j = 0$.

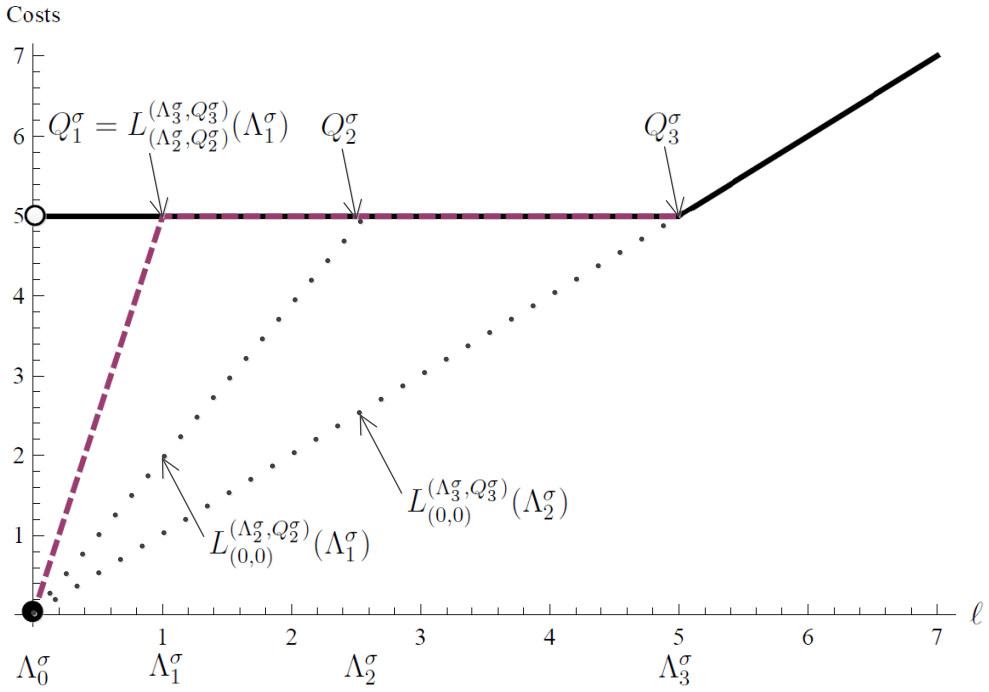

Figure 8.2: *The function $\tilde{K}$ from Example 8.9 and a σ-concavication (dashed).*

The following lemma collects several properties of a σ-concavication, which follow directly from its definition.

Lemma 8.10. *Let $(N, \tilde{K}, \lambda) \in \mathcal{E}$, and let σ be an ordering on N.*

(i) $\tilde{K}_\sigma^{conc}$ *is concave.*

(ii) $\tilde{K}_\sigma^{conc}(\ell) \leq \tilde{K}(\ell)$ *for all* $\ell \in [0, \lambda_N)$ *and* $\tilde{K}_\sigma^{conc}(\lambda_N) = \tilde{K}(\lambda_N)$.

(iii) If $\tilde{K}$ *is concave, then* $\tilde{K}_\sigma^{conc}(\Lambda_i^\sigma) = \tilde{K}(\Lambda_i^\sigma)$ *for all* $i \in \{0, 1, \ldots, n\}$.

(iv) $\tilde{K}_\sigma^{conc}(0) = 0$.

The following example illustrates the construction of a σ-concavication.

Example 8.9. Consider the single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ with $N = \{1, 2, 3\}$, $\lambda_1 = 1$, $\lambda_2 = 1.5$, $\lambda_3 = 2.5$, and elastic cost function $\tilde{K}$ equal to the function considered in Example 8.3, which is for convenience represented again in Figure 8.2. For the ordering σ with $\sigma^{-1} = (1, 2, 3)$, the σ-concavication $\tilde{K}_\sigma^{\text{conc}} : [0, 5] \to \mathbb{R}$ corresponding to attribute vector λ is given by

$$\tilde{K}_\sigma^{\text{conc}}(\ell) = \begin{cases} 5\ell & \text{if } \ell \in [0, 1]; \\ 5 & \text{if } \ell \in (1, 5]. \end{cases}$$

See Figure 8.2. For other orderings, the construction is similar.

If, however, we would change the situation by increasing one player's attribute by $\epsilon > 0$, then it is easily inferred from Figure 8.2 that for *every* ordering σ on N, the resulting σ-concavication $\tilde{K}_\sigma^{\text{conc}} : [0, 5+\epsilon] \to \mathbb{R}$ is given by $\tilde{K}_\sigma^{\text{conc}}(\ell) = \ell$ for all $\ell \in [0, 5+\epsilon]$. So, a small change in the attribute vector may lead to a large change in the σ-concavication. $\diamond$

The following example illustrates the construction of a σ-concavication for a more complicated situation.

Example 8.10. Consider the elastic single-attribute situation $(N, \tilde{K}, \lambda)$ with $N = \{1, 2, 3\}$, $\lambda_1 = 1$, $\lambda_2 = 4$, $\lambda_3 = 6$, and cost function $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$\tilde{K}(\ell) = \begin{cases} 4\ell & \text{if } \ell \in [0,1]; \\ 4 & \text{if } \ell \in (1,3]; \\ 2 + \ell \cdot 2/3 & \text{if } \ell \in (3,6]; \\ 6 + ((\ell-6)/4)^2 & \text{if } \ell \in (6,10]; \\ 7 & \text{if } \ell > 10; \end{cases}$$

This function and each of its σ-concavications are graphically represented in Figure 8.3. The function $\tilde{K}$ is elastic because

- for ℓ on $(0,1]$, $\tilde{K}(\ell)/\ell = 4$;
- for ℓ on $(1,3]$, $\tilde{K}(\ell)/\ell = 4/\ell$ is decreasing in ℓ and ranges from 4 to $4/3$;
- for ℓ on $(3,6]$, $\tilde{K}(\ell)/\ell = 2/3 + 2/\ell$ is decreasing in ℓ and ranges from $4/3$ to 1;
- for ℓ on $(6,10]$, $\tilde{K}(\ell)/\ell = 6/\ell + (\ell - 12 + 36/\ell)/16$ is decreasing[59] in ℓ and ranges from 1 to $7/10$;
- for $\ell > 10$, $\tilde{K}(\ell)/\ell = 7/\ell$ is decreasing in ℓ with a maximal value of $7/10$.

We illustrate the construction of a σ-concavication for two orderings. First, consider the ordering σ with $\sigma^{-1} = (2, 3, 1)$. So, player 2 is in first position, player 3 is in second position, and player 1 is in third position. Hence, $\Lambda_1^\sigma = 4$, $\Lambda_2^\sigma = 10$, and $\Lambda_3^\sigma = 11$. Obviously, $Q_3^\sigma = 7$, $Q_2^\sigma = 7$, and $Q_0^\sigma = 0$. Consider Q_1^σ, i.e., the largest $q \leq \tilde{K}(\Lambda_1^\sigma) = 4\frac{2}{3}$ such that $L_{(4,q)}^{(10,7)}(\ell) \leq \tilde{K}(\ell)$ for all $\ell \in [4, 10]$. Since the derivative of $6 + ((\ell-6)/4)^2$ evaluated at $\ell = 10$ is equal to 0.5, any $q > \tilde{K}(\Lambda_2^\sigma) - 0.5 \cdot (\Lambda_2^\sigma - \Lambda_1^\sigma) = 4$ would result in going above the graph of $\tilde{K}$. Yet, $q = 4$ would not take us above the graph of $\tilde{K}$, so $Q_1^\sigma = 4$.

Next, consider the ordering σ with $\sigma^{-1} = (3, 1, 2)$. So, player 3 is in first position, player 1 is in second position, and player 2 is in third position. Hence, $\Lambda_1^\sigma = 6$, $\Lambda_2^\sigma = 7$, $\Lambda_3^\sigma = 11$. Obviously, $Q_3^\sigma = 7$ and $Q_0^\sigma = 0$. Consider Q_2^σ, i.e., the largest $q \leq \tilde{K}(\Lambda_2^\sigma) = 6\frac{1}{16}$ such that $L_{(7,q)}^{(11,7)}(\ell) \leq \tilde{K}(\ell)$ for all $\ell \in [7, 11]$. For $q \in [5, 6\frac{1}{16}]$, standard optimization techniques

[59] Indeed, its derivative $(1 - 132/\ell^2)/16$ is negative for ℓ on $(6, 10]$.

