COORDINATION AND INVENTORY MANAGEMENT IN SUPPLY CHAINS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MANAGEMENT
SCIENCE AND ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Hongxia Xiong
March 2007

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas A. Weber Principal Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Warren H. Hausman

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Hau L. Lee
(Graduate School of Business)

Approved for the University Committee on Graduate Studies.

Abstract

This dissertation focuses on coordination and inventory management in supply chains. The three main chapters cover contract design, surplus extraction in multi-principal multi-agent supply chains, and stock positioning for distribution systems with service constraints. I review these three problems and summarize the contributions of this work in Chapter 1.

Chapter 2 is based on the paper "Contract Design in Multi-Principal Multi-Agent Supply Chains" which I co-wrote with my thesis advisor, Professor Thomas Weber. This chapter is concerned with incentive alignment issues for decentralized decision-making in a multi-principal multi-agent supply chain setting. A supply chain is coordinated if the aggregate payoffs of all firms are maximized. Most of the operations management literature studies parametric contracts to coordinate a single-principal single-agent supply chain. Prat and Rustichini (2003) provide a foundation to study a multi-principal multi-agent game. Applying their characterization of weakly truthful equilibria, we provide a constructive approach to finding a set of nonlinear transfer schedules that implement any socially efficient outcome as a weakly truthful equilibrium. Our two-step solution approach is to first solve a reduced contract design problem based on excess measures and next to transform the solution to actual transfer payments. Our approach can be applied to a very large class of many-to-many supply chain settings. Although the topic is motivated in the context of supply chain management, the methods developed in this work can be applied in many other economic applications such as political economy, industrial organization, and auction theory. In addition, our results integrate existing contractual schemes that contain ex-post provisions into a unified framework, and accommodate a broad range of other vertical contracts such as ties and requirement contracts.

Chapter 3 is based on "Cooperative Bargaining in Multi-Principal Multi-Agent Supply Chains," a paper also co-written with Professor Thomas Weber. Significant

flexibilities exist in the design of coordinating contracts in Chapter 2 [1]. However, the principals have a conflict of interest regarding who pays the agents less. As a natural follow-up study of Chapter 2, Chapter 3 addresses surplus extraction problem, i.e., how much surplus the principals can extract from the agents in the multi-principal multi-agent contracting game. We first show that the solutions to the reduced contract design problem can fully determine the in-equilibrium net payoffs of the principals. We then show that the set of attainable payoff vectors for the principals is convex. In the special case of common agency with one agent, we provide some necessary and some sufficient conditions under which each principal can obtain her maximum possible payoff. We finally discuss how to find the Pareto frontier by numerical implementation. Several numerical experiments are conducted to illustrate the applications of our results and provide insights on contracting in multi-principal multi-agent supply chains. We emphasize that the models and results we have developed can be applied to a wide range of application settings beyond supply chain management.

Finally, Chapter 4 is based on the paper "Stock Positioning in Distribution Systems with Service Constraints," which I co-wrote with Professor Özalp Özer. This chapter addresses inventory allocation in a one-warehouse multi-retailer distribution system subject to a service-level constraint at each retailer. We focus on a fill-rate type of service level, which is defined as the fraction of demand that is met from on-hand inventory at a retailer. The objective is to satisfy end customer demand while minimizing the inventory holding cost by optimally allocating inventory among the warehouse and multiple retailers. Because in practice distribution systems can be very large and computational improvements can be very important, we concentrate on developing easy-to-use and easy-to-describe heuristics and approximations. We first present an optimization algorithm, which provides a benchmark to evaluate the performance of our heuristics and approximations. We next develop the Triple-Search Heuristic, which is close-to-optimal, and the Newsvendor Heuristic, which is in closed-form, easy to use and easy to describe. We also present closed-form approximations such as the Normal Approximation, the Newsvendor Approximation, and the Distribution-Free Approximation. Finally, we present numerical experiments that evaluate the performance of our heuristics and approximations, as well as provide insights into distribution system design issues. In this, this chapter at least partially fills a significant gap between theory and practice on inventory management.

[1]Recently, Weber and Xiong (2006) have extended the results in Chapter 2 to find all socially efficient contracts in closed-form, allowing for both agent payoff externalities and payoff nonconcavities.

Acknowledgements

I wish to express my deep appreciation to Professor Thomas Weber for serving as my advisor, as well as for his invaluable guidance of this research. His ability to look at problems in new and innovative ways has been profoundly beneficial. I was fortunate to have the pleasure and privilege of working with him on my first two research projects on coordination in multi-principal multi-agent supply chains, and the experience was extremely fruitful for my progress as a researcher. I also greatly appreciate his always prompt, insightful, to-the-point advice and feedback.

I am indebted to Professor Warren Hausman and Professor Hau Lee for their willingness to serve on the reading committee. They have both served as the principal examples for my academic development. Their encouragement and inspiration will provide continuing motivation for my future work. Professor Hausman has been extremely generous with his advice and guidance in my academic and career development. I turned to him whenever I found myself in difficulty and he was unstinting in his help. Thanks are also due to Professor Pete Veinott for his insightful comments and generosity with his time. He not only showed me how to be devoted to one's work, but also gave me invaluable feedback during my job search.

I owe a great deal of gratitude to a number of friends and colleagues who have been generous with their friendship and their technical expertise. The list of these people is too long to be given in this limited space, but I will always remember you since you make my journey fun and special.

I am also indebted to Professor Senol Utku from Duke University who has been a source of much insight, encouragement, and good cheer. He has served as an extraordinary example of someone who is striving to make the world a better place, and I greatly appreciate knowing him.

My family has been especially valuable to me during my graduate study. In particular I would like to thank my parents Baoxiu Guo and Tangcheng Xiong for their unconditional love and support. I would also like to thank my eighteen-month-old son Bryan Xiao who has brought me abundant pride, inspiration, and joy. Finally, I would like to thank my husband Zhiwei (David) Xiao for his patience, encouragement, and love. He made countless sacrifices so I could pursue my academic goals. It is with love and gratitude that I dedicate this dissertation to him.

Stanford, CA
January 2007
H.X.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation studies two topics of supply chain management, coordination and inventory management. Although the research topics on coordination are motivated in a supply chain management setting, the results and approaches can be applied in other settings such as political economy, industrial organization, and auction theory.

The organization of the dissertation is as follows. Chapter 1 presents the motivation, reviews the problems, and summarizes our contributions. Chapters 2 and 3 treat the first topic on coordination in multi-principal multi-agent supply chains. The second topic, stock positioning in a distribution system with service constraints, is presented in Chapter 4, which can be read independently.

1.1 Motivation

Supply chains often consist of many participating firms, in part as a result of an increasing trend in outsourcing business processes. Coordination of such multi-principal multi-agent supply chains is important since it avoids aggregate efficiency losses across all participating firms, due to, for example, double marginalization. Chapter 2 focuses on construction of a set of nonparametric bilateral contracts between principals and agents that coordinate a two-echelon multi-principal multi-agent supply chain, while all involved firms maintain control over their own actions. Consider Coke and Pepsi and their common retailers who compete in a common consumer market. The question is how Coke and Pepsi, who act as principals, can noncooperatively design incentive-compatible contracts that will make the retailers take socially efficient actions. Chapter 2 tries to find answers to this question in a complete-information setting. In particular, we identify contract mechanisms that can be used to coordinate multi-principal multi-agent supply chains. To achieve this, we first solve a reduced contract design problem and then transform the solutions to find socially efficient pricing schedules. It turns out that the principals have significant flexibility in coordinating a multi-principal multi-agent supply chain. A related question is, how Coke and Pepsi can extract as much surplus as possible from their retailers while coordinating the supply chain. Chapter 3 aims to answer this question. We show that the principals' net payoff vector is on the Pareto frontier[1], which is found by optimizing a weighted summation of the net payoffs of Coke and Pepsi over all possible coordinating schemes.

In addition to coordination, maintaining the right inventory level at the right location in a multi-echelon supply chain is a fundamental problem in supply chain optimization. Tremendous progress in this subject has been made, but the formulations and the methodologies developed in multi-echelon production and distribution systems may be difficult to explain to non-mathematically oriented academics and practitioners. Furthermore, in practice, distribution systems can be very large, and computational improvement on optimal inventory allocation can be very important. Chapter 4 focuses on developing easy-to-describe, close-to-optimal and robust heuristics for efficient inventory management in multi-period, multi-product supply chains. We first present an optimization algorithm that can serve as a benchmark to evaluate the performance of the heuristics and approximations. Some of our approaches are based on solving newsvendor-type problems and thus can be easily implemented on

[1]Pareto frontier is with respect to all the principals, i.e., the principals appropriate surplus from the agents that cannot be increased for one principal without making another principal worse off.

a spreadsheet. Our easy-to-implement heuristics and approximations allow further insight into system design issues and enable managers to manage large-scale systems.

1.2 Overview

Before presenting the three studies, I provide an overview of each of the chapters below. In the beginning of each chapter we include further motivation, a literature review, and detailed descriptions of the problem environment. We discuss possible future work at the end of each chapter.

1.2.1 Contract Design in Multi-Principal Multi-Agent Supply Chains

Chapter 2 shows the design of non-parametric coordinating contracts in a multi-principal multi-agent supply chain in a complete-information setting. Maximizing the firms' aggregrate surplus requires both vertical coordination through contracts and implicit horizontal coordination without the use of anticompetitive practices. Prat and Rustichini (2003) characterize weakly truthful equilibria in multi-principal multi-agent games, which guarantee the noncooperative implementation of socially efficient outcomes. In a two-echelon multi-principal multi-agent supply chain, when all agents' actions are contractible, payoff externalities among agents are additive, and all gross payoffs are concave. Chapter 2 provides a construction of a set of non-parametric socially efficient contracts. More specifically, we first find solutions to an equivalent reduced contract design problem and then map the solutions to the original contract design problem by an outcome-contingent transformation. The flexibility in the design of the contracts depends substantially on how interdependent the payoffs are in the supply chain. In the case of additively separable payoffs, any efficient outcome can be implemented via a customizable outcome-contingent surplus-sharing contract. If payoffs are also concave, a simple affine contract can implement efficient outcomes. In the presence of nonadditive payoff interdependencies, principals need to propose requirements contracts which contain provisions with respect to an agent's actions for other principals. When payoffs in the supply chain are additive, virtually all known commercial contracts can be employed by almost all principals to coordinate a two-echelon multi-principal multi-agent supply chain. We find that the coordinating solutions that have been proposed in operations management literature, such as quantity-dependent pricing (for example, two-part tariff contracts and quantity-discount contracts) and royalty schemes (for example, buy-back contracts, revenue-sharing contracts, quantity-flexibility contracts, and sales-rebate contracts) can be unified into our nonparametric general contract design framework. Finally, our results also accommodate a broad range of other vertical contracts such as ties (discounts across products) and requirement contracts (such as resale price maintenance and exclusive dealing). From the point view of practical implementation of coordination through contract design, we find that affine contracts are generally not desirable to principals.

1.2.2 Surplus Extraction in Multi-Principal Multi-Agent Supply Chains

As a natural follow-up study of Chapter 2, Chapter 3 addresses surplus extraction problem, i.e., how much surplus the principals can extract from the agents in the multi-principal multi-agent contracting game. There is tremendous flexibility in designing coordinating contracts, as the solution to a reduced contract design problem is not unique ($M \times N$ unknowns with $M + N$ constraints, where M denotes the number of principals, and N denotes the number of agents), and the transformation from the solution to a reduced contract design problem to pricing schedules is not unique. One expects that the principals select a coordinating contract from multiple coordinating contracts to extract as much surplus as possible from the agents. To address the surplus extraction question, we first show that the solutions to the reduced contract-design problem completely determine the in-equilibrium net payoffs of the principals. We can thus find the principals' net payoff Pareto frontier by optimizing the sum of weighted principal payoffs over all possible coordinating contracts. We find that the attainable utility payoff set is convex.

When the Pareto frontier is unique or symmetric, we have also implicitly answered the questions of what coordinating contracts the principals may choose and what the in-equilibrium net payoffs are to the principals and the agents. In the special case of common agency, we find that it is possible that each principal can obtain her maximum possible payoff. For a general multi-principal multi-agent contracting game, we find the Pareto frontier of the principals' net payoffs by solving a linear programming problem. Finally, we conduct numerical experiments to gain insights on contract design, and surplus extraction in multi-principal multi-agent supply chains. In a capacity constrained example, we demonstrate that selecting a coordinating contract from a complete set of coordinating contracts is of advantage to the principals, i.e., they can actually extract all the system surplus. In Cournot oligopoly examples with common agency[2], we illustrate that in the case of product substitutes, the principals can obtain their maximum attainable payoffs but cannot extract all the system surplus. In the case of product complements, the principals cannot obtain their maximum possible payoffs but can extract all the system surplus. Although our results and models are motivated by the coordination of multi-principal multi-agent supply chains, they can be applied to multi-principal multi-agent games in other application settings such as political economy, industrial organization, and auction theory.

[2]Common agency refers to a special case when there is only one agent who contracts with all multiple principals.

1.2.3 Stock Positioning in Distributions Systems with Service Constraints

A transparent one-to-one relationship between a service-constrained model and a backorder-cost model does not exist for multi-echelon inventory systems. Hence, since the 1970s two separate streams of research have evolved: one dealing with backorder-cost problems, the other one dealing with service-constrained problems. Chapter 4 investigates how to maintain the right inventory level at the right location in a multi-echelon supply chain to deliver desired end-customer service levels while minimizing inventory holding costs. In particular, we study a continuous-review distribution system with one warehouse replenishing multiple retailers subject to fill-rate type service-level constraints, who then satisfy stochastic customer demands through positive on-hand inventory. We assume that unsatisfied demand is back-ordered at each location and that the warehouse applies a first-come-first-served inventory allocation strategy. Both the warehouse and the retailers follow a base-stock policy based on local inventory information. The warehouse replenishes from an outside source with an ample supply. Each shipment requires a leadtime but has no fixed costs. Each location incurs a holding cost. The data are assumed stationary and the horizon is infinite.

We provide an algorithm that optimally allocates stock to each location while ensuring that fill-rate service requirements at the retailers are satisfied. Because distribution systems can be very large in practice, and hence optimal solutions may not be tractable for large-scale systems, we next focus on developing close-to-optimal, robust, and easy-to-use heuristics and approximations. More specifically, we propose two heuristics to allocate stock in a distribution system: the triple-search heuristic and the newsvendor heuristic. The triple-search heuristic considers three feasible inventory allocation strategies and selects the best. Extensive numerical study (over 530 problem instances) has found that this heuristic's cost is on average 1.18% more than the best base-stock policy's cost. It is computationally fast, and hence amenable for real applications. The newsvendor heuristic solves $2J^3$ newsvendor problems to allocate inventory across the distribution system. Its cost is 18% more than that of the best base-stock policy. This heuristic is not as accurate as the triple-search heuristic, but it is computationally much faster, and simpler to use and describe. The computational efficiency enables one to analyze large-scale systems that manage thousands of SKUs. Finally, this heuristic also provides a step for developing a closed-form approximation later on.

Next we propose three closed-form approximations. The main purpose of these approximations is to predict the system's performance. Our tests show that all three approximations perform this task fairly well. Using these approximations we provide insights into stock positioning and *quantify*, for example, how logistic postponement or consolidation of retailers affects the distribution system's performance. Compared to the optimal algorithm and the heuristics, these approximations require much less data and are easier to describe to non-mathematically oriented students and practitioners.

1.3 Contributions

The significance of our research is that it provides a constructive approach for finding contracts that implement a socially efficient outcome in a two-echelon supply chain consisting of multiple upstream firms and multiple downstream firms with complete information. It also provides easy-to-understand and easy-to-implement means for inventory management in large-scale distribution systems. Our contributions are summarized as follows:

First, we tackle the coordination problem in a multi-principal multi-agent supply chain setting, which is, as far as we know, the first such effort in a supply chain management context.

[3] J is the number of retailers in a distribution system.

Second, we provide a general constructive approach to treat contract design in multi-principal multi-agent supply chains. Our two-step solution provides the most comprehensive way of understanding the contract design problem. For example, based on our contract design framework, Strulovici and Weber (2004) have characterized the Pareto coordinating contracts assuming payoff concavity, and Weber and Xiong (2006) have found all socially efficient contracts in closed-form allowing for agent payoff externalities and payoff nonconcavities.

Third, our treatment of contract design is nonparametric and thus our results integrate existing contractual schemes that contain ex-post provisions into a unified framework, and also accommodates a broad range of other vertical contracts such as ties and requirement contracts.

Fourth, we take one step further in contract design to address surplus extraction problem. By characterizing and computing the Pareto frontier of the principals' payoff set, we gain insights into what coordinating contracts the principals will possibly select when multiple such contracts exist.

Fifth, the theoretic framework and results on coordination have a wide application beyond supply chain management. For example, they can be applied to the analysis of multi-principal multi-agent games in fields such as political economy, industrial organization, and auction theory.

Sixth, the easy-to-describe and easy-to-implement heuristics and approximations provide means for inventory management of large-scale systems.

Finally, the combination of the results developed in coordination and inventory management has provided a basis to explore contract design among multiple firms that manage distribution systems.

Chapter 2

Contract Design in Multi-Principal Multi-Agent Supply Chains

2.1 Introduction

In most modern supply chains a number of different organizations ("firms") contribute to the making and selling of products and services. In the eyes of the end consumers some of these products and services may be substitutes and others complements, which naturally leads to payoff interdependencies, at least for the competing firms downstream in the supply chain. Payoff interdependencies can also exist further upstream as a result of the firms' interactions with intermediate component markets. Due to legal restrictions of "anticompetitive behavior," *horizontal* interactions between different upstream or different downstream firms are typically confined to noncooperative market transactions, limiting the possibilities for explicit interfirm coordination. In contrast to this, many *vertical* relations in a supply chain are governed by nonmarket contractual mechanisms which by their very nature allow a high degree of interfirm coordination. Coordination in a supply chain is important since it avoids efficiency losses due to double marginalization, which Spengler (1950) identified as a natural consequence of noncooperative behavior as long as the market price for end consumers reflects some monopoly power. Indeed, a supply chain is said to be "coordinated" if it maximizes the aggregate net payoffs of all firms involved. Our central *research question* is to *identify contractual mechanisms that can be used to coordinate multi-principal multi-agent supply chains.* The latter terminology suggests that supply chains – similar to firms (Jensen and Meckling 1976) – can be viewed as a nexus of contracts in which principals (as the designers of the contracts) propose appropriate individually rational and incentive-compatible mechanisms to their common agents. To capture some of the existing payoff externalities in supply chains our model allows for a multitude of principals and agents, which engage in bilateral vertical contracting.

2.1.1 Literature

Fuelled by an increasing trend to outsource certain productive activities, contract design in supply chains has attracted great interest from practitioners and scholars alike. Cachon (2003) provides an excellent survey of the extant literature. The purpose of contract design generally consists of specifying a contractual mechanism that coordinates a given supply chain while all involved firms maintain control over their own actions. Most of the available results pertain to two-echelon single-principal single-agent supply chains in which the agent takes a single-dimensional decision. Coordinating solutions that have thus been proposed, including buy-back contracts (Pasternack 1985), revenue-sharing contracts (Cachon and Lariviere 2002), quantity-flexibility contracts (Tsay 1999), sales-rebate contracts (Taylor 2002), and quantity-discount contracts (Jeuland and Shugan 1983, Moorthy 1987), generally consist of parameterized reward schedules relating to the agent's action (e.g., order quantity, price, effort), as long as the latter is observable.[1] Here, in addition to admitting

[1] In models with asymmetric information the agent's action may be hidden ("moral hazard") or either party may possess some private information ("hidden information"). We assume here that outcomes are contractible in the sense that they are both observable and verifiable by a third party. This presupposes a sufficiently high level of transparency in industries. It can be achieved (at least approximately) if monitoring costs are sufficiently low or if the contractual output can be sufficiently

multiple principals and multiple agents, which can be either upstream or downstream in a two-echelon supply chain, we adopt a largely nonparametric approach with multidimensional actions.

In addition to the literature on serial two-echelon supply chains, there is some work in operations management dealing with multiple upstream firms and a single downstream firm (an "assembly system"), or, conversely, one upstream firm and multiple downstream firms (a "distribution system"). Carr and Karmarkar (2003) study competition in multi-echelon supply chains with an assembly system structure. They apply price-only contracts to achieve quantity coordination, i.e., the production quantity of each supplier (upstream firm) equals that of the manufacturer (downstream firm) who uses the suppliers' outputs as its own input in fixed proportions. However, their quantity coordination contract cannot achieve channel coordination. In a similar spirit, Majumder and Srinivasan (2003) consider competing supply chains each with a single supplier and multiple buyers ("supply trees"). The authors show that it is possible to coordinate the individual supply trees using two-part tariffs (i.e., a linear pricing schedule in addition to a fixed franchise fee). Bernstein and Federgruen (2005) investigate a distribution system with competing retailers and random demand and determine certain coordinating price-discount contracts. To the best of our knowledge, we are the first to consider the contract design problem in a multi-principal multi-agent framework in an operations management setting. Indeed, as Cachon (2003) concludes, "[m]ore research is needed on how multiple suppliers compete for the affection of multiple retailers, i.e., additional emphasis is needed on many-to-one or many-to-many supply chain structures."

In economics there has been work on principal-agent game in several directions. Bernheim and Whinston (1986) investigate a game of "common agency" (containing multiple principals and a single agent) under complete information. They show coalition-proof self-enforcing equilibria can be obtained by refining the set of Nash

well specified and measured between parties.

equilibria. Their "weakly truthful equilibria" are guaranteed to exist and yield efficient outcomes. Segal (1999) considers a game with one principal and multiple agents in which the agents' payoffs are interdependent. The author shows that as a result of the agents' payoff externalities an efficient Nash equilibrium may not exist in this game. Thus, when considering coordinating supply chain contracts, we need to limit the structure of payoff interdependencies between agents to be separable in each agent's own and the other agents' actions. The more recent work by Prat and Rustichini (2003) studies a multi-principal multi-agent game (a "game played through agents"), much in the same vein as we do. They characterize pure-strategy equilibria and provide necessary and sufficient conditions for the existence of an equilibrium with an efficient outcome. Our approach somewhat generalizes the findings by Pratt and Rustichini and focuses constructively on the *design* of general coordinating contracts that can achieve an efficient outcome. In this we aim to bridge the gap between the operations management framework and the economics literature.

2.1.2 Outline

The rest of the chapter is organized as follows. In Section 2.2 we first introduce the general model and the underlying equilibrium concept. We then focus our discussion on weakly truthful equilibria which yield efficient outcomes and thus coordinating contracts. Based on Prat and Rustichini's (2003) work we characterize weakly truthful equilibria, guarantee existence, and show that our framework applies equally well to two-echelon supply chain situations with supplier control or with buyer control. In Section 2.3 we study efficient contract design. For this, it is sufficient to consider a reduced contract design problem that expresses the equivalent original contract design problem in terms of excess measures. In this framework we provide a number of necessary and sufficient conditions for coordinating contracts. While strongest results can be obtained under a minimum of payoff interdependencies, it is also possible to solve the contract design problem in the very general case, as long as all principals' and all agents' payoffs are concave. In Section 2.4 we discuss the application of the general method to the coordination of supply chains and compare the results to standard commercial contracts often used in practice. Finally, we conclude in Section 2.5 with a discussion of the results as well as directions for future research.

2.2 The Model

2.2.1 Preliminaries

Consider a setting in which principals can write outcome-contingent contracts with a number of different agents. Let $\mathcal{M} = \{1, \ldots, M\}$ and $\mathcal{N} = \{1, \ldots, N\}$ denote the corresponding sets of M principals and N agents. After each principal $m \in \mathcal{M}$ ("she") and each agent $n \in \mathcal{N}$ ("he") signs contracts with each other,[2] agents noncooperatively implement an action (or "outcome") $x \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. The outcome vector $x = (x_n^m)_{m,n=1}^{M,N}$ contains each agent n's individual action vector $x_n = (x_n^1, \ldots, x_n^M) \in \mathcal{X}_n \subset \mathbb{R}_+^{ML}$ which in turn is composed of M different L-dimensional actions x_n^m. His action set $\mathcal{X}_n$ is thereby a compact subset of $\mathbb{R}_+^{ML}$ which contains at least one point to allow for the possibility of inaction. In a supply chain context it is useful to think of a "trade" x_n^m as an L-dimensional vector of goods and services flowing from agent n to principal m.

Each principal m designs a mapping $t^m : \mathcal{X} \to \mathcal{T}^m$ from outcomes x to transfer payments $t_n^m(x)$ directed at each agent $n \in \mathcal{N}$. The choice of principal m's transfer payment domain $\mathcal{T}^m$ thereby accommodates constraints reflecting the relationships between principals and agents, and we assume that $\mathcal{T}^m = \mathcal{T}_1^m \times \cdots \times \mathcal{T}_N^m$, where each $\mathcal{T}_n^m$ is either $\{0\}$ or $\mathbb{R}_+$. If no contractual relationship exists between principal m and agent n (cf. also Remark 3), then the n-th component of $t^m(x)$ could be constrained to vanish, i.e., $\mathcal{T}_n^m = \{0\}$, otherwise $\mathcal{T}_n^m = \mathbb{R}_+$. We assume in addition that principal m's

[2]We allow for the possibility of "selective contracting" where some principals do not sign contracts with some agents.

Symbol	*Explanation*
$\mathcal{M} = \{1, \ldots, M\}$	Set of Principals
$\mathcal{N} = \{1, \ldots, N\}$	Set of Agents
$\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$	Set of Feasible Outcomes/Actions
$\mathcal{T} = \mathcal{T}^1 \times \cdots \times \mathcal{T}^M$, $\mathcal{T}^m = \mathcal{T}_1^m \times \cdots \times \mathcal{T}_N^m$	Set of Admissible Payments ($\mathcal{T}_n^m \in \{\{0\}, \mathbb{R}_+\}$)
$C(\mathcal{X}, \mathcal{T})$	Set of Continuous Functions $f : \mathcal{X} \to \mathcal{T}$
$x = (x^m)_{m=1}^M = (x_n)_{n=1}^N = (x_n^m)_{m,n=1}^{M,N}$	Feasible Outcome/Action ($x \in \mathcal{X}$)
$t = (t^m)_{m=1}^M = (t_n)_{n=1}^N = (t_n^m)_{m,n=1}^{M,N} = (t_n^m, t_{-n}^{-m})$	Transfer Schedule ($t \in C(\mathcal{X}, \mathcal{T})$)
$\Delta = (\Delta_n^m)_{m,n=1}^{M,N}$, $\Delta^m = \sum_n \Delta_n^m$, $\Delta^n = \sum_m \Delta_n^m$	Excess Transfer Schedule
$V^m/v^m/F^m$	Principal m's Net/Gross/Excess Payoff
$U_n/u_n/G_n$	Agent n's Net/Gross/Excess Payoff
$W = \sum_m V^m + \sum_n U_n$	Total Surplus

Table 2.1: Some Notation.

transfer to agent n can be separated into *direct* and *indirect* reward components.

ASSUMPTION 1 (TRANSFER SEPARABILITY) *For any outcome $x \in \mathcal{X}$, the transfer $t_n^m(x)$ from principal $m \in \mathcal{M}$ to agent $n \in \mathcal{N}$ is separable, i.e., it can be represented in the form*

$$t_n^m(x) = t_{n,n}^m(x_n) + t_{n,-n}^m(x_{-n}),$$

where $t_{n,n}^m \in C(\mathcal{X}_n, \mathcal{T}_n^m)$ and $t_{n,-n}^m \in C(\mathcal{X}_{-n}, \mathcal{T}_{-n}^m)$ are appropriate continuous functions mapping the outcome $x = (x_n, x_{-n})$ to agent n's direct and indirect reward respectively.

Assuming transfer separability renders each agent's optimal action at the margin independent of the compensation received by other agents. Each agent n cares about his action $x_n \in \mathcal{X}_n$ and the sum of all transfers he obtains in equilibrium. His net payoff is given by

$$U_n(x; t) = u_n(x) + \sum_{m \in \mathcal{M}} t_n^m(x), \tag{2.1}$$

where $u_n(x)$ is the payoff to agent n from the outcome $x = (x_n, x_{-n})$ when he takes action x_n and all other players implement x_{-n}.

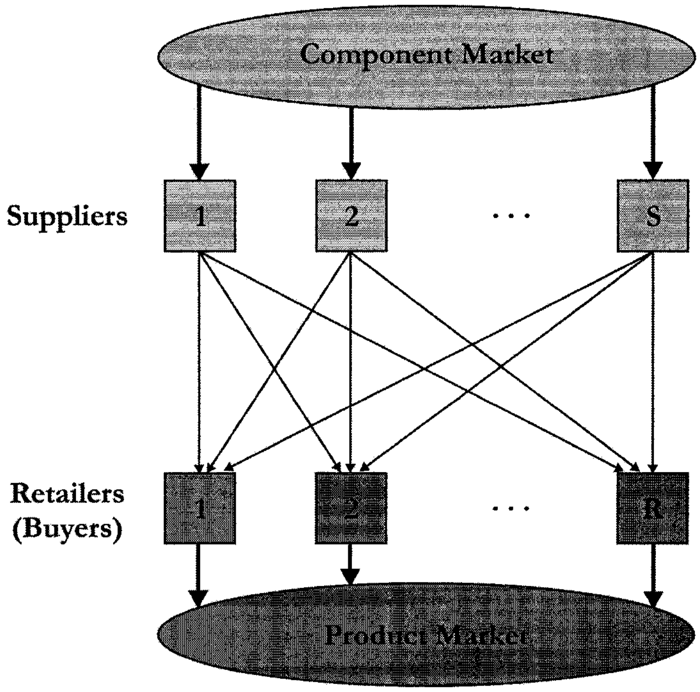

Figure 2.1: General Two-Echelon Supply Chain Setup.

ASSUMPTION 2 (AGENT PAYOFF SEPARABILITY) *Given any outcome $x \in \mathcal{X}$, each agent n's payoff can be written in the form*

$$u_n(x) = u_{n,n}(x_n) + u_{n,-n}(x_{-n}),$$

where $u_{n,n} \in C(\mathcal{X}_n, \mathbb{R})$ and $u_{n,-n} \in C(\mathcal{X}_{-n}, \mathbb{R})$ are appropriate functions representing agent n's self-generated payoff from his own action x_n and his externality payoff through the other agents' action x_{-n} respectively.

Agent payoff separability together with transfer separability (Assumption 1) implies that each principal, by changing one agent's individual payoff, does not change the actions implemented by other agents, since their marginal incentives are unconnected. When selecting optimal remuneration schemes (contracts) for the different agents, each principal cares about both her monetary payments *and* the agents' actions. Let $v^m(x)$ be principal m's *gross* payoff if action x is taken. If she offers the transfer schedule $t^m = (t_1^m, \ldots, t_N^m)$ and agents implement the outcome x, her *net* payoff is

$$V^m(x; t^m) = v^m(x) - \sum_{n \in \mathcal{N}} t_n^m(x). \tag{2.2}$$

Our modelling framework is general enough to accommodate both positive and negative transfer payments corresponding to what we refer to as *bottom-up* (principals downstream) or *top-down* (principals upstream) contracting (cf. Section 2.2.5). In the terminology chosen by Grossman and Hart (1986) we will also refer to bottom-up contracting as "buyer control" and to top-down contracting as "supplier control." In this context we consider as our leading example a two-echelon supply chain consisting of S suppliers and R retailers (buyers) with retailers buying products from their upstream suppliers (cf. Figure 2.1).[3] Depending on the balance of bargaining power in the supply chain, the retailers could act as either principals or agents. In bottom-up contracting when retailer (principal) m buys the quantity x_n^m of goods and services from supplier (agent) n, we can expect t_n^m to be positive and our framework exactly applies. In top-down contracting, when the supplier (principal) m sells a quantity vector $x_n^m > 0$ to retailer (agent) n, we might expect the transfer t_n^m from principal m to agent n to be negative, even though we earlier required that the payment domain $\mathcal{T}_n^m$ is a subset of $\mathbb{R}_+$. Nevertheless, we show in Section 2.2.5 that top-down contracting can be simply accommodated in the given framework by converting transfers into nonnegative payments ("quantity discounts" or "rebates") which are subtracted from a large enough transfer ("undiscounted wholesale price") in the opposite direction. Hence, the distinction between buying and selling, between bottom-up or top-down contracting, will prove insignificant for the results in this paper. In Section 2.4 we discuss applications of both types.

[3] If $\mathcal{S} = \{1, \ldots, S\}$ is the set of suppliers and $\mathcal{R} = \{1, \ldots, R\}$ is the set of retailers, then $(\mathcal{S}, \mathcal{R}) = (\mathcal{M}, \mathcal{N})$ in the case of top-down contracting and $(\mathcal{S}, \mathcal{R}) = (\mathcal{N}, \mathcal{M})$ in the case of bottom-up contracting.

2.2.2 Equilibrium Concept

The sequence of events is as follows: First, each principal offers her vector of transfer payment schedules to all agents simultaneously and noncooperatively. The transfer payment schedules are publicly announced to all agents.[4] Second, the agents noncooperatively implement their most preferred actions. A pure-strategy equilibrium of the two-stage game

$$\mathcal{G} = \{\{\mathcal{M}, \mathcal{N}\}, \{V^m(\cdot), U_n(\cdot)\}, \{C(\mathcal{X}, \mathcal{T}^m), \mathcal{X}_n\}\}$$

is a subgame-perfect Nash equilibrium in which all principals and agents use pure strategies.

DEFINITION 1 *A pure-strategy equilibrium of the game $\mathcal{G}$ is a pair $(\hat{t}, \hat{x}) \in C(\mathcal{X}, \mathcal{T}) \times \mathcal{X}$ in which (i) for every $n \in \mathcal{N}$ given any $t \in C(\mathcal{X}, \mathcal{T})$ it is*

$$\hat{x}_n(t) \in \arg\max_{x_n \in \mathcal{X}_n} U_n(x_n, \hat{x}_{-n}; t), \tag{2.3}$$

and (ii) for every $m \in \mathcal{M}$ given $\hat{t}^{-m} \in C(\mathcal{X}, \mathcal{T}^{-m})$ the relation

$$\hat{t}^m \in \arg\max_{t^m \in C(\mathcal{X}, \mathcal{T}^m)} V^m(\hat{x}(t^m, \hat{t}^{-m}); t^m) \tag{2.4}$$

holds.

We limit our analysis to pure-strategy equilibria of $\mathcal{G}$. If the equilibrium contracts achieve *coordination* of the supply chain, it implements by definition an efficient outcome (Cachon 2003).

DEFINITION 2 *The outcome $\hat{x} \in \mathcal{X}$ is efficient*[5] *if $W(\hat{x}) \geq W(x)$ for all $x \in \mathcal{X}$,*

[4]The case in which contracting is bilateral and each agent is only informed about his own contract terms with a particular principal is more delicate (cf. Segal and Whinston (2003) for an analysis of such a situation with one principal and n agents).

[5]By "efficient" we mean that the aggregate benefits of principals and agents (constituting the

where $W(x) = \sum_{n \in \mathcal{N}} u_n(x) + \sum_{m \in \mathcal{M}} v^m(x)$.

The set of all efficient outcomes for a given supply chain (i.e., given the principals' and agents' payoff functions) corresponds thus to the set of maximizers of the total surplus W on $\mathcal{X}$, which by the Weierstrass theorem (Bertsekas 1995, p. 540) exists and by Berge's (1963) maximum theorem is compact valued. In what follows, we assume that there is a consensus about which particular efficient outcome $\hat{x}$ is to be implemented. In other words, parties should be able to communicate about (i.e., coordinate on) the outcome. In the special case when all parties' payoff functions are strictly concave (cf. Assumption 3 below) and the set of implementable outcomes $\mathcal{X}$ is convex, there exists a *unique* efficient outcome $\hat{x}$.

Since we are considering a game with multiple principals, it is necessary to take into account the possibility of coalition formation among principals. Note that such coalitions may form noncooperatively, i.e., without binding contracts between the principals. Such coalitional games have first been considered by Von Neumann and Morgenstern (1944), who also coined the notion of a "stable equilibrium," which is such that all players want to join a coalition only if the resulting payoffs are not dominated by any coalitional deviation. Bernheim et al. (1987) introduce (via a recursive definition) the notion of a coalition-proof Nash equilibrium, which is self-enforcing among the members of any coalition. As Bernheim and Whinston (1986) demonstrate, to guarantee coalition proofness of a Nash equilibrium (in a game with multiple principals and one agent) it is sufficient to guarantee that there are no profitable coalitional deviations, which is achieved at what they term *truthful equilibria*. Prat and Rustichini (2003), while restricting attention to efficient outcomes, apply this notion to a game played through agents with multiple principals, analogous to the one considered here. It is the latter definition that we choose to adopt.