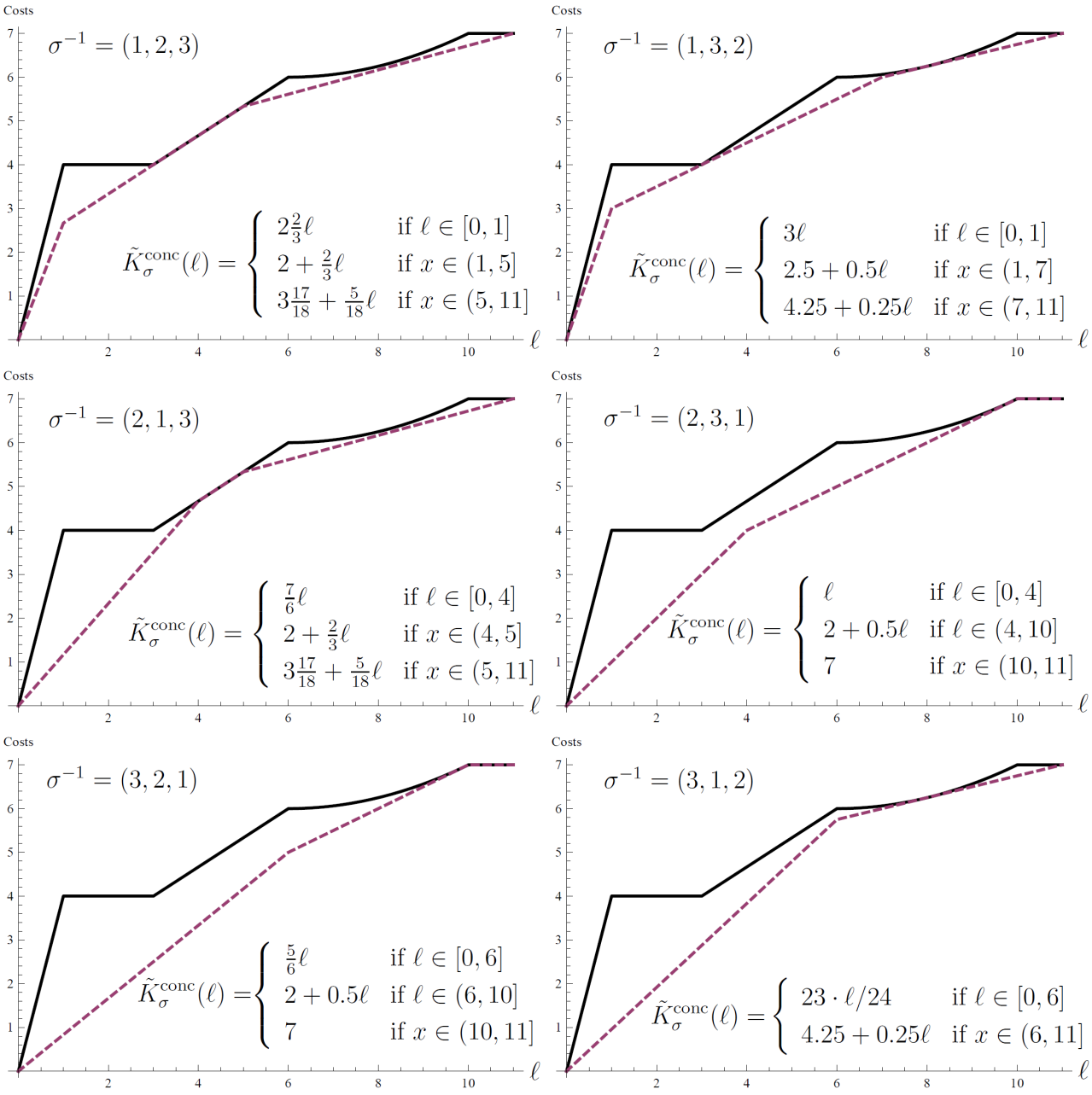

Figure 8.3: *The function $\tilde{K}$ from Example 8.10 and all its σ-concavications (dashed).*

reveal that $6 + ((\ell - 6)/4)^2 - L^{(11,7)}_{(7,q)}(\ell)$ as a function of ℓ on $[7, 11]$ has a unique minimizer $\ell^*(q) = 20 - 2q$. Consequently, $6 + ((\ell^*(q) - 6)/4)^2 = L^{(11,7)}_{(7,q)}(\ell^*(q))$ if and only if $q = 6$. Hence, $Q_2^\sigma = 6$. Finally, the line $L^{(11,7)}_{(7,6)}$ is not above $\tilde{K}$ on $[6, 7]$, so $Q_2^\sigma = L^{11,7)}_{(7,6)}(6) = 5.75$.

The construction of the σ-concavication for the other orderings is similar; see Figure 8.3. ◇

8.6.3 Two concavicated rules

Based on σ-concavicated situations, we will introduce two allocation rules on elastic single-attribute situations.

Definition 8.15. For an elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ and an ordering σ on N, we call $\varphi(\sigma) = (N, \tilde{K}^{\text{conc}}_\sigma, \lambda)$ the corresponding *σ-concavicated situation*.

We remark that although the domain of $\tilde{K}^{\text{conc}}_\sigma$ is $[0, \lambda_N]$ and not $\mathbb{R}_+$, this does not pose a problem for the two allocation rules that we define next, as they do not depend on the cost function beyond λ_N.

Definition 8.16. The *concavicated increasing marginal rule* $\mathscr{M}$ on $\mathcal{E}$ assigns to each elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ the allocation $\mathscr{M}(\varphi) = \sum_{\sigma \in \Theta(\varphi)} m^\sigma(N, c^{\varphi(\sigma)})/|\Theta(\varphi)|$.

Definition 8.17. The *concavicated average marginal rule* $\mathscr{A}$ on $\mathcal{E}$ assigns to each elastic single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ the allocation $\mathscr{A}(\varphi) = \sum_{\sigma \in \Pi(N)} m^\sigma(N, c^{\varphi(\sigma)})/|\Pi(N)|$.

The following lemma states that if the cost function is concave, then marginal allocations are unchanged by concavicated situations.

Lemma 8.11. *Let $\varphi = (N, \tilde{K}, \lambda) \in \mathcal{E}$ with concave $\tilde{K}$.*
(i) For all orderings σ on N and all $i \in N$, it holds that $m_i^\sigma(N, c^\varphi) = m_i^\sigma\left(N, c^{\varphi(\sigma)}\right)$.
(ii) $\mathscr{M}(\varphi) = \sum_{\sigma \in \Theta(\varphi)} m^\sigma(N, c^\varphi)/|\Theta(\varphi)|$.
(iii) $\mathscr{A}(\varphi) = \Phi(N, c^\varphi)$.

Proof. For Part (i), let $\sigma \in \Pi(N)$ and $i \in N$. by Part (iii) of Lemma 8.10, $\tilde{K}^{conc}(\Lambda^\sigma_{\sigma(i)}) - \tilde{K}^{conc}(\Lambda^\sigma_{\sigma(i)-1}) = \tilde{K}^{conc}(\lambda_{P_i^\sigma} + \lambda_i) - \tilde{K}^{conc}(\lambda_{P_i^\sigma})$, which implies that $m_i^\sigma(N, c^\varphi) = m_i^\sigma\left(N, c^{\varphi(\sigma)}\right)$.

Part (ii) and (iii) follow immediately from Part (i). □

The concavicated rules only evaluate a σ-concavication at the arguments Λ^σ_0, Λ^σ_1, …, Λ^σ_n. The intermediate behavior of the σ-concavication, however, is still important because it affects the corresponding game. Despite not taking into account this intermediate behavior, our concavicated rules still accomplish core allocations. Before proving this, we provide an illustration.

M	λ_M	$c^\varphi(M)$	$\sum_{i\in M}\mathscr{M}_i(\varphi)$	$\sum_{i\in M}\mathscr{A}_i(\varphi)$
$\{1\}$	1	1	$2\frac{2}{3}$	$1\frac{7}{72}$
$\{2\}$	4	$4\frac{2}{3}$	$2\frac{2}{3}$	$2\frac{5}{9}$
$\{3\}$	6	6	$1\frac{2}{3}$	$3\frac{25}{72}$
$\{1,2\}$	5	$5\frac{1}{3}$	$5\frac{1}{3}$	$3\frac{47}{72}$
$\{1,3\}$	7	$6\frac{1}{16}$	$4\frac{1}{3}$	$4\frac{32}{72}$
$\{2,3\}$	10	7	$4\frac{1}{3}$	$5\frac{65}{72}$
N	11	7	7	7

Table 8.2: *The game and concavicated marginal allocations in Example 8.11.*

σ^{-1}	$m_1^\sigma\left(N,c^{\varphi(\sigma)}\right)$	$m_2^\sigma\left(N,c^{\varphi(\sigma)}\right)$	$m_3^\sigma\left(N,c^{\varphi(\sigma)}\right)$
(1,2,3)	$2\frac{2}{3}$	$2\frac{2}{3}$	$1\frac{2}{3}$
(1,3,2)	3	1	3
(2,1,3)	$\frac{2}{3}$	$4\frac{2}{3}$	$1\frac{2}{3}$
(2,3,1)	0	4	3
(3,2,1)	0	2	5
(3,1,2)	$\frac{1}{4}$	1	$5\frac{3}{4}$
Sum	$6\frac{7}{12}$	$15\frac{1}{3}$	$20\frac{1}{12}$

Table 8.3: *The marginal allocations corresponding to all σ-concavicated situations in Example 8.11.*

Example 8.11. Reconsider the single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ from Example 8.10. The associated single-attribute game (N, c^φ) is described in Table 8.2. For any ordering σ on N, the cost function for the corresponding σ-concavicated situation $\varphi(\sigma) = (N, \tilde{K}_\sigma^{\text{conc}}, \lambda)$ is given in Figure 8.3, and the corresponding marginal allocation is described in Table 8.3.

The concavicated increasing marginal allocation $\mathscr{M}(\varphi)$ is given by $(2\frac{2}{3}, 2\frac{2}{3}, 1\frac{2}{3})$ because $\Theta(\varphi)$ only consists of the ordering σ with $\sigma^{-1} = (1,2,3)$. The concavicated average marginal allocation $\mathscr{M}(\varphi)$ is obtained by averaging all marginal vectors in Table 8.3, which results in $(1\frac{7}{72}, 2\frac{5}{9}, 3\frac{25}{72})$. Note that $\mathscr{A}$ is *not* the Shapley value of a "straightforwardly" derived game because each marginal allocation is based on a different σ-concavicated situation and thus on a different game.