---

producer system) are maximized *excluding* end consumers (on the product market) and further upstream suppliers (on the component market), whose benefits we consider as exogenous to the principal-agent system.

DEFINITION 3 *Principal m's transfer* $t^m \in C(\mathcal{X}, \mathcal{T}^m)$ *is weakly truthful relative to the outcome* $\hat{x} \in \mathcal{X}$ *if* $V^m(\hat{x}; t^m) \geq V^m(x; t^m)$ *for all* $x \in \mathcal{X}$.

If the principals' equilibrium transfers are all weakly truthful relative to the outcome $\hat{x}$ (not necessarily assumed to be efficient), then no principal would prefer to implement a different outcome with her transfer. Correspondingly, an equilibrium where all the principals' transfers are weakly truthful with respect to the same outcome must be self-enforcing, i.e., coalition-proof as desired.

DEFINITION 4 *A weakly truthful equilibrium (WTE) of the game* $\mathcal{G}$ *is a pair* $(\hat{t}, \hat{x})$ *that is a pure-strategy equilibrium of* $\mathcal{G}$ *with outcome* $\hat{x}$ *and in which the transfer* $\hat{t}^m$ *of every principal* $m \in \mathcal{M}$ *is weakly truthful relative to* $\hat{x}$.

The notion of weak truthfulness is directly related to supply chain coordination, since any outcome $\hat{x}$ that is part of a weakly truthful equilibrium must be efficient.

PROPOSITION 1 *An outcome* $\hat{x}$ *of a weakly truthful equilibrium* $(\hat{t}, \hat{x})$ *of* $\mathcal{G}$ *is efficient.*

Since we are interested in supply chain coordination, we limit our attention to weakly truthful equilibria, for which – it turns out – there exists a simple and useful characterization.

2.2.3 Characterization of Weakly Truthful Equilibria

As Bernheim and Whinston (1986) and subsequently Prat and Rustichini (2003) indicate, it is sufficient to consider nonnegative transfer payments when considering weakly truthful equilibria. Nonnegative transfer payments are natural in a bottom-up contracting situation with buyer control, as then typically positive payments are made for any goods flowing from suppliers to buyers. Nevertheless, by a simple change of variables it is possible to equivalently formulate the contracting problem with nonnegative transfers in a top-down contracting situation with supplier control, as is shown in Section 2.2.5. The following characterization of WTEs, which somewhat generalizes Prat and Rustichini (in the sense that we allow for separable agent payoff externalities), is later used to find contracts that implement a given efficient outcome $\hat{x}$.

THEOREM 1 (CHARACTERIZATION OF A WTE) *Under Assumptions 1–2 a pair* $(\hat{t}, \hat{x})$ *arises in a weakly truthful equilibrium if and only if (i) it includes no indirect payments (i.e.,* $\hat{t}^m_{n,-n} = 0$ *for all* $(m,n) \in \mathcal{M} \times \mathcal{N}$*) and (ii) the following three conditions are satisfied:*

$$v^m(\hat{x}) - \sum_{n \in \mathcal{N}} \hat{t}^m_{n,n}(\hat{x}_n) \geq v^m(x) - \sum_{n \in \mathcal{N}} \hat{t}^m_{n,n}(x_n), \tag{WT}$$

for every principal $m \in \mathcal{M}$ *and every outcome* $x \in \mathcal{X}$*;*

$$u_{n,n}(\hat{x}_n) + \sum_{m \in \mathcal{M}} \hat{t}^m_{n,n}(\hat{x}_n) \geq u_{n,n}(x_n) + \sum_{m \in \mathcal{M}} \hat{t}^m_{n,n}(x_n), \tag{AM}$$

for every agent $n \in \mathcal{N}$ *with arbitrary action* $x_n \in \mathcal{X}_n$*; and*

$$u_{n,n}(\hat{x}_n) + \sum_{i \in M} \hat{t}^i_{n,n}(\hat{x}_n) = \max_{x_n \in \mathcal{X}_n} \left\{ u_{n,n}(x_n) + \sum_{i \neq m} \hat{t}^i_{n,n}(x_n) \right\}, \tag{PM}$$

for every principal $m \in \mathcal{M}$ *and every agent* $n \in \mathcal{N}$*.*

The intuition of the equilibrium characterization in Theorem 1 is as follows: As a consequence of Assumptions 1–2 principals cannot influence agent n's actions through transfers to different agents $j \neq n$. It follows that $t_{n,-n}$ must vanish so that only direct reward components matter. The weak truthfulness requirement in Definition 3 can thus be rewritten in the form (WT). Given the principals' equilibrium transfer schedules $\hat{t}^1, \ldots, \hat{t}^M$, the agents implement a Nash equilibrium. In other words, the outcome $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_N)$ must be composed of element of the agents' respective best-response correspondences, which by Assumption 2 is equivalent to requiring that (AM) ("agent payoff maximization") holds. Finally, in equilibrium each principal chooses her transfer schedules such as to minimize the cost of implementing the outcome $\hat{x}$. In other words, since transfers are nonnegative by assumption, principal m has to pay agent n not more than this agent would obtain by implementing his otherwise optimal action given that $t_n^m = 0$ and all other principals' reward functions stay in place. This is principal m's "cost minimization" condition (PM) with respect to agent n.

REMARK 1 As a consequence of Theorem 1 under Assumptions 1–2 at any WTE the transfer from principal m to agent n depends only on agent n's action. Thus, in what follows we write $t_n^m(x_n)$ (instead of $t_{n,n}^m(x_n)$ or $t_n^m(x)$) to denote this transfer. Similarly we simply write $u_n(x_n)$ instead of $u_{n,n}(x_n)$ where possible.

2.2.4 Existence of a Weakly Truthful Equilibrium

Assumptions 1 and 2 ensure that externalities in the agents' payoffs do not actually influence their best actions. If one of these assumptions is not satisfied, then a WTE of $\mathcal{G}$ may not exist. Indeed, Segal (1999) demonstrates in a common-agency model with only one principal that there may not exist an efficient equilibrium when externalities between agents are not separable. If all players' payoffs are concave, then a solution to the system (WT),(AM),(PM) without indirect transfers does exist.

ASSUMPTION 3 (PAYOFF CONCAVITY) *Principal m's gross payoff $v^m \in C(\mathcal{X}, \mathcal{T}^m)$ and gent n's gross payoff $u_n \in C(\mathcal{X}_n, \mathbb{R})$ are concave for all $(m,n) \in \mathcal{M} \times \mathcal{N}$.*

Prat and Rustichini (2003, Theorem 8) guarantee the existence of a weakly truthful equilibrium under Assumptions 1–3. However, their existence proof, based on a Banach-space generalization of Farkas' Lemma (Aubin and Ekeland 1984, p. 144), is highly nonconstructive and therefore offers no particular insights as to *how* to actually *design* efficient contracts, the question of central practical importance. In Section 2.3 we provide a simple explicit equilibrium construction and thus resolve the question of existence in a satisfying direct way. We also note that the payoff concavity (Assumption 3) is not necessary for the existence of equilibria. In fact, some of our efficient contract designs in later sections do not depend on this assumption.

2.2.5 Standard Two-Echelon Supply Chain Modelling

At the end of Section 2.2.1 we have already pointed out that our model directly applies to bottom-up contracting and needs some slight modifications to accommodate top-down contracting. For this consider a two-echelon supply chain with suppliers (upstream firms) being the principals and the retailers (downstream firms) representing the agents. Note first that each supplier m's gross profit function v^m and each retailer n's gross payoff function u_n is by assumption continuous on the compact set of outcomes $\mathcal{X}$ and $\mathcal{X}_n$ respectively, so that these functions are Lipschitz there. The gross benefit (or revenue) that retailer n derives from x_n is $u_n(x_n)$. Each seller offers a menu of contracts to each buyer. In a top-down contracting game the transfer payment goes from the buyers to the sellers, i.e., from the agents to the principals. In this case we modify the payoffs of the buyers and the sellers as follows: First, choose for each agent n a reference point $\tilde{x}_n = (\tilde{x}_n^1, \ldots, \tilde{x}_n^M) \in \mathcal{X}_n$ and let

$$\tilde{u}_n(x_n) = u_n(x_n) - \sum_{m \in \mathcal{M}} w_n^m \Lambda_n^m (x_n^m - \tilde{x}_n^m), \tag{2.5}$$

where w_n^m are appropriate positive constants and each $\Lambda_n^m \in C(\mathbb{R}_+^L, \mathbb{R})$ is convex and such that

$$\left\{(x_n^m, x_n^{-m}) \in \mathcal{X}_n : \Lambda_n^m(x_n^m - \tilde{x}_n^m) \geq 0\right\} = \{\tilde{x}_n\}.$$

The functions Λ_n^m can be seen as functions that penalize deviations from the reference point. The constants w_n^m and functions Λ_n^m are chosen such that the slope around the reference point is large enough (in absolute value) so that

$$\tilde{x}_n = \arg \max_{x_n \in \mathcal{X}_n} \tilde{u}_n(x_n).$$

In other words, given no transfer from any principal, agent n chooses $\tilde{x}_n$ as his strictly preferred action. For instance, if $\tilde{x}_n^m$ is an interior local maximizer of u_n on $\mathcal{X}_n$, then one might choose $\Lambda_n^m(x_n^m - \tilde{x}_n^m) = \|x_n^m - \tilde{x}_n^m\|_2^2$. In that case, if in addition the maximizer $\tilde{x}_n$ is *global* on $\mathcal{X}_n$, then any positive constants $w_n^m > 0$ ensures that the maximizer becomes strict. Defining principal m's modified payoff by

$$\tilde{v}^m(x) = v^m(x) + \sum_{n \in \mathcal{N}} w_n^m \Lambda_n^m (x_n^m - \tilde{x}_n^m)$$

ensures strategic equivalence of the game $\tilde{\mathcal{G}}$ in the modified payoffs with the original game $\mathcal{G}$. Indeed, note first that any efficient outcome $\hat{x}$ of $\mathcal{G}$ is also an efficient outcome of $\tilde{\mathcal{G}}$ and vice-versa, since total surplus W remains unaffected by the changes. In addition, note that in the game with modified payoffs each agent n requires a strictly positive transfer payment to implement any action other than $\tilde{x}_n$. Relation (2.5) together with the convexity of Λ_n^m implies that $\Lambda_n^m(x_n^m)$ is nonnegative for all feasible x_n^m. If we then consider the modified transfer

$$\tilde{t}_n^m(x_n) = t_n^m(x_n) + w_n^m \Lambda_n^m (x_n^m - \tilde{x}_n^m) \geq 0,$$

which is admissible (i.e., nonnegative) if t_n^m is, then the principals' and the agents' net payoffs V^m and U_n are identical in $\tilde{\mathcal{G}}$ and $\mathcal{G}$. Note that for large enough w_n^m, despite the nonpositivity of t_n^m in the top-down contracting setting, the modified transfer $\tilde{t}_n^m$ is always nonnegative (and thus admissible) in equilibrium.

REMARK 2 If $\mathcal{X}_n = \mathbb{R}_+^{ML}$, then with $\tilde{x}_n = 0$ it is possible to choose $\Lambda_n^m(x_n^m - \tilde{x}_n^m) = x_n^m$. The resulting linearly augmented modified transfer

$$\tilde{t}_n^m(x_n) = t_n^m(x_n) + w_n^m x_n^m$$

can be interpreted as (negative) quantity discounts $t_n^m(x_n)$ being offered by principal m to agent n relative to an "expensive" wholesale price w_n^m.

REMARK 3 (MISSING LINKS) If a trading link between principal m and agent n in the two-tier supply chain is missing (e.g., due to restricted international trade), i.e., when $\mathcal{T}_n^m = \{0\}$, then agent n chooses his action x_n independently of principal m. More specifically, he selects an action $x_n = (\hat{x}_n^m, x_n^{-m})$ that maximizes his own payoffs, given the proposed transfers by principals with trading links to him. As a result, any efficient outcome $\hat{x}$ contains the component $\hat{x}_n^m(\hat{x}_n^{-m})$ so chosen; moreover, by principal m's cost minimization (PM) her transfer $\hat{t}_n^m$ will vanish in equilibrium. With this in mind, we can thus omit the special case of missing trading links in the subsequent discussion, as it can be accommodated naturally in our framework.

2.3 Efficient Contract Design

In this section we study how efficient outcomes can be implemented in a supply chain or other setting by an appropriate contract design. Our approach is nonparametric and we are basically looking for a set of (nonlinear) transfers $\hat{t} = [\hat{t}_n^m]$ that implement $(\hat{t}, \hat{x})$ as a WTE. To accomplish this, we first simplify the problem of solving the system (WT),(AM),(PM) by ignoring (PM) , which leads to a much simpler equivalent "reduced contract design problem" to (WT)and (AM). Any solution to the latter problem can then be mapped to a solution of the original problem by adding appropriate constant transfers, such that the resulting equilibrium payment schedules satisfy (PM) and the nonnegativity constraint. The flexibility in the design of the contract depends substantially on how interdependent the payoffs in the supply chain are. In the case of additively separable payoffs the flexibility is greatest. It turns out that any efficient outcome can be implemented via a very customizable outcome-contingent surplus-sharing contract (cf. Corollary 1). If in addition payoffs are concave (i.e., Assumption 3 holds), then it is possible to implement efficient outcomes using a simple affine contract. The situation becomes more complicated in the presence of nonadditive payoff interdependencies, for which we discuss contracting solutions in Section 2.3.3.

2.3.1 Contract Design – General Results

Based on the characterization of weakly truthful equilibria, we now turn to the practical problem of designing contracts that implement an efficient outcome.[6] For this we first consider a *reduced contract design problem*, the solution to which can be directly mapped to a solution to the original efficient contract design problem. Given an efficient outcome $\hat{x} \in \mathcal{X}$, let $F^m(x) = v^m(x) - v^m(\hat{x})$ denote principal m's *excess revenue* from implementing x instead of $\hat{x}$, and let $G_n(x_n) = u_n(\hat{x}_n) - u_n(x_n)$ denote agent n's *excess cost* from taking an action x_n instead of $\hat{x}_n$. If $t_n^m(x_n)$ represents a direct transfer from principal m to agent n contingent on his taking action $x_n \in \mathcal{X}_n$, then it is useful to consider the *excess transfer*,

$$\Delta_n^m(x_n) = t_n^m(x_n) - t_n^m(\hat{x}_n), \tag{2.6}$$

relative to agent n's efficient action $\hat{x}_n$. Thus, by only looking at deviations from the efficient outcome $\hat{x}$ we can rewrite conditions (WT) and (AM) equivalently in terms of excess measures, which yields

$$F^m - \sum_{n \in \mathcal{N}} \Delta_n^m \leq 0, \tag{WT'}$$

for any principal $m \in \mathcal{M}$, and

$$G_n - \sum_{m \in \mathcal{M}} \Delta_n^m \geq 0, \tag{AM'}$$

for any agent $n \in \mathcal{N}$. In other words, at any efficient equilibrium of the underlying game $\mathcal{G}$ there does not exist a more profitable outcome than $\hat{x}$ for any principal, and any agent's excess cost of implementing any action different from $\hat{x}_n$ outweighs the excess transfer he could obtain, which thus keeps him from deviating. We call the

[6] Any such outcome can be determined as a maximizer of W on $\mathcal{X}$ (cf. Definition 2).

problem of finding a matrix function

$$\Delta = [\Delta_n^m] \in C(\mathcal{X}, \mathbb{R}^{M\times N}),$$

whose elements satisfy the system (WT'),(AM') of $M+N$ inequalities on $\mathcal{X}$, the *reduced contract design problem.* Note that a solution to the reduced contract design problem is generally not unique and does not represent an excess transfer of the form (2.6) which by its very definition vanishes at $\hat{x}$. The following result describes a class of solutions up to a constant matrix (at $\hat{x}$), so that without loss of generality it is possible to restrict attention to solutions to (WT'),(AM') that also satisfy $\Delta(\hat{x}) = 0$, a necessary condition for representing excess transfers of the form (2.6).

PROPOSITION 2 *(i) Any solution Δ to the reduced contract design problem (WT'),(AM') satisfies*

$$\sum_{i\in\mathcal{M}} \Delta_n^i(\hat{x}_n) = \sum_{j\in\mathcal{N}} \Delta_j^m(\hat{x}_j) = 0, \tag{2.7}$$

for all $(m,n) \in \mathcal{M}\times\mathcal{N}$. (ii) If Δ is a solution to (WT'),(AM') and $\delta \in \mathbb{R}^{M\times N}$ is a constant matrix, then $\Delta + \delta$ also solves the reduced contract design problem if and only if

$$\sum_{i\in\mathcal{M}} \delta_n^i = \sum_{j\in\mathcal{N}} \delta_j^m = 0, \tag{2.8}$$

for all $(m,n) \in \mathcal{M}\times\mathcal{N}$.

Proposition 2 implies that the excess transfer $\bar{\Delta}(x) = \Delta(x) - \Delta(\hat{x})$ (satisfying $\bar{\Delta}(\hat{x}) = 0$) is a solution to the reduced contract design problem, if only Δ solves the system of inequalities (WT'),(AM').[7] Thus, in everything that follows *we only consider solutions to the reduced contract design problem that satisfy* $\Delta(\hat{x}) = 0$ without having to impose this condition as an extra constraint in the search for a solution. If such an excess transfer matrix Δ has been found, then as a consequence of Theorem 1

[7]Note that $\delta = -\Delta(\hat{x})$ satisfies (2.8) as a direct consequence of (2.7).

the function

$$t_n^m(x_n) = \Delta_n^m(x_n) + \alpha_n^m,$$

with $\alpha_n^m \in \mathbb{R}_+$ being some appropriate nonnegative constant, is a candidate equilibrium transfer from principal m to agent n. The constant α_n^m corresponds to the payment contingent on the equilibrium outcome, i.e., $t_n^m(\hat{x}_n) = \alpha_n^m$.

In addition to the reduction of the solution space to solutions that vanish at the efficient outcome, we also obtain that the set of solutions to the reduced contract design problem is convex.

PROPOSITION 3 *Given any two solutions Δ and $\tilde{\Delta}$ to the reduced contract design problem (WT'),(AM'), the convex combination $\lambda\Delta + (1-\lambda)\tilde{\Delta}$ is also a solution, for any $\lambda \in (0,1)$.*

We next state the key result for the design of efficient multi-principal multi-agent contracts, given *any* solution to the reduced contract design problem.

THEOREM 2 (EFFICIENT CONTRACT DESIGN) *Under Assumptions 1–2, if the excess transfer matrix $\Delta = [\Delta_n^m]$ solves the reduced contract design problem (WT'),(AM') implementing the efficient outcome $\hat{x}$, then a WTE of the multi-principal multi-agent game is given by $(\hat{t}, \hat{x})$ with*

$$\hat{t}_n^m(x_n) = \hat{\Delta}_n^m(x_n; \vartheta_n^m) - \min_{x_n \in \mathcal{X}_n} \hat{\Delta}_n^m(x_n; \vartheta_n^m) \tag{2.9}$$

and

$$\hat{\Delta}_n^m(x_n; \vartheta_n^m) = \Delta_n^m(x_n) + \vartheta_n^m(x_n)\left(G_n(x_n) - \sum_{i \in \mathcal{M}} \Delta_n^i(x_n)\right) \tag{2.10}$$

for all $(m,n) \in \mathcal{M} \times \mathcal{N}$ and arbitrary $\vartheta_n^m \in C(\mathcal{X}_n, [0,1])$ satisfying $\sum_{m \in \mathcal{M}} \vartheta_n^m = 1$.

The last result is essential for finding coordinating contracts implementing any given efficient outcome $\hat{x}$. It allows for the solution of a reduced contract design problem (WT'),(AM') instead of the original contract design problem based on (WT),(AM), and (PM). Any solution Δ of the reduced contract design problem can then be mapped to a solution of the original problem by the outcome-contingent transformation (3.6)–(3.7). The modified transfer matrix $\hat{\Delta} = [\hat{\Delta}_n^m]$ also satisfies the reduced contract design problem (WT'),(AM'). It is remarkable that under the transformation (3.6)–(3.7) the *constant nonnegative shifts*

$$\alpha_n^m(\vartheta_n^m) = -\min_{x_n \in \mathcal{X}_n} \hat{\Delta}_n^m(x_n; \vartheta_n^m) \geq 0 \tag{2.11}$$

of the modified excess transfers $\hat{\Delta}_n^m$ *correspond exactly to the amounts transferred from principal m to agent n in equilibrium.* These amounts generally depend on the outcome-contingent convex combination selected in (3.7). For each agent $n \in \mathcal{N}$ the principals are thereby in a conflict about who should pay him less, since the higher ϑ_n^m, the lower principal m's transfer to agent n in equilibrium. If for principal m the weight $\vartheta_n^m = 1$, her equilibrium transfer to agent n is indeed as small as possible, given the solution Δ to the reduced contract design problem (WT'),(AM').

If all gross payoffs in the system are concave and differentiable, then we can state sufficient conditions for a solution to the reduced contract design problem as follows:

PROPOSITION 4 *Let all principals' and all agents' payoff functions be twice differentiable. Under Assumptions 1–3 a twice differentiable excess transfer matrix Δ solves the reduced contract design problem (WT'),(AM') if*

$$\nabla^2 (v^m - \Delta^m) \leq 0 \quad \textbf{and} \quad \nabla(v^m - \Delta^m)(\hat{x}) = 0, \tag{WT''}$$

for any principal $m \in \mathcal{M}$, and

$$\nabla^2 (u_n + \Delta_n) \leq 0 \quad \textbf{and} \quad \nabla(u_n + \Delta_n)(\hat{x}_n) = 0, \tag{AM''}$$

for any agent $n \in \mathcal{N}$, whereby $\Delta^m = \sum_{j\in\mathcal{N}} \Delta_j^m$ and $\Delta_n = \sum_{i\in\mathcal{M}} \Delta_n^i$.

Conditions (WT"),(AM") can be easily checked, which contributes to the practical importance of Proposition 4. In what follows we sometimes use this result to obtain coordinating contracts under a variety of setups.

2.3.2 Contract Design with Additive Payoffs

When payoffs are additively separable, the incentives in a supply chain system can be easily disentangled, which leads to a simple implementation of any efficient outcome by a broad variety of contracts. As we show in Section 2.4, the results obtained here nest virtually all the extant results on supply chain coordination under complete information with a single supplier and a single buyer.

Assumption 2' (Agent Payoff Additivity) *Each agent n's gross payoff u_n is additive, i.e., it can be written in the form $u_n(x_n) = \sum_{m\in\mathcal{M}} \gamma_n^m(x_n^m)$ for all $x_n \in \mathcal{X}_n$, whereby $\gamma_n^m \in C(\mathbb{R}^L, \mathbb{R})$ for all $(m,n) \in \mathcal{M} \times \mathcal{N}$.*

Agents' gross payoffs are additive if their actions take place in different markets or pertain to noncompeting product lines. Note that we do not require that the agents' gross payoffs are separable with respect to each of the L components of an action relevant to any particular principal.

Assumption 4 (Principal Payoff Additivity) *Each principal m's gross payoff v^m is additive, i.e., it can be written in the form $v^m(x) = \sum_{n\in\mathcal{N}} \pi_n^m(x_n^m)$ for all $x \in \mathcal{X}$, whereby $\pi_n^m \in C(\mathbb{R}_+^L, \mathbb{R})$ for all $(m,n) \in \mathcal{M} \times \mathcal{N}$.*

Assumptions 2' and 4 imply that each principal m's excess payoff F^m and each agent n's excess cost G_n can be rewritten as $F^m = \sum_{j\in\mathcal{N}} F_j^m$ and $G_n = \sum_{i\in\mathcal{M}} G_n^i$ respectively, where $F_n^m(x_n^m) = \pi_n^m(x_n^m) - \pi_n^m(\hat{x}_n^m)$ and $G_n^m(x_n^m) = \gamma_n^m(\hat{x}_n^m) - \gamma_n^m(x_n^m)$. To achieve complete additive separation of the contract design problem, it is necessary to impose that no externalities are introduced through infeasibilities with respect to an agent's implementation of a joint action for several principals.[8]

[8]Such infeasibilities could arise in the case in which the agent is a supplier with limited production

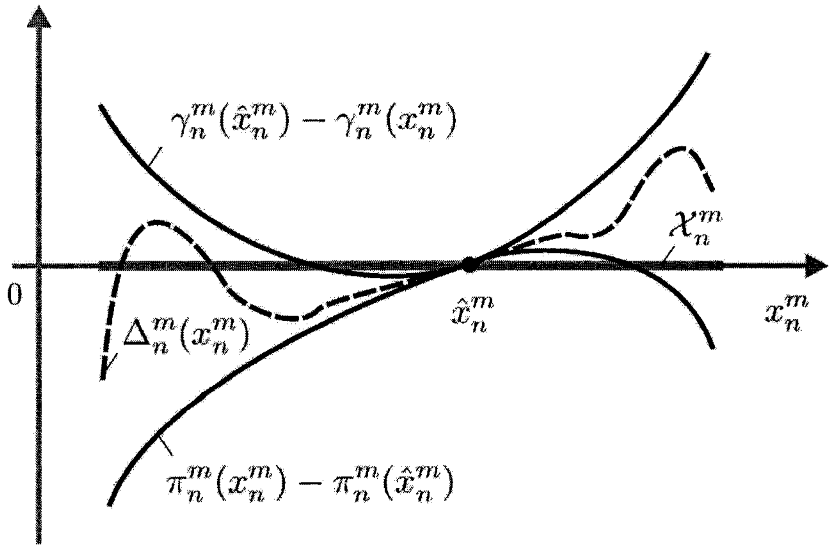

Figure 2.2: Efficient Additive Solution to the Reduced Contract Design Problem.

Assumption 5 (Outcome Decomposability) *Any agent $n \in \mathcal{N}$ can choose his actions independently for each principal, i.e., his set $\mathcal{X}_n$ of feasible actions can be written in the form $\mathcal{X}_n = \mathcal{X}_n^1 \times \cdots \times \mathcal{X}_n^M$, where each $\mathcal{X}_n^m$ is a convex compact subset of $\mathbb{R}_+^L$.*

Outcomes are decomposable if the feasibility of each action x_n^m can be determined in isolation, i.e., without taking agent n's choices pertaining to principals other than m into account. The following result characterizes efficient additive contracts:

Theorem 3 (Additive Solutions to the Reduced Contract Design Problem)

Under Assumptions 1,2',4,5 the excess transfer matrix Δ solves the reduced contract design problem (WT'),(AM') if and only if[9]

$$F_n^m(x_n^m) \leq \Delta_n^m(x_n^m) \leq G_n^m(x_n^m) \tag{2.12}$$

for all $x_n^m \in \mathcal{X}_n^m$ and all $(m,n) \in \mathcal{M} \times \mathcal{N}$.

capacity, so that when completing part of a job for one buyer (principal) not all of the production capacity can be dedicated to another buyer. By contrast, outcome decomposability requires that each agent can choose his actions independently for each principal.

[9]The "only if" part is with respect to all admissible excess payoffs F^m and G_n.

As long as the equilibrium excess transfers lie between the principals' excess profits and agents' excess costs, an efficient outcome can be implemented (cf. Figure 2.2). Hence, any convex combination of excess profits and excess costs solves the reduced contract design problem. Moreover, the convex combination itself can be made contingent upon the realization of the outcome. Thus, depending on the outcome, the principals might opt for a payoff based either more on costs or more on profits.

Corollary 1 (Efficient Surplus Sharing) *Under the assumptions of Theorem 3 the transfer matrix $\Delta(x;\theta)$ with*

$$\Delta_n^m(x_n^m;\theta_n^m) = \theta_n^m(x_n^m)\left(\pi_n^m(x_n^m) - \pi_n^m(\hat{x}_n^m)\right) + \left(1-\theta_n^m(x_n^m)\right)\left(\gamma_n^m(\hat{x}_n^m) - \gamma_n^m(x_n^m)\right)$$

solves the reduced contract design problem (WT'),(AM') for any matrix function $\theta = [\theta_n^m]$ with $\theta_n^m \in C(\mathcal{X}_n^m, [0,1])$.

For any solution $\Delta = [\Delta_n^m]$ to the reduced contract design problem specified in the above corollary, relation (3.6) in Theorem 2 yields the corresponding equilibrium transfer matrix $\hat{t} = [\hat{t}_n^m]$.

Corollary 2 (Additive Equilibrium Transfers) *Under the assumptions of Theorem 3 the only additively separable equilibrium contract implied by Theorem 2 implementing the efficient outcome $\hat{x}$ is given by the transfers*

$$\hat{t}_n^m(x_n^m) = \max_{x_n^m \in \mathcal{X}_n^m} \{\gamma_n^m(x_n^m)\} - \gamma_n^m(x_n^m). \qquad (2.13)$$

The additive equilibrium transfers in (2.13) correspond to *full cost compensation*, which is the only possible additive equilibrium contract that always coordinates an additively separable supply chain. Other contracts (many of which are used in practice, cf. Section 2.4) are thus not generally additive. This in itself is important: even though it is possible to coordinate a multi-principal multi-agent supply chain using additively separable contracts (full cost compensation) (Corollary 2), additive separability in the contracts is not necessary for coordination (the separability Assumptions 2',4,5 notwithstanding). As before, the selection of the outcome-contingent weights θ_n^m and ϑ_n^m influences the distribution of surplus within the supply chain. As an alternative to the efficient surplus sharing in Corollary 1, the reduced contract design problem may, under the additional assumption of payoff concavity (Assumption 3), also be coordinated starting with simple affine solutions to the reduced contract design problem. The following result provides sufficient conditions and is hereafter used to construct such contracts:

PROPOSITION 5 *Let all principals' and all agents' payoff functions be twice differentiable. Under Assumptions 1,2',3–5 a twice differentiable excess transfer matrix* Δ *solves the reduced contract design problem (WT'),(AM') if*

$$\nabla^2 \pi_n^m \leq \nabla^2 \Delta_n^m \leq -\nabla^2 \gamma_n^m \quad \textbf{and} \quad \nabla \Delta_n^m(\hat{x}_n^m) = \nabla \pi_n^m(\hat{x}_n^m), \tag{2.14}$$

for all $(m,n) \in \mathcal{M} \times \mathcal{N}$.

Starting with a general parameterized affine contract, its coefficients can be directly determined using Proposition 5.

COROLLARY 3 (EFFICIENT AFFINE CONTRACTS) *Under the assumptions of Proposition 5 the affine excess transfer matrix* Δ *with*

$$\Delta_n^m(x) = \langle \nabla \pi_n^m(\hat{x}_n^m), x_n^m - \hat{x}_n^m \rangle$$

solves the reduced contract design problem (WT'),(AM').

Affine contracts seem attractive in practice, since they correspond to simple two-part tariffs, essentially a constant wholesale price measured exactly in the agents' action units (say, "quantity delivered" or "sales") and a fixed transfer between the parties. It is possible to preserve affinity of solutions to the reduced contract design problem through the transformation (2.24) for all but one principal (since the weights ϑ_n^m sum up to one over m). It is important to note that even though affine contracts may be simpler in practice to implement, they are generally *not* desirable, since any principal is keen on capturing a share of the nonlinearity corresponding to out-of-equilibrium transfers, in order to reduce her in-equilibrium transfer. The design of the out-of equilibrium transfers is thus essential for surplus extraction from the agents. In Section 2.4 we examine some important standard supply chain contracting schemes within our framework.

2.3.3 Contract Design with Interdependent Payoffs

Consider now the reduced contract design problem (WT'),(AM') for the general case in which there exist payoff interdependencies between principals and in which agents' actions may not be decomposable, i.e., in which Assumptions 4 and 5 do not hold. We show that merely under Assumptions 1 through 3 there is an affine solution to the reduced contract design problem. Moreover, this solution can be easily determined provided the payoffs are differentiable at the efficient outcome under consideration.

THEOREM 4 (AFFINE SOLUTION TO THE REDUCED CONTRACT DESIGN PROBLEM) *Let all principals' and all agents' payoff functions be differentiable. Given any efficient outcome $\hat{x}$, under Assumptions 1–3 the affine excess transfer matrix $\Delta = [\Delta_n^m]$ with*

$$\Delta_n^m(x_n) = \langle \frac{\partial v^m(\hat{x})}{\partial x_n}, x_n - \hat{x}_n \rangle \qquad (2.15)$$

solves the reduced contract design problem (WT'),(AM') on $\mathcal{X}$ for all $(m,n) \in \mathcal{M}\times\mathcal{N}$.

The differentiability assumption in Theorem 4 can be relaxed, since by the Rademacher Theorem (Magaril-Il'yaev and Tikhomirov 2003, p. 160) payoff concavity (i.e., Assumption 3) already implies differentiability of the principals' and agents' payoffs

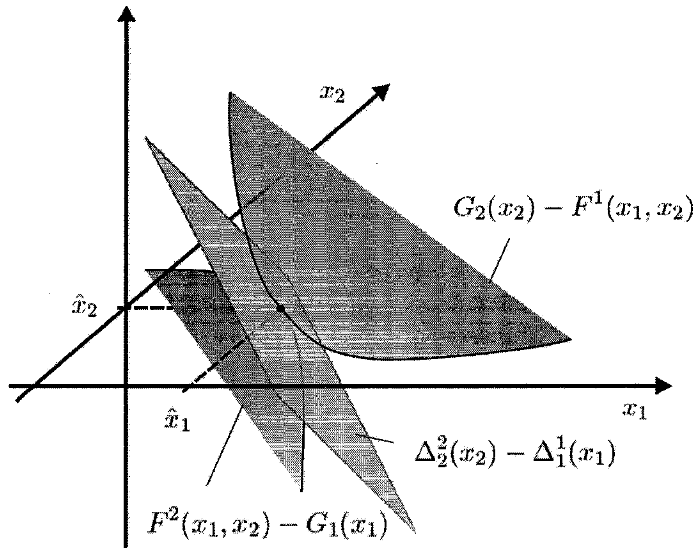

Figure 2.3: General Efficient 2 × 2 Contract. [Note: for visualization purposes x_1, x_2 have been collapsed to one dimension each.]

almost everywhere. Thus to guarantee that (2.15) is well defined, only the differentiability of the principals' payoffs at the efficient outcome $\hat{x}$ (is needed along a path to the boundary of $\mathcal{X}$ if $\hat{x} \in \partial\mathcal{X}$, so that $\mathcal{X}$ needs to be locally path-connected in that case).[10] The solution (2.15) to the reduced contract design problem in conjunction with Theorem 2 allows implementing any efficient outcome of the general multi-principal multi-agent game $\mathcal{G}$ as a WTE. Which particular WTE (in terms of the equilibrium transfer schedules) is chosen, influences the distribution of surplus in the supply chain. There is thus some interest in finding all possible solutions to the reduced contract design problem. In the simple case of two principals and two agents, a complete set of solutions to (WT'),(AM') can be easily obtained (cf. also Remark 4).

[10]Even if a principal's payoff is not differentiable at the efficient outcome, it is always possible to select an appropriate element of the corresponding subdifferential.

PROPOSITION 6 (GENERAL EFFICIENT $\mathbf{2} \times \mathbf{2}$ CONTRACTS) *Let $M = N = 2$. Under Assumptions 1–3 the excess transfer matrix Δ with*

$$\begin{aligned}\Delta_1^2(x_1;\theta_1) &= \theta_1(x_1)\left(\phi_1(x_1) - \phi_1(\hat{x}_1)\right) + \left(1 - \theta_1(x_1)\right)\left(G_1(x_1) - \Delta_1^1(x_1)\right),\\ \Delta_2^1(x_2;\theta_2) &= \theta_2(x_2)\left(\phi_2(x_2) - \phi_2(\hat{x}_2)\right) + \left(1 - \theta_2(x_2)\right)\left(G_2(x_2) - \Delta_2^2(x_2)\right),\end{aligned}$$

where (for $n \in \{1,2\}$)

$$\phi_n(x_n) = \max_{x_{-n}\in\mathcal{X}_{-n}} \left\{F^{-n}(x_n, x_{-n}) - \Delta_{-n}^{-n}(x_{-n})\right\},$$

the functions $\theta_n \in C(\mathcal{X}_n, [0,1])$ are arbitrary, and the functions Δ_n^n are chosen such that

$$F^2(x_1, x_2) - G_1(x_1) \leq \Delta_2^2(x_2) - \Delta_1^1(x_1) \leq G_2(x_2) - F^1(x_1, x_2), \tag{2.16}$$

for all $x \in \mathcal{X}$, solves the reduced contract design problem.