It is easy to infer from Table 8.2 that both $\mathscr{M}(\varphi)$ and $\mathscr{A}(\varphi)$ are stable allocations for (N, c^φ). ◇

Theorem 8.12. *Both $\mathscr{M}$ and $\mathscr{A}$ satisfy the coalitional rationality property on $\mathcal{E}$.*

Proof. Let $\varphi = (N, \tilde{K}, \lambda)$ be an elastic single-attribute situation, and let σ be any ordering on N. Then, by Part (i) of Lemma 8.6, the single-attribute game associated with the

σ-concavicated situation $\varphi(\sigma)$ is concave. By Part (ii) of Lemma 8.6, it follows that $m^\sigma\left(N, c^{\varphi(\sigma)}\right) \in \mathcal{C}\left(N, c^{\varphi(\sigma)}\right)$. By Part (ii) of Lemma 8.10, it holds that $\mathcal{C}\left(N, c^{\varphi(\sigma)}\right) \subseteq \mathcal{C}(N, c^\varphi)$. Hence, $m^\sigma\left(N, c^{\varphi(\sigma)}\right) \in \mathcal{C}(N, c^\varphi)$ as well. Since $\mathscr{M}(\varphi)$ and $\mathscr{A}(\varphi)$ are defined as averages of marginal allocations for games associated with concavicated situations, their coalitional rationality is immediate. □

Both $\mathscr{M}$ and $\mathscr{A}$ lack the benefit ordering property because, by Theorem 8.12, they prescribe the unique core allocation for the situation of Example 8.3, which described our impossibility result. However, both rules satisfy the relaxation BOC.

Theorem 8.13. *Both $\mathscr{M}$ and $\mathscr{A}$ satisfy the benefit ordering property under concavity.*

Proof. Let $\varphi = (N, \tilde{K}, \lambda)$ be any elastic single-attribute situation with concave $\tilde{K}$, and let σ be any ordering on N. By Part (i) of Lemma 8.10, $\tilde{K}^{conc}$ is concave. Hence, Lemma 8.11 applies. Benefit ordering under concavity of $\mathscr{M}$ then follows from Theorem 8.8 and from Part (ii) of Lemma 8.11, while benefit ordering under concavity of $\mathscr{A}$ follows from Theorem 8.9 and from Part (iii) of Lemma 8.11. □

The following example shows that both concavicated rules lack non-manipulability and continuity in attributes.

Example 8.12. Reconsider the single-attribute situation $\varphi = (N, \tilde{K}, \lambda)$ from Example 8.9. Here, $\mathscr{M}(\varphi) = (5, 0, 0)$, while $\mathscr{A}(\varphi) = \left(\frac{5}{3}, \frac{5}{3}, \frac{5}{3}\right)$. We remark that the smallest player does not get any benefits at all under $\mathscr{M}$, whereas $\mathscr{A}$ does not exhibit such extreme behavior.

By merging players 1 and 3 into player 4, we obtain the manipulation $\bar{\varphi} = (\bar{N}, \tilde{K}, \bar{\lambda})$ with $\bar{N} = \{2, 4\}$, $\bar{\lambda}_2 = 1.5$, and $\bar{\lambda}_4 = 3.5$. Then, $\mathscr{M}_2^{\text{conc}}(\bar{\varphi}) = 5$, $\mathscr{M}_4^{\text{conc}}(\bar{\varphi}) = 0$, $\mathscr{A}_2^{\text{conc}}(\bar{\varphi}) = 2.5$, and $\mathscr{A}_4^{\text{conc}}(\bar{\varphi}) = 2.5$. So, under both rules, players 1 and 3 have an incentive to merge, which means that neither satisfies non-manipulability.

By taking the sequence of attribute vectors $(\mathcal{L}^k)_{k=1}^\infty$ and corresponding situations $\varphi(k)$ as described in Example 8.9, we observe that $\lim_{k\to\infty} \mathscr{M}_1(\varphi(k)) = 1 \neq 5 = \mathscr{M}_1(\varphi)$ and $\lim_{k\to\infty} \mathscr{A}_1(\varphi(k)) = 1 \neq \frac{5}{3} = \mathscr{A}_1(\varphi)$. So, neither $\mathscr{M}$ nor $\mathscr{A}$ satisfy continuity in attributes. ◇

We conclude that both $\mathscr{M}$ and $\mathscr{A}$ satisfy CR, IR, and BOC, but lack NM, CA, and BO.

8.7 Conclusion

Table 8.4 presents an overview of our main results and a legend for our abbreviations. We have shown that $\mathscr{P}$ is the unique rule satisfying the combination of NM and CA. We have also shown that CR and BO are incompatible. At the same time, we have found two rules that satisfy the combination of CR and BOC: $\mathscr{M}$ and $\mathscr{A}$. Accordingly, if we desire to improve on the proportional rule with regard to the ordering of players' benefits, while keeping coalitional rationality intact, then these rules would be appealing solutions. Remarkably, $\mathscr{A}$ coincides

Rule	NM	CA	CR	IR	BO	BOC
$\mathscr{P}$	√	√	√	√	X	X
$\mathscr{S}$	X	√	X	√	X	√
$\mathscr{B}$	X	√	X	√	√	√
$\mathscr{M}$	X	X	√	√	X	√
$\mathscr{A}$	X	X	√	√	X	√

Table 8.4: *Overview of the various rules and their properties. Legend for rules: $\mathscr{P}$ is the proportional rule, $\mathscr{S}$ is the serial rule, $\mathscr{B}$ is the benefit-proportional rule, $\mathscr{M}$ is the concavicated increasing marginal rule, and $\mathscr{A}$ is the concavicated average marginal rule. Legend for properties: NM is non-manipulability, CA is continuity in attributes, CR is the coalitional rational property, IR is the individual rationality property, BO is the benefit ordering property, and BOC is the benefit ordering property under concavity.*

with the Shapley value — one of the most celebrated solutions for cooperative games — when the cost function is concave. Yet, $\mathscr{A}$ remedies the possible non-stability of the Shapley value when the cost function is merely elastic.

We close out by providing three directions for future research. A first direction would be on alternative concavications. Indeed, the collection of functions described in Definition 8.1 are not the *only* concave functions that fit under an elastic function. Although we believe that the procedure of Definition 8.1 is compelling because of the property described in Part (*iii*) of Lemma 8.10 and because, from a computational perspective, it merely requires the determination of $|N|$ straight line segments, future research may look for alternative concavications that retain these nice properties while being continuous in the attribute vector.

A second direction for future research is on other fairness properties and other allocation rules. There are, of course, many other allocation cost rules possible for elastic single-attribute situations beyond the ones that we considered. The decreasing serial rule proposed in de Frutos (1998) might be an interesting one. At the same time, we have not exhausted the list of reasonable fairness criteria. For example, population monotonicity (cf. Sprumont, 1990) might be interesting.

A third and final possible research direction is on a study of the properties exhibited by rules on a restricted domain. Indeed, all the properties we have considered deal with the domain of elastic single-attribute situations. If we would restrict the domain of a rule to a specific class of situations (e.g., whose cost function represents the optimal costs in an Erlang loss model) then it is possible that a rule would exhibit specific behavior on that restricted domain.

—Essentially, all models are wrong, but some are useful.

George Box

9 Conclusion

9.1 Game over?

In this thesis, we studied the cost allocation problem that arises in stochastic inventory and queueing systems with pooled resources. Our general research problem, described in Section 1.3, was to identify (existence of) stable allocations of the total costs—stable in the sense that all service providers are motivated to join forces, i.e., such that no subset of players has an incentive to split off and form a separate pooling group. To do this, we formulated new classes of resource pooling games and derived structural properties of these games.

In this concluding chapter, we tie things together by discussing key results, insights, and common threads: We first describe and illustrate the main results by means of examples in Section 9.2, subsequently summarize a number of important lessons and surprises in Section 9.3, and close out with future research directions in Section 9.4.

9.2 Description and illustration of the main results

Tables 9.1 and 9.2 present an overview of the main stability results for Chapters 4, 5, and 6 under optimized and fixed numbers of resources, respectively. To illustrate these results, we revisit the two examples from the introductory chapter. We will first discuss and illustrate results for optimized numbers of resources and subsequently deal with results for fixed numbers of resources.

	Queueing problem	Inventory problem
No waiting allowed	Proportional allocation stable	Proportional allocation stable
Waiting possible	Stable allocation need not exist	Proportional allocation stable

Table 9.1: *Main results of Chapters 4, 5, and 6 under optimized numbers of resources.*

	Queueing problem	Inventory problem
No waiting allowed	Stable allocation exists	Stable allocation exists
Waiting possible	Stable allocation exists	Pipeline allocation stable

Table 9.2: *Main results of Chapters 4, 5, and 6 under fixed numbers of resources.*

Example 9.1. (Optimized numbers of resources.) Reconsider the two airlines from Example 1.1. This setting can be accurately modeled using the Erlang loss game with optimized base stock levels from Chapter 4. For this model, we showed in Section 4.5.3 that the allocation of expected costs proportional to the players' demand rates is always stable and reachable through a population monotonic allocation scheme (PMAS). Note that this concerns a proportional allocation of the *costs*, not the benefits. And that the allocation is in proportion to the *demand rates*, not in proportion to the individual optimal costs, individual optimal base stock levels, or anything else.

Stability means that if player A faces λ_A demands per year and player B faces λ_B demands per year (in expectation, of course), then both players are guaranteed to become better off in expecation if they join forces and player A would pay a fraction $\lambda_A/(\lambda_A + \lambda_B)$ of the total costs and player B would pay a fraction $\lambda_B/(\lambda_A + \lambda_B)$.