Payoff concavity (Assumption 3) is not necessary for the existence of excess transfers Δ_1^1, Δ_2^2 that satisfy (2.16). Nevertheless, if Assumption 3 holds, then the existence of *affine* excess transfers $\Delta_n^n(x_n) = \langle a_n, x_n - \hat{x}_n\rangle$ for $n \in \{1,2\}$ and appropriate constants $a_n \in \mathbb{R}^L$ (for $L \geq 1$) is guaranteed by the separating hyperplane theorem (Berge 1963, p. 163).

COROLLARY 4 (EFFICIENT PARTIALLY AFFINE $\mathbf{2} \times \mathbf{2}$ CONTRACTS) *If under the assumptions of Proposition 6 the principals' payoffs are differentiable at the efficient outcome $\hat{x} \in \operatorname{int}\mathcal{X}$, then the excess transfers*

$$\Delta_1^1(x_1) = \langle \frac{\partial v^1(\hat{x})}{\partial x_1}, x_1 - \hat{x}_1\rangle \quad \textit{and} \quad \Delta_2^2(x_2) = \langle \frac{\partial v^2(\hat{x})}{\partial x_2}, x_2 - \hat{x}_2\rangle$$

together satisfy (2.16) for all $x \in \mathcal{X}$.

Note that it is possible to equivalently restate Proposition 6 and, correspondingly, Corollary 4 so as to obtain analogous affine expressions for Δ_1^2 and Δ_2^1 instead of Δ_1^1 and Δ_2^2, effectively mirroring the form of "direct transfers" (from principal m to agent $n = m$) and "cross-transfers" (from principal m to agent $n \neq m$) in the results (cf. footnote 15).

REMARK 4 Building on the results obtained here, Strulovici and Weber (2004) provide a general representation of *any* solution of the reduced contract design problem (WT'),(AM') in terms of outcome-contingent convex combinations of elements of an extremal basis for this problem, which contains $M + 1$ excess transfers.

2.4 Supply Chain Coordination

We now show how the general results obtained in Sections 2.2 and 2.3 can be used to construct coordinating contracts in two-tier multi-principal multi-agent supply chains (cf. Figure 2.1). From our earlier discussions it is clear that the coordinating vertical contracts are such that the principals' transfers to the agents directly depend on the suppliers' multi-dimensional actions. For instance, in top-down contracting (cf. Section 2.2.5) a supplier's transfer to a retailer generally depends on the product quantity ordered (quantity-dependent pricing), possibly applying discounts across products and orders ("generalized tying"), and may contain clauses on the retailer's pricing policy (resale price maintenance) as well as provisions pertaining to actions for other suppliers (such as exclusive dealing). In addition, some of the contractual provisions may be contingent on the resolution of a random variable (such as demand), e.g., in royalty schemes. Although ex-post provisions can be accommodated in our framework, we emphasize that at least in the absence of renegotiation all the contractual terms are constructed based on all parties' *expected* payoffs. We thus obtain *certainty-equivalent contracts* which are ex-ante coordinating and may without loss of generality contain ex-post provisions.[11] However, ex-post provisions that result in the same expected payoffs (or expected utilities in the presence of risk aversion) are all equivalent, and thus we cannot expect specific ex-post design statements from our theory, but are nevertheless able to integrate existing contractual schemes that contain ex-post provisions (such as quantity-flexibility contracts) into a unifying framework.

In order to connect our approach to the existing literature on single-principal single-agent contracting, we first examine the different contractual provisions mentioned above and show how these generalize to multi-principal multi-agent environments. We then tackle a number of issues specific to multi-principal multi-agent environments generated by principal payoff externalities and non-decomposable agent actions.

2.4.1 Standard Contractual Provisions

In his excellent review of vertical contracting Katz (1989) outlines six functions of vertical contracts including quantity-dependent pricing, ties, royalty schemes, requirements contracts, resale price restraints, and resale customer restraints. We show how our framework can accommodate each of these functions and examine practical complete-information contracts that are often used in supply chain settings (Cachon 2003).

As our standard example we consider in this subsection a supply chain consisting of N suppliers and M retailers who each sell all N end products in geographically dispersed markets. In line with our main arguments so far we assume a bottom-up contracting situation in which the retailers act as principals. The suppliers incur two types of costs, the cost of capacity, and the cost of production after capacity is

[11]If $\tilde{\omega}$ is a random variable and $\tau_n^m(x_n, \tilde{\omega})$ is an ex-post contingent contract between principal m and agent n, then our theory makes statements about the certainty-equivalent contract $t_n^m(x_n) = E\tau_n^m(x_n, \tilde{\omega})$.

available. For simplicity we assume that the cost of capacity is increasing linearly in the installed capacity and that the cost of production is linear in the quantity produced. The suppliers install component capacity x_n^m at a unit cost $\kappa_n > 0$ prior to the realization of random demand $\tilde{D}_n^m$ which is distributed on the support $\mathbb{R}_+$ with the cumulative distribution function H_n^m. In our complete-information setup it is natural to assume that both x_n^m and κ_n are known to the retailers. Upon observing demand realizations the retailers place orders with the component suppliers, who then produce the ordered components at a unit cost of $c_n > 0$ and deliver as much of the orders as possible given their capacity constraints. The retailers and the suppliers thereby trade according to the supply contracts signed *ex ante*, i.e., before the demand has realized. The retailers then sell the products to the end consumers at fixed prices $p_n^m \geq c_n + \kappa_n$. Unmet demand is lost without additional stock-out penalty and the unsold inventory has no salvage value. Let $s(x_n^m)$ denote the expected sales of product n (made from component n) through retailer m,

$$s(x_n^m) = E \min \left\{ \tilde{D}_n^m, x_n^m \right\} = x_n^m - \int_0^{x_n^m} H_n^m(y)dy.$$

Retailer m's gross payoff can be expressed in the additively separable form

$$v^m(x) = \sum_{n \in \mathcal{N}} p_n^m s(x_n^m),$$

while supplier n's gross payoff is given by

$$u_n(x_n) = - \sum_{m \in \mathcal{M}} \left(c_n s(x_n^m) + \kappa_n x_n^m \right)$$

and is additively separable as well. The system surplus $W = \sum_{m \in \mathcal{M}} v^m + \sum_{n \in \mathcal{N}} u_n$ is strictly concave, and we obtain the classic newsvendor solution,

$$\hat{x}_n^m = (H_n^m)^{-1} \left(\frac{p_n^m - (c_n + \kappa_n)}{p_n^m - c_n} \right),$$

as the unique efficient outcome maximizing total surplus in the supply chain system. Assumptions 1,2',4,5 are satisfied, so that by Corollary 1 (Efficient Surplus Sharing) we obtain that for any $\theta_n^m \in C(\mathbb{R}_+, [0,1])$ the excess transfer

$$\Delta_n^m(x_n^m;\theta_n^m) = \theta_n^m(x_n^m)F_n^m(x_n^m) + (1-\theta_n^m(x_n^m))G_n^m(x_n^m)$$

solves the reduced contract design problem (WT'),(AM'), where

$$F_n^m(x_n^m) = p_n^m\left(s(x_n^m) - s(\hat{x}_n^m)\right)$$

are the excess profits and

$$G_n^m(x_n^m) = c_n\left(s(x_n^m) - s(\hat{x}_n^m)\right) + \kappa_n\left(x_n^m - \hat{x}_n^m\right)$$

are the excess costs for each retailer m and product n respectively. The in-equilibrium transfers α_n^m, which define the only additively separable equilibrium contracts, $\hat{t}_n^m = G_n^m + \alpha_n^m$, can be obtained from Corollary 2,

$$\alpha_n^m = (c_n s(\hat{x}_n^m) + \kappa_n \hat{x}_n^m) - \min_{x_n^m \in \mathcal{X}_n^m}\{c_n s(x_n^m) + \kappa_n x_n^m\} = c_n s(\hat{x}_n^m) + \kappa_n \hat{x}_n^m.$$

These nonnegative equilibrium payments from principal m compensate agent n for the largest possible excess payoff (in this case zero) he could achieve by not implementing the efficient outcome; he thus obtains exactly his total cost, and the principal is able to extract all surplus from the transaction.

Below we compare various standard "commercial contracts" $\xi_n^m(x_n)$ proposed by different authors and used in practice in single-principal single-agent supply chains to the coordinating multi-principal multi-agent equilibrium contracts $\hat{t}_n^m(x_n)$ in Theorem 2. It is clear that the standard commercial contracts are thereby typically of the form $\xi_n^m(x_n^m)$, i.e., payments from principal m to agent n typically depend only on actions of agent n for principal m. As can be seen from Corollary 2, the only commercial contract of this separable form which coordinates the supply chain (by implementing an efficient outcome as a WTE) is $\xi_n^m(x_n^m) = G_n^m + \alpha_n^m$, which corresponds to the direct cost compensation discussed above.

Nevertheless, by singling out one principal, say, principal M, in relation (3.7) of Theorem 2 and setting $\vartheta_n^m = 0$ for all $m \in \{1, \ldots, M-1\}$, it is possible to implement any solution of the reduced contract design problem (WT'),(AM') in equilibrium for all principals, except for principal M, who is allocated the entire beneficial "nonlinearity" in (3.7) through $\vartheta_n^M = 1$. In this way it is possible to extend a number of commercial contracts ξ_n^m, well-known for coordinating one-to-one supply chains, to the multi-principal multi-agent setting.[12] In the following discussions of frequently used commercial contracts we thus focus unless otherwise stated on the first $M-1$ principals and on solutions of the reduced contract design problem. The resulting WTEs can then be obtained by applying Theorem 2.

Quantity-Dependent Pricing

Nonlinear transfer schedules can serve both to price discriminate in situations with asymmetric information (screening contracts) and to coordinate a supply chain. In our complete-information setting we naturally limit our attention to coordination. Oren et al. (1982) demonstrated the powerful role of two-part tariffs in quantity-dependent pricing. In fact, as a direct consequence of convex analysis, as long as a nonlinear pricing schedule is concave it can be represented as the lower envelope of an indexed family of two-part tariffs. Jeuland and Shugan (1983) show that nonlinear pricing schemes can coordinate a channel, and Moorthy (1987) points out that indeed this

[12]In line with our theoretical developments we present any commercial contract ξ_n^m in terms of excess measures, relative to the transfer at the efficient outcome. We then compare it with the general surplus sharing contract in Corollary 1. As long as the weights θ_n^m in that contract are not outcome-contingent, it is possible to equivalently state ξ_n^m in terms of nonexcess measures and obtain $\check{\xi}_n^m(x_n^m) = \xi_n^m(x_n^m) + \beta_n^m$ with $\beta_n^m = \theta_n^m \pi_n^m(\hat{x}_n^m) - (1-\theta_n^m)\gamma_n^m(\hat{x}_n^m) - \alpha_n^m$.

can be achieved using simple two-part tariffs. Our findings confirm that Moorthy's intuition carries over to multi-principal multi-agent environments (cf. also Remark 4).

Two-Part Tariff Contract. The transfer from principal m to agent n contingent on expected sales $s(x_n^m)$ is

$$\xi_n^m(x_n^m) = w_n^m \left(s(x_n^m) - s(\hat{x}_n^m)\right) + \alpha_n^m.$$

From the efficient surplus sharing mechanism in Corollary 1 we obtain that with $w_n^m = p_n^m$ and $\theta_n^m = 1$ it is $\Delta_n^m + \alpha_n^m = \xi_n^m$.

Quantity-Discount Contract. Consider the commercial contract

$$\xi_n^m(x_n^m) = w_n^m(x_n^m) \left(s(x_n^m) - s(\hat{x}_n^m)\right) + \alpha_n^m.$$

From the efficient surplus sharing mechanism in Corollary 1 we obtain

$$w_n^m(x_n^m) = \theta_n^m p_n^m - (1 - \theta_n^m) \left(c_n + \frac{\kappa_n \left(x_n^m - \hat{x}_n^m\right)}{s(x_n^m) - s(\hat{x}_n^m)} \right).$$

for any appropriate $\theta_n^m \in C(\mathcal{X}_n^m, [1/2, 1])$. Note that for $x_n^m < \hat{x}_n^m$ the price $w_n^m(x_n^m) < p_n^m$ is monotonically decreasing in x_n^m (progressive discounts). Moreover, close to $\hat{x}_n^m$, we have $w_n^m(x_n^m) \approx (2\theta_n^m - 1)$, so that we restrict θ_n^m to values in $[1/2, 1]$.

Ties

Given that each agent's action is L-dimensional, it is possible for $L > 1$ that a principal m's equilibrium transfers are not additively separable in the different components, $x_{n,1}^m, \ldots, x_{n,L}^m$, of agent n's action.[13] In that case, the compensation for different components of an agent's action is linked, which amounts to a (generalized) tying arrangement. Tying arrangements in this sense may arise naturally as a consequence of (anti-) complementarities in an agent's cost structure, as can be easily

[13] If the transfer is twice differentiable, this corresponds to the situation in which $\partial^2 t_n^m / \partial x_{n,j}^m \partial x_{n,l}^m \neq 0$ for some $j, l \in \{1, \ldots, L\}$ with $j \neq l$.

seen from the coordinating contract $[\hat{\Delta}_n^m] = [G_n^m]$ in Corollary 2: tying occurs whenever $\partial^2 G_n^m / \partial x_{n,j}^m \partial x_{n,l}^m \neq 0$ for some $j, l \in \{1, \ldots, L\}$ with $j \neq l$. It also may be induced by non-additively-separable principal payoffs. We emphasize that tying arrangements under these circumstances are efficient in the sense of maximizing overall surplus of the supply chain.

Royalty Schemes

In contracts with royalty schemes, the transfer payment between a supplier and a buyer is a function of the buyer's sales in the final goods market rather than based on the amount of intermediary goods exchanged. In Section 2.4.1 we have already dealt with some royalty schemes based on sales. Let us now discuss a number of additional commercial contracts with royalty schemes that are commonly found in supply chains, and that – as we show – can be used in multi-principal multi-agent supply chains (except for one principal, as mentioned at the beginning of this section). These contracts are prevalent in markets for goods with a relatively short shelf life, such as periodicals, baked goods, or current car models.

Pay-Back Contract. One possibility for a supplier to coordinate a one-to-many supply chain recognized by Pasternack (1985) is the pay-back contract,[14] in which retailer m pays supplier n an amount of w_n^m per unit purchased, plus b_n^m per unit remaining at the end of the season in order to incentivize the supplier's capacity investment. We thus obtain

$$\xi_n^m(x_n^m) = w_n^m \left(s(x_n^m) - s(\hat{x}_n^m)\right) + b_n^m \left[(x_n^m - \hat{x}_n^m) - (s(x_n^m) - s(\hat{x}_n^m))\right] + \alpha_n^m.$$

[14]The original setup is in terms of top-down contracting. For ease of exposition it is framed here as a bottom-up contracting situation.

Comparing coefficients of the above commercial contract with the efficient surplus sharing mechanism in Corollary 1, we obtain

$$\theta_n^m = 1 - \frac{b_n^m}{\kappa_n} \quad \text{and} \quad w_n^m = p_n^m + b_n^m \left(1 - \frac{p_n^m - c_n^m}{\kappa_n}\right),$$

provided that $b_n^m \in [0, \kappa_n]$. As noted at the beginning of the section, the retailers are able under this scheme not only to coordinate the supply chain but also to extract all surplus from the suppliers.

Revenue-Sharing Contract. Cachon and Lariviere (2002) discuss revenue sharing as a way to coordinate supply chains in which reliable revenue monitoring is feasible, such as in the market for video rentals. Under such a revenue-sharing scheme retailer m pays supplier n an amount of w_n^m per unit capacity installed, plus a fraction $\varphi_n^m \in [0,1]$ of his revenue. The resulting commercial contract is of the form

$$\xi_n^m(x_n^m) = w_n^m \left(x_n^m - \hat{x}_n^m\right) + \varphi_n^m p_n^m \left(s(x_n^m) - s(\hat{x}_n^m)\right) + \alpha_n^m.$$

By comparing coefficients with the efficient surplus sharing contract in Corollary 1 we obtain

$$\theta_n^m = \frac{\varphi_n^m p_n^m - c_n}{p_n^m - c_n} \quad \text{and} \quad w_n^m = \frac{(1 - \varphi_n^m) p_n^m \kappa_n}{p_n^m - c_n},$$

provided that $\varphi_n^m \in [c_n/p_n^m, 1]$. As before, in equilibrium the suppliers' net payoffs are zero and the supply chain is coordinated.

Quantity-Flexibility Contract. Tsay (1999) studies supply chain coordination with quantity-flexibility contracts. Under a quantity-flexibility contract mechanism, retailer m pays supplier n an amount of w_n^m per unit purchased and compensates the supplier for unused capacity up to a fraction $\rho_n^m \in [0,1]$ of total capacity installed.

We thus obtain the following formulation of the commercial contract:

$$\begin{aligned}\xi_n^m(x_n^m) &= \kappa_n\left(\min\{x_n^m - s(x_n^m), \rho_n^m x_n^m\} - \min\{\hat{x}_n^m - s(\hat{x}_n^m), \rho_n^m \hat{x}_n^m\}\right) \\ &\quad + w_n^m\left(s(x_n^m) - s(\hat{x}_n^m)\right) + \alpha_n^m. \end{aligned} \tag{2.17}$$

By comparing coefficients with the efficient surplus sharing contract in Corollary 1 we obtain thus

$$\theta_n^m(x_n^m) = \frac{(w_n^m - c_n)s(x_n^m) - \kappa_n x_n^m + \kappa_n \int_{(1-\rho_n^m)x_n^m}^{x_n^m} H_n^m(y)dy}{(p_n^m - c_n)s(x_n^m) - \kappa_n x_n^m},$$

and

$$c_n - \frac{\kappa_n \bar{x}_n^m\left(1 - \rho_n^m H_n^m\left((1-\rho_n^m)\bar{x}_n^m\right)\right)}{s(\bar{x}_n^m)} \leq w_n^m \leq p_n^m - \frac{\kappa_n \rho_n^m \bar{x}_n^m}{s(\bar{x}_n^m)},$$

provided that $\rho_n^m \in [\frac{p_n^m s(\bar{x}_n^m)}{\kappa_n \bar{x}_n^m}, 1]$, where $\bar{x}_n^m$ is the upper bound of the capacity supplier n can install for retailer m. As before, in equilibrium the suppliers' net payoffs are zero and the supply chain is coordinated.

Sales-Rebate Contract. Taylor (2002) considers a sales-rebate contract for supply chain coordination, under which retailer m pays supplier n an amount of w_n^m per unit purchased and an extra rebate r_n^m per unit sold above a threshold q_n^m. The resulting (ex-ante) commercial contract is given by

$$\xi_n^m(x_n^m) = \alpha_n^m + \begin{cases} w_n^m\left(s(x_n^m) - s(\hat{x}_n^m)\right), & \text{if } x_n^m \leq q_n^m, \\ w_n^m\left(s(x_n^m) - s(\hat{x}_n^m)\right) + r_n^m\left((\hat{x}_n^m - x_n^m) - \int_{x_n^m}^{\hat{x}_n^m} H_n^m(y)dy\right), & \text{otherwise.} \end{cases}$$

By comparing coefficients with the efficient surplus sharing contract in Corollary 1 we obtain

$$\theta_n^m(x_n^m) = \frac{(w_n^m - c_n)s(x_n^m) - \kappa_n x_n^m + r_n^m\left(x_n^m - q_n^m - \int_{q_n^m}^{x_n^m} H_n^m(y)dy\right)}{(p_n^m - c_n)s(x_n^m) - \kappa_n x_n^m},$$

with

$$c_n + \frac{\kappa_n \bar{x}_n^m}{s(\bar{x}_n^m)} \leq w_n^m \leq p_n^m - \frac{r_n^m(\bar{x}_n^m - q_n^m)(1 - F_n^m(q_n^m))}{s(\bar{x}_n^m)}$$

for $x_n^m > q_n^m$, and $\theta_n^m = 1$ with $w_n^m = p_n^m$ otherwise. As before, in equilibrium the suppliers' net payoffs are zero and the supply chain is coordinated.

Requirements Contracts

We have already seen that coordinating agreements in a multi-principal multi-agent supply chain are likely to make principal m's equilibrium transfer $\hat{t}_n^m$ to agent n contingent on not only agent n's action x_n^m for principal m but also his actions x_n^{-m} for the other principals. Indeed, as pointed out earlier, the coordinating contracts in Theorem 2 generally exhibit this property at least for one principal (except in the special case of full cost compensation in Corollary 2). Bilateral contracts containing provisions that affect an agent's payoff with respect to his behavior across different principals are generally termed *requirements contracts.* In the extreme, requirements contracts could involve *exclusive dealing* arrangements, in which certain agents exclusively trade with certain principals and are compensated accordingly. In cases where agents each implement multiple actions, such as capacity orders and pricing decisions, requirements contracts may also include *resale price restraints* (e.g., to achieve a price maintenance level) or *resale customer restraints* (e.g., to guarantee territoriality and thus restrict agent competition detrimental to overall supply chain profit). As a direct consequence of Theorem 2, to implement efficient outcomes in multi-principal multi-agent supply chains, at least one requirements contract is generally unavoidable, except possibly under additive payoffs and decomposable actions.

2.4.2 Nonstandard Contractual Provisions

Instead of adapting available commercial contracts to our framework as in the last subsection, we now tackle issues specific to *multi*-principal *multi*-agent environments.

For this, we first relax Assumption 5 and consider the case in which the agents' actions are interdependent with respect to principals, in the sense that the feasibility of each action x_n^m cannot be determined in isolation. That is, for agent n the choice of x_n^m depends on his actions chosen for principals other than m. Next, we relax Assumptions 2' and 4 and discuss contract design in a supply chain with payoff externalities.

Interdependent Actions

Let us consider a supply chain consisting of two upstream component suppliers and two retailers who each sell one end product in geographically dispersed markets. Again, in line with our main argument, we assume a bottom-up contracting scenario in which the retailers act as principals. The retailers and the suppliers trade according to supply contracts signed *ex ante*, i.e., before the uncertain demand is realized. Each supplier has a limited installed capacity to be allocated to the retailers. For simplicity we normalize the total capacity at each supplier to one. Thus, if x_n^m denotes supplier n's capacity allocation to retailer m, it is $x_1^1 + x_1^2 = 1$ and $x_2^1 + x_2^2 = 1$. The costs to install the capacities are assumed sunk, i.e., supplier n's gross payoff is given by $u_n = 0$. We further suppose that additional capacity is a desirable resource which both retailers prefer to have more of, and that the lead time for production is long, so that suppliers need to set production quantities *before* observing the demand realizations. Once the quantity x_n^m of product n is produced and delivered to retailer m, the retailer sells the product to the end consumers at a fixed price p_n. As in the last subsection, the random demand $\tilde{D}_n^m$ has the cumulative distribution function H_n^m, with mean μ_n^m. To simplify the algebra in this example, we assume that $\tilde{D}_n^m$ is uniformly distributed on $[0,1]$. Let $s(x_n^m)$ denote the expected sales of product n (made from component n) through retailer m, i.e., $s(x_n^m) = E\min\left\{\tilde{D}_n^m, x_n^m\right\}$. Retailer m's gross payoff can then be expressed in the form

$$v^m(x) = \sum_{n \in \mathcal{N}} p_n^m s(x_n^m),$$

while supplier n's gross payoff is given by $u_n(x_n) = 0$. The system surplus $W = \sum_{m\in\mathcal{M}} v^m + \sum_{n\in\mathcal{N}} u_n$ is strictly concave, and we obtain a unique efficient outcome, $\hat{x}_1^1 = \hat{x}_1^2 = \hat{x}_2^1 = \hat{x}_2^2 = 1/2$, maximizing total surplus in the supply chain system. Since Assumption 5 is not satisfied, Corollary 1 (Efficient Surplus Sharing) cannot be applied. Instead, we apply Proposition 6 and Corollary 4 to obtain the excess transfers that solve the reduced contract design problem (WT'),(AM'). By Corollary 4 we have

$$\Delta_1^1 = p_1\left(\frac{x_1^1}{2} - \frac{1}{4}\right) \quad \text{and} \quad \Delta_2^2 = p_2\left(-\frac{x_2^1}{2} + \frac{1}{4}\right),$$

and from Proposition 6 we obtain

$$\Delta_1^2 = \theta_1 p_1\left(-(x_1^1)^2 + \frac{x_1^1}{2}\right) - (1-\theta_1)p_1\left(\frac{x_1^1}{2} - \frac{1}{4}\right),$$

and

$$\Delta_2^1 = \theta_2 p_2\left(-(x_2^1)^2 + \frac{3}{2}x_2^1 - \frac{1}{2}\right) + (1-\theta_2)p_2\left(\frac{x_2^1}{2} - \frac{1}{4}\right),$$

for $\theta_n \in (\mathcal{X}_n, [0,1])$. For simplicity we assume that the weights θ_n are not outcome-contingent, (i.e., are constant), and we let $\vartheta_n^1 = 0$ and $\vartheta_n^2 = 1$ for $n \in \{1,2\}$. The in-equilibrium transfers α_n^m which define the equilibrium contracts, can be obtained from Theorem 2. We thus have

$$\alpha_1^1 = \alpha_1^2 = \frac{p_1}{4}, \quad \alpha_2^1 = \frac{p_2(1+\theta_2)}{4}$$

and

$$\alpha_2^2 = \begin{cases} \frac{p_2(1-\theta_2)}{4}, & \text{if } 0 \le \theta_2 \le \frac{1}{2}, \\ \frac{p_2}{16\theta_2}, & \text{if } \frac{1}{2} < \theta_2 \le 1. \end{cases}$$

These nonnegative equilibrium payments from principal m compensate agent n for the largest possible excess payoff (in this case nonzero) he could achieve by not implementing the efficient outcome. In equilibrium, the net payoffs to retailer 1 and supplier 1 are

$$V^1 = \frac{(p_1 + p_2) - \theta_2 p_2}{4} \quad \text{and} \quad U_1 = \frac{p_1}{2},$$

and the net payoffs to retailer 2 and supplier 2 are

$$V^2 = \begin{cases} \frac{(p_1+p_2)+\theta_2 p_2}{4}, & \text{if } 0 \le \theta_2 \le \frac{1}{2}, \\ (\frac{p_1+2p_2}{4} - \frac{1}{16\theta_2})p_2, & \text{if } \frac{1}{2} < \theta_2 \le 1, \end{cases} \quad U_2 = \begin{cases} \frac{p_2}{2}, & \text{if } 0 \le \theta_2 \le \frac{1}{2}, \\ \frac{p_2(1+\theta_2)}{4} + \frac{p_2}{16\theta_2}, & \text{if } \frac{1}{2} < \theta_2 \le 1, \end{cases}$$

respectively. In equilibrium, each supplier's net payoff is strictly positive. When the actions are interdependent, the retailers (principals) are not able to extract all surplus from the transaction. This result is in contrast to the decomposable action case in which the net payoffs to the suppliers (agents) are zero and the retailers (principals) extract all the surplus. It is worth noting that since retailer 2 shares the nonlinearity of the excess cost ($\vartheta_n^2 = 1$), in equilibrium, the net payoff of retailer 2 is strictly larger than that of retailer 1. Moreover, both retailers as well as supplier 2's net payoffs are dependent on the choice of θ_2, which illustrates that surplus sharing in equilibrium depends on θ_2.

Interdependent Payoffs

Let us now consider two cases of payoff interdependencies in a supply chain consisting of two component suppliers and two retailers. In the first case the retailers each sell (in geographically dispersed markets) two end products which can be either substitutes or complements, i.e., a demand externality (or "product externality") exists in each market. In the second case the retailers sell their end products in a common market as imperfect substitutes, which again as a result of a demand externality on the common market ("market externality") naturally entails a payoff interdependency between the retailers.

Again, in line with our main argument, we assume a bottom-up contracting scenario in which the retailers act as principals. The retailers and the suppliers trade according to the supply contracts signed *ex ante*. Given the contracts, supplier n with infinite capacity decides on how much to produce and to deliver to retailer m, i.e., x_n^m. For simplicity we assume that the production cost is linear in the quantity produced. Upon receiving the delivered products, retailer m sells the end product n to the end consumers at a retail price p_n^m to clear the market. Given a unit production cost $c_n > 0$, supplier n's gross payoff is given by

$$u_n(x_n) = -\sum_{m\in\mathcal{M}} c_n x_n^m.$$

Product Externality. The two retailers each sell two end products in geographically dispersed markets. We consider a linear demand of the form

$$p_n^m(x^m) = a_n^m - b_n^m x_n^m + d^m x_{-n}^m,$$

where $a_n^m > 0, b_n^m > |d^m|$. The constants d^m can be either positive or negative depending on the products being either substitutes or complements. When the markets are geographically dispersed, retailer m's gross payoff can be expressed as

$$v^m(x) = \sum_{n\in\mathcal{N}} p_n^m(x^m)x_n^m.$$

The system surplus $W = \sum_{m\in\mathcal{M}} v^m + \sum_{n\in\mathcal{N}} u_n$ is strictly concave, and we obtain

$$\hat{x}_n^m = \frac{(a_n^m - c_n)b_{-n}^m + (a_{-n}^m - c_{-n})d^m}{2(b_n^m b_{-n}^m - (d^m)^2)},$$

as the unique efficient outcome maximizing total surplus in the supply chain system. Assumptions 1–3 are satisfied, so that by Theorem 4 (Affine Solution to the Reduced Contract Design Problem) we obtain that the excess transfer

$$\Delta_n^m(x_n^m) = (a_n^m - 2b_n^m \hat{x}_n^m + 2d^m x_{-n}^m)(x_n^m - \hat{x}_n^m) = c_n(x_n^m - \hat{x}_n^m)$$

solves the reduced contract design problem (WT'),(AM'), where the second equality follows from the system optimality condition. The in-equilibrium transfers α_n^m, which define the equilibrium contracts, can be obtained from Theorem 2. For simplicity we let $\vartheta_n^1 = 0$ and $\vartheta_n^2 = 1$ for $n \in \{1, 2\}$. We thus have

$$\alpha_n^m = c_n \hat{x}_n^m \quad \text{for} \quad m \in \{1,2\} \quad \text{and} \quad n \in \{1,2\}.$$

These nonnegative equilibrium payments from retailer m compensate supplier n for the largest possible excess payoff (in this case zero) he could achieve by not implementing the efficient outcome; he thus obtains exactly his total cost and the principal is able to extract all surplus from the transaction. Moreover, the payoffs are independent of the sign of d^m. That is, no matter whether the products are substitutes or complements, the retailers extract all the surplus.

Market Externality. The retailers sell end products in a common market. We consider a linear demand of the form

$$p_n^m(x) = a_n^m - b_n^m x_n^m - \omega_n x_n^{-m},$$

where $a_n^m > 0, b_n^m > \omega_n > 0$. Retailer m's gross payoff can be expressed as

$$v^m(x) = \sum_{n \in \mathcal{N}} p_n^m(x) x_n^m.$$

The system surplus $W = \sum_{m \in \mathcal{M}} v^m + \sum_{n \in \mathcal{N}} u_n$ is strictly concave, and we obtain

$$\hat{x}_n^m = \frac{(a_n^m - c_n) b_n^{-m} - (a_n^{-m} - c_n)\omega_n}{2(b_n^m b_n^{-m} - (\omega_n)^2)},$$

as the unique efficient outcome maximizing total surplus in the supply chain system. We apply Proposition 6 and Corollary 4 to obtain the excess transfers that solve the reduced contract design problem (WT'),(AM'). To keep the algebra in this example simple, we let $\theta_n = 0$ in Proposition 6. We thus obtain

$$\Delta_n^m = c_n(x_n^m - \hat{x}_n^m) + \omega_n(\hat{x}_n^{-m} x_n^m - \hat{x}_n^m x_n^{-m})$$

for $m \in \{1, 2\}$ and $n \in \{1, 2\}$. The in-equilibrium transfers α_n^m which define the equilibrium transfers, can be obtained from Theorem 2. For simplicity we let $\vartheta_n^1 = 0$ and $\vartheta_n^2 = 1$ for $n \in \{1, 2\}$. We thus have

$$\alpha_n^m = c_n \hat{x}_n^m + \omega_n \hat{x}_n^m \bar{x}_n^{-m}$$

for $m \in \{1, 2\}$ and $n \in \{1, 2\}$, where $\bar{x}_n^{-m}$ is the upper bound of the production quantity from agent n to a principal other than m. These nonnegative equilibrium payments from principal m compensate agent n for the largest possible excess payoff (in this case zero) he could achieve by not implementing the efficient outcome; because of the competition between the two retailers in the end markets, he obtains more than his total cost and both principals are thus unable to extract all surplus from the transaction under these contracts.

2.5 Discussion

Supply chains with many participating firms are a ubiquitous reality. It may therefore seem surprising that the coordination of such supply chains has received virtually no attention in the existing literature. This lack of attention is most likely not due to oversight but to the technical difficulties of supply chain coordination, which have been partially overcome by recent advances in the economics literature. Illustrating the additional complications, Cachon's (2003) postulate that "[e]ach firm in a supply chain must execute a precise set of actions to achieve optimal supply chain performance" acquires a new meaning in multi-principal multi-agent supply chains, since maximizing the firms' overall surplus requires not only vertical coordination through contracts but also implicit horizontal coordination, however without the use of anti-competitive practices. The noncooperative nature of the principals' contract design introduces a number of technical difficulties, and this may account for the current lack of research in multi-principal multi-agent supply chain contracting, a significant gap we hope to fill at least partially: as long as all agents' actions are contractible, payoff externalities between agents are additive, and all gross payoffs are concave, we have provided a set of contracts that coordinate any multi-principal multi-agent supply chain. More specifically, starting from a solution to a "reduced contract design problem" (Theorem 4) we have shown that it is possible to obtain coordinating contracts that allocate surplus to the different participants in the supply chain by assigning appropriate outcome-contingent weights in the simple transformation (Theorem 2) leading to coordinating bilateral contracts in terms of (nonlinear) transfer payment schedules. We emphasize that in contrast to most of the available results, even in the literature on single-principal single-agent supply chain contracting, our approach is entirely nonparametric: weighting functions are used solely to select particular elements from the set of coordinating contracts.

In addition to filling a void in the theoretical literature on supply chain contracting, our results have a number of interesting practical implications, of which we stress only two. *First*, virtually all known (and used) commercial contracts can be employed by (almost all) principals to coordinate multi-principal multi-agent supply chains, at least if all payoffs are additively separable and the agents' actions are not constrained across principals. The results on single-principal single-agent supply chains are thus naturally nested in our more general framework. However, we have shown that to coordinate the supply chain in many cases at least one principal (any one with positive weight ϑ_n^m in Theorem 2) needs to propose a requirements contract which contains provisions with respect to an agent's actions for other principals. *Second*, affine contracts – even though easy to write down – are generally not desirable to principals: for any given agent n, the *more* a principal is able to promise *out of equilibrium*, the *less* she has to pay this agent *in equilibrium*! In other words, in-equilibrium transfers are dramatically related to the out-of-equilibrium contract design. However, since in order to enable supply chain coordination the total amount of out-of-equilibrium promises is limited for each agent (to G_n), each principal has a vested interested to capture as much of these feasible promises as possible. As a result, the allocation of the (outcome-contingent) weights ϑ_n^m (which sum to one over all $m \in \mathcal{M}$) is most likely subject to negotiation, despite the fact that the actual equilibrium implementation is by construction noncooperative and is thus not in conflict with anticompetitive clauses.

The results obtained here are limited in the sense that they do not allow for non-additive externalities between agents. As pointed out before, under such externalities an efficient equilibrium may not exist and thus it may be impossible to fully coordinate the supply chain; nevertheless, the situation with agent externalities remains important in practice (e.g., in top-down contracting where suppliers control a number of retailers that sell on a common market) and thus presents a promising direction for further research. Another limitation, and thus an opportunity for further research, lies in our nonexhaustive answer to the question of selecting "appropriate" coordinating contracts out of the set of feasible coordinating contracts. Depending on the situation this selection might be guided by aspects such as revenue extraction or practical implementability. For instance, under customary compliance regimes in a certain industry some contracts might be preferred to others; or, some contracts (e.g., the ones without requirements clauses) may allow for simpler monitoring given the industry specifics. The last points us to another major research direction related to multi-principal multi-agent contracting in which we foresee much activity: the relaxation of the full information (or full contractibility) assumption, allowing for moral hazard and/or hidden information in the vertical relationships.

2.6 Appendix A: Proofs

Proof of Proposition 1. Since $\hat{t}^m$ is weakly truthful with respect to the equilibrium outcome $\hat{x}$ we obtain by Definition 3 that for all $x \in \mathcal{X}$:

$$v^m(\hat{x}) - \sum_{n \in \mathcal{N}} \hat{t}_n^m(\hat{x}) \geq v^m(x) - \sum_{n \in \mathcal{N}} \hat{t}_n^m(x). \tag{2.18}$$

By (2.3) we have that

$$u_n(\hat{x}, \hat{t}) + \sum_{m \in \mathcal{M}} \hat{t}_n^m(\hat{x}) \geq u_n(x, \hat{t}) + \sum_{m \in \mathcal{M}} \hat{t}_n^m(x) \tag{2.19}$$

for all $x \in \mathcal{X}$. Hence, by summing up relation (2.18) over all $m \in \mathcal{M}$ and relation (2.19) over all $n \in \mathcal{N}$ we obtain that $W(\hat{x}) \geq W(x)$ for all $x \in \mathcal{X}$. Thus, the outcome $\hat{x}$ is efficient by Definition 2. ■

Proof of Theorem 1. Using the separability Assumptions 1–2 the corresponding relation (2.3) can be equivalently rewritten in the form (AM). Principal m, given the other principals' transfer vector $\hat{t}^{-m}$, can induce agents to implement any outcome $x \in \mathcal{X}$, if only she promises each agent n a transfer that is larger than the difference the agent would obtain by implementing his otherwise preferred action. In other words, to persuade agent n implement the outcome x_n which is part of her desired overall outcome $x = (x_1, \ldots, x_N)$, principal m's transfer $\hat{t}_{n,n}^m(x_n)$ to agent n as a reward for action x_n needs to satisfy

$$\begin{aligned}
\hat{t}_{n,n}^m(x_n) \;&\geq\; \max_{\kappa_n \in \mathcal{X}_n} \left\{ u_n(\kappa_n, x_{-n}) + \sum_{i \neq m} \hat{t}_n^i(\kappa_n, x_{-n}) \right\} - \left(u_n(x) + \sum_{i \neq m} \hat{t}_n^i(x) \right) \\
&= R_n^m - \left(u_{n,n}(x_n) + \sum_{i \neq m} \hat{t}_{n,n}^i(x_n) \right),
\end{aligned}$$

where

$$R_n^m = \max_{\kappa_n \in \mathcal{X}_n} \left\{ u_{n,n}(\kappa_n) + \sum_{i \neq m} \hat{t}_{n,n}^i(\kappa_n) \right\}$$

is agent n's highest *incremental* (i.e., over and above what he obtains by simply free-riding on the other agents' actions) reward without principal m. As a consequence, any equilibrium outcome $\hat{x}$ needs to maximize principal m's net payoff,

$$\begin{aligned} V^m(x, \hat{t}^m) &= v^m(x) - \sum_{n \in \mathcal{N}} \left(\hat{t}_{n,n}^m(x_n) + \hat{t}_{n,-n}^m(x_{-n}) \right) \\ &\geq v^m(x) - \sum_{n \in \mathcal{N}} \left(R_n^m + \hat{t}_{n,-n}^m(x_{-n}) - \left(u_{n,n}(x_n) + \sum_{i \neq m} \hat{t}_{n,n}^i(x_n) \right) \right), \end{aligned}$$

so that

$$\hat{x} \in \arg\max_{k \in \mathcal{X}} \left\{ v^m(x) + \sum_{n \in \mathcal{N}} \left(u_{n,n}(x_n) - \hat{t}_{n,-n}^m(x_{-n}) + \sum_{i \neq m} \hat{t}_{n,n}^i(x_n) \right) \right\}. \tag{2.20}$$

Let us now consider principal m's equivalent cost minimization problem. Given an outcome $\check{x}$ principal m solves

$$\min_{t^m \in C(\mathcal{X}, \mathcal{T}^m)} \sum_{n \in \mathcal{N}} t_n^m(\check{x}), \tag{2.21}$$

subject to

$$t_{n,n}(\check{x}_n) + u_{n,n}(\check{x}_n) + \sum_{i \neq m} t_{n,n}^i(\check{x}_n) \geq t_{n,n}(x_n) + u_{n,n}(x_n) + \sum_{i \neq m} t_{n,n}^i(x_n), \tag{2.22}$$

for all $x_n \in \mathcal{X}_n$. This immediately implies that indirect payments to any agent are never optimal, i.e., necessarily $\hat{t}_{n,-n}^m = 0$ for all $n \in \mathcal{N}$. In addition, one can verify that any solution $\hat{t}^m$ to the cost minimization problem (2.21)–(2.22) is such that

$$\hat{t}_{n,n}^m(\check{x}_n) = R_n^m - \left(u_{n,n}(\check{x}_n) + \sum_{i \neq m} \hat{t}_{n,n}^i(\check{x}_n) \right)$$

and

$$\hat{t}^m_{n,n}(\check{x}_n) \leq R^m_n - \left(u_{n,n}(\check{x}_n) + \sum_{i \neq m} \hat{t}^i_{n,n}(x_n) \right)$$

for all $x_n \in \mathcal{X}_n$. By replacing $\check{k}$ with the equilibrium outcome $\hat{x}$ we obtain

$$u_{n,n}(\hat{x}_n) + \sum_{i \in \mathcal{M}} \hat{t}^i_{n,n}(\hat{x}_n) = R^m_n$$

and

$$u_{n,n}(x_n) + \sum_{i \in \mathcal{M}} \hat{t}^i_{n,n}(x_n) \leq R^m_n$$

for all $x_n \in \mathcal{X}_n$. The last two relations together are equivalent to (PM).

$\Leftarrow$: Let $(\hat{t}, \hat{x}) \in C(\mathcal{X}, \mathcal{T}) \times \mathcal{X}$ be a pair that satisfies (WT), (AM), (PM) and be such that there are no indirect transfer payments, i.e., $\hat{t}^m_{n,-n} = 0$ for all $(m,n) \in \mathcal{M} \times \mathcal{N}$. The inequality (WT) implies by that $\hat{t}^m$ is weakly truthful for any principal $m \in \mathcal{M}$. Given Assumptions 1–2 it is clear that (AM) is equivalent to (2.3), i.e., given $\hat{t}$ the action $\hat{x}_n$ is a best response for any agent $n \in \mathcal{N}$. By summing up (AM) over all $n \in \mathcal{N}$ and adding (WT) we obtain

$$\begin{aligned} &v^m(\hat{x}) + \sum_{n \in \mathcal{N}} \left(u_{n,n}(\hat{x}_n) - \hat{t}^m_{n,-n}(\hat{x}_{-n}) + \sum_{i \neq m} \hat{t}^i_{n,n}(\hat{x}_n) \right) \\ &\geq v^m(x) + \sum_{n \in \mathcal{N}} \left(u_{n,n}(x_n) - \hat{t}^m_{n,-n}(x_{-n}) + \sum_{i \neq m} \hat{t}^i_{n,n}(x_n) \right) \end{aligned}$$

for all principals $m \in \mathcal{M}$ and outcomes $x \in \mathcal{X}$. The latter inequality is equivalent to (2.20), which in turn implies (2.4). Thus, the pair $(\hat{t}, \hat{x})$ must be a WTE of the game $\mathcal{G}$. ■

Proof of Proposition 2. (i) At the efficient outcome $\hat{x}$ we have that $F^m(\hat{x}) = G_n(\hat{x}_n) = 0$ for all $(m,n) \in \mathcal{M} \times \mathcal{N}$. Hence, the matrix $\Delta(\hat{x}) = [\Delta^i_j(\hat{x}_j)]$ must be such that

$$\sum_{i\in\mathcal{M}} \Delta_n^i(\hat{x}_n) \leq 0 \leq \sum_{j\in\mathcal{N}} \Delta_j^m(\hat{x}_j),$$

for all $(m,n) \in \mathcal{M} \times \mathcal{N}$. Assume that there exists an index $\hat{m} \in \mathcal{M}$ such that $\sum_{j\in\mathcal{N}} \Delta_j^{\hat{m}}(\hat{x}_j) > 0$. By separately summing up all the rows and summing up all the columns of $\Delta(\hat{x})$ we obtain

$$\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{M}} \Delta_n^i(\hat{x}_n) \leq 0 < \sum_{m\in\mathcal{M}}\sum_{j\in\mathcal{N}} \Delta_j^m(\hat{x}_j),$$

a contradiction. As a result $\sum_{j\in\mathcal{N}} \Delta_j^{\hat{m}}(\hat{x}_j) = 0$ for all $m \in \mathcal{M}$. We can show in an analogous manner that necessarily $\sum_{i\in\mathcal{M}} \Delta_n^i(\hat{x}_n) = 0$, whence relation (2.7) obtains. (ii) $\Rightarrow$: The continuous matrix function Δ satisfies (WT'),(AM'). Since in addition (2.7) holds at $\hat{x}$ and δ is a constant matrix by assume, it must be true that

$$\sum_{i\in\mathcal{M}} \delta_n^i \leq 0 \leq \sum_{j\in\mathcal{N}} \delta_j^m$$

for all $(m,n) \in \mathcal{M} \times \mathcal{N}$. Thus, as in part (i) we can conclude that relation (2.8) necessarily holds. $\Leftarrow$: Substituting (2.8) into the reduced contract design inequalities for $\Delta + \delta$ yields the same inequalities as if we had set $\delta = 0$. Since Δ solves the reduced contract design problem (WT'),(AM') by assumption, the matrix function $\Delta + \delta$ constitutes also a solution to the reduced contract design problem. ■

Proof of Proposition 9. For any $\lambda \in (0,1)$ we have that $\lambda\left(F^m - \sum_{n\in\mathcal{N}} \Delta_n^m\right) + (1-\lambda)\left(F^m - \sum_{n\in\mathcal{N}} \tilde{\Delta}_n^m\right) \leq 0 \leq \lambda\left(G_n - \sum_{m\in\mathcal{M}} \Delta_n^m\right) + (1-\lambda)\left(G_n - \sum_{m\in\mathcal{M}} \tilde{\Delta}_n^m\right)$ for any two solutions $\Delta = [\Delta_n^m]$ and $\tilde{\Delta} = [\tilde{\Delta}_n^m]$ to the reduced contract design problem (WT'),(AM'). ■

Proof of Theorem 2. Let $\hat{x} \in \arg\max_{k \in \mathcal{X}} W(x)$ be a given efficient outcome. We first rewrite (PM) in terms of excess measures which yields

$$\hat{t}_n^m(\hat{x}_n) = \max_{x_n \in \mathcal{X}_n} \left\{ -G_n(x_n) + \sum_{i \neq m} \Delta_n^i(x_n) \right\}. \tag{PM'}$$

Suppose that we have found an excess transfer matrix $\Delta = [\Delta_n^m]$ that solves the reduced contract design problem (WT'),(AM'). To prove that the pair $(\hat{t}, \hat{x})$ with $\hat{t}_n^m(x_n)$ as defined in (3.6)–(3.7) constitutes a WTE, it is by Theorem 1 sufficient to show that $\hat{\Delta}_n^m(x_n)$ satisfies the system (WT'), (AM') and (PM'). Consider first (WT'). Since Δ satisfies (AM'), we have by (3.7) that $\hat{\Delta}_n^m \geq \Delta_n^m$, so that

$$F^m - \sum_{n \in \mathcal{N}} \hat{\Delta}_n^m \leq F^m - \sum_{n \in \mathcal{N}} \Delta_n^m \leq 0.$$

Thus, $\hat{\Delta}_n^m(x_n)$ satisfies (WT'). To show that it also satisfies (AM'), we simply note that

$$\sum_{m \in \mathcal{M}} \hat{\Delta}_n^m = G_n,$$

since $\sum_{m \in \mathcal{M}} \vartheta_n^m = 1$ on $\mathcal{X}_n$ for all $n \in \mathcal{N}$. Then (PM') becomes

$$\hat{t}_n^m(\hat{x}_n) = \max_{x_n \in \mathcal{X}_n} \left\{ \left(-G_n(x_n) + \sum_{i \in \mathcal{M}} \hat{\Delta}_n^i(x_n) \right) - \hat{\Delta}_n^m(x_n) \right\} = \max_{x_n \in \mathcal{X}_n} \left\{ -\hat{\Delta}_n^m(x_n) \right\}.$$

In other words, with $\alpha_n^m = \hat{t}_n^m(\hat{x}_n)$,

$$\hat{t}(x_n) = \hat{\Delta}(x_n) + \alpha_n^m = \hat{\Delta}(x_n) - \min_{x_n \in \mathcal{X}_n} \hat{\Delta}(x_n) \geq 0,$$

for all $x_n \in \mathcal{X}_n$ and $(m, n) \in \mathcal{M} \times \mathcal{N}$. The equilibrium transfer matrix $\hat{t} = [\hat{t}_n^m]$ has nonnegative entries and is thus an element of $C(\mathcal{X}, \mathcal{T})$. Hence, we have shown that given any efficient outcome $\hat{x}$, the pair $(\hat{t}_n^m, \hat{x})$ with $\hat{t}$ as defined in (3.6) constitutes a WTE. This concludes our proof. ∎

Proof of Proposition 4. By (WT") the real-valued function $F^m - \Delta^m$ is concave for any $m \in \mathcal{M}$. By (AM") the real-valued function $G_n - \Delta_n$ is convex for any $n \in \mathcal{N}$. Moreover, since $\Delta(\hat{x}) = 0$, we have that $(F^m - \Delta^m)(\hat{x}) = (G_n - \Delta_n)(\hat{x}) = 0$ for all $(m, n) \in \mathcal{M} \times \mathcal{N}$. In other words, $\mathcal{X} \times \{0\}$ and $\mathcal{X}_n \times \{0\}$ are supporting hyperplanes for the graphs of $F^m - \Delta^m$ and $G_n - \Delta_n$ respectively. ■

Proof of Theorem 3. $\Rightarrow$: Consider the outcome $\bar{x}(x_n^m) = (x_n^m, \hat{x}_{-n}^{-m}) \in \mathcal{X}$ which is obtained from $\hat{x}$ by replacing $\hat{x}_n^m$ with a feasible $x_n^m \in \mathbb{R}_+^L$. Clearly, $F^m(\bar{x}(x_n^m)) = F_n^m(x_n^m)$ and $\Delta^m(\bar{x}(x_n^m)) = \Delta_n^m(x_n^m)$, so that by (WT')

$$F_n^m(x_n^m) \leq \Delta_n^m(x_n^m) \tag{2.23}$$

for all $x_n^m \in \mathbb{R}_+^L$ such that $\bar{x}(x_n^m) \in \mathcal{X}$. By Assumption 5 we have that $\bar{x}(x_n^m) \in \mathcal{X}$ if and only if $x_n^m \in \mathcal{X}_n^m$. Hence, the inequality (2.23) holds for any $x_n^m \in \mathcal{X}_n^m$ and any $(m, n) \in \mathcal{M} \times \mathcal{N}$. In a completely analogous manner one can show that $\Delta_n^m \leq G_n^m$ for all $(m, n) \in \mathcal{M} \times \mathcal{N}$. $\Leftarrow$: From (2.12) it follows trivially (by summation) that both (WT') and (AM') are satisfied, i.e., that Δ solves the reduced contract design problem. ■

Proof of Corollary 1. We simply need to verify that relation (2.12) holds. But this follows immediately since $F_n^m \leq \theta_n^m F_n^m + (1 - \theta_n^m) G_n^m \leq G_n^m$ for any admissible θ_n^m with values in $[0, 1]$, and

$$\Delta_n^m(x_n^m; \theta_n^m) = \theta_n^m(x_n^m) F_n^m(x_n^m) + (1 - \theta_n^m(x_n^m)) G_n^m(x_n^m),$$

which completes the proof. ■

Proof of Corollary 2. Under Assumptions 1,2',4,5 Theorem 3 characterizes all additive solutions to the reduced contract design problem (WT'),(AM'), so that Corollary 1 describes all additive solutions. From Theorem 2 we obtain

$$\hat{t}_n^m(x_n;\theta_n^1,\ldots,\theta_n^M,\vartheta_n^m) = \hat{\Delta}_n^m(x_n;\theta_n^1,\ldots,\theta_n^M,\vartheta_n^m) - \min_{x_n\in\mathcal{X}_n}\hat{\Delta}(x_n;\theta_n^1,\ldots,\theta_n^M,\vartheta_n^m),$$

where

$$\hat{\Delta}_n^m(x_n;\theta_n^1,\ldots,\theta_n^M,\vartheta_n^m) = \Delta_n^m(x_n^m;\theta_n^m) + \vartheta_n^m(x_n)\sum_{i\in\mathcal{M}}\left(G_n^i(x_n^i) - \Delta_n^i(x_n^i;\theta_n^i)\right), \quad (2.24)$$

given any solution $\Delta_n^m(\cdot;\theta_n^m)$ to the reduced contract design problem (WT'),(AM') and any $\theta_n^m, \vartheta_n^m \in C(\mathcal{X}_n,[0,1])$ with $\sum_{m\in\mathcal{M}}\vartheta_n^m = 1$. From Theorem 3 the equilibrium transfers $\hat{t}_n^m$ only depend on x_n^m (generally) only if the modified excess transfers $\hat{\Delta}_n^m$ in (2.24) depend exclusively on the outcome x_n^m that directly relates principal m and agent n. But this can only be achieved in general by setting $\Delta_n^m = G_n^m$, i.e., by setting $\theta_n^m = 0$ for all $(m,n)\in\mathcal{M}\times\mathcal{N}$. Thus, $[\hat{\Delta}_n^m] = [G_n^m]$ which yields the additively separable WTE described by the equilibrium transfers $\hat{t}_n^m$ in (2.13) (which are also nonnegative for all $(m,n)\in\mathcal{M}\times\mathcal{N}$). ■

Proof of Proposition 5. The claim follows essentially from Proposition 4. Indeed, under Assumptions 4–5 we obtain that (WT") is equivalent to

$$\nabla^2\pi_n^m \le \nabla^2\Delta_n^m \quad \textbf{and} \quad \nabla\Delta_n^m(\hat{x}_n^m) = \nabla\pi_n^m(\hat{x}_n^m) \quad (2.25)$$

for all $(m,n)\in\mathcal{M}\times\mathcal{N}$. Similarly, Assumptions 2' and 5 imply that (AM") becomes equivalent to

$$\nabla^2\Delta_n^m \le -\nabla^2\gamma_n^m \quad \textbf{and} \quad \nabla\Delta_n^m(\hat{x}_n^m) = -\nabla\gamma_n^m(\hat{x}_n^m) \quad (2.26)$$

for all $(m,n) \in \mathcal{M} \times \mathcal{N}$. Note that at any interior efficient outcome $\hat{x}$ we have necessarily that $\nabla\pi_n^m(\hat{x}_n^m) - \nabla\gamma_n^m(\hat{x}_n^m) = 0$ which is consistent with the equalities in (2.25)–(2.26). Thus, combining (2.25)–(2.26) we obtain from Proposition 4 that relation (2.14) is sufficient for the excess transfer matrix Δ to be a solution to the reduced contract design problem (WT'),(AM'). ■

Proof of Corollary 3. It is sufficient to verify that Δ_n^m satisfies condition (2.14) in Proposition 5. For this, note first that $\nabla\langle\nabla\pi_n^m(\hat{x}_n^m), x_n^m - \hat{x}_n^m\rangle = \nabla\pi_n^m(\hat{x}_n^m)$ as required. Furthermore, it is $\nabla^2\langle\nabla\pi_n^m(\hat{x}_n^m), x_n^m - \hat{x}_n^m\rangle = 0$ and, by Assumption 3, $\nabla^2\pi_n^m \leq 0 \leq -\nabla^2\gamma_n^m$, which completes the proof. ■

Proof of Theorem 4. Let us first show that the affine excess transfer $\Delta = [\Delta_n^m]$ with Δ_n^m as defined in (2.15) solves (WT'). For this, note that (with $\partial v^m/\partial x_n = \partial F^m/\partial x_n$)

$$F^m(x) - \sum_{n\in\mathcal{N}} \Delta_n^m(x_n) = F^m(x) - \sum_{n\in\mathcal{N}} \langle \frac{\partial F^m(\hat{x})}{\partial x_n}, x_n - \hat{x}_n\rangle \tag{2.27}$$

vanishes at the efficient outcome, i.e., for $x = \hat{x}$. Moreover, by Assumption 3 the last expression is concave. If it is maximized at an outcome $\bar{x}$, then it has to satisfy

$$\frac{\partial F^m(\bar{x})}{\partial x_n} = \frac{\partial F^m(\hat{x})}{\partial x_n}, \tag{2.28}$$

for all $n \in \mathcal{N}$. But (2.28) holds for $\bar{x} = \hat{x}$, so that by concavity of (2.27) we obtain that $\hat{x}$ is a *global* maximizer of the left-hand side of (WT'). Hence, the inequality (WT') is satisfied on $\mathcal{X}$ for all $m \in \mathcal{M}$. Since total surplus W is maximized at the efficient outcome $\hat{x}$, we have that

$$G_n'(\hat{x}_n) = \sum_{m\in\mathcal{M}} \frac{\partial F^m(\hat{x})}{\partial x_n}$$

for all $n \in \mathcal{N}$. As a result, $\hat{x}_n$ is a critical point of left-hand side of the inequality (AM') which at this point also vanishes. Since the left-hand side of (AM') is by Assumption 3 convex, it is globally maximized at $\hat{x}_n$. Hence, the inequality (AM') holds on $\mathcal{X}_n$ for all $n \in \mathcal{N}$. ■

Proof of Proposition 6. Let us start by rewriting the reduced contract design problem (WT'),(AM') equivalently in the form[15]

$$F^1(x_1, x_2) - \Delta_1^1(x_1) \leq \Delta_2^1(x_2) \leq G_2(x_2) - \Delta_2^2(x_2), \tag{2.29}$$

and

$$F^2(x_1, x_2) - \Delta_2^2(x_2) \leq \Delta_1^2(x_1) \leq G_1(x_1) - \Delta_1^1(x_1), \tag{2.30}$$

for all $x \in \mathcal{X}$. Note first that since (2.29) holds for all $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ it can be equivalently restated as

$$\phi_2(x_2) = \max_{x_1 \in \mathcal{X}_1} \left\{ F^1(x_1, x_2) - \Delta_1^1(x_1) \right\} \leq \Delta_2^1(x_2) \leq G_2(x_2) - \Delta_2^2(x_2), \tag{2.31}$$

which is independent of x_1. Thus, as long as

$$F^1(x_1, x_2) - \Delta_1^1(x_1) \leq G_2(x_2) - \Delta_2^2(x_2) \tag{2.32}$$

holds for all $x \in \mathcal{X}$, it is clear that (2.31) is satisfied if and only if for a given x_1 the number $\Delta_1^2(x_1)$ lies somewhere between the right-hand and left-hand side of (2.32). The latter is the case (for a continuous Δ_2^1) if and only if we choose $\Delta_2^1(\cdot; \theta_2)$ as in Proposition 6 for some continuous function θ_2 defined on $\mathcal{X}_2$ with values in $[0, 1]$. In a completely analogous manner we can show the necessity and sufficiency of $\Delta_1^2(\cdot; \theta_1)$

[15] Another equivalent set of inequalities is given by $F^1 - \Delta_2^1 \leq \Delta_1^1 \leq G_1 - \Delta_1^2$ and $F^2 - \Delta_1^2 \leq \Delta_2^2 \leq G_2 - \Delta_2^1$ which leads to the different but equivalent formulation of Proposition 6 mentioned after Corollary 4.

as in Proposition 6, as long as

$$F^2(x_1, x_2) - \Delta_2^2(x_2) \leq G_1(x_1) - \Delta_1^1(x_1) \tag{2.33}$$

for all $x \in \mathcal{X}$. Combining (2.32)–(2.33) the sum $\Delta_2^2 - \Delta_1^1$ needs to satisfy

$$F^2(x_1, x_2) - G_1(x_1) \leq \Delta_2^2(x_2) - \Delta_1^1(x_1) \leq G_2(x_2) - F^1(x_1, x_2)$$

for all $x \in \mathcal{X}$. Since $\hat{x}$ is by hypothesis an interior (global) maximizer of total surplus W we have that there can never be any excess surplus, i.e.,

$$W(x) - W(\hat{x}) = F^1(x_1, x_2) + F^2(x_1, x_2) - G_1(x_1) - G_2(x_2) \leq 0 \tag{2.34}$$

for all $x \in \mathcal{X}$, which implies that (2.33) can always be satisfied pointwise (i.e., without the continuity requirement on Δ_n^m), i.e., locally. By Assumption 3 we have that $F^2 - G_1$ is concave and $G_2 - F^1$ is convex on $\mathcal{X}$. Hence, the separating hyperplane theorem (together with the fact that $F^2 - G^1 = G_2 - F^1$ at $\hat{x}$) implies that there exist constants a^1 and a^2 such that the continuous (and affine) functions $\Delta_n^n(x_n) = \langle a_n, x_n - \hat{x} \rangle$ for $n \in \{1, 2\}$ satisfy (2.34) on $\mathcal{X}$. A simple application of Proposition 2 concludes our proof by ensuring that the obtained solution to the reduced contract design problem vanishes at the efficient outcome. ■

Proof of Corollary 4. By taking $\Delta_n^n(x_n) = \langle a_n, x_n - \hat{x}_n \rangle$ (cf. the proof of Proposition 6) we obtain from (2.16) that $a_n = \partial F^n(\hat{x})/\partial x_n = \partial v^n(\hat{x})/\partial x_n$ for $n \in \{1, 2\}$. ■

2.7 Appendix B: Comparison with Available Results

In this subsection we provide two supply chain coordination examples related to the current literature on supply chain coordination. One is a multi-principal one-agent supply chain example considered by Carr and Karmarkar (2003). The other is a one-principal multi-agent supply chain example considered by Majumder and Srinivasan (2003). Although these two works have different research focuses from this paper, we apply our contract design method to their setting to demonstrate how supply chain coordination can be achieved.

B.1 Multiple Principals and One Agent

Carr and Karmarkar (2003) consider an assembly structure example consisting of two component suppliers and one manufacturer. They apply price-only contracts to achieve quantity coordination, that is, the production quantity of each supplier (upstream firm) equal to that of the manufacturer (downstream firm). It is well known that price-only contracts with double marginalization cannot achieve channel coordination. We apply our contract design method to achieve channel coordination in this assembly system. In illustrating the idea in Section 2.2.5 we assume a top-down contracting situation in which the suppliers act as principals. The suppliers incur the cost of production, which is linear in the quantity produced. Upon observing the contract terms the manufacturer places orders with the component suppliers who then produce the ordered components at a unit cost of $c^m > 0$. The manufacturer and the suppliers thereby trade according to the supply contracts signed *ex ante*, i.e., before the production has happened. After receiving the components from the two suppliers, the manufacturer then produces with these components at an extra unit production cost c_1 and sells the product to the end consumers at a retail price to clear the market. In our setting the production quantity is naturally and automatically coordinated.

Let x be the production quantity of the manufacturer, who acts as an agent for the two suppliers. His gross payoff is given by $u_1 = [(a - bx) - c_1]x$. Supplier m's gross payoff is $v^m = -c^m x$. The system surplus $W = v^1 + v^2 + u_1$ is strictly concave, and we obtain the production quantity

$$\hat{x} = \arg\max_{x \in \mathcal{X}} \{(a - bx)x - (c_1 + c^1 + c^2)x\} = \frac{a - c_1 - c^1 - c^2}{2b}$$

as the unique efficient outcome maximizing total surplus in the supply chain system, which is larger than the production quantity $(a - c_1 - c^1 - c^2)/(6b)$ as obtained by Carr and Karmarkar (2003, Proposition 1). Consequently, the efficient system surplus $\hat{W} = (a - c_1 - c^1 - c^2)/(4b)$ is larger than their system surplus $\tilde{W} = 5(a - c_1 - c^1 - c^2)/(36b)$.

To accommodate *top-down contracting*, we apply a linearly augmented modified transfer with an "expensive" wholesale price w_n^m. Correspondingly, we define

$$\tilde{u}_1 = [(a - bx) - c_1]x - (w_1^1 + w_1^2)x, \quad \text{and} \quad \tilde{v}^m = -c^m x + w_1^m x$$

for $m \in \{1, 2\}$. Assumptions 1,2',3–5 are satisfied and by Corollary 3 the affine excess transfer

$$\tilde{\Delta}_1^m = \frac{d\tilde{v}^m}{dx}(x - \hat{x}) = (w_1^m - c^m)(x - \hat{x})$$

solves the reduced contract design problem. The in-equilibrium transfers α_n^m which define the equilibrium contracts, can be obtained from Theorem 2. For simplicity we let $\vartheta_1^1 = 0$ and $\vartheta_1^2 = 1$. We thus obtain

$$\tilde{\alpha}_1^1 = (w_1^1 - c^1)\hat{x} \quad \text{and} \quad \tilde{\alpha}_1^2 = \left(w_1^2 - (a - b\hat{x} - c_1 - c^1)\right)\hat{x}.$$

The transfers in the original system are shifted to $t_n^m = \tilde{t}_n^m - w_n^m x_n^m$. In this example, the actual in-equilibrium transfers are given by

$$\alpha_1^1 = -c^1\hat{x} \quad \text{and} \quad \alpha_1^2 = -(a - b\hat{x} - c_1 - c^1)\hat{x}.$$

These negativity of the equilibrium payments means that the actual transfers are from the agent (manufacturer) to the principals (suppliers). The net in-equilibrium payoff to the manufacturer is zero and the principals are able to extract all the surplus from the transaction. However, the sharing of the surplus between suppliers 1 and 2 depends on the choice of ϑ_1^1. When $\vartheta_1^1 = 0$, supplier 1's net payoff is zero and all the surplus is extracted by supplier 2.

B.2 One Principal and Multiple Agents

Majumder and Srinivasan (2003) consider a multi-echelon supply chain system where some members in the chain procure complementary components from more than one supplier. The authors study the effect of contract leadership and coordination and show that two-part tariffs can coordinate the whole supply chain. Without loss of generality, we consider a simple assembly structure consisting of two suppliers and one manufacturer. In line with our main argument we assume a bottom-up contracting scenario in which the manufacturer acts as a principal. The manufacturer and the suppliers trade according to the supply contracts signed *ex ante.* Given the contracts, supplier n with infinite capacity decides on how much to produce and to deliver to the manufacturer, i.e., x_n. We assume that the production cost is quadratic in the quantity produced. Upon receiving the delivered components, the manufacturer produces an end product at a unit production cost c^1 and sells the end product to the end consumers at a retail price p to clear the market, with $p = a - bx$ and a is big enough to keep system surplus nonnegative. Supplier n's gross payoff is given by

$$u_n(x_n) = -\left(\frac{a_n(x_n)^2}{2} + b_n x_n\right).$$

The manufacturer's gross payoff can be written as

$$v^1(x) = (a - bx)x - c^1 x.$$

The (strictly concave) system surplus $W(x) = v^1(x) + u_1(x) + u_2(x)$ is maximized at

$$\hat{x}_1 = \hat{x}_2 = \hat{x} = \frac{a - c^1 - b_1 - b_2}{2b + a_1 + a_2},$$

as the unique efficient outcome. Assumptions 1–3 are satisfied, so that from Theorem 4 (Affine Solution to the Reduced Contract Design Problem) we obtain that the excess transfer described by

$$\Delta_1^1(x) = \Delta_2^1(x) = (a - 2b\hat{x} - c^1)(x - \hat{x})$$

solves the reduced contract design problem (WT'),(AM'). The in-equilibrium transfers α_n^m can be obtained from Theorem 2. Since $\sum_{m \in \mathcal{M}} \vartheta_n^m = 1$ we let $\vartheta_n^1 = 1$ and obtain

$$\Delta_n^1(x) = G_n(x) = u_n(\hat{x}) - u_n(x) \quad \text{and} \quad \alpha_n^1 = -u_n(\hat{x}) = \frac{a_n(\hat{x})^2}{2} + b_n\hat{x}$$

for $n \in \{1, 2\}$. In equilibrium, supplier n obtains exactly his total cost and the manufacturer is thus able to extract all surplus from the transaction.

Chapter 3

Surplus Extraction In Multi-Principal Multi-Agent Supply Chains

3.1 Introduction

Recently, Weber and Xiong (2006) [W&X][1] have discussed the noncooperative design of coordinating full-information contracts in multi-principal multi-agent supply chains. The principals correspond to either buyers or sellers in a supply chain, and they propose contracting solutions. Allowing for arbitrary continuous payoff functions with externalities, they have provided a nonparametric characterization of such contracts in terms of weakly truthful equilibria (WTEs) of the underlying contracting game, in which first principals noncooperatively propose a set of contracts to agents, and then agents noncooperatively implement an outcome.[2] Using this characterization, W&X provide a set of closed-form contracting equilibria that coordinate a

[1]Chapter 3 and W&X both study efficient contract design in multi-principal multi-agent supply chains. The difference is that W&X address agent payoff externalities explicitly.

[2]The concept of weak truthfulness has been introduced in a seminal paper by Bernheim and Whinston (1986). It has been used in a multi-principal multi-agent setting (without any externalities) by Prat and Rustichini (2003).

two-echelon supply chain implementing any given efficient outcome as a WTE of the contracting game. If the efficient outcome is unique, they show that it can be implemented strongly in the sense that in an equilibrium, agents strictly prefer to take coordinating actions over all other available actions.

Since multiple contract schemes exist to implement an efficient outcome, a natural follow-up question is, how the principals can extract as much surplus as possible from the agents. The focus of this chapter is to answer the question. This *surplus extraction* problem has been first investigated by Strulovici and Weber (2004) in a setting without payoff externalities between participants of the supply chain. Here we show that the set of attainable equilibrium payoff vectors for principals is convex, and that maximum surplus is extracted from the agents on that set's Pareto frontier.

The rest of the chapter is organized as follows: In Section 3.2 we review the efficient contract design solutions developed in W&X. In Section 3.3 we propose a characterization on the Pareto frontier of the set of the principals' attainable payoff vectors. In Section 3.4 we describe how to compute the Pareto frontier. In Section 3.5 we discuss the application of our general method to the coordination of a differentiated Cournot Oligopoly. Finally, we conclude in Section 3.6 with a discussion of the results as well as directions for future research.

3.2 Efficient Contract Design

Using essentially the same setup as in W&X, we consider a setting in which principals can write outcome-contingent contracts with a number of different agents. Let $\mathcal{M} = \{1, \ldots, M\}$ and $\mathcal{N} = \{1, \ldots, N\}$ denote the corresponding sets of M principals and N agents. After each principal $m \in \mathcal{M}$ ("she") and each agent $n \in \mathcal{N}$ ("he") sign contracts with each other, agents noncooperatively implement an action (or "outcome") $x \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. The outcome vector $x = (x_n^m)_{m,n=1}^{M,N}$ contains each agent n's individual action vector $x_n = (x_n^1, \ldots, x_n^M) \in \mathcal{X}_n \subset \mathbb{R}_+^{ML}$, which, in turn, is composed of M different L-dimensional actions x_n^m. The action set $\mathcal{X}_n$ is a compact subset of $\mathbb{R}_+^{ML}$ that contains at least one point to allow for the possibility of inaction.

In the first stage of the *contracting game*, each principal m designs a mapping $t^m : \mathcal{X} \to \mathbb{R}_+^N$ from outcomes x to nonnegative transfer payments $t_n^m(x)$ directed at each agent $n \in \mathcal{N}$.[3] These (generally nonlinear) payment schedules are proposed noncooperatively by the principals to the agents. In stage two, each agent n implements an action x_n. Agents can exert externalities on each other, i.e., each agent n cares about the implemented outcome $x = (x_n, x_{-n})$ and the sum of all the transfers he obtains in equilibrium. His net payoff is given by

$$U_n(x;t) = u_n(x) + \sum_{m \in \mathcal{M}} t_n^m(x), \tag{3.1}$$

where $u_n(x)$ is the gross payoff to agent n from the outcome $x = (x_n, x_{-n})$ when he takes action x_n and all other players implement x_{-n}.