Population monotonicity means that if a player C would join and the resulting new costs would still be shared in proportion to demand rates, then players A and B would get better off. Moreover, players A, B, and C would get better off if a player D would join. And so on. This yields a snowball effect that stimulates growth of the pool.

None of the above results depend on the numerical values of the demand rates or cost parameters; these are general results that follow from structural characteristics of the Erlang loss model. ◇

This example illustrated the result that, when each coalition uses an optimal number of resources, proportional allocations "work" in the $(S-1, S)$ inventory model with lost sales. This is also true in several other models with optimized numbers of resources, but not all. As shown in Sections 4.5.3 and 6.5.2, proportional allocations are also stable and PMAS-reachable in the $M/G/s/s$ queueing model and in the $(S-1, S)$ inventory model with backlogging, respectively. However, as shown in Section 5.5, the proportional allocation is merely close to stable in the $M/M/s$ queueing model. Table 9.1 provides a summary.

There are two important common threads through our models with optimized numbers of resources. First, all of them featured Poisson demand processes, which ensures that proportional allocations are easily implemented via a simple process for cost realizations, as shown in Section 6.5.4. Second, stability of a proportional allocation is guaranteed only when players are associated with a single attribute and the optimal cost function satisfies the elasticity property—a property that only the $M/M/s$ queueing model lacked.

Example 9.2. (Fixed numbers of resources.) Reconsider the service departments from Example 1.2. This setting can be accurately modeled using the Erlang delay game with fixed numbers of servers that we studied in Chapter 5. For this model, we showed in Section 5.4 that if all players have the *same* resources-to-demand ratio, then a proportional allocation "works". However, if the characteristics of a player cannot be adequately captured in a single real number, e.g., when a player brings in a large number of resources but a relatively low demand rate, then a proportional cost allocation may fall outside the core.

Nevertheless, a stable allocation of expected costs is always guaranteed to exist, irrespective of how many players there are, how many servers everyone brings in, or what their demand rates are. Given this result, certain allocation rules from cooperative game theory, such as the nucleolus, are then guaranteed to transform the cost figures of all coalitions into a stable allocation. ◇

Under fixed numbers of resources, this example illustrated the result that that a stable allocation always exists in the $M/M/s$ queueing model. A stable allocation was also guaranteed to exist for the $M/G/s/s$ queueing model and for the $(S-1,S)$ inventory model with lost sales. We went beyond existence in Section 6.4 for the $(S-1,S)$ inventory model with backlogging, where we showed that an allocation based on pipeline stock realizations is stable. Table 9.2 provides a summary.

In Chapter 7, we considered resource pooling games derived from a more general inventory model featuring non-stationary, dynamic demand processes. Despite their inherent complexity, we proved that these games still always admit a stable allocation.

In Chapter 8, we analyzed allocation rules on the class of elastic single-attribute situations, which includes, e.g., the setting of Example 9.1. We showed that the proportional rule is axiomatically characterized by continuity and non-manipulability (see Section 8.2). Proportional allocations do come with the downside that a smaller player might reap more benefits than a bigger player (see Section 8.3). To avoid this issue, we introduced concavicatied marginal rules in Section 8.6 and proved that, when the underlying cost function is concave, they accomplish a stable allocation that assigns larger benefits to larger players.

All these are elegant, arguably beautiful, results from a theoretical perspective. Yet, they also have an important practical implication: for the settings considered, resource pooling arrangements are not only beneficial for the society as a whole, but can also be supported by a stable cost allocation. This means that more collaborations could take place in practice.

Currently, there are many places where resources are not shared (e.g., because people are afraid a shared resource will not available at a time they need it) but this thesis has shown that as long as an adequate cost allocation rule is used, resource pooling is often a valid solution that makes everyone better off. This insight, when combined with the appealing cost allocation mechanisms that we devised, should pave the way for sustainable collaborations in practice.

9.3 Lessons and surprises

The following list provides a collection of lessons learned and surprises faced. Although they are based on the scientific content of this thesis, they are written from a personal perspective: interspersed are opinions, recollections, and musings.

Proportional allocations are often stable: At the start, the identification of an allocation rule satisfying a number of requirements (stability, population monotonicity, etcetera) seemed like a daunting task, especially given the complexity of the underlying queueing or inventory systems. Surprisingly, however, a very simple proportional allocation turned out to satisfy all the desired properties in many models! It was not obvious that such an allocation would always keep every possible coalition happy, but in most models the optimal cost function satisfied the essential elasticity property. The importance of this property was nicely described in the work of Özen et al. (2011), who independently and simultaneously derived some of the same results as I did. Interestingly, the expected cost allocations prescribed by the proportional rule coincide with many common practices, e.g., charging a fixed fee per flight hour for participation in an aircraft component pool (assuming that all costs are fully shared and that component failure rates per flight hour are the same across players). Thus, the results in this thesis provide support for these flight-hour charges and other similar pricing schemes from a game theoretical perspective.

Integrality makes life difficult: If only it would always be possible to set resources levels at arbitrary real numbers. Alas, resources can typically only be varied in discrete amounts, and the difference between placing one or two spare parts in a warehouse or between employing one or two service engineers in a repair shop can be massive. This integrality complicated the analysis substantially, as we often wanted to construct a queueing system with 1.5 servers in our proofs, but 1.5 is not an admissible number of servers with a well-defined waiting time. As a result, we frequently found ourselves in deep analytical waters, as we had to dive into extensions of the Erlang loss and delay functions. Under optimized numbers of servers, we were still able to prove balancedness of Erlang loss games, but Erlang delay games turned out to be non-balanced in general, in contrast to $M/M/1$ games. The key lesson is that analysis of games often becomes

substantially more complex when the optimization domain is the set of integers rather than the set of real numbers.

Difference between M/G/s/s and M/M/s: Under optimized numbers of servers, Erlang loss games turned out to exhibit different properties than Erlang delay games: The former class of games always had a non-empty core, whereas the latter class contained counterexamples. The sole culprit in the $M/M/s$ setting was the integrality requirement on the number of servers. Surprisingly, this integrality requirement is also present in the $M/G/s/s$ setting, but somehow doesn't pose an adverse enough effect there. Why does this integrality requirement affect the two settings differently, especially when there exists a tight link between the Erlang delay and Erlang loss functions? I could offer some mathematical tautologies as an explanation (e.g., saying that the optimal cost function in the Erlang delay model is not elastic, or that the linear interpolation of the Erlang delay function is not subhomogeneous of degree zero), but that doesn't really advance our understanding. I still find this difference quite peculiar, and I am not even sure what is more surprising: the fact that everything did work for $M/G/s/s$, or the fact that it didn't for $M/M/s$.

Seemingly unrelated problems are related after all: At first glance, a single-period problem seems far removed from an infinite-horizon problem. Nevertheless, the problem of finding optimal base stock levels in the infinite-horizon spare parts model with backlogging corresponds to the single-period newsvendor problem, at least from a mathematical point of view. But that is a known correspondence. What was more surprising, at least to me, was that the spare parts games with backlogging under fixed base stock levels are closely linked to linear production games via the pipeline stock realizations at an arbitrary point in time. This perspective offered a very simple, intuitive, and constructive proof of balancedness. This easy proof superseded an earlier, nonconstructive proof with complex inequalities involving the exponential function's Taylor expansion that gave no insight at all. The key lesson here is that spotting a connection with a known problem can often lead to a nicer proof and new insights.

When demands are correlated, information matters: While working on Chapter 7, at some point I had both a counterexample and a proof for the statement that joining late is never beneficial. Clearly, that is impossible, so something was wrong. After a careful check, I discovered the mistake in my analysis: I had not properly kept track of the way information on previous demands is revealed. Such information is essential if demands are correlated between periods and between players. This inspired me to extend the model with observation possibilities. Interestingly, if I had not made the initial mistake, I might never have thought of this feature. This serves to illustrate that research is filled with mistakes, surprises, and unexpected detours.

The concavicated average marginal allocation works: In Chapter 8, we built up concavicated allocation rules for elastic single-attribute situations from marginal allocations. For a certain class of marginal allocations, we found that the benefits of the players are always ordered in the same way as their attributes. This also held for the average of all marginal allocations. But it didn't hold for all marginal allocations! In this light, it was quite a surprise that the average did work.

And finally, two more general musings on research and writing based on my own experiences; basically, two philosophies that I embraced as my PhD project progressed.

Formal models are limited, but powerful: While focusing on deep, mathematical structures, it is easy to lose track of what we are really trying to accomplish. Our ultimate goal was not to develop cost allocation mechanism that can be applied in practice directly, but to generate insights that improve our understanding of cooperation in stochastic service systems. We did that by formulating and analyzing formal models, which admittedly are nothing more than tales. Tales, to borrow terminology from Rubinstein (2012), that provide intellectual entertainment and useful insights, but tales nonetheless. To spin nice tales, we often adopted a model that was mathematically tractable at the expense of some realism. For example, we commonly assumed symmetric and linear cost parameters, free transshipments, identical service requirements, and Poisson demand processes; we did not concern ourselves with the difficulty of estimating parameters from data; and we disregarded the issue that resource pooling may lead to an additional administrative burden. These are limitations, but they were conscious choices; we never intended for our models to capture reality perfectly. Our goal was to expand our knowledge. And by simplifying, we can grasp the world around us. In this way, the limitations of our models are their greatest strength.

A simple example is worth a thousand theorems: Writing this thesis reinforced my belief that simple examples can already convey most of the information with only a fraction of the difficulty. I like how illustrative examples keep us from getting lost in a morass of abstractions and that they can actually present results in themselves, e.g., an impossibility. Moreover, I consider examples, stories, colloquial language, and anecdotes to be more friendly to a reader than complex, general models and dry, impenetrable theorems. Funky examples are concrete and can create an emotional investment that makes it easier to generate, understand, and remember new scientific knowledge. Basically, as I progressed in my PhD project, I started clamoring for more fun and more examples in research papers. Nevertheless, we would never have good examples without underlying theorems. A theorem provides a result that an example can actually illustrate, and that is invaluable.

9.4 Game on!

In the concluding sections of each chapter, we already discussed chapter-specific directions for future research, usually on the analysis of more general models with less symmetry assumptions. Next to this, there are several other, more generic future research directions, all motivated by the work in this thesis.

Risk averse players: Throughout this entire thesis, we assumed that players are risk neutral. That is, they are only interested in their *expected* costs. In reality, players often have a certain degree of risk averseness. Risk averse players would value not only a reduction in their expected costs but also a reduction in the *variance* thereof. Consequently, any decision taken by a coalition would also have to take this variance into account. To tackle resource pooling between risk averse players, the work of Suijs et al. (1999) may provide guidance.

Adjustable server speeds: Currently, two streams of queueing games have been considered in the literature: Erlang loss and Erlang delay games (which feature multiple servers in parallel with exogenously given speeds) and $M/M/1$ games (which feature a single server with an adjustable speed). It would also be interesting to study queueing games that include both aspects, i.e., multiple servers and adjustable speed. The different speeds could represent, e.g., the efficiency of alternative types of equipment. Any coalition would then simultaneously optimize the number of servers and their service rates over a certain domain, taking into account different operating costs for each of the alternative speeds. It would be interesting to analyse the resulting cost allocation problem.

Multi-item setting with setup costs: Throughout this thesis, we considered models with a *single* item. We do not mean to suggest that capital goods such as airplanes are only comprised of a single component, of course. The idea underlying our choice for single-item models is that resource pools for multiple items can often be simply considered as the sum of separate single-item games. This is indeed true when there are no economic dependencies across items. However, in the presence of high setup costs for maintenance activities, the multi-item perspective becomes relevant. As described in Section 3.2.1, several authors have used cooperative games to analyze sharing of joint setup costs, though in the context of economic lot sizing problems with deterministic demands. For stochastically deteriorating components, the synchronization of maintenance activities to reduce setup costs has been considered in Dekker et al. (1997), Zhu et al. (2012), and references therein. However, they did not study the resulting cost allocation problem; there is a research opportunity here.

Empirical validation: This thesis has dealt solely with mathematical models. Consequently, all aspects of real life that could not be formally captured as part of an

inventory or queueing model were left out. This includes the human factor. To assess the quality and usability of this thesis' models, assumptions, and solutions, the proposed cost allocation mechanisms should be tested in real life. For example, by attempting to implement them in practice and subsequently observing all the resistance and problems that arise. Although such empirical validation may be time consuming because it will likely involve the participation of several companies, it can result in valuable feedback.

Separate pooling provider: Incorporating the OEM or a third-party pooling provider into the model is also an interesting extension. Especially in competitive environments where it is important that private information does not fall in the hands of competitors, the involvement of a trustworthy third party may be essential for the implementation of a pool. This third party would be provided with all necessary information and would advise everyone how much stock to carry, without revealing competitor's information directly. Of course, this leads to the question of which fraction of the pooling benefits such a third-party pooling provider would be entitled to.

Contract design: Another interesting direction is to relax the assumption that binding agreements can be made. In case all players can select their resource levels and pooling policies for themselves, we can run into the problem of how everyone can be motivated to act such that the central optimal solution is achieved. Such coordination might be achieved via incentive-compatible contracts, though it is unclear how to design those contracts for our resource pooling settings. We remark that, after players have independently chosen resource levels, the remaining problem can be seen as a cooperative game with fixed numbers of resources. We refer the interested reader to Anupindi et al. (2001) and Özen et al. (2008), who hybridized non-cooperative and cooperative approaches in a newsvendor context. Finally, one can also think about a more detailed investigation of contracts that are offered by an OEM or a third-party pooling provider to its customers.

Games for other settings: This thesis showed the value of the cooperative game theoretical approach for research pooling situations. However, resource pooling only makes sense as long as there are actually some resources around. And our planet's resources are finite. This ties in to some of the upcoming problems that humanity is facing this century. For example, the looming ecological disasters chronicled in, e.g., Catton (1982) and Diamond (2006) and the colonization of outer space that is required to tap into the vast resources available there, as advocated in, e.g., Szklarski (2011). The perspective of cooperative game theory may yield useful insights for international collaboration on these upcoming problems.

Bibliography

P. Alfredsson and J. Verrijdt. Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science*, 45(10):1416–1431, 1999.

S. Anily and M. Haviv. The cost allocation problem for the first order interaction joint replenishment model. *Operations Research*, 55(2):292–302, 2007.

S. Anily and M. Haviv. Cooperation in service systems. *Operations Research*, 58(3):660–673, 2010.

S. Anily and M. Haviv. Homogeneous of degree one games are balanced with applications to service systems. Working Paper, School of Business, Tel Aviv University, 2011.

ANP. Rampenzenders komen internetcapaciteit tekort. Last accessed May 17, 2013. Published online on February 8, 2011.

R. Anupindi, Y. Bassok, and E. Zemel. A general framework for the study of decentralized distribution systems. *Manufacturing and Service Operations Management*, 3(4):349–368, 2001.

K. Aronis, I. Magou, R. Dekker, and G. Tagaras. Inventory control of spare parts using a Bayesian approach: A case study. *European Journal of Operational Research*, 154(3): 730–739, 2004.

K.J. Arrow, T. Harris, and J. Marschak. Optimal inventory policy. *Econometrica*, 19(3): 250–272, 1951.

K.S. Azoury. Bayes solution to dynamic inventory models under unknown demand distribution. *Management Science*, 31(9):1150–1160, 1985.

R. Banker. Equity considerations in traditional full cost allocation practices: An axiomatic perspective. In S. Moriarity, editor, *Joint cost allocations*, pages 110–130. University of Oklahoma Press, 1981.

S. Benjaafar. Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research*, 87(2):375–388, 1995.

J.W. Bertrand and J.C. Fransoo. Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, 22 (2):241–264, 2002.

O. Bondareva. Certain applications of the methods of linear programming to the theory of cooperative games (in Russian). *Problemy Kibernetiki*, 10:119–139, 1963.

M.A.A. Boon, R.D. van der Mei, and E.M.M. Winands. Applications of polling systems. *Surveys in Operations Research and Management Science*, 16(2):67–82, 2011.

P. Borm, H. Hamers, and R. Hendrickx. Operations research games: A survey. *TOP*, 9(2): 139–199, 2001.

S. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

M. Braglia and M. Frosolini. Virtual pooled inventories for equipment-intensive industries: An implementation in a paper district. *Reliability Engineering & System Safety*, 112:26–37, 2013.

E. Brockmeyer, H.L. Halstrøm, and A. Jensen. *The life and works of A.K. Erlang.* Transactions of the Danish Academy of Technical Sciences, 1948.

G. Cachon and S. Netessine. Game theory in supply chain analysis. In D. Simchi-Levi, S.D. Wu, and M. Shen, editors, *Supply Chain Analysis in the eBusiness Era.* Kluwer, 2004.

J.M. Calabrese. Optimal workload allocation in open networks of multiserver queues. *Management Science*, 38(12):1792–1802, 1992.

W.R. Catton. *Overshoot: The ecological basis of revolutionary change.* University of Illinois Press, 1982.

A.L. Cauchy. *Cours d'analyse de l'Ecole Royale Polytechnique, Part 1: Analyse Algebrique.* Paris, 1821.

X. Chen. Inventory centralization games with price-dependent demand and quantity discount. *Operations Research*, 57(6):1394–1406, 2009.

X. Chen and J. Zhang. Duality approaches to economic lot-sizing games. Working Paper, New York University, 2008.

X. Chen and J. Zhang. A stochastic programming duality approach to inventory centralization games. *Operations Research*, 57(4):840–851, 2009.

S.S. Chiu and R.C. Larson. Locating an n-server facility in a stochastic environment. *Computers & Operations Research*, 12(6):509–516, 1985.

Y. Chun. The proportional solution for rights problems. *Mathematical Social Sciences*, 15 (3):231–246, 1988.

M.A. Cohen, N. Agrawal, and V. Agrawal. Winning in the Aftermarket. *Harvard Business Review*, 84(5):129–138, 2006.

R.B. Cooper. *Introduction to queueing theory*. North-Holland, 1981.

C.J. Corbett and K. Rajaram. A generalization of the inventory pooling effect to nonnormal dependent demand. *Manufacturing & Service Operations Management*, 8(4):351–358, 2006.

F. Cruijssen, M. Cools, and W. Dullaert. Horizontal cooperation in logistics: Opportunities and impediments. *Transportation Research Part E*, 43(2):129–142, 2007.

I. Curiel. *Cooperative game theory and applications: Cooperative games arising from combinatorial optimization problems*. Springer, 1997.

I. Curiel, H. Hamers, and F. Klijn. Sequencing games: A survey. In P. Borm and H. Peters, editors, *Chapters in game theory: in honor of Stef Tijs*, pages 27–50. Kluwer Academic Publishers, 2002.

A.M. De Bruin, R. Bekker, L. van Zanten, and G.M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2010.

M.A. de Frutos. Decreasing serial cost sharing under economies of scale. *Journal of Economic Theory*, 79(2):245–275, 1998.