When selecting optimal remuneration schemes (contracts) for the different agents, each principal cares about both her monetary payments *and* the agents' actions. Let $v^m(x)$ be principal m's *gross* payoff if action x is taken. If she offers the transfer schedule $t^m = (t_1^m, \dots, t_N^m)$ and agents implement the outcome x, her *net* payoff is

$$V^m(x;t^m) = v^m(x) - \sum_{n \in \mathcal{N}} t_n^m(x). \tag{3.2}$$

The only critical assumption we make in this paper is that all payoff functions V^m and U_n are *continuous*. Together with the compactness of $\mathcal{X}$ this guarantees the

[3]It is not critical that transfer payments be nonnegative, only that there is a known lower bound, reflecting the fact that the principals cannot inflict infinite out-of-equilibrium punishments on their agents. We assume here for sake of discussion that the worst each principal can do to an agent is not to pay him at all.

existence of an efficient outcome,

$$\hat{x} \in \arg\max_{x \in \mathcal{X}} W(x), \tag{3.3}$$

so that the supply chain can actually be coordinated.[4] The function

$$W(x) = \sum_{m \in \mathcal{M}} V^m(x) + \sum_{n \in \mathcal{N}} U_n(x)$$

corresponds to the total surplus, i.e., the sum of all payoffs in the supply chain. In order to use the results of W&X regarding the characterization of the set of equilibrium contracts, it is useful to formulate the payoffs in terms of excess measures relative to the payoffs that would be obtained at an efficient outcome $\hat{x}$. For simplicity we assume here that the efficient outcome $\hat{x}$ is actually unique, in order to avoid coordination issues related to the selection of the efficient outcome itself. Principal m's *excess revenue* from implementing x instead of $\hat{x}$ is

$$F^m(x) = v^m(x) - v^m(\hat{x})$$

and agent n's *excess cost* from taking an action x_n instead of $\hat{x}_n$, given the other agents' out of equilibrium actions x_{-n} and in-equilibrium actions $\hat{x}_{-n}$, is

$$G_n(x) = u_n(\hat{x}) - u_n(x).$$

Similarly, the *excess transfer*

$$\Delta_n^m(x) = t_n^m(x) - t_n^m(\hat{x}),$$

[4] We use the following definition of (uniform) continuity: *a function $g : \mathcal{X} \to \mathbb{R}$ is continuous on $\mathcal{X}$ if, given any $x \in \mathcal{X}$, for any $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon) > 0$ such that $\hat{x} \in \{\xi \in \mathcal{X} : \|\xi - x\| < \delta\} \Rightarrow |g(\hat{x}) - g(x)| < \varepsilon$, where $\|\cdot\|$ is a given norm on the Euclidean space $\mathbb{R}^{MNL}$.* Note that according to this definition *any* function is continuous when $\mathcal{X}$ is *discrete.*

is paid from principal m to agent n as the difference in compensation for helping to implement an outcome x instead of the efficient outcome $\hat{x}$. W&X characterize WTEs for the multi-principal multi-agent game and show that any solution $\Delta = [\Delta_n^m] \in C(\mathcal{X}, \mathbb{R}^{M\times N})$ to a *reduced contract-design problem* of the form

$$F^m(x) - \sum_{j\in\mathcal{N}} \Delta_j^m(x) \leq 0 \leq G_n(x) - \sum_{i\in\mathcal{M}} \Delta_n^i(x), \tag{R}$$

for all $(m, n) \in \mathcal{M} \times \mathcal{N}$ and all $x \in \mathcal{X}$ can be directly mapped to a WTE of the original efficient contract-design problem. In fact, if we consider the solution set

$$\mathcal{R} = \left\{\Delta \in C(\mathcal{X}, \mathbb{R}^{M\times N}) : \Delta \text{ solves (R)}\right\}$$

then $\Delta \in \mathcal{R}$ if and only if there are functions $\varphi^m, \gamma_n \in C(\mathcal{X}, \mathbb{R}_+)$ such that

$$\begin{cases} \sum_{j\in\mathcal{N}} \Delta_j^m(x) &= F^m(x) + \varphi^m(x), \\ \sum_{i\in\mathcal{M}} \Delta_n^i(x) &= G_n(x) - \gamma_n(x), \end{cases} \tag{3.4}$$

for all $(m, n) \in \mathcal{M} \times \mathcal{N}$ and all $x \in \mathcal{X}$, and

$$\sum_{i\in\mathcal{M}} \varphi^i(x) + \sum_{j\in\mathcal{N}} \gamma_j(x) = \Omega(x) \tag{3.5}$$

on $\mathcal{X}$, where

$$\Omega(x) = W(\hat{x}) - W(x)$$

corresponds to the *total surplus deficit* in the supply chain (Theorem 2 in W&X).[5] In particular, one can show that the set $\mathcal{R}$ is convex (W&X, Lemma 3). From our preceding remarks it is clear that $\mathcal{R} \neq \emptyset$ and that it typically contains more than one element, since the $N + M$ inequalities in (R) are generally not enough to pin down an $N \times M$ solution matrix Δ. If $|\mathcal{R}| > 1$, the principals can (and in practical situations

[5]Minus the total surplus deficit, $-\Omega$, corresponds to the *total excess surplus*, which by the definition of the efficient outcome $\hat{x}$ in (3.3) must be nonpositive, so that $\Omega \geq 0$.

do) have different preferences regarding which solution to choose, as the distribution of surplus may vary substantially among different solutions. In our modelling framework below, the $\Delta \in \mathcal{R}$ is chosen so that the corresponding vector of in-equilibrium net payoffs to the principals lies on the Pareto frontier. Given this choice, all gains from coordination are exhausted and the principals appropriate surplus from the agents that cannot be increased for one principal without making another principal worse off.

The reduced contract design problem (R) embodies the fact that each agent maximizes his payoffs given the actions of all other agents, and that each principal prefers the efficient outcome $\hat{x}$ given the transfer schedules offered by the other principals. In addition, the WTE requires that, in equilibrium, principals minimize their expenditures in the sense that each principal m pays an agent n exactly the difference between what this agent could gain by ignoring principal m (and obtaining an amount of zero from her) to implement his most preferred action, and the cost for agent n of choosing the efficient action for principal m. This additional "principal cost minimization" requirement can be written in terms of our excess measures as follows:

$$\hat{t}_n^m(\hat{x}) = \max_{x_n \in \mathcal{X}_n} \left\{ -G_n(x_n, \hat{x}_{-n}) + \sum_{i \neq m} \hat{\Delta}_n^i(x_n, \hat{x}_{-n}) \right\}. \tag{PM}$$

We denote by

$$\mathcal{C} = \{\Delta \in C(\mathcal{X}, \mathbb{R}^{M \times N}) : \Delta \text{ solves (R) and (PM) } \} \subset \mathcal{R}$$

the set of solutions to the complete contract design problem, consisting of (R) combined with (PM).[6] The following key result maps any $\Delta \in \mathcal{R}$ to a solution of the complete contract-design problem $\Delta \in \mathcal{C}$, which results in a WTE of the overall contracting game.

[6]Recall from W&X that (R) and (PM) are equivalent to the characterization of a WTE.

PROPOSITION 7 (EFFICIENT CONTRACT DESIGN; W&X, THEOREM 3) *If the excess transfer matrix* $\Delta = [\Delta_n^m] \in \mathcal{R}$ *solves the reduced contract-design problem (R) implementing the efficient outcome* $\hat{x}$*, then a WTE of the multi-principal multi-agent game is given by* $(\hat{t}, \hat{x})$ *with*

$$\hat{t}_n^m(x; \vartheta_n^m, \theta_n) = \hat{\Delta}_n^m(x; \vartheta_n^m, \theta_n) - \min_{x_n \in \mathcal{X}_n} \left\{ \hat{\Delta}_n^m(x_n, \hat{x}_{-n}; \vartheta_n^m + \theta_n, \vartheta_n^m \theta_n / (\vartheta_n^m + \theta_n)) \right\} \tag{3.6}$$

and

$$\hat{\Delta}_n^m(x; \vartheta_n^m, \theta_n) = \Delta_n^m(x) + \vartheta_n^m(x)(1 - \theta_n(x)) \left(G_n(x) - \sum_{i \in \mathcal{M}} \Delta_n^i(x) \right) \in \mathcal{C} \tag{3.7}$$

for all $(m, n) \in \mathcal{M} \times \mathcal{N}$*,* $\theta_n \in C(\mathcal{X}_n, [0, \bar{\theta}_n])$ *with*

$$\bar{\theta}_n = \sup \left\{ \hat{\theta}_n \in [0, 1] : \min_{i \in \mathcal{M}} \hat{t}_n^i(\hat{x}; \vartheta_n^i, \hat{\theta}_n) \geq 0 \right\}$$

and arbitrary $\vartheta_n^m \in C(\mathcal{X}_n, [0, 1])$ *satisfying* $\sum_{i \in \mathcal{M}} \vartheta_n^i(x) \equiv 1$.

The function θ_n can be interpreted as the "strength" of the implementation from the agents' point of view. For $\theta_n = 0$, which is always feasible, the implementation of the efficient outcome is weak in the sense that all agents are indifferent to any particular action $x_n \in \mathcal{X}_n$. Since a strong implementation represents an added cost to the principals, which can be made arbitrarily small by reducing θ_n, we assume here without any loss in generality for our discussion that $\theta_n = 0$. Then the *constant nonnegative shifts*

$$\alpha_n^m(\vartheta_n^m) = - \min_{x_n \in \mathcal{X}_n} \hat{\Delta}_n^m(x_n, \hat{x}_{-n}; \vartheta_n^m) \geq 0 \tag{3.8}$$

of the modified excess transfers $\hat{\Delta}_n^m$ *correspond exactly to the amounts transferred from principal* m *to agent* n *in equilibrium.* As discussed in W&X, these amounts generally depend on the outcome-contingent convex combination selected in (3.7). Generally, for each agent $n \in \mathcal{N}$ the principals are in conflict about who should pay him less, since the higher ϑ_n^m, the lower principal m's transfer to agent n in equilibrium. If for principal m the weight $\vartheta_n^m = 1$, her equilibrium transfer to agent n is indeed as small as possible, given the solution Δ to the reduced contract-design problem (R). This conflict of interest is not resolved as part of the contracting game; it is a consequence of the multiplicity of coordinating equilibria. The multiplicity can be restricted to the Pareto frontier of the principals' payoff set, which is the focus of Section 3.3.

3.3 The Pareto Frontier of the Principals' Payoff Set

3.3.1 Primitives

We first establish that the set $\mathcal{V}$ of attainable payoff vectors for the principals is convex.

The Payoff Attainability Set $\mathcal{V}$

Based on Proposition 7 we can represent the set $\mathcal{C}$ of solutions to the complete contract-design problem as the range of a suitable transformation T of solutions to the reduced contract-design problem (R). More specifically, if for any

$$\vartheta = [\vartheta_n^m] \in \Theta = \{\theta = [\theta_n^m] \in C(\mathcal{X}, \mathbb{R}^{M\times N}) : \sum_{i\in\mathcal{M}} \vartheta_n^i(x) \equiv 1\}$$

we set

$$T(\Delta, \vartheta) = \left[\Delta_n^m(x) + \vartheta_n^m(x)\left(G_n(x) - \sum_{i\in\mathcal{M}} \Delta_n^i(x)\right)\right],$$

then it is

$$\mathcal{C} = \{T(\Delta, \vartheta) : (\Delta, \vartheta) \in \mathcal{R} \times \Theta\} = T(\mathcal{R}, \Theta).$$

In other words, the range of $T(\cdot, \Theta)$ over its domain $\mathcal{R}$ constitutes the solution set $\mathcal{C}$ of the contract-design problem (R),(PM). Interestingly, this range does not depend on being able to use all elements of Θ. The following result establishes that in fact any one element is enough to represent all elements of $\mathcal{C}$ using elements of $\mathcal{R}$.

PROPOSITION 8 (INVARIANCE) *For any $\vartheta \in \Theta$, we have $T(\mathcal{R}, \vartheta) = T(\mathcal{R}, \Theta) = \mathcal{C}$.*

We can therefore, without any loss in generality, choose $\vartheta \in \Theta$, such that $\vartheta_n^m = 1/M$. For any $\Delta \in \mathcal{R}$ we denote the corresponding element of $\mathcal{C}$ by $T(\Delta)$, dropping the explicit dependence on ϑ. Proposition 8 ensures that $T(\mathcal{R}) = \mathcal{C}$.

Consider now the set $\mathcal{V}$ of attainable payoffs for the principals, where

$$\mathcal{V}=\left\{(\hat{V}^1(\Delta),\ldots,\hat{V}^M(\Delta)):\hat{V}^m(\Delta)=v^m(\hat{x})+\sum_n \min_{x_n\in\mathcal{X}_n}\{T_n^m(\Delta(x_n,\hat{x}_{-n}))\},\Delta\in\mathcal{R},m\in\mathcal{M}\right\},$$

and

$$T_n^m(\Delta(x)) = \Delta_n^m(x) + \frac{1}{M}\left(G_n(x) - \sum_{i\in\mathcal{M}} \Delta_n^i(x)\right)$$

for all $x \in \mathcal{X}$. The following key result establishes that the set of all attainable equilibrium payoff vectors for the principals is convex.

PROPOSITION 9 (CONVEXITY) *The (nonempty) set $\mathcal{V}$ is convex.*

Remarkably, the convexity of $\mathcal{V}$ is independent of any assumptions on the payoff functions (other than continuity). It is a consequence of the convexity of $\mathcal{R}$, the set of solutions to the reduced contract-design problem (R), together with the concavity induced by the minimization in expression (3.8) for the in-equilibrium transfers from principals to agents.

To further describe the set $\mathcal{V}$ using the primitives of the contracting problem, we provide the following bounds on any given principal's equilibrium payoffs.

PROPOSITION 10 (PRINCIPAL PAYOFF BOUNDS) *The set $\mathcal{V}$ is bounded; any given principal m's equilibrium net payoff $\hat{V}^m = V^m(\hat{x})$ in a WTE implementing the efficient outcome $\hat{x}$ satisfies*

$$\sum_n \min_{x_n \in \mathcal{X}_n} \{F^m(x_n, \hat{x}_{-n})\} \leq \hat{V}^m - v^m(\hat{x}) \leq \sum_n \min_{x_n \in \mathcal{X}_n} \{\Omega(x_n, \hat{x}_{-n}) + F^m(x_n, \hat{x}_{-n})\},$$

where $\Omega(x) = W(\hat{x}) - W(x)$.

The lower bound for $\hat{V}^m$ is achieved when the slack variables $\varphi^m(x_n, \hat{x}_{-n})$ and $\gamma_n(x_n, \hat{x}_{-n})$ in (3.4) vanish for all (m, n). Similarly, the upper bound is tight when $\varphi^i(x_n, \hat{x}_{-n}) \equiv 0$ for all $i \neq m$.

3.3.2 Special Case: Common Agency

We now consider the important special case when $N = 1$, so that there is a single agent that contracts with all M principals, a situation referred to as *common agency.* In that setting it is possible to sharpen the general results and obtain explicit expressions. In particular, we discuss the interesting and somewhat counterintuitive phenomenon that all principals may be able to simultaneously obtain their maximum possible payoffs, which we therefore term a *win-win scenario.*

LEMMA 1 (MAXIMUM PRINCIPAL PAYOFF) *Under common agency, principal m's maximum attainable equilibrium payoff is*

$$\bar{V}^m = v^m(\hat{x}) + \min_{x_1 \in \mathcal{X}_1} \left\{ G_1(x_1) - \sum_{i \neq m} F^i(x_1) \right\}.$$

The intuition for this result is as follows. It is clear that principal m's maximum payoff $\bar{V}^m$ is attained at any outcome

$$\xi^m \in \arg \min_{x_1 \in \mathcal{X}_1} \left\{ G_1(x_1) - \sum_{i \neq m} F^i(x_1) \right\}$$

with an excess transfer $\Delta = [\Delta_1^m] \in \mathcal{R}$ such that $\Delta_1^m(\xi^m) = G_1(\xi^m) - \sum_i F^i(\xi^m)$ and $\Delta_1^i(\xi^m) = F^i(\xi^m)$ for $i \neq m$. This implies that necessarily

$$G_1(\xi^m)-\sum_{i\in\mathcal{M}}\Delta_1^i(\xi^m)=G_1(\xi^m)-\sum_{i\neq m}\Delta_1^i(\xi^m)-\Delta_1^m(\xi^m)=G_1(\xi^m)-\sum_{i\neq m}F^i(\xi^m)-\Delta_1^m(\xi^m)=0,$$

and hence

$$T_1^m(\Delta(\xi^m)) = \Delta_1^m(\xi^m) = G_1(\xi^m) - \sum_{i\neq m} F^i(\xi^m).$$

By the cost-minimization condition (PM) we know that principal m's in-equilibrium transfer to agent n in (3.8) depends on that agent's best action in the (hypothetical) situation when principal m "removes" herself from the game (i.e., sets all her transfers for all possible outcomes to their minimum value, zero). Naturally, the agent's best action ξ^m in principal m's absence is generally different from his equilibrium action. This leads to a surprising result: if, for all $m \neq i$, the agent's action ξ^m (determining principal m in-equilibrium transfer to the agent) does not influence the agent's action ξ^i (determining principal i's in-equilibrium transfer to the agent), then the principals can coordinate their contract profile in a way that allows all principals to obtain their maximum possible payoffs. Correspondingly, a *win-win scenario* is such that

$$\hat{V}^m = \bar{V}^m$$

for all $m \in \mathcal{M}$, i.e., each principal obtains her maximum attainable payoff. The following result provides necessary and sufficient conditions for a win-win scenario to exist under common agency.

LEMMA 2 (WIN-WIN SCENARIO) *Let $N = 1$ (common agency). (i) A win-win scenario exists if*

$$F^m(x) \geq \Omega(\xi^m) + F^m(\xi^m) \tag{3.9}$$

for all $x \in \mathcal{X} \setminus \{\xi^m\}$ and all $m \in \mathcal{M}$. (ii) If a win-win scenario exists, then (3.9) *must hold for all* $x \in \{\xi^1, \ldots, \xi^M\} \setminus \{\xi^m\}$ *and all* $m \in \mathcal{M}$.

Intuitively, a win-win scenario exists if the excess payoff of any principal m (evaluated at "appropriate" outcomes $x \neq \xi^m$) exceeds the excess payoff of that principal (evaluated at ξ^m) by at least the total supply chain deficit (also evaluated at ξ^m). Condition (3.9) cannot be satisfied for all $x \in \mathcal{X} \setminus \{\xi^m\}$ if ξ^m is not either an efficient outcome or an isolated point. The sufficient condition of part (i) of Lemma 2 is therefore of practical use only if the action set is discrete. An example in Section 3.5 demonstrates that a win-win scenario can be obtained under less stringent conditions.

3.4 Computing the Pareto Frontier

Since the set of the principals' attainable payoff vectors is convex and bounded by Propositions 9 and 10, in equilibrium the principals select points of the Pareto frontier

$$\mathcal{P} = \{v \in \mathcal{V} : v \leq \hat{v} \in \mathcal{V} \ \Rightarrow \ v = \hat{v}\}.$$

Let $V(\Delta) = (V^1(\Delta), \ldots, V^M(\Delta))$ denote the payoff vectors obtained by a solution $T(\Delta) \in \mathcal{C}$ to the complete contract design problem for some $\Delta \in \mathcal{R}$. Then, if $\hat{V}$ is an element of the Pareto frontier (or its closure $\bar{\mathcal{P}}$), there exists a vector $\lambda = (\lambda^1, \ldots, \lambda^M)$ with $\sum_{i=1}^M \lambda^i = 1$ such that

$$\hat{V} = \sup_{\Delta \in \mathcal{R}} \{\lambda \cdot V(\Delta)\}. \tag{3.10}$$

We now discuss the numerical implementation of obtaining a Pareto frontier. Recall that

$$\hat{V}^m(\Delta) = v^m(\hat{x}) + \sum_n \min_{x_n \in \mathcal{X}_n} \{T_n^m(\Delta(x_n, \hat{x}_{-n}))\},$$

and

$$T_n^m(\Delta(x)) = \Delta_n^m(x) + \frac{1}{M}\left(G_n(x) - \sum_{i \in \mathcal{M}} \Delta_n^i(x)\right).$$

Finding the Pareto frontier of the principals' in-equilibrium net payoffs amounts to solving the following problem:

$$\begin{array}{ll} \max\limits_{[q_n^m],[\Delta_n^m]} & \sum_{i,j=1}^{M,N} \lambda^i q_j^i \\ \text{s.t.} & \sum_{i=1}^{M} \Delta_n^i + M(q_n^m - \Delta_n^m) \leq G_n, \\ & \sum_{i=1}^{M} \Delta_n^i \leq G_n, \\ & \sum_{j=1}^{N} \Delta_j^m \geq F^m, \\ & (m,n) \in \{1,\ldots,M\} \times \{1,\ldots,N\}, \end{array} \tag{3.11}$$

for $\lambda \in \mathbb{R}_{++}^M$ in the unit M-simplex. The variables $q_n^m = -\alpha_n^m$ in problem (3.11) can be interpreted in terms of amounts transferred from agent n to principal m, for instance as discounts off a large enough base amount (e.g., a wholesale price), guaranteeing that the actual transfers t_n^m remain nonnegative (cf. footnote 3).

Problem (3.11) is fundamentally a variational problem, since the optimization variables Δ_n^m correspond to (continuous) functions mapping the set of feasible outcomes $\mathcal{X}$ to real numbers. Nonetheless, by discretizing the set $\mathcal{X}$, the variational problem (3.11) can be converted to a linear program. More specifically, if we let a represent a vector of minimum elements of $\mathcal{X}$ for each coordinate of $\mathbb{R}^{MNL}$, and b the corresponding vector of the maxima, then $\mathcal{X} \subset [a,b]$. For the numerical solution of (3.11) we restrict attention to the finite grid

$$\hat{\mathcal{X}}_K = \left\{ \hat{x} : \hat{x} = a + \frac{k \cdot (b-a)}{K} \in \mathcal{X} + \varepsilon, k \in \{0,\ldots,K\}^{MNL}, \text{ for some } \varepsilon \in (-\tfrac{1}{2K}, \tfrac{1}{2K}]^{MNL} \right\},$$

for a sufficiently large integer $K \geq 1$ (the *grid-size*). Note that the grid $\hat{\mathcal{X}}_K$ can also be used when $\mathcal{X}$ is discrete. In nondegenerate cases the set $\mathcal{X}_K$ contains on the order of $(K+1)^{MNL}$ discretization points $\hat{x}[k]$ for

$$k \in \mathcal{K} = \left\{ \hat{x} \in \{0,\ldots,K\}^{MNL} : a + \frac{\hat{x} \cdot (b-a)}{K} \in \hat{\mathcal{X}}_K \right\}$$

. Instead of Δ_n^m we can then consider $\delta_n^m[k] = \Delta_n^m(\hat{x}[k])$ for all $x \in \mathcal{K}$. With this (3.11) becomes a linear programming problem that can be solved using standard optimization software such as Matlab. Varying p in increments results in a (discretized) approximate Pareto frontier.

3.5 Supply Chain Contracting Examples

We now study several examples to illustrate how our framework can be applied to study issues of surplus extraction while coordination is implemented in a common-agency supply chain setting.

3.5.1 Constrained Capacity

Consider a supply chain consisting of one upstream component supplier and two retailers who each sell one end product in their respective, geographically dispersed markets.[7] We assume that the retailers act as principals, i.e., initiate the contracts. The supplier has only a limited installed capacity that can be allocated to the retailers, which we normalize to one for simplicity. Thus, if x_1^m denotes the supplier's capacity allocation to retailer m, we assume that $x_1^1 + x_1^2 = 1$. The supplier's cost for installing the capacity is assumed sunk, i.e., her gross payoff is $u_1 = 0$. We further suppose that additional capacity is a desirable resource of which both retailers prefer to have more, and that the lead time for production is not negligible, so that the supplier needs to set production quantities *before* observing demand realizations. The sequence of events is as follows: The retailers and the supplier sign supply contracts before the uncertain demand is realized. The supplier decides on how much capacity to allocate to each retailer and also how much to produce for each retailer. The retailers and the supplier trade according to supply contracts signed, and the quantity x_1^m is produced and delivered to retailer m. Then random demand is realized and each retailer sells

[7]This example is similar to the interdependent action example in W&X (2006). While they consider a two-supplier two-retailer supply chain and apply an efficient partially affine contract to coordinate, here for simplicity, we consider only a one-supplier two-retailer supply chain.

the product to the end consumers at a fixed price p. To simplify the algebra in this example, we assume that the random demand $\tilde{D}_1^m$ is uniformly distributed on $[0, 1]$ and $p = 1$. The gross payoff to retailer m is $v^m(x^m) = p(-(x_1^m)^2 + 3x_1^m/2)$ as assumed in Chapter 2. Applying efficient partially affine contracts to obtain the excess transfers that solve the reduced contract design problem (R), we obtained in Chapter 2 that in equilibrium, the net payoffs to retailer 1, retailer 2, and supplier 1 are $\hat{V}^1 = \hat{V}^2 = \frac{1}{4}$ and $U_1 = \frac{1}{2}$. In equilibrium, the supplier's net payoff is strictly positive. Therefore, with efficient partially affine contracts, the retailers (principals) cannot extract all the surplus from the transaction. In contrast, when we apply our methods in Sections 3.3 and 3.4, the Pareto frontier can be found be solving the linear programming problem defined in (3.11) with Matlab; we find that the Pareto frontier has one point with $(\hat{V}^1, \hat{V}^2) = (0.5, 0.5)$.

This example illustrates that by applying our framework, the retailers' equilibrium payoffs are on the Pareto frontier of their attainable payoff set. We can design coordinating contracts to maximize the retailers' attainable equilibrium payoffs over all the possible coordinating contracts. Implicitly, this also answers the question of how the retailers can select from multiple coordinating contracting profiles.

3.5.2 Supply Chain Contracting in a Cournot Oligopoly

We now illustrate how the manufacturers in a multi-principal multi-agent Cournot oligopoly, where $M \geq 2$ manufacturers (principals) supply differentiated goods to $N \geq 1$ retailers, select a coordinating contracting profile. A similar version of this example is discussed in W&X to illustrate how to design coordinating contracts in a multi-principal multi-agent supply chain. As noted in W&X, "[t]his is an archetypical supply chain contracting problem of which many practical instances can be observed (such as Coke and Pepsi supplying their products to a retail chain)." Instead of considering a symmetric system as in W&X, in our approach here the manufacturers and the retailers can be asymmetric. Furthermore, their coordinating contracts based on surplus sharing are only a subset of all possible coordinating contracts. They only answer the question of *how* to find contracting profiles that coordinate the supply chain. The natural follow-up questions are, first, how the manufacturers can extract as much surplus as possible from the system through coordination contract design, and, second, how the surplus can be divided among the manufacturers. We aim to answer these two questions here, and begin with a review of the assumptions. Assume that each manufacturer m produces its goods at a marginal cost $c^m \geq 0$. Each retailer n sells the quantities x_n^m it orders from manufacturer m on a common market at the price

$$p^m(x) = \mu^m - \sum_{j \in \mathcal{N}} \left(x_j^m + \beta \sum_{i \neq m} x_j^i \right).$$

The constant $\beta \in (-1/(M-1), 1)$ indicates the degree to which the products are substitutes ($\beta \geq 0$) or complements ($\beta \leq 0$), and the constant $\mu^m > c^m$ defines the market potential. Since this is a top-down contracting problem, we introduce base wholesale prices w paid from the retailers to the manufacturers and assume that $w^m \geq c^m$. Manufacturer m's gross payoff (before discounts) is thus

$$v^m(x; w) = w^m \sum_{j \in \mathcal{N}} x_j^m - c^m \left(\sum_{j \in \mathcal{N}} x_j^m \right)^{q^m},$$

where $q^m > 0$ indicates the degree of economy of scale. To simplify the computations in our numerical experiments we assume that $q^1 = q^2 = q$. The retailer's gross payoff (before discounts) is

$$u_n(x; w) = \sum_{i \in \mathcal{M}} (p^i(x) - w^i) x_n^i.$$

Naturally, total surplus $W(x) = \sum_{i \in \mathcal{M}} v^i(x; w) + \sum_{j \in \mathcal{N}} u_j(x; w)$ is independent of w and is maximized at the unique efficient outcome $\hat{x}$. Without loss of generality, we assume that the retailers' action set is restricted to all nonnegative order quantities not exceeding twice the maximally plausible "monopolistic" order quantity $\mu^i - c^i$, which would never be exceeded: That is, the retailers' feasible action set $[0, \mu^1 - c^1] \times$

	$K=1$			$K=2$	$K=4$	$K=8$	$K=16$
	$p<0.5$	$p=0.5$	$p>0.5$				
$\hat{V}^1$	0	0.7208	1.0026	0.9401	0.9401	0.9401	0.9401
$\hat{V}^2$	1.0026	0.2818	0	0.0026	0.0026	0.0026	0.0026
$\hat{U}$		0		0.0599	0.0599	0.0599	0.0599

Table 3.1: Sensitivity of Pareto Frontier w.r.t. Discretization.

$\cdots \times [0, \mu^M - c^M]$ is closed and bounded, i.e., compact, as required by the framework outlined in Section 3.2.

We concentrate the numerical experiments on common agency with two manufacturers and one retailer. In these situations, we assume that the marginal production costs are constant, i.e., $q^m = 1$. We consider $\mu^1 = 3$ and $\mu^2 = 1$ to account for different market potentials for the two products, and $c^1 = 1.0, c^2 = 0.5$ to account for different unit production costs. We use Matlab to solve the linear programming problems in (3.11) to obtain the Pareto frontiers, and hence we are able to characterize the Pareto frontiers of the manufacturers' attainable payoff sets.

Common Agency

In this subsection, we focus on the case of common agency with $M = 2$ and $N = 1$.

The Pareto Frontier

We first study how different *discretizations* of the feasible action sets impact the Pareto frontiers and hence the surplus extraction by the two manufacturers. To do this, we fix $\beta = 0.2$, $w^1 = w^2 = 5$, and vary p^1 from 0.1 to 0.9 with an increment of 0.1. Let K represent the grid-size over $[0, \mu^i - c^i]$.

Table 3.1 shows our numerical results of the Pareto frontier with respect to different degrees of discretization. We observe that as long as $K > 1$, the Pareto frontier has one point with $\hat{V}^1 = 0.9401$ and $\hat{V}^2 = 0.0026$. Therefore, for the rest of study with $N = 1$, we fix $K = 2$.

Next we study how *the base wholesale price* w impacts the Pareto frontier. As discussed in W&X, we need to "transform a top-down contracting situation (with

	$w^1 = 1, w^2 = 0.5$			$w^2 = 1$				$w^1 = 2$			
				$w^1 = 1.9$			$w^1 = 2$	$w^2 = 0.5$			$w^2 = 0.6$
	$p < 0.5$	$p = 0.5$	$p > 0.5$	$p < 0.5$	$p = 0.5$	$p > 0.5$		$p < 0.5$	$p = 0.5$	$p > 0.5$	
$\hat{V}^1$	0	0.0013	0.0026	0.9	0.9018	0.9026	0.9401	0.9375	0.9386	0.9401	0.9401
$\hat{V}^2$	0.0026	0.0013	0	0.0026	0.0008	0	0.0026	0.0026	0.0015	0	0.0026
$\hat{V}^1 + \hat{V}^2$		0.0026			0.9026		0.9427		0.9401		0.9427
$\hat{U}_1$		1			0.1		0.0599		0.0625		0.0599

Table 3.2: Sensitivity of Pareto Frontier w.r.t. Base Wholesale Prices.

arbitrary bounded transfers paid from agents to principals) into our standard bottom-up contracting framework (with nonnegative transfers paid from principals to agents) by interpreting outcome-contingent variations of a transfer schedule from an agent to a principal as a nonnegative variable discount that the principal offers to the agent, off a sufficiently high outcome-contingent 'base transfer schedule' from that the agent to the principal." The questions are, first, how high does the base wholesale price need to be; and, second, how does the base wholesale price impact the division of surplus between the manufacturers and the retailer. The experiments that follow aim to answer these two questions. In this set of experiments, we fix $\beta = 0.2$, $K = 2$, and vary p^1 from 0.1 to 0.9 with an increment of 0.1.

Table 3.2 shows our numerical results. When $w = c$, i.e., the base wholesale prices are equal to unit production costs, the Pareto frontier is a straight line. The total payoffs to the two manufacturers are 0.0026 and the payoff to the retailer is 1. When we fix $w^2 = \mu^2 = 1$ for $w^1 \geq 2$, the Pareto frontier has one point independent of w^1 when w^1 is large enough (≥ 2). The total payoffs to the two manufacturers are 0.9427 and the payoff to the retailer is 0.0599. When we fix $w^1 = 2$, for $w^2 \geq 0.6$, the Pareto frontier has one point independent of w^2 when w^2 is large enough (≥ 0.6). The total payoffs to the two manufacturers are 0.9427 and the retailer's payoff is 0.0599. Thus, we demonstrate that when $w^1 \geq 2/3\mu^1, w^2 \geq 0.6\mu^2$, the base wholesale prices are large enough. On the one hand, the manufacturers can use the vehicle of w to extract as much payoff as possible from the system. On the other hand, the manufacturers are unable to extract the full surplus, no matter how large w is, when the products are substitutes.

We next study how *the degree to which the products are substitutes or complements* impacts the Pareto frontier, with $\beta > 0$ accounting for substitutes and $\beta < 0$ accounting for complements. In this set of experiments, we fix $K = 2$, $w^1 = 3, w^2 = 1$, and vary p^1 from 0.1 to 0.9 with an increment of 0.1.

When $\beta > 0$, i.e., when the two products are substitutes, the in-equilibrium net payoff to the retailer is strictly positive and nondecreasing in β, while the in-equilibrium net payoff to the manufacturers are nonincreasing in β. We also observe that the Pareto frontier has one point. The percentage of the surplus of the manufacturers over the total system surplus is 93.75%, and thus the manufacturers cannot extract all the surplus. The division of surplus between the manufacturers and the retailer is independent of the vector λ in (3.10). Therefore, we conjecture that the reason is that both the manufacturers can obtain their maximum attainable payoffs. As shown in Lemma 1, the maximum attainable equilibrium payoff of manufacturer m is $\bar{V}^m = v^m(\hat{x}) + \min_{x_1 \in \mathcal{X}_1} \left\{ G_1(x_1) - \sum_{i \neq m} F^i(x_1) \right\}$. This is actually what happens in this example. Take $\beta = 0.2$ for example. With $\mu^1 = w^1 = 3$, $\mu^2 = w^2 = 1$, and $c^1 = 1$ and $c^2 = 0.5$, we have $\hat{x} = (0.9896, 0.0521)$ and $\xi^1 = (0, \frac{\mu^2 - c^2}{2}), \xi^2 = (\frac{\mu^1 - c^1}{2}, 0)$. Hence, by Lemma 1, $\bar{V}^1 = \hat{v}^1 + G_1(\xi^1) - F^2(\xi^1) = 0.9401$ and $\bar{V}^2 = \hat{v}^2 + G_1(\xi^2) - F^1(\xi^2) = 0.0026$. In our numerical study, we do have $\hat{V}^1 = \bar{V}^1 = 0.9401$ and $\hat{V}^2 = \bar{V}^2 = 0.0026$. Therefore, in our example, each manufacturer is able to obtain her maximum attainable payoff independent of λ in (3.10). We further demonstrate that the necessary conditions in Lemma 2 are satisfied. That is, when each manufacturer obtains her maximum attainable payoff, it is $F^m(\xi^i) \geq \Omega(\xi^m) + F^m(\xi^m)$. In this example, we have $F^1(\xi^2) - (\Omega(\xi^1) + F^1(\xi^1)) = 1.0599 > 0$ and $F^2(\xi^1) - (\Omega(\xi^2) + F^2(\xi^2)) = 0.1224 > 0$. Therefore, the retailer's action ξ^1 does not influence the retailer's action ξ^2. It is possible that the manufacturers can coordinate their actions in such a way that each manufacturer obtains her maximum possible payoff, and thus, the Pareto surplus division between the manufacturers and the retailer is independent of λ. The Pareto frontier has one point. Unfortunately, this example cannot demonstrate that the sufficient conditions in Lemma 2 are satisfied.

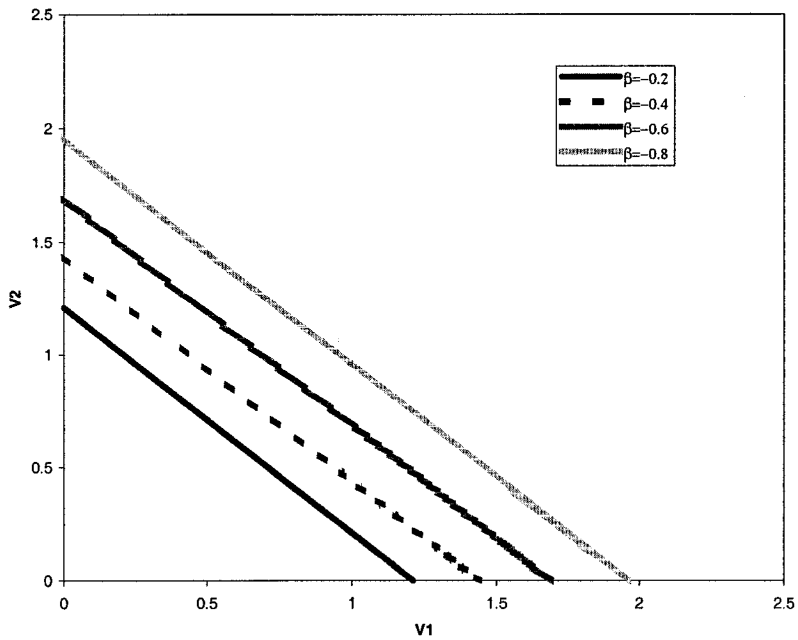

Figure 3.1: Sensitivity of Pareto Frontier w.r.t. β when $\beta < 0$ (Complements).