M.A. de Frutos. Coalitional manipulations in a bankruptcy problem. *Review of Economic Design*, 4(3):255–272, 1999.

R. Dekker, R. Wildeman, and F. van der Duyn Schouten. A review of multi-component maintenance models with economic dependence. *Mathematical Methods of Operations Research*, 45(3):411–435, 1997.

J.M. Diamond. *Collapse: How societies choose to fail or succeed*. Penguin, 2006.

J. Drechsel and A. Kimms. Cooperative lot sizing with transshipments and scarce capacities: Solutions and fair cost allocations. *International Journal of Production Research*, 49(9): 2643–2668, 2011.

M. Dror and B.C. Hartman. Allocation of gains from inventory centralization in newsvendor environments. *IIE Transactions*, 37(2):93–107, 2005.

M. Dror and B.C. Hartman. Shipment consolidation: Who pays for it and how much? *Management Science*, 53(1):78–87, 2007.

M. Dror and B.C. Hartman. Survey of cooperative inventory games and extensions. *Journal of the Operational Research Society*, 62(4):565–580, 2011.

M. Dror, L.A. Guardiola, A. Meca, and J. Puerto. Dynamic realization games in newsvendor inventory centralization. *International Journal of Game Theory*, 37(1):139–153, 2008.

M. Dror, B.C. Hartman, and W. Chang. The cost allocation issue in joint replenishment. *International Journal of Production Economics*, 135(1):242–254, 2012.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. The inventory problem: II. Case of unknown distributions of demand. *Econometrica*, 20(3):450–466, 1952.

M.E. Dyer and L.G. Proll. On the validity of marginal analysis for allocating servers in $M/M/c$ queues. *Management Science*, 23(9):1019–1022, 1977.

G.D. Eppen. Note—effects of centralization on expected costs in a multi-location newsboy problem. *Management Science*, 25(5):498–501, 1979.

G.D. Eppen and A.V. Iyer. Improved fashion buying with Bayesian updates. *Operations Research*, 45(6):805–819, 1997.

G.D. Eppen and L. Schrage. Centralized ordering policies in a multi-warehouse system with lead times and random demand. In *Multi-level production/inventory control systems: Theory and practice*, pages 51–67. North-Holland, 1981.

A.K. Erlang. Løsning af nogle problemer fra sandsynlighedsregningen af betydning for de automatiske telefoncentraler. *Electroteknikeren*, 13:5–13, 1917. Translation: Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. In: E. Brockmeyer, H.L. Halstrøm, and A. Jensen, editors, *The Life and Works of A.K. Erlang*, pp. 138–155. Transactions of the Danish Academy of Technical Sciences, 1948.

G.J. Feeney and C.C. Sherbrooke. The $(S-1, S)$ inventory policy under compound Poisson demand. *Management Science*, 12(5):391–411, 1966.

M.G. Fiestras-Janeiro, I. García-Jurado, A. Meca, and M.A. Mosquera. Cooperative game theory and inventory management. *European Journal of Operational Research*, 210(3): 459–466, 2011.

M.G. Fiestras-Janeiro, I. García-Jurado, A. Meca, and M.A. Mosquera. Cost allocation in inventory transportation systems. *TOP*, 20(2):397–410, 2012.

M. Fisher and A. Raman. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research*, 44(1):87–99, 1996.

A.A. Fredericks. Congestion in blocking systems—A simple approximation technique. *Bell System Technical Journal*, 59(6):805–827, 1980.

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

M.D. García-Sanz, F.R. Fernández, M.G. Fiestras-Janeiro, I. García-Jurado, and J. Puerto. Cooperation in Markovian queueing models. *European Journal of Operational Research*, 188(2):485–495, 2008.

Y. Gerchak and D. Gupta. On apportioning costs to customers in centralized continuous review inventory systems. *Journal of Operations Management*, 10(4):546–551, 1991.

R.P. Gilles. *The cooperative game theory of networks and hierarchies*. Springer, 2010.

D.B. Gillies. *Some theorems on n-person games*. PhD thesis, Princeton University, 1953.

D.B. Gillies. Solutions to general non-zero-sum games. In A. Tucker and R. Luce, editors, *Contribution to the theory of games IV, Volume 40 of Annals of Mathematics Studies*, pages 47–85. Princeton University Press, 1959.

P. González and C. Herrero. Optimal sharing of surgical costs in the presence of queues. *Mathematical Methods of Operations Research*, 59(3):435–446, 2004.

M. Gopaladesikan, N.A. Uhan, and J. Zou. A primal-dual algorithm for computing a cost allocation in the core of economic lot-sizing games. *Operations Research Letters*, 40(6): 453–458, 2012.

J. Grahovac and A. Chakravarty. Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. *Management Science*, 47(4):579–594, 2001.

D. Granot and G. Sošić. A three-stage model for a decentralized distribution system of retailers. *Operations Research*, 51(5):771–784, 2003.

M. Guajardo, M. Ronnqvist, A.M. Halvorsenb, and S.I. Kallevik. Inventory management of spare parts in an energy company. Working Paper, Norwegian School of Economics, 2012.

L.A. Guardiola, A. Meca, and J. Puerto. Production-inventory games and PMAS-games: Characterizations of the Owen point. *Mathematical Social Sciences*, 56(1):96–108, 2008.

L.A. Guardiola, A. Meca, and J. Puerto. Production-inventory games: A new class of totally balanced combinatorial optimization games. *Games and Economic Behavior*, 65(1):205–219, 2009.

G. Hadley and T.M. Whitin. A family of dynamic inventory models. *Management Science*, 8 (4):458–469, 1962.

S.S. Hamlen, W.A. Hamlen Jr., and J.T. Tschirhart. The use of core theory in evaluating joint cost allocation schemes. *The Accounting Review*, 52(3):616–627, 1977.

A. Harel. Convexity properties of the Erlang loss formula. *Operations Research*, 38(3):499–505, 1990.

B.C. Hartman and M. Dror. Cost allocation in continuous-review inventory models. *Naval Research Logistics*, 43(4):549–561, 1996.

R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer, 2003.

S. He, J. Zhang, and S. Zhang. Polymatroid optimization, submodularity, and joint replenishment games. *Operations Research*, 60(1):128–137, 2012.

D.L. Jagerman. Some properties of the Erlang loss function. *Bell System Technical Journal*, 53(3):525–551, 1974.

A.A. Jagers and E.A. van Doorn. On the continued Erlang loss function. *Operations Research Letters*, 5(1):43–46, 1986.

A.A. Jagers and E.A. van Doorn. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review*, 33(2):281–282, 1991.

A. Janssen, J. van Leeuwaarden, and B. Zwart. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Advances in Applied Probability*, 40 (1):122–143, 2008.

A. Janssen, J. van Leeuwaarden, and B. Zwart. Refining square-root safety staffing by expanding Erlang C. *Operations Research*, 59(6):1512–1522, 2011.

T. Jin and Y. Tian. Optimizing reliability and service parts logistics for a time-varying installed base. *European Journal of Operational Research*, 218(1):152–162, 2012.

S. Karlin. Dynamic inventory policy with varying stochastic demands. *Management Science*, 6(3):231–258, 1960.

F. Karsten and R. Basten. Pooling of spare parts between multiple users: how to share the benefits? *European Journal of Operational Research*, forthcoming, 2013.

F. Karsten, M. Slikker, and G.J. van Houtum. Analysis of resource pooling games via a new extension of the Erlang loss function. BETA Working Paper 344, Eindhoven University of Technology, 2011a.

F. Karsten, M. Slikker, and G.J. van Houtum. Resource pooling and cost allocation among independent service providers. BETA Working Paper 352, Eindhoven University of Technology, 2011b.

F. Karsten, M. Slikker, and G.J. van Houtum. Inventory pooling games for expensive, low-demand spare parts. *Naval Research Logistics*, 59(5):311–324, 2012.

F. Karsten, M. Slikker, M. Akan, and A. Scheller-Wolf. Inventory pooling games in a stochastic dynamic environment. Working Paper, 2013a.

F. Karsten, M. Slikker, and P. Borm. Allocation rules on elastic single-attribute situations. Working Paper, 2013b.

E. Kemahlıoğlu-Ziya and J.J. Bartholdi III. Centralizing inventory in supply chains by using Shapley value to allocate the profits. *Manufacturing & Service Operations Management*, 13(2):146–162, 2011.

P. Keskinocak. Corporate high flyers. *OR/MS Today*, 26(6):22–25, 1999.

J. Kilpi, J. Töyli, and A. Vepsäläinen. Cooperative strategies for the availability service of repairable aircraft components. *International Journal of Production Economics*, 117(2): 360–370, 2009.

P.J. Kolesar and L.V. Green. Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management*, 7(3):282–293, 1998.

K.O. Kortanek, D.N. Lee, and G.G. Polak. A linear programming model for design of communications networks with time varying probabilistic demands. *Naval Research Logistics Quarterly*, 28(1):1–32, 1981.

W.A. Kosmala. *A friendly introduction to analysis.* Prentice Hall, 2nd edition, 2004.

A.A. Kranenburg. *Spare parts inventory control under system availability constraints.* PhD thesis, Eindhoven University of Technology, 2006.

A.A. Kranenburg and G.J. van Houtum. A new partial pooling structure for spare parts networks. *European Journal of Operational Research*, 199(3):908–921, 2009.

K.R. Krishnan. The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates. *IEEE Transactions on Communications*, 38(9):1314–1316, 1990.

A. Kukreja, C.P. Schmidt, and D.M. Miller. Stocking decisions for low-usage items in a multilocation inventory system. *Management Science*, 47(10):1371–1383, 2001.

M.A. Lariviere and J.A. van Mieghem. Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing & Service Operations Management*, 6(1):23–40, 2004.

M. Leng and M. Parlar. Analytic solution for the nucleolus of a three-player cooperative game. *Naval Research Logistics*, 57(7):667–672, 2010.

R.P. McLean and W.W. Sharkey. An approach to the pricing of broadband telecommunications services. *Telecommunication Systems*, 2(1):159–184, 1993.

A. Meca, I. García-Jurado, and P. Borm. Cooperation and competition in inventory games. *Mathematical Methods of Operations Research*, 57(3):481–493, 2003.

A. Meca, J. Timmer, I. García-Jurado, and P. Borm. Inventory games. *European Journal of Operational Research*, 156(1):127–139, 2004.

A. Meca, L.A. Guardiola, and A. Toledo. p-additive games: A class of totally balanced games arising from inventory situations with temporary discounts. *TOP*, 15(2):322–340, 2007.

B.L. Miller. A queueing reward system with several customer classes. *Management Science*, 16(3):234–245, 1969.

L. Montrucchio and M. Scarsini. Large newsvendor games. *Games and Economic Behavior*, 58(2):316–337, 2007.

L. Montrucchio, H. Norde, U. Özen, M. Scarsini, and M. Slikker. Cooperative newsvendor games: A review. In T.M. Choi, editor, *Handbook of Newsvendor Problems*, pages 137–162. Springer, 2012.

M.A. Mosquera, I. García-Jurado, and M.G. Fiestras-Janeiro. A note on coalitional manipulation and centralized inventory management. *Annals of Operations Research*, 158 (1):183–188, 2008.

H. Moulin and S. Shenker. Serial cost sharing. *Econometrica*, 60(5):1009–1037, 1992.

A. Müller, M. Scarsini, and M. Shaked. The newsvendor game has a nonempty core. *Games and Economic Behavior*, 38(1):118–126, 2002.

G.R. Murray and E.A. Silver. A Bayesian analysis of the style goods inventory problem. *Management Science*, 12(11):785–797, 1966.

M. Nagarajan and G. Sošić. Game-theoretic analysis of cooperation among supply chain agents: Review and extensions. *European Journal of Operational Research*, 187(3):719–745, 2008.

J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

J.J. Neale and S.P. Willems. Managing inventory in supply chains with nonstationary demand. *Interfaces*, 39(5):388–399, 2009.

H. Norde and H. Reijnierse. A dual description of the class of games with a population monotonic allocation scheme. *Games and Economic Behavior*, 41(2):322–343, 2002.

H. Norde, U. Özen, and M. Slikker. Setting the right incentives for global planning and operations. Working Paper, 2011.

B. O'Neill. A problem of rights arbitration from the Talmud. *Mathematical Social Sciences*, 2(4):345–371, 1982.

G. Owen. On the core of linear production games. *Mathematical programming*, 9(1):358–370, 1975.

U. Özen, J. Fransoo, H. Norde, and M. Slikker. Cooperation between multiple newsvendors with warehouses. *Manufacturing & Service Operations Management*, 10(2):311–324, 2008.

U. Özen, H. Norde, and M. Slikker. On the convexity of newsvendor games. *International Journal of Production Economics*, 133(1):35–42, 2011.

U. Özen, M.I. Reiman, and Q. Wang. On the core of cooperative queueing games. *Operations Research Letters*, 39(5):385–389, 2011.

U. Özen, N. Erkip, and M. Slikker. Stability and monotonicity in newsvendor situations. *European Journal of Operational Research*, 218(2):416–425, 2012a.

U. Özen, G. Sošić, and M. Slikker. A collaborative decentralized distribution system with demand forecast updates. *European Journal of Operational Research*, 216(3):573–583, 2012b.

C. Palm. Analysis of the Erlang traffic formulae for busy-signal arrangements. *Ericsson Technics*, 5:39–58, 1938.

F. Papier and U.W. Thonemann. Queuing models for sizing and structuring rental fleets. *Transportation Science*, 42(3):302–317, 2008.

C. Paterson, G. Kiesmuller, R. Teunter, and K. Glazebrook. Inventory models with lateral transshipments: A review. *European Journal of Operational Research*, 210(2):125–136, 2011.

B. Peleg and P. Sudhölter. *Introduction to the theory of cooperative games*. Springer, 2nd edition, 2007.

N.C. Petruzzi and M. Dada. Information and inventory recourse for a two-market, price-setting retailer. *Manufacturing & Service Operations Management*, 3(3):242–263, 2001.

C. Pinçe and R. Dekker. An inventory model for slow moving items subject to obsolescence. *European Journal of Operational Research*, 213(1):83–95, 2011.

D. Pinedo. Een scootmobiel kun je ook delen. NRC Handelsblad, August 7, 2012.

J.B. Popović. Decision making on stock levels in cases of uncertain demand rate. *European Journal of Operational Research*, 32(2):276–290, 1987.

W. Poundstone. *Prisoner's dilemma: John von Neumann, game theory and the puzzle of the bomb*. Anchor, 1993.

W.B. Powell. *Approximate dynamic programming: Solving the curses of dimensionality*. Wiley, 2011.

M. Restrepo, S.G. Henderson, and H. Topaloglu. Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, 12(1):67–79, 2009.

L.W. Robinson. A comment on Gerchak and Gupta's "On apportioning costs to customers in centralized continuous review inventory system". *Journal of Operations Management*, 11(1):99–102, 1993.

S.M. Ross. *Introduction to probability models*. Academic press, 9th edition, 2007.

A. Rubinstein. *Economic fables*. Open Book Publishers, 2012.

J. Sánchez-Soriano, M.A. Lopez, and I. Garcıa-Jurado. On the core of transportation games. *Mathematical Social Sciences*, 41(2):215–225, 2001.

H.E. Scarf. Bayes solutions of the statistical inventory problem. *The Annals of Mathematical Statistics*, 30(2):490–508, 1959.

H.E. Scarf. Inventory theory. *Operations Research*, pages 186–191, 2002.

D. Schmeidler. The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics*, 17(6):1163–1170, 1969.

L.S. Shapley. A value for n-person games. In H. Kuhn and A. Tucker, editors, *Contribution to the theory of games II, Volume 28 of Annals of Mathematics Studies*, pages 307–317. Princeton University Press, 1953.

L.S. Shapley. On balanced sets and cores. *Naval Research Logistics Quarterly*, 14:453–460, 1967.

L.S. Shapley. Cores of convex games. *International Journal of Game Theory*, 1(1):11–26, 1971.

L.S. Shapley and M. Shubik. Quasi-cores in a monetary economy with nonconvex preferences. *Econometrica*, 34(4):805–827, 1966.

L.S. Shapley and M. Shubik. On market games. *Journal of Economic Theory*, 1:9–25, 1969.

R. Sidel. Banks join the do-it-yourself craze. The Wall Street Journal. Last accessed May 17, 2013. Published online on May 15, 2012.

E.A. Silver. Operations research in inventory management: A review and critique. *Operations Research*, 29(4):628–645, 1981.

M. Slikker, J. Fransoo, and M. Wouters. Joint ordering in multiple news-vendor situations: A game-theoretical approach. BETA Working Paper 64, Eindhoven University of Technology, 2001.

M. Slikker, J. Fransoo, and M. Wouters. Cooperation between multiple news-vendors with transshipments. *European Journal of Operational Research*, 167(2):370–380, 2005.

D.R. Smith and W. Whitt. Resource sharing for efficiency in traffic systems. *Bell System Technical Journal*, 60(1):39–55, 1981.

C. Snijders. Axiomatization of the nucleolus. *Mathematics of Operations Research*, 20(1): 189–196, 1995.

Y. Sprumont. Population monotonic allocation schemes for cooperative games with transferable utility. *Games and Economic Behavior*, 2(4):378–394, 1990.

S. Stidham Jr. Analysis, design, and control of queueing systems. *Operations Research*, 50 (1):197–216, 2002.

P. Sudhölter. Axiomatizations of game theoretical solutions for one-output cost sharing problems. *Games and Economic Behavior*, 24(1):142–171, 1998.

J. Suijs, P. Borm, A. De Waegenaere, and S. Tijs. Cooperative games with stochastic payoffs. *European Journal of Operational Research*, 113(1):193–205, 1999.

C. Szklarski. Space is our only hope, says Hawking. Last accessed May 17, 2013. Published online on November 19, 2011.

J. Timmer and W. Scheinhardt. How to share the cost of cooperating queues in a tandem network? In *Conference Proceedings of the 22nd International Teletraffic Congress*, pages 1–7. IEEE, 2010.

J.T. Treharne and C.R. Sox. Adaptive inventory control for nonstationary demand and partial information. *Management Science*, 48(5):607–624, 2002.

W. van den Heuvel, P. Borm, and H. Hamers. Economic lot-sizing games. *European Journal of Operational Research*, 176(2):1117–1130, 2007.

G.J. van Houtum and W.H.M. Zijm. On the relationship between cost and service models for general inventory systems. *Statistica Neerlandica*, 54(2):127–147, 2000.

W. van Jaarsveld and R. Dekker. Finding optimal policies in the $(S-1, S)$ lost sales inventory model with multiple demand classes. Econometric Institute Report 2009-14, Erasmus University Rotterdam, 2009.

W. van Jaarsveld and R. Dekker. Estimating obsolescence risk from demand data to enhance inventory control—a case study. *International Journal of Production Economics*, 133(1): 423–431, 2011.