Figure 3.1 shows the sensitivity of the Pareto frontier with respect to the degree of complements. When $\beta < 0$, the Pareto frontiers are straight lines with 45 degree slope. This is because when the products are complements, the principals can extract all the surplus, and we have $\hat{V}^1+\hat{V}^2 = \hat{W}$. The total surplus of the two manufacturers is increasing in the absolute value of β. The more the products are complements, the higher the system surplus and hence the higher the total surplus distributed to the manufacturers.

Finally, we study how a different *production cost structure* (namely, through economies of scale) influences the Pareto frontier. We study two cases with $\beta = 0.2$ (substitutes) and $\beta = -0.2$ (complements). Here we set $K = 4$ and vary p^1 from 0.1 to 0.9 with an increment of 0.1. Figures 3.2 and 3.3 show the Pareto frontier varying with respect to q for substitutes and complements respectively. We observe that for substitutes, the Pareto frontiers have one point, and the manufacturers cannot extract all the surplus except for $q^m = 0.25$. For complements, the Pareto frontiers are straight lines, and the manufacturers can extract all the surplus.

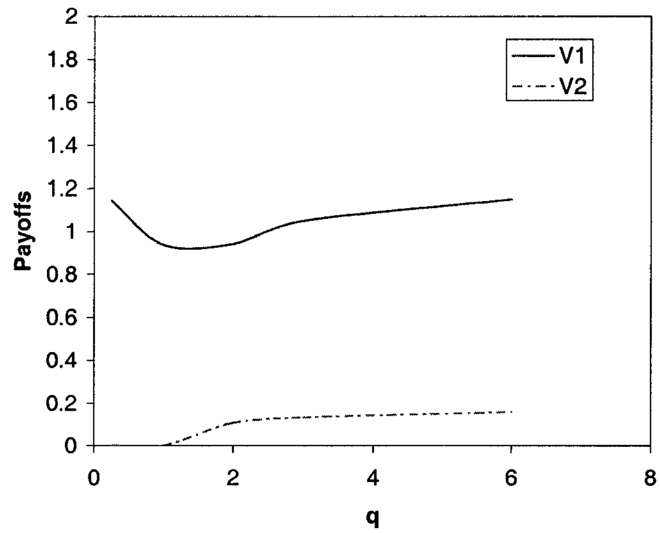

Figure 3.2: Sensitivity of Pareto Frontier w.r.t. q for Substitutes.

In summary, when the products are complements, the Pareto frontier is a straight line, with the manufacturers extracting all the system surplus.

3.6 Discussion

W&X provide a set of closed-form contracting equilibria that coordinate a two-echelon supply chain implementing any given efficient outcome as a WTE of the contracting game. In this work we provide an answer to the question of surplus extraction, i.e., among all the coordinating contracts, what contracts would enable the principals to extract as much as surplus as possible from the agents. To achieve that, we first show that the in-equilibrium net payoffs of the principals can be completely determined by the solutions to the reduced contract-design problem (R) since the operator mapping solutions of (R) to solutions of the full contract design problem (R),(PM) can be

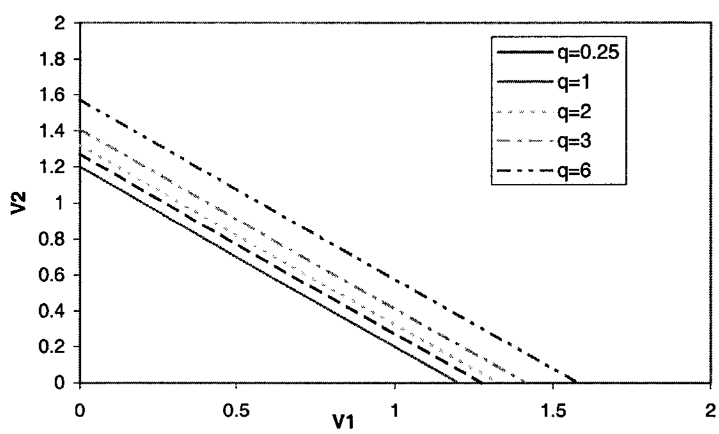

Figure 3.3: Sensitivity of Pareto Frontier w.r.t. q for Complements.

made independent of the parametrization in W&X. We then derive the primitives of the principals' payoff attainability set $\mathcal{V}$ and demonstrate that $\mathcal{V}$ (generated by all coordinating WTEs of the contracting game) is convex. We also show that it is possible that in a win-win scenario every principal gets what she can possibly hope for, provided that all the other principals cooperate.

By numerical study we can obtain the Pareto frontier for general multi-principal multi-agent supply chains by solving linear programming problems. In our numerical example of Cournot Oligopoly, we also implicitly address the question of surplus division among the principals and the agents. When the products are substitutes, the Pareto frontier has only one point, which is the mostly likely outcome of cooperative pre-play communication. When the products are complements, the Pareto frontiers are straight lines with 45 degree slope.

In a general case, the division of surplus between principals and agents in a coordinated supply chain depends on which particular profile of equilibrium contracts is selected by the principals. To fully resolve the multiplicity problem[8], one needs to address the question of *surplus distribution* among principals after it has been extracted

[8]As suggested by Martimort (2006), this *multiplicity problem* is a common feature of multi-contracting problems and therefore leads naturally to a coordination problem for principals.

from agents, which is a worthwhile future research challenge. Since the principals' payoff attainability set is convex, in future research one can apply cooperative bargaining theory to arrive at predictions of which principal-efficient allocation(s) are the most likely outcome of cooperative pre-play communication.

3.7 Appendix: Proofs

Proof of Proposition 8. Fix any $\check{\Delta} \in \mathcal{R}$ and any $\vartheta, \check{\vartheta} \in \Theta$. To prove the claim, we only need to show that there exists a $\Delta \in \mathcal{R}$ such that $T(\Delta, \vartheta) = T(\check{\Delta}, \check{\vartheta})$. Since ϑ is in Θ and thus the element of a simplex, the nonnegative constant

$$\rho_n(\vartheta, \check{\vartheta}) = \min_{m \in \{i : \vartheta_n^i > 0\}} \{\check{\vartheta}_n^m / \vartheta_n^m\}$$

is well defined for any $\vartheta, \check{\vartheta} \in \Theta$. Thus, by setting

$$\Delta = [\Delta_n^m] = \left[\check{\Delta}_n^m + \left(\check{\vartheta}_n^m - \rho_n(\vartheta, \check{\vartheta})\vartheta_n^m\right)\left(G_n(x) - \sum_{i \in \mathcal{M}} \check{\Delta}_n^i(x)\right)\right]$$

we can guarantee that $\Delta \in \mathcal{R}$, i.e., Δ solves (R). First we show that Δ satisfies the second inequality in (R), corresponding to the agents' payoff maximization problem. For this, we notice that by the construction of Δ one obtains that

$$\sum_{i \in \mathcal{M}} \Delta_n^i = \sum_{i \in \mathcal{M}} \check{\Delta}_n^i + (1 - \rho_n)\left(G_n(x) - \sum_{i \in \mathcal{M}} \check{\Delta}_n^i(x)\right)$$

for all $n \in \mathcal{N}$, and

$$G_n(x) - \sum_{i \in \mathcal{M}} \Delta_n^i(x) = \rho_n\left(G_n(x) - \sum_{i \in \mathcal{M}} \check{\Delta}_n^i(x)\right) \geq 0$$

The inequality holds because by assumption $\check{\Delta} \in \mathcal{R}$, i.e., $\check{\Delta}$ indeed satisfies the second inequality in (R), so that $G_n(x) - \sum_{i \in \mathcal{M}} \check{\Delta}_n^i(x) \geq 0$ (by its definition, $\rho_n \geq 0$). Next we show that Δ also satisfies the first inequality in (R). By the definition of ρ_n we have $\check{\vartheta}_n^m \geq \rho_n \vartheta_n^m$ and hence $\Delta_n^m \geq \check{\Delta}_n^m$. Since by assumption $\check{\Delta} \in \mathcal{R}$ and thus $\check{\Delta}$ satisfies the first inequality in (R), it is

$$F^m(x) - \sum_{j \in \mathcal{N}} \Delta_j^m \leq F^m(x) - \sum_{j \in \mathcal{N}} \check{\Delta}_j^m(x) \leq 0$$

for all $m \in \mathcal{M}$ and all $x \in \mathcal{X}$. Therefore, Δ_n^m also satisfies the first inequality (R).

Finally, we have that

$$\begin{aligned} T(\Delta, \vartheta) &= \left[\Delta_n^m + \vartheta_n^m \left(G_n - \sum_{i \in \mathcal{M}} \Delta_n^i\right)\right] \\ &= \left[\check{\Delta}_n^m + \left(\check{\vartheta}_n^m - \rho_n \vartheta_n^m\right) \left(G_n(x) - \sum_{i \in \mathcal{M}} \check{\Delta}_n^i(x)\right) + \vartheta_n^m \rho_n \left(G_n - \sum_{i \in \mathcal{M}} \hat{\Delta}_n^i\right)\right] \\ &= \left[\check{\Delta}_n^m + \check{\vartheta}_n^m \left(G_n - \sum_{i \in \mathcal{M}} \check{\Delta}_n^i\right)\right] \\ &= T(\check{\Delta}, \check{\vartheta}), \end{aligned}$$

which concludes our proof. ∎

Proof of Proposition 9. Given any $\Delta \in \mathcal{R}$, principal m's in-equilibrium net payoff is given by

$$\hat{V}^m(\Delta) = v^m(\hat{x}) + \sum_n \min_{x_n \in \mathcal{X}_n} \{T_n^m(\Delta(x_n, \hat{x}_{-n}))\}.$$

The mapping $\hat{V}^m : \mathcal{R} \to \mathbb{R}$ is concave. Indeed, consider $\Delta, \hat{\Delta} \in \mathcal{R}$. Then by the convexity of $\mathcal{R}$ the convex combination $\lambda\Delta + (1-\lambda)\hat{\Delta}$ is also in $\mathcal{R}$ for any $\lambda \in (0,1)$.

Furthermore,

$$\begin{aligned}
\hat{V}^m(\lambda\Delta + (1-\lambda)\hat{\Delta}) &= v^m(\hat{x}) + \sum_n \min_{x_n \in \mathcal{X}_n} \left\{ T_n^m(\lambda\Delta(x_n, \hat{x}_{-n})) + (1-\lambda)\hat{\Delta}(x_n, \hat{x}_{-n})) \right\} \\
&= v^m(\hat{x}) + \sum_n \min_{x_n \in \mathcal{X}_n} \left\{ \lambda T_n^m(\Delta(x_n, \hat{x}_{-n})) + (1-\lambda) T_n^m(\hat{\Delta}(x_n, \hat{x}_{-n})) \right\} \\
&\geq \lambda \hat{V}^m(\Delta) + (1-\lambda)\hat{V}^m(\hat{\Delta}),
\end{aligned}$$

so that

$$\hat{V}(\Delta), \hat{V}(\hat{\Delta}) \in \mathcal{V} \quad \Rightarrow \quad \lambda \hat{V}^m(\Delta) + (1-\lambda)\hat{V}^m(\hat{\Delta}) \leq V(\lambda\Delta + (1-\lambda)\hat{\Delta}) \in \mathcal{V}$$

for all $\lambda \in (0,1)$, which implies convexity of $\mathcal{V}$.

■

Proof of Proposition 10. Fix any $m \in \mathcal{M}$. Principal m's payoffs cannot decrease if she obtains $\vartheta_n^m = 1$ for all agents $n \in \mathcal{N}$. Fixing the corresponding $\vartheta \in \Theta$, we obtain that

$$T_n^m(\Delta, \vartheta) = \Delta_n^m + \left(G_n - \sum_{i \in \mathcal{M}} \Delta_n^i \right) \geq T_n^m(\Delta, \hat{\vartheta})$$

for any $\hat{\vartheta} \in \Theta$. Hence,

$$\begin{aligned}
\hat{V}^m(\Delta) - v^m(\hat{x}) &= \sum_n \min_{x_n \in \mathcal{X}_n} \{ T_n^m(\Delta(x_n, \hat{x}_{-n})) \} \\
&= \sum_n \min_{x_n \in \mathcal{X}_n} \left\{ \Delta_n^m(x_n, \hat{x}_{-n}) + \left(G_n(x_n, \hat{x}_{-n}) - \sum_{i \in \mathcal{M}} \Delta_n^i(x_n, \hat{x}_{-n}) \right) \right\} \\
&= \min_{x \in \mathcal{X}} \left\{ \sum_{n \in \mathcal{N}} \Delta_n^m(x_n, \hat{x}_{-n}) + \sum_{n \in \mathcal{N}} \left(G_n(x_n, \hat{x}_{-n}) - \sum_{i \in \mathcal{M}} \Delta_n^i(x_n, \hat{x}_{-n}) \right) \right\} \\
&= \min_{x \in \mathcal{X}} \left\{ F^m(x_n, \hat{x}_{-n}) + \varphi^m(x_n, \hat{x}_{-n}) + \sum_{n \in \mathcal{N}} \gamma_n(x_n, \hat{x}_{-n}) \right\}
\end{aligned}$$

Since the slack variables φ^m and γ_n are nonnegative, it is

$$\hat{V}^m(\Delta) - v^m(\hat{x}) \geq \min_{x \in \mathcal{X}} \{F^m(x_n, \hat{x}_{-n})\}.$$

On the other hand, using relations (3.4) and (3.5) we have that

$$\begin{aligned}
\hat{V}^m(\Delta) - v^m(\hat{x}) &= \min_{x \in \mathcal{X}} \left\{ F^m(x_n, \hat{x}_{-n}) + \varphi^m(x_n, \hat{x}_{-n}) + \Omega(x_n, \hat{x}_{-n}) - \sum_{i \in \mathcal{M}} \varphi^i(x_n, \hat{x}_{-n}) \right\} \\
&= \min_{x \in \mathcal{X}} \left\{ \Omega(x_n, \hat{x}_{-n}) + F^m(x_n, \hat{x}_{-n}) - \sum_{i \neq m} \varphi^i(x_n, \hat{x}_{-n}) \right\} \\
&\leq \min_{x \in \mathcal{X}} \{ \Omega(x_n, \hat{x}_{-n}) + F^m(x_n, \hat{x}_{-n}) \},
\end{aligned}$$

which concludes our proof. ■

Proof of Lemma 1. In a WTE, the net payoff to principal m under common agency can by virtue of (3.8) be written in the form

$$\hat{V}^m(\Delta) = v^m(\hat{x}) + \min_{x_1 \in \mathcal{X}_1} T_1^m(\Delta(x_1))$$

with

$$T_1^m(\Delta(x)) = \Delta_1^m(x) + \frac{1}{M}\left(G_1(x) - \sum_{i \in \mathcal{M}} \Delta_1^i(x) \right),$$

where $\Delta \in \mathcal{R}, m \in \mathcal{M}$. Since $\Delta \in \mathcal{R}$, Δ_1^m solves (R), so that

$$\Delta_1^m \leq G_1 - \sum_{i \neq m} \Delta_1^i \leq G_1 - \sum_{i \neq m} F^i$$

The first inequality is by (R) equivalent to $\sum_i \Delta_1^i \leq G_1$, and the second inequality is equivalent to $-\Delta_1^m \leq -F^m$. Therefore,

$$\begin{aligned}\hat{V}^m(\Delta) - v^m(\hat{x}) &= \min_{x_1 \in \mathcal{X}_1} \left\{ \Delta_1^m(x_1) + \left(G_1(x_1) - \sum_{i \in \mathcal{M}} \Delta_1^i(x_1) \right) /M \right\} \\ &\leq \min_{x_1 \in \mathcal{X}_1} \Delta_1^m(x_1) \\ &\leq \min_{x_1 \in \mathcal{X}_1} \left\{ G_1(x_1) - \sum_{i \neq m} F^i(x_1) \right\},\end{aligned}$$

which concludes our proof. ■

Proof of Lemma 2. (i) When (3.9) is satisfied, we can set

$$\Delta_1^m(\xi^m) = G_1(\xi^m) - \sum_i F^i(\xi^m) = \Omega(\xi^m) + F(\xi^m) \geq F^m(\xi^m),$$

and $\Delta_1^i(\xi^m) = F^i(\xi^m)$ for all $m \in \mathcal{M}$ and all $i \neq m$. At all other points $x_1 \neq \xi^m$ for all $m \in \mathcal{M}$, we let $\Delta_1^m(x_1) = F_1^m(x_1)$. Clearly $\Delta_1^m(x_1)$ satisfies the first inequality in (R). Since $G_1(x_1) - \sum_{i \in \mathcal{M}} \Delta_1^i(x_1) = G_1(x_1) - \sum_{i \in \mathcal{M}} F^i(x_1) = \Omega(x_1) \geq 0$, $\Delta_1^m(x_1)$ satisfies the second inequality in (R). And thus $\Delta_1^m \in \mathcal{R}$. Next, we observe that

$$\Delta_1^m(x_1) + \left(G_1(x_1) - \sum_i \Delta_1^i(x_1) \right) /M \geq \Delta_1^m(x_1) = F^m(x_1) \geq \Delta_1^m(\xi^m) = \Omega(\xi^m) + F^m(\xi^m).$$

The first inequality following from the second inequality in (R), the second inequality holds by the construction of $\Delta_1^m(x_1)$, and the last inequality is implied by (3.9). Therefore, the above construction of Δ enables the agent to choose action ξ^m when principal m removes herself from the contracting game for all $m \in \mathcal{M}$. By Lemma 1 each principal obtains her maximum attainable payoff in equilibrium. (ii) By Lemma 1, principal m's maximum possible payoff is obtained at $\xi^m = \arg\min_{x_1 \in \mathcal{X}_1} \left\{ G_1(x_1) - \sum_{i \neq m} F^i(x_1) \right\}$ with

$$\Delta_1^m(\xi^m) = G_1(\xi^m) - \sum_{i \neq m} F^i(\xi^m), \tag{3.12}$$

$$\Delta_1^i(\xi^m) = F^i(\xi^m), \tag{3.13}$$

$$G_1(\xi^m) - \sum_{i \in \mathcal{M}} \Delta_1^i(\xi^m) = 0 \tag{3.14}$$

for all $i \neq m$. Since by assumption, principal m gets her maximum attainable payoff, then it is actually in the agent's best interests to take the action ξ^m when principal m removes herself from the contracting game, i.e., at any points x_1 other than ξ^m,

$$\begin{aligned}
T_1^m(\Delta(x_1)) &= \Delta_1^m(x_1) + \left(G_1(x_1) - \sum_i \Delta_1^i(x_1) \right) /M \\
&\geq T_1^m(\Delta(\xi^m)) \\
&= \Delta_1^m(\xi^m) + \left(G_1(\xi^m) - \sum_{i \in \mathcal{M}} \Delta_1^i(\xi^m) \right) /M \\
&= \Delta_1^m(\xi^m) \\
&= G_1(\xi^m) - \sum_{i \neq m} F^i(\xi^m)
\end{aligned} \tag{3.15}$$

for $m \in \mathcal{M}$. In particular, (3.15) must hold for $x_1 = \xi^i$, $i \neq m$, i.e.,

$$\begin{aligned}
T_1^m(\Delta(\xi^i)) &= \Delta_1^m(\xi^i) + \left(G_1(\xi^i) - \sum_i \Delta_1^i(\xi^i) \right) /M = \Delta_1^m(\xi^i) = F^m(\xi^i) \\
&\geq \Delta_1^m(\xi^m) = \Omega(\xi^m) + F^m(\xi^m)
\end{aligned}$$

for all $m \in \mathcal{M}$ and all $i \neq m$, which concludes our proof. ■

Chapter 4

Stock Positioning for Distribution Systems with Service Constraints

4.1 Introduction

The present chapter addresses inventory allocation in a *distribution* system subject to service-level constraints. The objective is to satisfy end customer demand while minimizing the inventory holding cost by optimally allocating inventory among the warehouse and several retailers. In the retail industry, service levels are among the most important performance metrics. In a recent survey (Chain Store Age 2002), respondents rank customer fill rate as the number-one metric used for inventory management. Cohen, Desphande and Wang (2003) provide various examples from two different industries, namely semiconductor and military supply chains, to illustrate the practical importance of service constraints in multi-echelon systems.

In particular, we study a distribution system with one warehouse replenishing multiple non-identical J retailers. Stochastic demands arrive at the retailers only. Each retailer satisfies demand through on-hand inventory, if there is any. Otherwise, unsatisfied demand is backordered at each location. The warehouse allocates inventory to the retailers on a first come first served basis. Inventory in the distribution system is reviewed continuously. Each location follows a base-stock policy, that is, a one-for-one replenishment policy based on local inventory information. The warehouse replenishes from an outside source with ample supply. Each shipment requires a leadtime but no fixed costs. Each location incurs a holding cost. The retailers are subject to service-level constraints. We focus mainly on fill-rate type service-level constraints, e.g., the fraction of demand met from on-hand inventory. The data is assumed stationary and the time horizon is infinite. The objective is to minimize the total average holding cost subject to service-level requirements at the retailers.

A distribution system with service-level constraints at the retailers is a fundamental structure for many supply chains. We provide an algorithm that provides an optimal base-stock level at each location while ensuring that fill-rate service-level requirements are satisfied. Distribution systems can be very large in practice. For example, GM service parts distribution center manages 5 Million parts across 8000 dealers (see Cohen et al 2003 for other examples). Hence, optimal solutions may not be tractable for large-scale systems. Computational improvements can thus be very important. We establish monotonicity results and bounds on base-stock levels that help to improve the computational requirement for the algorithm. These bounds also help to develop heuristics that are computationally very fast.

We propose two heuristics to determine base-stock levels in a distribution system. The triple-search heuristic (TSH) considers three feasible stock positioning strategies and selects the best. An extensive numerical study (over 530 tested problem instances) has found that this heuristic's cost is on average 1.18% more than the best base-stock policy's cost. It is computationally fast, and hence amenable for real applications. The newsvendor heuristic (NVH) solves $2J$ newsvendor problems to allocate inventory across the distribution system. The idea is based on decomposing the distribution system into J two-location serial systems by restricting the warehouse to keep separate stocks for each retailer. Next the NVH solves a single newsvendor problem per location of the decomposed system. The heuristic's cost is 18% more than that of the best base-stock policy in our numerical study. This NVH is not as accurate as the TSH, but it is computationally much faster and simpler to use and describe. The only advanced knowledge required to apply this heuristic is the newsvendor formulation taught in any operations management course. The computational efficiency also enables one to analyze large-scale systems that manage thousands of SKUs. Finally, the NVH also provides a step for developing a closed-form approximation, i.e., the newsvendor approximation.

Next, we propose three closed-form approximations. The main purpose of these approximations is to predict the system's performance. Our tests show that all the three approximations perform this task fairly well. Using these approximations we provide insights into stock positioning and quantify, for example, how logistic postponement or consolidation of retailers affect the distribution system's performance. Compared to the optimal solution and the heuristics, these approximations require much less data and are easier to describe to practitioners.

An obvious one-to-one relationship between a service-constrained model and a backorder-cost models does not exist for multi-echelon inventory systems (see, for example, Boyaci and Gallego 2001 for a discussion on serial systems). Hence, since the 1970s two separate streams of research have evolved: one dealing with backorder-cost problems and another one dealing with service-constrained problems. For example, several authors study serial systems with backorder costs (Clark and Scarf 1960, Gallego and Zipkin 1990, Chen and Zheng 1994, Shang and Song 2003, and references therein). Yet, several other researchers also study serial systems with service constraints (Boyaci and Gallego 2001, Axsater 2003 and Sobel 2003). Researchers have been addressing service-constrained models and backorder-cost models separately because (i) the service-constrained systems have significant practical relevance, (ii) yet they are computationally and analytically difficult to deal with even for a *single* location model, and (iii) they cannot be addressed by a simple conversion to backorder-cost models. Diks, de Kok and Lagodimos (1996) provide a comprehensive review of *service-constrained* inventory models. Next we briefly review the literature on service-constrained models.

The multi-echelon literature studies four fundamental topologies: single location, locations in serial, assembly systems, and distribution (arborescent) systems.[1] The literature on service-constrained inventory control problems mainly studies a *single location* problem. Cohen, Kleindorfer and Lee (1988a) develop a greedy heuristic to minimize expected costs subject to a fill-rate type service-level constraint for a periodic-review inventory system with fixed costs, priority demand classes, and lost sales. Bashyam and Fu (1998) propose a gradient-based simulation method to compute (s, S) inventory control policies in systems with random leadtime and a service constraint. Agrawal and Seshadri (2000) consider a fill-rate constrained (Q, r) inventory problem. They develop distribution-free bounds for the policy parameters and provide an algorithm to obtain those policy parameters.

Another group of researchers studies *multiple-locations* in which *only* the last location faces a service constraint. Boyaci and Gallego (2001) develop an optimal algorithm to obtain base-stock levels for a *serial* system. This algorithm could be computationally complex for a system with more than two stages. Hence, they also provide efficient heuristics. Sobel (2003) presents formulas for the fill-rate of a periodic-review serial system that follows base-stock policies. Cohen, Kleindorfer and Lee (1988b) and Cheng, Ettl, Lin and Yao (2002) consider *assembly* systems and a *configure-to-order* environment, respectively. They provide algorithms that allocate inventory across the systems to satisfy the service constraints for the final products.

Relatively little work exists in the literature on distribution systems subject to service requirements at various demand points. Rosenbaum (1981) provides one of the earliest works that considers service-level relationships in a continuous review distribution system. She develops a heuristic to determine the service level a customer receives given a combination of warehouse and retailer service levels. Schwarz,

[1] Axsäter (1993) and Federgruen (1993) provide a comprehensive review for the extensive *backorder-cost* inventory models.

Deuermeyer and Badinelli (1985) consider continuous review replenishment policies. The authors approximate the retailer fill-rates and expected warehouse delay and maximize system fill-rate subject to an investment constraint on system safety stock. We are unaware of any work that provides an algorithm to optimally determine base-stock levels in a distribution system subject to service requirements at each retailer. Note that a service-constrained distribution system is substantially different from a service-constrained serial system and a backorder-cost distribution system. In this chapter, we build on the recent developments in multi-echelon inventory systems and address this fundamental supply chain structure. Throughout the paper we provide additional references.

The rest of the chapter is organized as follows. In Section 4.2, we describe the model and provide some properties of optimal base-stock levels. Using these properties, we propose an algorithm to obtain optimal base-stock levels for a distribution system subject to service-level constraints. In Section 4.3, we develop two heurisitics: the triple-search heuristic and the newsvendor heuristic. In Section 4.4, we provide three closed-form approximations. In Section 4.5, we conduct an extensive numerical study, and report the accuracy of the heuristics and the performance of the approximations. In Section 4.6, we investigate the relationship between the backorder-cost models and the service-constrained models. In Section 4.7, we conclude the paper with a discussion on how these heuristics and approximations contribute to theory and practice.

4.2 The Model

We consider a distribution system with single warehouse supplying multiple retailers. Let J denote the number of retailers. The warehouse is indexed by $j = 0$, and the retailers are indexed by $j = \{1, \ldots, J\}$. All locations are allowed to carry inventory. The local holding cost at retailer j is h_j per unit of on-hand inventory. Due to value added operations, holding inventory at the retailers is more expensive than holding it in the warehouse, and hence, $h_j \geq h_0$. Stochastic demands arrive at the retailers. Demand is satisfied through on-hand inventory, if there is any. Otherwise, unsatisfied demand is backordered at each location. Only the retailers are subject to service-level constraints. Inventory is reviewed and replenished continuously. Each location follows a local base-stock policy. The warehouse replenishes from a source with ample supply. Whenever the inventory position[2] at location $j \geq 0$ falls below s_j, it orders from its immediate upstream location to bring its inventory position back to s_j. The warehouse satisfies retailers' orders on a first come first served basis[3]. Shipments from an upstream location arrive at the downstream location after time L_j. Demand at each retailer j is independent and follows a Poisson process $\{\mathcal{D}_j(t), t \geq 0\}$ with rate λ_j. Hence, the warehouse's demand process is Poisson with rate $\lambda_0 = \sum_{j>0} \lambda_j$.

The state of the system is defined by B_j, the backorders, and by I_j, the on-hand inventory at each location $j = \{0, \ldots, J\}$. We let B_{0j} denote the number of backorders for retailer j at the warehouse. Following a top-down approach, we have

$$\begin{aligned} B_0 &= [D_0 - s_0]^+, \\ I_0 &= [s_0 - D_0]^+, \\ B_j &= [(B_{0j}(s_0) + D_j) - s_j]^+, \\ I_j &= [s_j - (B_{0j}(s_0) + D_j)]^+. \end{aligned}$$

where D_0 and D_j are lead-time demand at the warehouse and retailer j. $B_{0j} + D_j$ can be interpreted as the effective demand at retailer j. Because B_{0j} and D_j are independent, $B_{0j} \mid B_0$ is binomial with parameters B_0 and $\theta_j \equiv \frac{\lambda_j}{\lambda_0}$.

The type of service-level constraint we consider here is the limit probability of

[2]Inventory position equals on-hand inventory plus inventory on order minus backorders.

[3]This is also called a local control system in which each location's decision is only based on its local inventory information. The local control has an advantage of decentralized management among different organizations or within different departments in a large organization. In addition, it simplifies the analysis and it is a fair policy which is commonly used in practice.

having *positive* on-hand inventory at retailer j, that is,

$$Pr(B_{0j}(s_0) + D_j < s_j).$$

Because we assume Poisson demand processes, this limit probability is equal to the long-run fraction of demand that sees a positive inventory level at retailer j, which is also defined as fill rate.

The objective is to minimize the total average inventory holding cost subject to fill-rate type service-level constraints.

$$\begin{aligned} \min_{s_0, s_1, \ldots, s_J} \quad & h_0 E[s_0 - D_0]^+ + \textstyle\sum_{j>0} h_j E[s_j - (B_{0j}(s_0) + D_j)]^+ \\ \text{subject to} \quad & Pr(B_{0j}(s_0) + D_j < s_j) \geq \beta_j, \quad \forall j \in \{1, \ldots, J\}, \end{aligned} \tag{4.1}$$

where β_j represents a pre-specified fill-rate for retailer j. Here, we ignore the pipeline inventory holding cost because this cost is a constant in the steady-state.

Given the base-stock level s_0 at the warehouse, the distribution of B_{0j} can be obtained. Hence, given s_0, the distribution system separates into J single-location problems subject to the service-level constraints as follows.

$$\begin{aligned} \min_{s_j} \quad & h_j E[s_j - (B_{0j}(s_0) + D_j)]^+ \\ \text{subject to} \quad & Pr(B_{0j}(s_0) + D_j < s_j) \geq \beta_j \end{aligned} \tag{4.2}$$

for $j \in \{1, \ldots, J\}$.

These single-location problems are simpler to solve than the original problem in (4.1) because B_{0j} depends only on s_0, and $h_j E[s_j - (B_{0j}(s_0) + D_j)]^+$ is increasing in s_j. Hence, the optimal base-stock level at retailer j is the minimum one that satisfies its service-level constraint. We define

$$s_j^*(s_0) \equiv \min\{s_j : Pr(B_{0j}(s_0) + D_j < s_j) \geq \beta_j\}. \tag{4.3}$$

4.2.1 Properties of Optimal Base-Stock Levels

Here we provide some monotonicity results and discuss the analytical relationship between the fill rates and the base-stock levels. These results enable us to obtain bounds on the optimal base-stock levels, and to develop heuristics that improve the computational time required to obtain the optimal cost and base-stock levels at each location.

The following lemma shows that increasing the warehouse base-stock level provides higher fill-rates at all the retailers. We defer all the proofs to the appendix.

LEMMA 1 $Pr(B_{0j}(s_0) + D_j < s_j) \leq Pr(B_{0j}(s_0 + 1) + D_j < s_j)$ *for any* $j > 0$.

Next we investigate the impact of moving one unit of inventory from the warehouse to a retailer, a location closer to end customers. The next lemma proves that moving one unit of stock from the warehouse to retailer j does indeed increase the fill rate at retailer j for a *distribution* system.

LEMMA 2 $Pr(B_{0j}(s_0) + D_j < s_j + 1) \geq Pr(B_{0j}(s_0 + 1) + D_j < s_j)$ *for any* $j > 0$.

Let s_j^l denote the minimum base-stock level necessary to satisfy the fill-rate requirement at retailer j when the warehouse has *infinite* supply, that is,

$$s_j^l \equiv s_j^*(\infty) \equiv \min\{s_j : Pr(D_j < s_j) \geq \beta_j\}. \tag{4.4}$$

Similarly, we define s_j^u as the minimum base-stock level necessary to satisfy the fill-rate at retailer j when the warehouse holds zero inventory, that is,

$$s_j^u \equiv s_j^*(0) \equiv \min\{s_j : Pr(B_{0j}(0) + D_j < s_j) \geq \beta_j\}, \tag{4.5}$$

$B_{0j}(0)$ has Poisson distribution with mean $\lambda_j L_0$ because $Pr(B_{0j} = m) = \sum_{n \geq m} Pr(B_{0j} = m \mid B_0 = n) Pr(B_0 = n) = \frac{(\lambda_j L_0)^m}{m!} e^{-\lambda_j L_0}$. Hence, $B_{0j}(0) + D_j$ has Poisson distribution with mean $\lambda_j(L_0 + L_j)$.

Note that it is computationally straightforward to obtain s_j^l and s_j^u, i.e., to increase s_j from 0 until (4.4) or (4.5) is satisfied. Finally, we define

$$s_0^*(s_1, ..., s_J) = \min\{s_0 : Pr(B_{0j}(s_0) + D_j < s_j) \geq \beta_j, \forall j > 0\}. \tag{4.6}$$

Next we use Lemmas 1 and 2 to establish a relationship between the warehouse base-stock level s_0 and the retailer base-stock level s_j's.

Proposition 1 *For all $j \geq 1$, we have*

1. $s_0^*(s_1, ..., s_j + 1, ..., s_J) \leq s_0^*(s_1, ..., s_j, ..., s_J)$,
2. $s_j^*(s_0 + 1) \leq s_j^*(s_0)$,
3. $s_j^l \leq s_j^*(s_0^*) \leq s_j^u$,
4. $s_j^*(s_0) \leq s_j^*(s_0 + 1) + 1$.

The first and second parts of the proposition demonstrate that the warehouse stock and the retailer stocks are complementary for satisfying the service-level constraints at the retailers. To satisfy a given service requirement β_j, when the warehouse carries more inventory, the retailer j optimally carries less inventory. Similarly, if any retailer carries additional inventory, the warehouse's optimal base-stock level will be kept at the same level or reduced. The third part shows that the optimal base-stock level at any retailer is bounded. These bounds are very easy to compute. The fourth part implies that a unit of stock closer to the end customer provides a better protection against demand uncertainty.

Next, we obtain an upper bound s_0^u for the optimal base-stock level at the warehouse. From Part 1 of Proposition 1, we know that the less inventory the retailers carry, the more stock is needed at the warehouse. Hence, by setting the base-stock level s_j at each retailer j equal to its lower bound s_j^l, we obtain the upper bound s_0^u

as follows.

$$s_0^u \equiv \min\{s_0 : Pr(B_{0j}(s_0) + D_j < s_j^l) \geq \beta_j, \forall j > 0\}. \tag{4.7}$$

PROPOSITION 2 *It is* $s_0^* \leq s_0^u$.

Note that the above properties play an important role in establishing a fast algorithm to obtain the optimal base-stock levels and the resulting cost. They also enable us to develop efficient heuristics in addition to providing insights into stock positioning in a distribution system.