E. van Outeren. Eerlijk zullen we alles delen, ook het strooizout. NRC Handelsblad, December 23, 2010.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

L. Ward. *See you in a hundred years: Discover one young family's search for a simpler life ... Four seasons of living in the year 1900*. Random House Publishing Group, 2008.

W. Whitt. Partitioning customers into service groups. *Management Science*, 45(11):1579–1592, 1999.

W. Whitt. The Erlang B and C formulas: problems and solutions. Unpublished class notes. Available online: `http://www.columbia.edu/~ww2040/ErlangBandCFormulas.pdf`, 2002.

R.I. Wilkinson. Theories for toll traffic engineering in the USA. *Bell System Technical Journal*, 35(2):421–514, 1956.

R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

H. Wong, D. Cattrysse, and D. van Oudheusden. Stocking decisions for repairable spare parts pooling in a multi-hub system. *International Journal of Production Economics*, 93: 309–317, 2005.

H. Wong, G.J. van Houtum, D. Cattrysse, and D. van Oudheusden. Multi-item spare parts systems with lateral transshipments and waiting time constraints. *European Journal of Operational Research*, 171(3):1071–1093, 2006.

H. Wong, D. van Oudheusden, and D. Cattrysse. Cost allocation in spare parts inventory pooling. *Transportation Research Part E*, 43(4):370–386, 2007.

D. Xu and R. Yang. A cost-sharing method for an economic lot-sizing game. *Operations Research Letters*, 37(2):107–110, 2009.

Y. Yu, S. Benjaafar, and Y. Gerchak. Capacity Pooling and Cost Sharing among Independent Firms in the Presence of Congestion. Working paper, University of Minnesota, 2007.

Y. Yu, S. Benjaafar, and Y. Gerchak. Capacity Sharing and Cost Allocation among Independent Firms in the Presence of Congestion. Working paper, University of Minnesota, 2009.

Y. Zeng, J. Li, and X. Cai. Economic lot-sizing games with perishable inventory. In *Conference Proceedings of the 8th International Conference on Service Systems and Service Management*, pages 1–5. IEEE, 2011.

J. Zhang. Cost allocation for joint replenishment models. *Operations Research*, 57(1):146–156, 2009.

J. Zhao. The relative interior of the base polyhedron and the core. *Economic Theory*, 18(3): 635–648, 2001.

Q. Zhu, H. Peng, and G.J. van Houtum. A condition-based maintenance policy for multi-component systems with a high maintenance setup cost. BETA Working Paper 400, Eindhoven University of Technology, 2012.

P.H. Zipkin. *Foundations of inventory management*. McGraw-Hill, 2000.

Summary of Resource Pooling Games

Consider a number of neighboring companies whose primary business processes are all based around a certain type of resource. These resources (which could be inventories for retailers, repairmen or spare parts for maintenance organizations, production equipment for manufacturing firms, medical staff for hospital departments, and so on) are meant to fill demands of arriving customers. Each company has its own, fixed customer base, but there is uncertainty: customers for each company arrive according to a stochastic process, and their service times are generally unpredictable. For example, if the companies are maintenance organizations, each responsible for maintaining trains and railway infrastructure in different geographical regions, then demands for their key resources (repairmen and spare parts) are driven by unpredictable equipment failures, and failure occurrences and repair completion times cannot be predicted with certainty in advance.

Due to this inherent uncertainty, the number of available resources at a company will fluctuate over time, and a customer may even arrive to a company to find all resources occupied. In the maintenance example, sometimes the shelf with spare parts is empty or all repairmen are busy. In that case, the customer demand cannot be served immediately, and the shortage may result in costly delays, emergency procedures, lost revenue, or contractual fines. Meanwhile, a neighboring company may very well have the required resource available. If companies would be acting separately, then the end result would be one company with a shortage and another with a surplus—a clear mismatch. If, however, companies would have joined forces and be sharing their resources, which is referred to as resource pooling, then they can help each other out in case of a shortage. Moreover, they may be able to re-optimize their total number of resources, which may lead to additional savings. Altogether, such collaborations offer the opportunity to benefit from large (statistical) economies of scale.

Nevertheless, independent companies or players are usually not interested in the benefits of pooling for society as a whole, but in the consequences for their own bottom line. Whether or not collaboration is beneficial for every individual player depends on who pays for the shared resources. Hence, a key question is how the collective costs of the pooled system are allocated amongst the players, and whether or not this is done in a fair way. This cost allocation problem can be daunting when multiple players are involved.

That's where this thesis comes in. To shed some light on this problem, we follow the methodology of quantitative modeling and apply concepts from cooperative game theory. Assuming that binding agreements and side payments are allowed, we simplify reality into the mathematical model of a cooperative game, which is essentially a list of the costs that any coalition (subgroup of players) would have to pay when acting separately. When those costs represent the total resource and shortage costs of setting up a pooled inventory or queueing system, we obtain resource pooling games.

For these games, we examine (existence of) stable allocations of the costs of the pooled system with *all* players. Here, "stable" means that no subset of players is allocated more costs than they would face when acting separately. So, if all players collaborate and the total resource and shortage costs of their resource pool is split according to a stable allocation, then no subset of players has an incentive to split off and act separately—all players will be motivated to join forces.

To determine the cost figures for all coalitions from input parameters (such as the customer arrival or demand rate, service or lead times, resource costs, and shortage costs), we represent the pooled systems as inventory or queueing models and determine their performance via mathematical formulas from inventory theory or queueing theory.

Whereas previous literature on cooperative inventory/queueing games focused on restrictive newsvendor and $M/M/1$ models, this thesis extends the analysis to multi-period inventory systems and multi-server queueing systems: we study games derived from pooling of $M/G/s/s$ queues (i.e., Erlang loss systems), $M/M/s$ queues (i.e., Erlang delay systems), and $(S-1,S)$ inventory systems. Assuming that players face identical shortage costs, we formulate two game variants for each of these settings: one in which the total number of resources for any coalition is optimized and one in which each player brings a fixed number of resources to every coalition. We also study games derived from inventory pooling in a finite-horizon model with non-stationary, stochastic demands with arbitrary correlations over time.

For all of these games, we examine the existence of a stable allocation. We prove that, except for the Erlang delay game with optimized numbers of resources, the existence of a stable allocation is always guaranteed. In the process of proving structural results for our games, we run into a technical obstacle: the integrality in the number of resources complicates the analysis. We tackle this by deriving and exploiting several new analytical properties of (extensions of) the classic Erlang loss and Erlang delay functions and of the cost function in the $(S-1,S)$ inventory model.

Additionally, for the Erlang loss game and the $(S-1,S)$ game, both under optimized numbers of resources, we *identify* a simple allocation that is stable: the one proportional to players' demand rates. In proving this, we use that the per-demand optimal cost is elastic (i.e., non-increasing in the total demand served) and that every player is characterized by a single attribute (his demand rate) only.

In the final chapter of the thesis, we study general situations in which every player has a single attribute and the underlying cost function is elastic. For the resulting class of games, we look beyond stability and propose a new fairness property called benefit ordering, which says that larger players should reap larger benefits than small players. We prove that this property is incompatible with stability in general. Afterwards, we relax our fairness properties into compatible ones, introduce several new cost sharing mechanisms, and show whether or not they satisfy (relaxations of) stability and benefit ordering in general.

As an illustration of the types of models and results in the thesis, we conclude this summary with brief description of one model—the Erlang loss game with optimized numbers of servers—and one theorem for that model. Formally, we have a set of players N, and customer arrivals for any player $i \in N$ are governed by a Poisson process with rate λ_i. Customers are served by servers (on average, service takes one unit of time) and resource costs for servers amount to h per server per unit of time. If a customer finds no free server upon arrival, he is lost and a penalty cost p is incurred. So, if any coalition of players $M \subseteq N$ would set up an optimal Erlang loss system for their joint arrival rate $\lambda_M = \sum_{i \in M} \lambda_i$, then their long-term average costs per unit time would be $c(M) = \min_{s \in \{0,1,\ldots\}} \{hs + B(s, \lambda_M)\lambda_M p\}$, where $B(s, \lambda_M)$ is the probability that an arriving customer is lost when the system has s servers and faces arrival rate λ_M.

Analysis of (an extension of) the Erlang loss function B yields the following theorem: if all players collaborate, facing total costs $c(N)$, then a relatively simple proportional allocation, which assigns $x_i = c(N)\lambda_i/\lambda_N$ to every player $i \in N$, is stable. In other words, $\sum_{i \in M} x_i \leq c(M)$ for all coalitions $M \subseteq N$. This means that resource pooling, at least under this allocation mechanism for the Erlang loss setting, is beneficial for all parties involved. It should be stressed that this is only one theorem for one model, included for illustration only. Many other results and models, some with a similar flavor, are contained in this thesis as well.

About the author

Frank Karsten was born in Heerlen, a city in the south of the Netherlands, on June 19, 1984. At the age of 16, he completed his pre-university education (gymnasium) cum laude at the Bernardinuscollege in Heerlen. After that, he received a Bachelor's degree in Industrial Engineering and Management Science and a Master's degree in Operations Management and Logistics. Both were obtained from the Eindhoven University of Technology.

Meanwhile, he successfully competed in the professional tournament circuit of the strategic card game Magic: the Gathering. In 2009, Frank was inducted into the Magic Pro Tour Hall of Fame as one of the game's foremost analytical minds and writers.

Subsequently, he started a PhD project on operations research and game theory at the Eindhoven University of Technology, advised by Marco Slikker and Geert-Jan van Houtum. On November 28, 2013, Frank defends his PhD thesis.