4.2.2 Optimal Solution

The total average holding cost in (4.1) is not necessarily convex in s_0. Hence, we search over all possible values of s_0 to find an optimal solution. To simplify this search, we use the results on the upper bound for the optimal base-stock level at the warehouse s_0^u from the previous subsection. Recall also from (4.2) and (4.3) that given a base-stock level s_0 for the warehouse, the optimal base-stock levels for the retailers are obtained by solving independent single location problems. Let c^* be the *optimal* total average holding cost and $s^* = (s_0^*, s_1^*, ..., s_J^*)$ be the corresponding base-stock levels. The following algorithm is used to obtain c^* and s^*.

SET $c^* =$ large number

FOR $s_0 = 0$ to s_0^u

  SET $c \leftarrow h_0 E[I_0]$

  FOR $j = 1, ..., J$

    $s_j^*(s_0) \leftarrow \min\{s_j : Pr(B_{0j}(s_0) + D_j < s_j) \geq \beta_j\}$

    $c \leftarrow c + h_j E[s_j^*(s_0) - (B_{0j}(s_0) + D_j)]^+$

  END

  IF $c^* > c$ THEN $c^* \leftarrow c$; $s_0^* \leftarrow s_0$; $s_j^* \leftarrow s_j^*(s_0)$

END

The computational complexity of this algorithm is $O(J^2)$ for fixed retailer demand rates λ_j's. In the rest of the paper, we present and evaluate heuristics and approximations for the above service constrained distribution system.

4.3 Heuristics

4.3.1 Triple-Search Heuristic (TSH)

This heuristic considers three possible base-stock levels for the warehouse. Recall that once s_0 is set, the heuristic solves J single-location optimization problems as in (4.2).

The heuristic first sets $s_0 = 0$. Recall from the discussion after (4.5) that $B_{0j}(0) + D_j$ has Poisson distribution with mean $\lambda_j(L_0 + L_j)$. Hence, the resulting single-location optimization problems are of newsvendor type. The optimal retailer base-stock levels and corresponding costs can be computed using a simple spreadsheet. Note that this strategy stocks all inventory at the retailers and no inventory at the warehouse.

Next, the heuristic sets $s_0 = E[D_0]$. In other words, it allocates zero safety stock to the warehouse and simply keeps enough inventory to satisfy the expected total replenishment orders from all retailers.

Finally, the heuristic sets $s_0 = s_0^u$, as s_0^u is defined in (4.7). Recall from Proposition 2 that this base-stock level is the maximum quantity ever needed for the warehouse. In other words, the heuristic allocates maximum inventory to the warehouse and the minimum necessary inventory to the retailers to satisfy the fill-rate requirements.

The TSH considers the above three feasible ways to allocate inventory to the distribution system and chooses the one that yields the lowest cost. Gallego, Özer and Zipkin (2003) also propose a heuristic that considers three possible ways to allocate base-stock to the warehouse. The resulting optimization problems in their backorder cost model paper and in our service constrained model, e.g., (4.1) and (4.2), are related but different. The computational complexity of this heuristic is $O(J)$. Hence, it is computationally much faster than the optimization algorithm given in Section 4.2.

4.3.2 Newsvendor Heuristic (NVH)

Here we propose a heuristic that obtains base-stock levels for each of the J locations by solving newsvendor problems. To do so, we first *decompose* the distribution system into J two-location serial systems by restricting the warehouse to keep separate safety stock for each retailer. We let D_{0j} and s_{0j} denote demand and the base-stock level at the sub-warehouse in the resulting jth two-location serial system for which the optimization problem is given by

$$\begin{aligned} \min_{s_{0j}, s_j} \quad & h_0 E[s_{0j} - D_{0j}]^+ + h_j E[s_j - (D_j + (D_{0j} - s_{0j})^+)]^+ \\ \text{subject to} \quad & Pr((D_{0j} - s_{0j})^+ + D_j < s_j) \geq \beta_j. \end{aligned} \tag{4.8}$$

Next we define a backorder cost $b_j = \frac{\beta_j}{1-\beta_j} h_j$ for each retailer j. Using these backorder costs, we convert each of the two-location service-constrained serial systems to two-location serial systems with backorder costs at the retailers, i.e.,

$$\min_{s_{0j}, s_j} \quad h_0 E[s_{0j} - D_{0j}]^+ + h_j E[s_j - (D_j + (D_{0j} - s_{0j})^+)]^+ + b_j E[(D_j + (D_{0j} - s_{0j})^+ - s_j)]^+. \tag{4.9}$$

To obtain the *optimal* base-stock levels for the above serial system requires one to solve a recursive algorithm (Gallego and Zipkin 1999). In particular,

$$C_j(s) = E\{(h_j - h_0)(s - D_j)^+ + (b_j + h_0)(s - D_j)^-]\}, \tag{4.10}$$

$$s_j^e = \arg\min_s C_j(s),$$

$$C_{0j}(s) = E\{h_0(s - D_{0j})^+ + C_j(\min[s - D_{0j}, s_j^e])\}, \tag{4.11}$$

$$s_{0j}^{e*} = \arg\min_s C_{0j}(s).$$

Note that the *last* location's optimal base-stock level is obtained by solving the newsvendor problem in (4.10). Hence, the optimal base-stock level for the last location of the resulting two-location serial system is given by

$$s_j^e = \min\{s : Pr(D_j < s) \geq \frac{b_j + h_0}{b_j + h_j}\}. \tag{4.12}$$

However, to obtain the upstream location's optimal base-stock level requires one to solve the recursive equation in (4.11). Instead, we solve a newsvendor problem. To do so, instead of charging the holding cost rate h_0 to excess inventory in the first stage and adding the cost of the second stage C_j, we charge the approximate holding cost rate

$$h_{0j} = \frac{L_0}{L_0 + L_j} h_0 + \frac{L_j}{L_0 + L_j} h_j, \tag{4.13}$$

to excess inventory. The idea is based on adding the holding cost as the product goes through the warehouse and the retailer without delay and then dividing by the total leadtime that it spends before reaching the end customer. We charge this approximate holding cost to any excess inventory in the upstream echelon that faces demand uncertainty over the leadtime $L_0 + L_j$ and the penalty cost b_j. The resulting problem has a newsvendor-type cost structure of

$$\tilde{C}_{0j}(s) = \min_s E[h_{0j}(s - \tilde{D}_j)^+ + b_j(s - \tilde{D}_j)^-], \tag{4.14}$$

$$s_{0j}^e = \arg\min_s \tilde{C}_{0j}(s) \tag{4.15}$$

where $\tilde{D}_j$ is Poisson with rate $\lambda_j(L_0 + L_j)$. The optimal *echelon* base-stock level is given by

$$s_{0j}^e = \min\{s : Pr(\tilde{D}_j < s) \geq \frac{b_j}{b_j + h_{0j}}\}. \tag{4.16}$$

Substituting b_j and h_{0j} into (4.12) and (4.16), we obtain

$$s_j^e = \min\left\{s : Pr(D_j < s) \geq 1 - \frac{1-\beta_j}{1+\frac{h_0}{h_j-h_0}}\right\}, \quad (4.17)$$

$$s_{0j}^e = \min\left\{s : Pr(\tilde{D}_j < s) \geq \frac{1}{1+(\frac{1}{\beta_j}-1)(1-\frac{L_0}{L_j+L_0}\frac{1}{1+\frac{h_0}{h_j-h_0}})}\right\}. \quad (4.18)$$

Given these *echelon* base-stock levels[4], one can obtain the local base-stock level at the sub-warehouse j as $s_{0j} = \max(s_{0j}^e - s_j^e,\ 0)$. Similarly, the base-stock level at retailer j is $s_j^{NH} = \min(s_{0j}^e, s_j^e)$. Zipkin (2000, pg. 306) provides a discussion on the equivalence of these local and echelon base-stock levels for serial systems.

In summary, the newsvendor heuristic sets the warehouse base-stock level as $s_0^{NVH} = \sum_{j>0} s_{0j}$ and the retailer base-stock levels as s_j^{NVH} for $j > 0$. The following proposition shows that the retailer base-stock levels obtained by the newsvendor heuristic are no less than the lower bounds for the *optimal* base-stock levels s_j^* at each retailer.

PROPOSITION 3 *$s_j^{NVH} \geq s_j^l$ for $j > 0$.*

Note that the base-stock levels s_j^{NVH} are solutions of simple newsvendor problems. Hence, this heuristic for the *distribution* system can be implemented with a simple spreadsheet. The NVH is inspired by the heuristic developed for a *serial* system with a *backorder* cost by Gallego and Özer (2005).

[4]For a two-location serial system, the local base-stock level of the downstream location is the same as its echelon base-stock level. Hence, with abuse of notation, we use s_j^e to denote both base-stock levels.

4.4 Approximations

4.4.1 Newsvendor Approximation (NVA)

Here we approximate customer demand at each retailer with a Normal distribution and hence simplify the newsvendor heuristic. This approximation yields closed-form solutions for the resulting base-stock levels and the total average holding cost.

Recall that s_j^e is the optimal solution to the problem in (4.10). We approximate D_j with a normal distribution with mean $\lambda_j L_j$ and standard deviation $\sqrt{\lambda_j L_j}$.[5] Then the base-stock level at retailer j can be written as

$$s_j^{Ne} = \lambda_j L_j + \Phi^{-1}(\psi_j)\sqrt{\lambda_j L_j}, \quad (4.19)$$

$$\psi_j \equiv 1 - \frac{1-\beta_j}{1+\frac{h_0}{h_j - h_0}}. \quad (4.20)$$

Similarly, recall that s_{0j}^e is the solution to the problem in (4.14). We also approximate $\tilde{D}_j$ with a normal distribution with mean $\lambda_j(L_0 + L_j)$ and standard deviation $\sqrt{\lambda_j(L_0 + L_j)}$. Then the echelon base-stock level at sub-warehouse j is

$$s_{0j}^{Ne} = \lambda_j(L_0 + L_j) + \Phi^{-1}(\eta_j)\sqrt{\lambda_j(L_0 + L_j)}, \quad (4.21)$$

$$\eta_j \equiv \frac{1}{1 + (\frac{1}{\beta_j} - 1)(1 - \frac{L_0}{L_j + L_0}\frac{1}{1+\frac{h_0}{h_j - h_0}})}. \quad (4.22)$$

The resulting total average inventory holding cost for the distribution system is

$$c = \sum_{j>0}\left[\frac{1}{1-\beta_j}h_j - \frac{L_0}{L_0 + L_j}(h_j - h_0)\right]\phi(z_j)\sqrt{\lambda_j(L_0 + L_j)}, \quad (4.23)$$

where $z_j = \Phi^{-1}(\eta_j)$.

The approximate base-stock level at the warehouse is $s_0^{NVA} = \sum_{j>0} \max(s_{0j}^{Ne} -$

[5] A Poisson distribution with parameter λ is approximately normal for large λ. The approximating normal distribution has mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$. It is typically the case that such approximations are less accurate in the tails of the distribution.

s_j^{Ne}, 0). The base-stock level at retailer j is $s_j^{NVA} = \min(s_{0j}^{Ne}, s_j^{Ne})$.

Next we illustrate how to use such a closed-form approximation to gain transparent insights on key factors affecting distribution system design and stock positioning. In the following discussion, we assume that the parameters satisfy $\frac{b_j}{b_j+h_{0j}} = \eta_j > 0.5$ for $j > 0$. In other words, the resulting safety stocks at the sub-warehouses are always nonnegative.

Effect of Cost Parameters Consider the role of holding costs in stock positioning. Note from (4.19) and (4.20) that as $\frac{h_0}{h_j}$ decreases, ψ_j decreases and the base-stock level at retailer j decreases. Similarly, note from (4.21) and (4.22) that as $\frac{h_0}{h_j}$ decreases, η_j increases, the echelon base-stock level, and hence the stocking amount at the warehouse increases. As $h_0 \to 0$, $s_j \to s_j^l$. In this case, large amounts of inventory can be carried at the warehouse and hence retailer j only needs to protect itself against demand uncertainty within the L_j time units, as also shown in Part 2 of Proposition 1. As $\frac{h_0}{h_j}$ decreases, η_j increases, when $\eta_j > 0.50$ by assumption, z_j increases and $\phi(z_j)$ decreases, hence c decreases as can be seen from (4.23). This observation is consistent with our intuition that delaying the value-added operations from the warehouse to the retailers reduces the total average holding cost.

Note from (4.17) and (4.18) that when *none* of the retailers add any value to the final product, that is, $h_j = h_0$ for $j > 0$, then for the newsvendor heuristic, we have $s_j^e = \infty, s_{0j}^e = \min\{s : Pr(\tilde{D}_j < s) \geq \beta_j\} = s_j^u$. $s_{0j} = \max\{s_{0j}^e - s_j^e, 0\} = 0$, and hence $s_0^{NVH} = 0$. $s_j^{NVH} = \min\{s_{0j}^e, s_j^e\} = \min\{s_j^u, \infty\} = s_j^u$. Similarly, for the newsvendor approximation, we obtain $s_0^{NVA} = 0$, $s_j^{NVA} = s_j^u$.

Effect of Lead Times Consider the effect of leadtimes on stock positioning and the total average cost. The base-stock level at retailer j increases with its own leadtime L_j but is independent of the warehouse leadtime L_0. Fixing L_0, the system stocking amount also increases in $L_0 + L_j$. Keeping the total leadtimes from the outside supplier to the retailers $L_0 + L_j$ fixed, the base-stock level at retailer j decreases with the warehouse leadtime L_0. On the other hand, the warehouse stock level increases in L_0. When $\eta_j > 0.5$, the overall system cost in (4.23) decreases in L_0, which corresponds to the benefits of logistic postponement. A careful look at the approximate cost also suggests that locating the warehouse closer to a retailer with a larger demand rate achieves greater system cost reduction (because the system cost in (4.23) is proportional to the square root of $\lambda_j(L_0 + L_j)$).

Effect of Service Levels Finally, (4.19) and (4.21) show that both the retailer base-stock levels and the system stock level increase in the required service level β_j. The warehouse base-stock level can be increasing or decreasing in β_j. However, when $\eta_j > 0.5$, the average cost is always increasing with β_j.

The above closed-form solutions reveal in a transparent way how stock positioning in a distribution system is affected by (i) the ratio of the value-added operations, i.e., the echelon holding costs at the warehouse and at the retailers, $\frac{h_0}{h_j}$, (ii) the warehouse and retailer leadtimes, and (iii) the target fill-rates β_j. Note also that the impact of changes in system parameters can be quantified using these closed-form approximations. For example, leadtimes can be reduced by a new processing or tracking technology. The approximation enables one to carry out back of the envelope analysis to address the impact of a new technology on the distribution system cost and stocks. Through a numerical study, we will illustrate later that such analysis yields essentially the same results as one would obtain by using the optimization algorithm in Section 4.2.

4.4.2 Normal Approximation (NA)

Here we apply a normal approximation directly to the original problem in (4.1). In particular, we approximate the leadtime demand D_j by a normal distribution with mean $\lambda_j L_j$ and variance $\lambda_j L_j$. We let ϕ denote the standard normal density function, Φ the standard normal cumulative distribution function, Φ^0 the standard normal complementary cumulative distribution function, Φ^1 the standard normal loss function, and Φ^2 the standard normal second-order loss function. The normal approximation at the warehouse yields

$$\begin{aligned} E[B_0] &= \Phi^1(z_0)\sqrt{\lambda_0 L_0}, && (4.24) \\ E[B_0^2] &= 2\Phi^2(z_0)\lambda_0 L_0 + E[B_0], && (4.25) \\ E[I_0] &= \Phi^1(-z_0)\sqrt{\lambda_0 L_0}, \end{aligned}$$

where $z_0 = (s_0 - \lambda_0 L_0)/\sqrt{\lambda_0 L_0}$.

Recall that $\theta_j = \frac{\lambda_j}{\lambda_0}$. For each retailer j,

$$\begin{aligned} E[B_{0j}] &= \theta_j E[B_0], \\ V[B_{0j}] &= \theta_j(1-\theta_j)E[B_0] + (\theta_j)^2 V[B_0], \end{aligned}$$

where $E[B_0]$ and $V[B_0]$ can be obtained from (4.24) and (4.25). Then $B_{0j} + D_j$, the effective demand at retailer j, can be approximated by a normal distribution with mean and variance

$$\begin{aligned} \hat{\mu}_j &= E[B_{0j}] + \lambda_j L_j, \\ \hat{\sigma}_j^2 &= V[B_{0j}] + \lambda_j L_j, \end{aligned}$$

because B_{0j} and D_j are independent. Therefore, the normal approximation for retailer j yields

$$E[I_j] = \Phi^1(-z_j)\hat{\sigma}_j = z_j\sigma_j + \sigma_j \int_{z=z_j}^{\infty} (z - z_j)\phi(z)dz,$$

where z_j solves $\Phi(z_j) = \beta_j$. Given s_0, the base-stock level at retailer j can be approximated by

$$s_j(s_0) = \hat{\mu}_j + z_j\hat{\sigma}_j.$$

The total average inventory holding cost reduces to a function of s_0 only, which can be written as

$$c(s_0) = h_0 \Phi^1(-z_0)\sqrt{\lambda_0 L_0} + \sum_{j>0} h_j \Phi^1(-z_j)\hat{\sigma}_j.$$

We optimize this function numerically. In our numerical study, after obtaining the optimal base-stock levels, we truncate up the fractional values. Note that this approximation, though easy, requires one to optimize over s_0 to allocate inventory among the warehouse and the retailers. On the other hand, the newsvendor approximation does not require one to carry out any optimization at the retailers.

4.4.3 Distribution-Free Approximation (DFA)

For a distribution system with service-level constraints, we set $b_j = \frac{\beta_j}{1-\beta_j} h_j$ for $j > 0$ and $b_0 = \sum_{j=1}^{J} \theta_j b_j$. We regard the warehouse as a single location with Poisson demand of rate $\lambda_0 L_0$, unit overage cost h_0, and unit underage cost b_0; each retailer j as a single location with Poisson demand of rate $\lambda_j L_j$, unit overage cost h_j, and unit underage cost b_j. Applying a distribution-free bound (Gallego and Moon 1993) to these single location problems, we obtain closed-form expressions for the base-stock levels at the warehouse and the retailers as follows.

$$\begin{aligned} s_j &= \lambda_j L_j + \frac{1}{2}\sqrt{\lambda_j L_j}\left(\sqrt{\frac{\beta_j}{1-\beta_j}} - \sqrt{\frac{1}{\beta_j} - 1}\right) \quad \text{for } j > 0\ , \\ s_0 &= \lambda_0 L_0 + \frac{1}{2}\sqrt{\lambda_0 L_0}\left(\sqrt{\frac{b_0}{h_0}} - \sqrt{\frac{h_0}{b_0}}\right). \end{aligned}$$

Furthermore, we obtain a closed-form solution for the total average holding cost as

$$c^* = \sqrt{\lambda_0 L_0}\sqrt{h_0 \sum_{j=1}^{J} \theta_j h_j \frac{\beta_j}{1-\beta_j}} + \sum_{j>0} h_j \sqrt{\lambda_j L_j}\sqrt{\frac{\beta_j}{1-\beta_j}}.$$

Note that none of the above equations require one to estimate demand distributions.

4.5 Numerical Study

This section reports the performance of our heuristics and approximations. We also provide some insights into stock positioning issues for a distribution system with service-level constraints. First, we compare the optimal solution to our heuristics and report the percentage error $\varepsilon_i = \frac{c_i - c*}{c*}$ for $i = \{\text{TSH, NVH}\}$ for 530 problem instances. The parameters are chosen to reflect a wide range of situations such as short and long leadtimes, cheap and expensive holding costs, low and high customer-service levels.

We study two sets of experiments with local holding cost $h_j = 1$ for $j > 0$. In the first set of experiments the retailers are identical. This set includes two subsets, for both we consider $L_0 \in \{0.10, 0.90\}$ with $L_j = 1 - L_0$ for $j > 0$ to address systems with different degrees of risk pooling at the warehouse, and $h_0 \in \{0.3, 0.9\}$, which corresponds to adding 30% and 90% of the value to the item at the warehouse. First, we test 80 instances with $J \in \{2, 4, 8, 16, 32\}$; $\lambda_0 = \{16, 64\}$; $\beta_j \in \{90\%, 97.5\%\}$. Second, we address the impact of different service levels on the optimal solutions and the heuristics. To do so, we consider 40[6] additional experiments with $\beta_j \in \{20\%, 30\%, 50\%, 90\%, 95\%, 99.9\%\}$, $J \in \{2, 4\}$; $\lambda_0 = \{16\}$.

In the second set of experiments the retailers are non-identical. We consider $L_0 \in \{0.10, 0.25\}$, $h_0 \in \{0.3, 0.9\}$, $J \in \{2, 4, 8, 16, 32\}$, $\lambda_0 \in \{16, 64\}$. There are 40 possible combinations. For each combination we generate 10 instances by generating randomly L_j and β_j according to independent uniform distributions. The ranges for these two parameters are $L_j \in [0.10, 0.25]$ and $\beta_j \in [90\%, 97.5\%]$. We also test the heuristics for $h_0 = h_j = 1.0$. Table 4.4 also includes 10 non-identical retailer experiments for $h_0 = h_j = 1.0$ cases with $L_0 = 0.10$ and L_j and β_j randomly generated from the uniform distributions described above. We study 410 instances of non-identical retailer case for each heuristic. Note that unequal leadtimes have the same effect as unequal demand rates at each retailer. Similarly, the relative importance of holding costs at each retailer is different because of the randomly generated service

[6]Note that 8 experiments with $\beta_j = 90\%$ have been included in the first subgroup.

level requirements.

In total, we have tested each heuristic for 530 experiments with all possible parameters that are critical for a distribution system.

4.5.1 Performance of Heuristics

Table 4.1 reports the performance of the triple search (TSH) and the newsvendor (NVH) heuristics for the first subgroup of identical retailer experiments. Table 4.2 reports the performance of TSH and NVH for the second subgroup of identical retailer experiments when changing β_j. Table 4.3 summarizes the performance for the second set of non-identical retailer cases. A close examination of the averages shown in this table reveals the environment for which each heuristic performs better. Table 4.4 provides specific experiment instances for non-identical retailer cases with $J = 2$. Table 4.5 provides experiments with non-identical retailers with $J = 4$.

For the **TSH**, the average error over 530 experiments is 1.18% and the standard deviation is 4.87%. The TSH provides *optimal* solutions for 378 experiments. The performance of the triple search heuristic is better under shorter leadtime and lower holding costs at the warehouse. The average error term for the TSH is 0.37% for $L_0 = 0.1$ among 270 experiments and 1.02% for $L_0 = 0.25$ among 200 experiments. Keeping everything else equal, either increasing the number of retailers J or decreasing the demand rate λ_0 reduces the error terms and the error deviation. In addition, the error terms decrease in the retailer service levels. When $\beta_j = 99.9\%$, the TSH always yields the optimal base-stock levels, and the worst case occurs when $\beta_j = 20\%, J = 2, h_0 = 0.90, L_0 = 0.90$. On the other hand, with $L_0 = 0.10$, the TSH obtains the optimal base-stock levels for all the 20 experiments when β_j changes from 20% to 99.9%. Overall, the TSH yields solutions very close to the optimal solutions.

For the **NVH**, the average error term and the standard deviation are 18.16% and 17.76%, respectively. The NVH provides close-to-optimal solutions when the warehouse leadtimes are small. When $L_0 = 0.1$, the average error term over 270 experiments is 9.85%. However, the error gap increases when the warehouse leadtimes are large. This increase is because the distribution system gains significantly from risk pooling. However, by assuming that the warehouse keeps stocks separately for each retailer, the NVH disregards the risk pooling effect. In terms of the impact of retailer service levels, based on the 40 experiments, there is no obvious pattern on how the error terms changes with respect to the retailer service levels for the NVH. We can however observe that the error terms are the smallest when $\beta_j = 99.9\%$. Note also that the NVH does not necessarily guarantee that fill-rate requirements are satisfied. However, we encountered only one infeasible case out of the 530 experiments.

Computational Complexity. The optimization algorithm for obtaining optimal base-stock levels and the backorder-cost heuristic take the longest time (less than 60 minutes) followed by the triple search heuristic (less than 1 minute) using a Pentium IV processor. The newsvendor heuristic requires the shortest time (less than a few seconds).

4.5.2 Performance of Approximations

We also investigate how well the approximations predict the performance of the system dynamics. To do so, we study whether changing the system parameters under these approximations has the same effect on the system cost as under the optimal base-stock policy. We use regression to compare the approximation costs to the actual optimal costs. For the purpose of comparison we vary one parameter at a time and keep all the other parameters fixed. For example, to see how well the Newsvendor Approximation (NVA) predicts the influence of relative magnitude of warehouse leadtimes L_0 on the total average cost, we fix $L_0 + L_j = 1.0$, $\lambda_0 = 16, J = 2, \beta_j = 90\%, h_0 = 0.50$ and change only L_0. Next, we calculate the coefficient of determination R^2 of the newsvendor approximation costs and the optimal costs.

In Table 4.6 we report the coefficients of determination R^2 for the Newsvendor Approximation(NVA), the Normal Approximation(NA), and the Distribution-Free

Approximation (DFA). We observe that the resulting R^2s are all close to 1. The performance of NVA is slightly better than NA, which is better than DFA. Hence, our approximate solutions are consistent with the optimal solutions in the sense that approximate solutions change in the same direction and relative magnitude as the optimal solutions when changing a system parameter.

4.5.3 System Design Issues

The simplicity and the consistency of the closed-form approximations allow one to evaluate different system design strategies without solving an optimization problem or applying an optimization algorithm. In this section, we illustrate how a distribution system manager can apply these closed-form approximations to evaluate system design strategies. All the three approximations yield similar results. For simplicity of computation in the following discussion we focus only on the newsvendor approximation.

Demand Aggregation

Suppose a distribution system manager is able to aggregate demand from several retailers to one retailer. She is interested in evaluating how the aggregation reduces the total system inventory costs. Consider a distribution system with eight retailers and the following system parameters: $h_0 = 0.5, h_j = 1, L_0 = 0.5$ and $L_j = 1$ for $j > 0$. This distribution system's approximation inventory holding cost is 15.92 (resp., the optimal cost is 16.41). When the demand can be aggregated to be satisfied from two retailers, the approximation system inventory holding cost reduces to 7.96 (resp., the optimal cost reduces to 8.02). The percentage decrease in the approximation inventory cost due to demand aggregation is 50% (resp., 51.13%).

This analysis can be easily extended to carry out sensitivity analysis on the benefits of demand aggregation as in Figure 4.1.

We observe that the relative benefits (percentage change) of stage consolidation

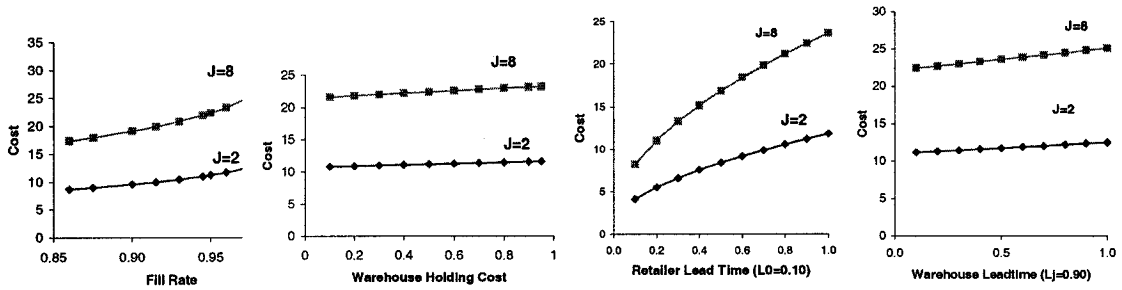

Figure 4.1: Impact of Demand Aggregation.

is robust with respect to the changes in system parameters β_j, h_0, L_0 and L_j. For example, when $\beta_j = 0.90$, the total average inventory costs for J=8 and J=2 are 19.10 and 9.55 respectively, which is a 50% cost reduction. When $\beta_j = 0.99$, the total average inventory costs for J=8 and J=2 are 28.83 and 14.42 respectively, which is also a 50% cost reduction. This observation is consistent with the cost function in (4.23). Keeping other system parameters fixed, $c(J)$ is in proportion to $\sqrt{J}$. Hence, $\frac{c(J=2)}{c(J=8)} = \sqrt{2/8} = 0.50$.

Note that the manager can quantify the cost benefits of demand aggregation fairly accurately by using the closed-form newsvendor approximation instead of having to solve the optimization algorithm in Section 4.2 or knowing anything about convolutions or Poisson distribution. The only advance knowledge required to obtain these results is the concept of "bell-shaped" demand distribution (Normal distribution), which is taught in any level operations courses.

Logistic Postponement

A manager can also estimate the cost benefits of a logistic postponement strategy using the newsvendor approximation. Consider, for example, a distribution system with $J = 2, \lambda_0 = 16, h_0 = 0.50, h_j = 1.0$. What is the inventory cost reduction (percentage change) due to changing $L_0 = 0.1$ to $L_0 = 0.9$ while keeping $L_0 + L_j = 1$, i.e., locating the warehouse much closer to the retailers? Approximately, the average inventory holding cost reduces from 22.39 to 14.27, which is a 36.27% cost reduction. This analysis can be extended to study the sensitivity of the benefits of logistic postponement to other system parameters, as illustrated in Figure 4.2.

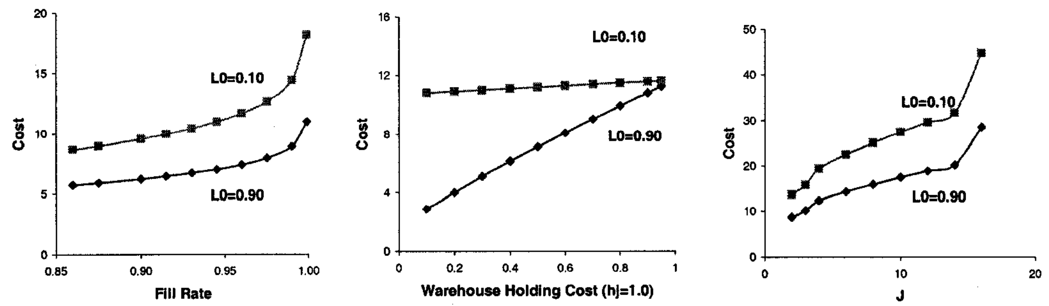

Figure 4.2: Impact of Logistics Postponement.

We observe that the cost benefit of logistic postponement increases in the fill rates. In particular, the percentage cost reduction due to logistic postponement across different fill-rate requirements ranges from 33.71% for $\beta_j = 0.86$ to 39.57% for $\beta_j = 0.999$. This observation suggests that when the retailers have different service-level requirements, it is more cost effective to locate the warehouse closer to the retailers with higher service-level requirements. We test this conjecture through a stylized numerical example. Consider a distribution system with two retailers. Retailer 1 has a fill rate of 95% and retailer 2 has a fill-rate of 99%. Before the postponement strategy is implemented, $L_0 = 0.10$ and $L_1 = L_2 = 0.90$. The total inventory-related cost is 12.80. Two options exist to implement logistic postponement. One is to locate the warehouse closer to retailer 1, in which case $L_1 = 0.10$ and $L_2 = 0.40$, which results in a system cost of 9.48, a 25.96% cost reduction. The other option is to locate the warehouse closer to retailer 2 in which case $L_1 = 0.40$ and $L_2 = 0.10$, which results in a system cost of 6.41, a 49.97% cost reduction.

As for the warehouse holding cost, we observe that the cost benefits of logistic postponement decreases in h_0 while keeping $h_j = 1.0$. For $h_0 = 0.10$, the inventory cost reduction due to postponement is 73.47% while for $h_0 = 0.95$ the cost reduction is

3.27%. This result suggests that logistic postponement is most cost effective when the relative value added at the warehouse (the upstream stage) is low. Finally, consistent with our observation from Subsection 4.5.3, the relative benefit (percentage change) of logistic postponement is insensitive to the change of number of retailers. The percentage change remains 36.27% across different J values.

Delaying Value-Added Operations to Retailers

A manager can shift value-added activities from the warehouse to the retailers through process redesign or re-engineering. Next, we quantify the benefits of such strategies. To do so, we keep $h_j = 1.0$ and change h_0 from 0.90 to 0.50. We plot the results in Figure 4.3.

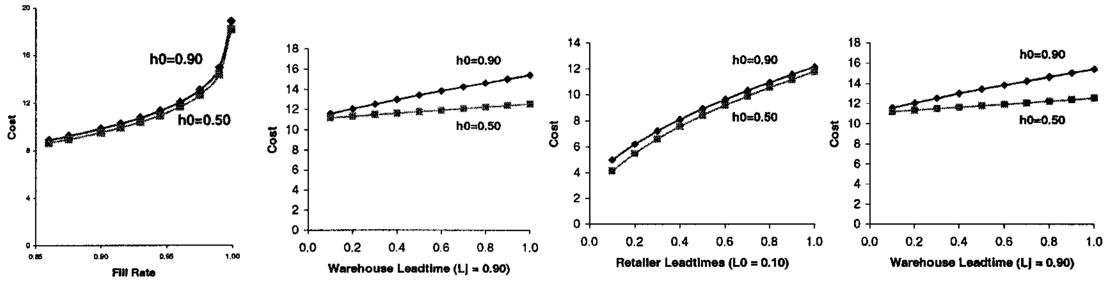

Figure 4.3: Impact of Delaying Value-Added Operations to Retailers.

We observe that as the fill rates increase, the cost reduction due to delaying value-added operations is increasing from 2.95% for $\beta_j = 0.86$ to 3.72% for $\beta_j = 0.99$. This result indicates that it is more cost effective to delay the value-added operations to the retailers with higher fill-rate requirements. The cost benefits are insensitive to the change of J. For instance, the cost reduction is 3.28% for both $J = 2$ and $J = 32$.

Now consider the impact of warehouse leadtime and retailer leadtimes on the strategy of delaying value-added operations to the retailers. We first keep $L_j = 0.9$ and let L_0 change from 0.1 to 1.0. The cost reduction (percentage change) increases from 3.28% to 38.53%. Next, we fix $L_0 = 0.1$ and change L_j from 0.1 to 1.0. The cost reduction (percentage change) decreases from 17.56% to 2.98%.

Finally consider the impact of logistic postponement on the strategy of delaying value-added operations to retailers. To do so, we fix $L_0 + L_j = 1.0$ and change L_0 from 0.10 to 1.0. Our numerical tests indicate that significant cost benefits of delaying value-added operations to the retailers can be achieved when combined with the logistic postponement strategy. For instance, when $L_0 = 0.10$, the approximate system inventory holding cost is approximately 11.57 for $h_0 = 0.90$ and 11.19 for $h_0 = 0.50$. Hence, delaying value-added operations to the downstream stages reduces cost by 3.28%. On the other hand, when $L_0 = 1.0$, the cost reduction is 38.53%, *i.e.*, $(10.71 - 6.59)/6.59$.

Increasing Service Levels at the Retailers

Here we investigate how the retailer service-level requirements affect the system inventory cost. To do so, we carry out some sensitivity analysis on cost changes when one increases the customer fill-rates from 95% to 99%. We plot the results in Figure 4.4. First, consistent with our intuition, we observe that the system costs increase

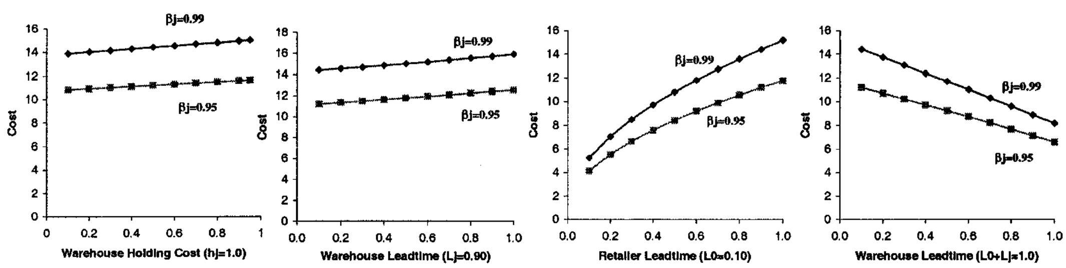

Figure 4.4: Impact of Fill-Rate Requirements.

in the retailer service-level requirements. Second, we observe that the percentage increases are relatively insensitive to all the other system parameters. For example, the percentage increases vary from 28.42% for $h_0 = 0.1$ to 29.17% for $h_0 = 0.95$; the percentage increase stays at 28.8% when we change J from 2 to 32; and fixing $L_j = 0.90$ and changing L_0 from 0.10 to 1.0, the percentage increase is decreasing from 28.78% to 26.77%. Note that the analysis is based on the cost ratio, not on the absolute cost value. The absolute values might be different.

Comparing System Design Strategies

Finally we illustrate how one can apply the Newsvendor Approximation to compare system design alternatives under budget constraints. Consider a scenario when a manager needs to increase the customer fill rates from 95% to 99.5% for a distribution system with $J = 8$, $\lambda_0 = 16$, $h_0 = 0.5$, $h_j = 1.0$, $\beta_j = 95\%$, $L_0 = 0.1$ and $L_j = 0.90$. The manager would like to keep the total holding cost at least the same while increasing the fill rates. She has an option of locating the warehouse closer to the retailers while keeping $L_0 + L_j = 1.0$. The Newsvendor Approximation suggests a postponement strategy with $L_0 = 0.69$ and $L_j = 0.31$ to achieve 99.5% fill-rates at the retailers without incurring additional holding cost, i.e., the retailer leadtimes L_j need to be reduced from 0.90 to at least 0.31, or equivalently, L_0 needs to be increased from 0.10 to 0.69.

4.6 Connection with the Backorder-Cost Model

A service-constrained model can be equivalently written as a backorder-cost model only for a single location problem with one-for-one replenishment. To be more exact, one can obtain the optimal base-stock level for the service-constrained model from a backorder-cost model by setting the penalty cost $b = \frac{\beta}{1-\beta}h$, where h is the unit holding cost [7] (Zipkin 2000). However, Boyaci and Gallego (2001) conclude that a similar conversion does not yield a one-to-one relationship for a multi-echelon serial system. Van Houtum and Zijm (2000) present an overview of possible relations between a backorder-cost model and a service-constrained model.

Similarly, an alternative solution methodology to obtain the base-stock levels for the service-constrained distribution system is to convert the problem to one with

[7] For discrete demand distributions, the optimal base-stock level for the service-constrained model is the optimal base stock level of the backorder-cost model plus 1.

backorder costs at each retailer, henceforth the backorder-cost (BC) method, and develop solution methodologies for the resulting backorder cost case. In particular, one can set the backorder cost b_j at retailer j as

$$b_j = \frac{\beta_j}{1-\beta_j} h_j \tag{4.26}$$

and convert the service-constrained problem into a distribution system with backorder costs at the retailers. The exact solutions discussed, for example in Axsäter (1990), can be used to obtain the optimal base-stock levels.

Next we investigate the connection between the service-constrained model and the backorder-cost method for a distribution system. Our numerical study indicates that the backorder-cost model leads to insufficient fill rates at the retailers. However, for a slightly different type of service-level constraint - limit probability of *nonnegative* inventory (PONI) [8], that is, $Pr(B_{0j}(s_0) + D_j \leq s_j)$, the backorder-cost method can guarantee that the service-level requirements at the retailers are satisfied. We refer the optimization problem in (4.1) that uses the PONI type service-level constraints as the PONI model. The algorithm, heuristics and approximations discussed throughout this chapter can be also applied to the PONI model. The following proposition characterizes the relationship between the PONI model and the backorder-cost model.

PROPOSITION 4 *The base-stock levels obtained by the backorder-cost method are feasible for the PONI model. The total average inventory costs calculated from the backorder-cost model are strictly larger than the optimal total average costs obtained by the PONI model.*

This proposition proves that the backorder-cost method results in (1) feasible base-stock levels for the PONI model and (2) strictly higher total system costs for the PONI model. In other words, using the backorder-cost method never yields an

[8]Boyaci and Gallego (2001) also uses a PONI service-level constraint to study the relationship between a service-constrained model and a backorder-cost model in a serial system and concludes that the backorder-cost heuristic incurs significant costs.

optimal solution for the PONI model. Note that Boyaci and Gallego (2001) illustrate by numerical experiments that using the backorder-cost model can be costly for a PONI-constrained serial system. Here, we prove analytically that a backorder-cost method is costly for a *distribution* system.

A better understanding of the relationship between the service-constrained models and the backorder-cost model requires a numerical study. Table 4.7 reports the performance of the backorder-cost method for both the fill-rate constrained model and the PONI model for 400 non-identical retailer experiments outlined in Section 4.5. In most cases, the backorder-cost method leads to insufficient fill rates for the fill-rate constrained model, and incurs significantly high costs for the PONI constrained model. This result also supports the need for developing methodologies for inventory management in service-constrained distribution systems.

4.7 Discussion

A distribution system with service-level constraints at each retailer is a fundamental structure for many supply chains. We provide an algorithm that optimally determines base-stock levels at each location. Often distribution systems can be very large in practice. Hence, computational improvements can be very important. We establish monotonicity results and bounds on the base-stock levels that help improve the computational requirement for the optimization algorithm. These bounds also help us to develop two heuristics and three approximations.

A heuristic or an approximation is appealing if it passes all or some of the following tests: (1) Is it close to optimal? (2) Is it simple to describe and use? (3) Can it be used to test system design issues accurately? (4) Is it robust? A similar discussion can be found in Zipkin (2000, pg 205). Focusing narrowly on the first criterion leads to a gap between practice and theory. However, realizing this gap, researchers have started to develop easy-to-use and robust heuristics and approximations that are insightful (see, for example, Lee, Billington and Carter 1993, Gallego, Özer and Zipkin 2003, Cohen, Deshpande and Wang 2003, and references therein). Next we clarify separately the contribution of each of our heuristics and approximations along these dimensions and their contribution as a whole.

The TSH is the most accurate heuristic with an average error term around 1%. This result suggests that considering three important stock levels at the warehouse yields close-to-optimal solutions. The longest computational time required to solve a problem instance considered here was less than one minute (with a Pentium IV processor). Hence, the TSH is amenable for application. The NVH is computationally the cheapest heuristic because it is based on solving $2J$ number of newsvendor problems. This heuristic yields an average error term of 18%, which is not as small as that of the TSH. However, it is computationally much faster than the TSH. The longest time required to solve a problem instance was less than a few seconds. It also passes tests 2, 3, and 4 very well. For example, the only advanced knowledge required to use this heuristic is the newsvendor model, which is taught in any operations management course. This heuristic also provides a step towards establishing the newsvendor approximation which is in closed-form. Finally, we develop two additional closed-form approximations that require less data and reveal important relationships between stock-positioning and fill-rate requirements. These approximations are robust and simple to describe. They can be used to address system design issues. Note that the distribution-free approximation does not require one to estimate demand distributions, which are often very difficult to obtain for a system without demand history. Hence, all the three approximations pass tests 2, 3 and 4. These approximations also help us to understand the drivers of system dynamics in a transparent way.

Given the above discussion, the heuristics and approximations presented in this chapter contribute to both research and practice in various ways. Some provide very efficient, computational means to obtain close-to-optimal solutions. Others are robust, easy-to-use and can provide insights into system design issues. Given the computational requirements for optimal algorithms, close-to-optimal and fast heuristics plus simple approximations can enable better management of large-scale systems. Hence, the results of this chapter also provide a step towards bridging the gap between theory and practice. A large-scale system may look like an arborescent system with more than two levels. Such systems can be decomposed into two-level distribution systems by judiciously allocating decoupling inventories. Next the decoupled distribution systems can be addressed using the methodologies developed in this chapter. However, it is impossible to obtain the optimality gap (or the error term) of an allocation strategy for such an arborescent system because the optimal base-stock allocation is unknown as of today. The system manager can perhaps compare the cost of any base-stock allocation strategy to the existing system's performance.

4.8 Appendix A: Proofs

Proof of Lemma 1. We first show that $Pr(B_{0j}(s_0) < s_j) \leq Pr(B_{0j}(s_0+1) < s_j)$.

$$\begin{aligned}
Pr(B_{0j}(s_0+1) < s_j) &= E_{D_0}[Pr(B_{0j}(s_0+1) < s_j) \mid [D_0 - (s_0+1)]^+] \\
&= Pr(D_0 \leq s_0) + Pr(D_0 = s_0+1) \\
&\quad + E_{D_0}[Pr(B_{0j}(s_0+1) < s_j) \mid D_0 > s_0+1] \\
&= Pr(D_0 \leq s_0) + Pr(D_0 = s_0+1) \\
&\quad + \sum_{i=1}^{\infty} Pr(\sum_{j=1}^{i} X_j < s_j) Pr(D_0 = s_0+1+i) \\
&\geq Pr(D_0 \leq s_0) + Pr(D_0 = s_0+1) Pr(X_1 < s_j) \\
&\quad + \sum_{i=1}^{\infty} Pr(\sum_{j=1}^{i+1} X_j < s_j) Pr(D_0 = s_0+1+i) \\
&= Pr(D_0 \leq s_0) + Pr(D_0 = s_0+1) Pr(X_1 < s_j) \\
&\quad + \sum_{m=2}^{\infty} Pr(\sum_{j=1}^{m} X_j < s_j) Pr(D_0 = s_0+m) \\
&= E_{D_0}[Pr(B_{0j}(s_0) < s_j) \mid [D_0 - s_0]^+] \\
&= Pr(B_{0j}(s_0) < s_j),
\end{aligned}$$

where X_j represents an independent Bernolli random variable with parameter θ_j. The first and second equalities are by definition of a conditional expectation. The third equality results from the relation between a binomial random variable and a Bernoulli random variable. The inequality follows because $Pr(X_1 < s_j) \leq 1$ and $Pr(\sum_{j=1}^{i} X_j < s_j) \geq Pr(\sum_{j=1}^{i+1} X_j < s_j)$. Given this result, we write

$$\begin{aligned} Pr(B_{0j}(s_0) + D_j < s_j) &= E_{D_j}[Pr(B_{0j}(s_0) + D_j < s_j) \mid D_j] \\ &= \sum_{i=0}^{s_j} Pr(B_{0j}(s_0) < s_j - i)Pr(D_j = i) \\ &\leq \sum_{i=0}^{s_j} Pr(B_{0j}(s_0+1) < s_j - i)Pr(D_j = i) \\ &= Pr(B_{0j}(s_0+1) + D_j < s_j), \end{aligned}$$

the inequality follows from the above. ■

Proof of Lemma 2. We first show that $Pr(B_{0j}(s_0) < s_j + 1) \geq Pr(B_{0j}(s_0+1) < s_j)$. Let $B' \equiv B_0(s_0)$, $B \equiv B_0(s_0+1)$, and X_j represents an independent Bernolli random variable with parameter θ_j. We have

$$\begin{aligned} Pr(B_{0j}(s_0) < s_j + 1) &= E_{B'} Pr(s_j + 1 - \sum_{j=1}^{B'} X_j > 0 \mid B') \\ &= Pr(B' = 0) + Pr(B' = 1) + \sum_{i=2}^{\infty} Pr(s_j + 1 - \sum_{j=1}^{i} X_j > 0) Pr(B' = i) \\ &= Pr(B = 0) + \sum_{i=2}^{\infty} Pr(s_j - \sum_{j=1}^{i-1} X_j + 1 - X_i > 0) Pr(B' = i) \\ &\geq Pr(B = 0) + \sum_{i=2}^{\infty} Pr(s_j - \sum_{j=1}^{i-1} X_j > 0) Pr(B' = i) \end{aligned}$$

$$\begin{aligned} &= Pr(B = 0) + \sum_{i=2}^{\infty} Pr(s_j - \sum_{j=1}^{i-1} X_j > 0) Pr(B = i - 1) \\ &= \sum_{m=0}^{\infty} Pr(s_j - \sum_{j=1}^{m} X_j > 0) Pr(B = m) \\ &= Pr(B_{0j}(s_0 + 1) < s_j). \end{aligned}$$

The first equality follows because $B_{0j} \mid B'$ has a binomial distribution. The second equality is by definition of a conditional expectation. The third equality is from

$Pr(B=0) = Pr(D_0 - s_0 - 1 < 0) + Pr(D_0 - s_0 - 1 = 0) = Pr(B' = 0) + Pr(B' = 1)$. The inequality is by the definition of Bernoulli random variables. The fourth equality is because for $i \geq 2$, $Pr(B' = i) = Pr([D_0 - s_0]^+ = i) = Pr(D_0 = s_0 + i) = Pr(D_0 - s_0 - 1 = i - 1) = Pr([D_0 - s_0 - 1]^+ = i - 1) = Pr(B = i - 1)$. Given this result, we write

$$\begin{aligned}
Pr(B_{0j}(s_0) + D_j < s_j + 1) &= E_{D_j}[Pr(B_{0j}(s_0) + D_j < s_j + 1) \mid D_j] \\
&= \sum_{i=0}^{s_j} Pr(B_{0j}(s_0) < s_j + 1 - i) Pr(D_j = i) \\
&\geq \sum_{i=0}^{s_j} Pr(B_{0j}(s_0 + 1) < s_j - i) Pr(D_j = i) \\
&= Pr(B_{0j}(s_0 + 1) + D_j < s_j).
\end{aligned}$$

The inequality follows from the above. ■

Proof of Proposition 1. For a particular j and base-stock levels s_j, let $s_0^{-j} = \min\{s_0 : Pr(B_{0i}(s_0) + D_i < s_i) \geq \beta_i, \forall i \neq j\}$, $s_{0j} = \min\{s_0 : Pr(B_{0j}(s_0) + D_j < s_j) \geq \beta_j\}$, and $s'_{0j} = \min\{s_0 : Pr(B_{0j}(s_0) + D_j < s_j + 1) \geq \beta_j\}$. Clearly, $s_{0j} \geq s'_{0j}$. By the definition of s_0^* in (4.6),

$$s_0^*(s_1, ..., s_j + 1, ...s_J) = \max\{s_0^{-j}, s'_{0j}\} \leq \max\{s_0^{-j}, s_{0j}\} = s_0^*(s_1, ..., s_j, ..., s_J). \quad (4.27)$$

To prove Part 2, We first note that

$$Pr(B_{0j}(s_0 + 1) + D_j < s_j^*(s_0)) \geq Pr(B_{0j}(s_0) + D_j < s_j^*(s_0)) \geq \beta_j. \quad (4.28)$$

The first inequality follows from Lemma 1. The second inequality is by definition of $s_j^*(s_0)$ given in (4.3). Again, from the definition of $s_j^*(s_0 + 1)$ and (4.28), it must be true that $s_j^*(s_0 + 1) \leq s_j^*(s_0)$. Otherwise, $s_j^*(s_0 + 1) > s_j^*(s_0)$, which contradicts the definition of $s_j^*(s_0 + 1)$.

Recall that $s_j^l = s_j^*(\infty)$ and $s_j^l = s_j^*(0)$. Part 3 follows directly from Part 2.

We prove Part 4 by a contradiction argument. Assume that $s_j^*(s_0) > s_j^*(s_0+1)+1$. From the definition of $s_j^*(s_0)$, we have

$$Pr(B_{0j}(s_0) + D_j < s_j^*(s_0+1)+1)) < \beta_j. \tag{4.29}$$

From Lemma 2, we have $Pr(B_{0j}(s_0)+D_j < s_j^*(s_0+1)+1) \geq Pr(B_{0j}(s_0+1)+D_j < s_j^*(s_0+1)) \geq \beta_j$, the second inequality is by definition of $s_j^*(s_0)$. These inequalities contradict (4.29), which concludes our proof. ■

Proof of Proposition 2 follows from Proposition 1

Proof of Proposition 3. We first prove that $s_j^e \geq s_j^l$. This relation follows by the way s_j^e is set in (4.17) and by the definition of s_j^l in (4.4) because $1 - \frac{1-\beta_j}{1+\frac{h_0}{h_j-h_0}} \geq 1-(1-\beta_j) = \beta_j$. The inequality holds in equality only when $h_0 = 0$, that is, when the warehouse is used for rerouting of incoming inventory.

We next show that $s_{0j}^e \geq s_j^u$. Similarly, this relation follows by the way s_{0j}^e is set in (4.18) and by the definition of s_j^u in (4.5) because $\frac{1}{1+(\frac{1}{\beta_j}-1)(1-\frac{L_0}{L_j+L_0}\frac{1}{1+\frac{h_0}{h_j-h_0}})} \geq \frac{1}{1+(\frac{1}{\beta_j}-1)} = \beta_j$. The inequality holds in equality only when $L_0 = 0$, that is, when the warehouse is located close to the outside supplier with sufficient supply. Finally,

$$s_j^{NH} = \min(s_{0j}^e, s_j^e) \geq \min(s_{0j}^e, s_j^l) \geq \min(s_j^u, s_j^l) \geq s_j^l.$$

The first two inequalities follow from our previous proofs. ■

Proof of Proposition 4. The backorder cost method solves the following optimization problem.

$$\min_{s_0,s_1,\ldots,s_j} \quad h_0E[s_0-D_0]^+ + \sum_{j>0} h_jE[s_j-(B_{0j}(s_0)+D_j)]^+ + \sum_{j>0} b_jE[(B_{0j}(s_0)+D_j)-s_j]^+ \tag{4.30}$$

We first show that by solving the backorder-cost model and obtaining the base-stock levels, one can obtain a *feasible* solution for the PONI-constrained model. Let $(\hat{s}_0, \hat{s}_1, ..., \hat{s}_J)$ denote the optimal base-stock levels for the backorder-cost model with b_j's defined as in (4.26), and let $(\tilde{s}_0, \tilde{s}_1, ..., \tilde{s}_J)$ denote the base-stock levels for the

PONI-constrained model.

Given a fixed s_0, by definition

$$\tilde{s}_j(s_0) = \min\{s_j : Pr(B_{0j}(s_0) + D_j \le s_j) \ge \beta_j\}.$$

Given s_0, the optimization problem in (4.30) can be simplified to solve the following J single-location problems.

$$\min_{s_j} \ h_j E[s_j - (B_{0j}(s_0) + D_j)]^+ + b_j E[(B_{0j}(s_0) + D_j) - s_j]^+$$

for $j = 1, ..., J$. Hence,

$$\hat{s}_j(s_0) = \min\{s_j : Pr(B_{0j}(s_0) + D_j \le s_j) \ge \frac{b_j}{b_j + h_j} = \beta_j\}. \tag{4.31}$$

The last equality is by the definition of b_j in (4.26). Given the same base-stock level s_0 at the warehouse, the effective demand at retailer j, $B_{0j}(s_0) + D_j$, has the same distribution in both models. Therefore, $\hat{s}_j(s_0) = \tilde{s}_j(s_0)$ for all $j > 0$. Note that by setting s_0 to $\hat{s}_0$ in (4.31), we obtain $Pr(B_{0j}(\hat{s}_0) + D_j \le \hat{s}_j) \ge \beta_j$. This result shows that base-stock levels $(\hat{s}_0, \ldots, \hat{s}_J)$ yield a feasible solution for the PONI-constrained model. Hence, the backorder-cost model yields a solution that costs at least as much as the solution obtained by the PONI-constrained model. Next we show that the backorder-cost model *never* yields an optimal solution for the PONI model.

Plugging $\tilde{s}_j(s_0)$ into the objective function in (4.1), we obtain

$$\begin{aligned} f(s_0) &\equiv h_0 E[s_0 - D_0]^+ + \sum_{j=1}^{J} h_j E[\tilde{s}_j(s_0) - (B_{0j} + D_j)]^+ \\ &= h_0 E[s_0 - D_0]^+ + \sum_{j=1}^{J} h_j E[\hat{s}_j(s_0) - (B_{0j} + D_j)]^+. \end{aligned}$$

The equality follows because $\hat{s}_j(s_0) = \tilde{s}_j(s_0)$ for all $j > 0$ as we described in the above discussion. Similarly, given the same s_0, plugging $\hat{s}_j(s_0)$ into the objective function of (4.30) yields

$$g(s_0) \equiv f(s_0) + \sum_{j=1}^{J} b_j E[(B_{0j} + D_j) - \hat{s}_j(s_0)]^+.$$

Let $p(s_0) \equiv \sum_{j=1}^{J} b_j E[(B_{0j} + D_j) - \hat{s}_j(s_0)]^+$, and thus we have $g(s_0) = f(s_0) + p(s_0)$. Now we prove that $g^* \equiv \min_{s_0} g(s_0)$ is larger than $f^* \equiv \min_{s_0} f(s_0)$. Since by definition $\tilde{s}_0 \equiv \arg\min f(s_0)$ and $\hat{s}_0 \equiv \arg\min g(s_0)$, then

$$g^* = g(\hat{s}_0) = f(\hat{s}_0) + p(\hat{s}_0) > f(\hat{s}_0) \geq f(\tilde{s}_0) = f^*.$$

The first inequality follows because $p(s_0)$ is strictly positive, and the second inequality follows because by definition $f(s_0)$ achieves its minimum at $\tilde{s}_0$. Therefore, the optimal total average holding cost of the backorder-cost model is *strictly* larger than the optimal cost obtained by solving the PONI-constrained model. ■

4.9 Appendix B: Tables

	$L_0 = 0.1, L_j = 0.9$							
	Optimal		TSH			NVH		
J	s_0, s_j	$c*$	s_0, s_j	c_{TSH}	%	s_0, s_j	c_{NVH}	%
2	1,12	9.04	1,12	9.04	0.00	2,12	9.64	6.64
4	1,7	13.12	1,7	13.12	0.00	0,8	16.13	22.94
8	0,5	24.18	0,5	24.18	0.00	0,5	24.18	0.00 5
16	0,3	32.37	0,3	32.37	0.00	0,3	32.37	0.00
32	0,2	48.52	0,2	48.52	0.00	0,2	48.52	0.00
	$\beta_j = 90\%, h_0 = 0.3$							
2	1,12	9.17	1,12	9.17	0.00	0,13	10.13	10.47
4	1,7	13.24	1,7	13.24	0.00	0,8	16.13	21.83
8	0,5	24.18	0,5	24.18	0.00	0,5	24.18	0.00
16	0,3	32.37	0,3	32.37	0.00	0,3	32.37	0.00
32	0,2	48.52	0,2	48.52	0.00	0,2	48.52	0.00
	$\beta_j = 90\%, h_0 = 0.9$							
2	1,14	12.9	1,14	12.9	0.00	2,14	13.52	4.81
4	0,9	20.05	0,9	20.05	0.00	0,9	20.05	0.00
8	0,6	32.05	0,6	32.05	0.00	0,6	32.05	0.00
16	0,4	48.07	0,4	48.07	0.00	0,4	48.07	0.00
32	0,3	80.06	0,3	80.06	0.00	0,3	80.06	0.00
	$\beta_j = 97.5\%, h_0 = 0.3$							
2	1,14	13.02	1,14	13.02	0.00	0,15	14.03	7.76
4	0,9	20.05	0,9	20.05	0.00	0,9	20.05	0.00
8	0,6	32.05	0,6	32.05	0.00	0,6	32.05	0.00
16	0,4	48.07	0,4	48.07	0.00	0,4	48.07	0.00
32	0,3	80.06	0,3	80.06	0.00	0,3	80.06	0.00
	$\beta_j = 97.5\%, h_0 = 0.9$							
	$L_0 = 0.9, L_j = 0.1$							
2	18,3	5.32	18,3	5.32	0.00	22,3	6.68	25.56
4	18,2	7.32	18,2	7.32	0.00	28,2	10.51	43.58
8	13,2	12.61	13,2	12.61	0.00	32,2	19.69	56.15
16	21,1	16.41	21,1	16.41	0.00	32,2	35.68	117.43
32	14,1	29.31	14,1	29.31	0.00	64,1	45.32	54.62
	$\beta_j = 90\%, h_0 = 0.3$							
2	15,4	6.98	18,3	7.71	10.46	18,4	9.64	38.11
4	18,2	9.71	18,2	9.71	0.00	20,3	15.44	59.01
8	13,2	13.14	13,2	13.14	0.00	16,3	23.77	80.90
16	21,1	20.42	21,1	20.42	0.00	16,2	31.78	55.63
32	14,1	30.09	14,1	30.09	0.00	0,2	48.52	61.25
	$\beta_j = 90\%, h_0 = 0.9$							
2	19,4	7.63	19,4	7.63	0.00	24,4	9.27	21.49
4	18,3	11.24	18,3	11.24	0.00	28,3	14.48	28.83
8	19,2	15.64	19,2	15.64	0.00	32,3	27.68	76.98
16	14,2	29.13	14,2	29.13	0.00	48,2	40.48	38.96
32	9,2	57.01	9,2	57.01	0.00	64,2	77.28	35.56
	$\beta_j = 97.5\%, h_0 = 0.3$							
2	19,4	10.54	19,4	10.54	0.00	20,5	13.43	27.42
4	18,3	13.63	18,3	13.63	0.00	20,4	19.43	42.55
8	19,2	18.54	19,2	18.54	0.00	24,3	31.04	67.42
16	14,2	29.92	14,2	29.92	0.00	16,3	47.76	59.63
32	9,2	57.06	9,2	57.06	0.00	32,2	78.24	37.12
	$\beta_j = 97.5\%, h_0 = 0.9$							

Table 4.1: Optimal and Heuristic Solutions: Identical Retailers ($\lambda_0 = 16$).

	$L_0 = 0.1, L_j = 0.9$							
	Optimal		TSH			NVH		
β_j	s_0, s_j	$c*$	s_0, s_j	c_{TSH}	%	s_0, s_j	c_{NVH}	%
20%	1 , 6	0.94	1 , 6	0.94	0.00	0 , 7	1.33	41.49
30%	0 , 7	1.33	0 , 7	1.33	0.00	0 , 8	2.23	67.67
50%	1 , 8	2.69	1 , 8	2.69	0.00	0 , 9	3.42	27.14
90%	1 , 12	9.04	1 , 12	9.04	0.00	2 , 12	9.64	6.64
95%	1 , 13	10.95	1 , 13	10.95	0.00	2 , 13	11.56	5.57
99.9%	1 , 18	20.86	1 , 18	20.86	0.00	2 , 18	21.49	3.02
	$h_0 = 0.3, J = 2$							
20%	0 , 3	1.39	0 , 3	1.39	0.00	0 , 3	1.39	0.00
30%	4 , 3	2.54	4 , 3	2.54	0.00	0 , 4	3.13	23.23
50%	3 , 4	4.28	3 , 4	4.28	0.00	0 , 5	5.64	31.78
90%	1 , 7	13.12	1 , 7	13.12	0.00	0 , 8	16.13	22.94
95%	1 , 8	16.96	1 , 8	16.96	0.00	4 , 8	18.37	8.31
99.9%	0 , 12	32	0 , 12	32	0.00	0 , 12	32	0.00
	$h_0 = 0.3, J = 4$							
20%	1 , 6	1.06	1 , 6	1.06	0.00	0 , 7	1.33	25.47
30%	0 , 7	1.33	0 , 7	1.33	0.00	0 , 7	1.33	0.00
50%	1 , 8	2.81	1 , 8	2.81	0.00	0 , 9	3.42	21.71
90%	1 , 12	9.17	1 , 12	9.17	0.00	0 , 13	10.13	10.47
95%	1 , 13	11.07	1 , 13	11.07	0.00	0 , 14	12.06	8.94
99.90%	1 , 18	20.98	1 , 18	20.98	0.00	0 , 19	22	4.86
	$h_0 = 0.9, J = 2$							
20%	0 , 3	1.39	0 , 3	1.39	0.00	0 , 3	1.39	0.00
30%	0 , 4	3.13	0 , 4	3.13	0.00	0 , 4	3.13	0.00
50%	3 , 4	5.19	3 , 4	5.19	0.00	0 , 5	5.64	8.67
90%	1 , 7	13.24	1 , 7	13.24	0.00	0 , 8	16.13	21.83
95%	1 , 8	17.08	1 , 8	17.08	0.00	0 , 9	20.05	17.39
99.90%	0 , 12	32	0 , 12	32	0.00	0 , 12	32	0.00
	$h_0 = 0.9, J = 4$							
	$L_0 = 0.9, L_j = 0.1$							
20%	9 , 2	0.59	13 , 1	0.73	23.73	14 ,1	0.95	61.02
30%	11 , 2	1.09	15 , 1	1.19	9.17	14 , 2	2.08	90.83
50%	14 , 2	2.08	14 , 2	2.08	0.00	18 , 2	3.5	68.27
90%	18, 3	5.32	18, 3	5.32	0.00	22 , 3	6.68	25.56
95%	17, 4	6.83	24, 3	7.29	6.73	22 , 4	8.65	26.65
99.9%	21, 6	12.32	21, 6	12.32	0.00	28, 6	14.48	17.53
	$h_0 = 0.3, J = 2$							
20%	9 , 1	0.9	9 ,1	0.9	0.00	12 ,1	1.7	88.89
30%	11 , 1	1.4	11 ,1	1.4	0.00	16 ,1	3.03	116.43
50%	15 , 1	2.69	15 ,1	2.69	0.00	20 ,1	4.33	60.97
90%	18 , 2	7.32	18 ,2	7.32	0.00	28 ,2	10.51	43.58
95%	16, 3	10.34	16 ,3	10.34	0.00	28 ,3	14.48	40.04
99.9%	23, 4	16.96	23 ,4	16.96	0.00	32,5	23.68	39.62
	$h_0 = 0.3, J = 4$							
20%	9 , 2	0.64	13 , 1	1.26	96.88	8 , 3	1.02	59.38
30%	4 , 5	1.18	0 , 7	1.33	12.71	10 , 3	1.83	55.08
50%	11 , 3	2.36	14 , 2	2.87	21.61	12 , 3	2.97	25.85
90%	15 , 4	6.98	18 , 3	7.71	10.46	18 , 4	9.64	38.11
95%	17 , 4	8.75	14 , 6	9.92	13.37	18 , 5	11.62	32.80
99.90%	21 , 6	16.33	21 , 6	16.33	0.00	24 , 7	21.04	28.84
	$h_0 = 0.9, J = 2$							
20%	9, 1	0.95	9 , 1	0.95	0.00	4 , 2	1.08	13.68
30%	11, 1	1.6	11 , 1	1.6	0.00	8 , 2	2.44	52.50
50%	10, 2	3.49	15 , 1	3.79	8.60	8 , 3	5.12	46.70
90%	18, 2	9.71	18 , 2	9.71	0.00	20 , 3	15.44	59.01
95%	16, 3	11.81	16 , 3	11.81	0.00	20 , 4	19.43	64.52
99.90%	23, 4	22.14	23 , 4	22.14	0.00	32 , 5	34.24	54.65
	$h_0 = 0.9, J = 4$							

Table 4.2: Impact of Fill Rates: Identical Retailers ($\lambda_0 = 16$).

		TSH		NVH	
	# Experiments	mean	STDEV	mean	STDEV
$J = 2$	80	1.25%	2.86%	10.08%	5.30%
$J = 32$	80	0.53%	0.81%	14.29%	5.76%
$\lambda_0 = 16$	200	0.47%	1.61%	11.83%	6.53%
$\lambda_0 = 64$	200	0.98%	1.84%	15.73%	6.07%
$L_0 = 0.10$	200	0.44%	1.31%	10.43%	5.09%
$L_0 = 0.25$	200	1.02%	2.05%	17.13%	6.21%
$h_0 = 0.30$	200	0.52%	1.46%	12.91%	4.14%
$h_0 = 0.90$	200	0.94%	1.97%	14.65%	8.27%

Table 4.3: Summary Statistics for Heuristic Solutions: Non-identical Retailers.

L	β	Optimal		TSH			NVH		
L_1,L_2	β_1,β_2	s_0,s_1,s_2	c*	s_0,s_1,s_2	c_{TSH}	$\%\epsilon$	s_0,s_1,s_2	c_{NVH}	$\%\epsilon$
0.231, 0.209	0.966, 0.955	4, 6, 6	7.96	4, 6, 6	7.96	0.00	7, 6, 5	8.33	4.65
0.243, 0.239	0.972, 0.969	5, 6, 6	8.18	5, 6, 6	8.18	0.00	7, 6, 6	8.99	9.90
0.181, 0.121	0.940, 0.911	4, 5, 4	6.09	4, 5, 4	6.09	0.00	6, 5, 4	7.06	15.93
0.169, 0.135	0.935, 0.918	6, 4, 4	6.06	6, 4, 4	6.06	0.00	6, 5, 4	7.05	16.34
0.229, 0.131	0.965, 0.916	4, 6, 4	6.62	4, 6, 4	6.62	0.00	6, 6, 4	7.59	14.65
0.217, 0.227	0.958, 0.963	4, 6, 6	7.94	4, 6, 6	7.94	0.00	6, 6, 6	8.93	12.47
0.250, 0.250	0.975, 0.975	6, 6, 6	8.49	6, 6, 6	8.49	0.00	8, 6, 6	9.19	8.24
0.192, 0.159	0.946, 0.929	5, 5, 4	6.26	5, 5, 4	6.26	0.00	7, 5, 4	7.06	12.78
0.140, 0.145	0.920, 0.922	5, 4, 4	5.78	5, 4, 4	5.78	0.00	6, 4, 4	6.22	7.61
0.226, 0.104	0.963, 0.902	5, 6, 3	6.42	5, 6, 3	6.42	0.00	7, 6, 3	7.23	12.62
$\lambda_0 = 16$, $h_0 = 0.30$, $L_0 = 0.25$									
0.122, 0.242	0.911, 0.971	7, 8, 15	11.88	6, 9, 15	11.92	0.34	4, 10, 17	12.97	9.18
0.121, 0.236	0.911, 0.968	5, 9, 15	11.22	8, 8, 14	12.05	7.40	3, 10, 17	12.25	9.18
0.204, 0.145	0.952, 0.923	5, 13, 10	10.49	8, 12, 9	11.31	7.82	2, 15, 12	11.51	9.72
0.164, 0.111	0.932, 0.905	8, 10, 7	9.73	8, 10, 7	9.73	0.00	2, 13, 10	9.93	2.06
0.245, 0.202	0.972, 0.951	5, 16, 13	13.31	8, 15, 12	14.13	6.16	2, 18, 15	14.33	7.66
0.123, 0.232	0.911, 0.966	7, 8, 14	11.21	7, 8, 14	11.21	0.00	3, 10, 17	12.33	9.99
0.223, 0.187	0.962, 0.944	6, 14, 12	12.45	6, 14, 12	12.45	0.00	2, 16, 14	12.54	0.72
0.129, 0.127	0.914, 0.913	5, 9, 9	8.53	8, 8, 8	9.34	9.50	2, 11, 11	9.54	11.84
0.223, 0.171	0.961, 0.936	6, 14, 11	12.00	6, 14, 11	12.00	0.00	2, 16, 13	12.09	0.75
0.123, 0.176	0.912, 0.938	7, 8, 11	10.01	6, 9, 11	10.05	0.40	4, 10, 13	11.09	10.79
$\lambda_0 = 64$, $h_0 = 0.90$, $L_0 = 0.10$									
0.100, 0.185	0.900, 0.942	1, 4, 5	6.15	1, 4, 5	6.15	0.00	0, 4, 6	6.17	0.33
0.129, 0.221	0.914, 0.961	1, 4, 6	6.63	1, 4, 6	6.63	0.00	0, 5, 7	7.62	14.93
0.188, 0.172	0.944, 0.936	1, 5, 5	6.56	1, 5, 5	6.56	0.00	0, 6, 6	7.54	14.94
0.153, 0.234	0.926, 0.967	1, 5, 6	7.33	1, 5, 6	7.33	0.00	0, 5, 7	7.34	0.14
0.223, 0.212	0.962, 0.956	1, 6, 6	7.94	1, 6, 6	7.94	0.00	0, 7, 6	7.94	0.00
0.126, 0.229	0.913, 0.964	1, 4, 6	6.60	1, 4, 6	6.6	0.00	0, 5, 7	7.58	14.85
0.207, 0.177	0.953, 0.939	2, 5, 5	7.36	1, 6, 5	7.36	0.00	0, 6, 6	7.36	0.00
0.146, 0.102	0.923, 0.901	2, 4, 3	5.46	2, 4, 3	5.46	0.00	0, 5, 4	5.47	0.18
0.114, 0.155	0.907, 0.927	1, 4, 5	6.28	1, 4, 5	6.28	0.00	0, 5, 5	6.29	0.16
0.122, 0.125	0.911, 0.912	1, 4, 4	5.47	1, 4, 4	5.47	0.00	0, 5, 5	6.45	17.92
$\lambda_0 = 16$, $h_0 = 1.0$, $L_0 = 0.10$									

Table 4.4: Instances for Non-identical Retailers with $J = 2$.

L	β	Optimal		TSH			NVH		
L_1,L_2, L_3,L_4	β_1,β_2, β_3,β_4	s_0,s_1,s_2, s_3,s_4	c^*	s_0,s_1,s_2 , s_3,s_4	c_{TSH}	$\%\epsilon$	s_0,s_1,s_2 , s_3,s_4	c_{NVH}	$\%\epsilon$
0.190, 0.158, 0.210, 0.191	0.945, 0.929, 0.955, 0.946	6, 3, 3, 4, 3	10.5	6, 3, 3, 4, 3	10.5	0.00	7, 4, 3, 4, 4	12.85	22.38
0.186, 0.154, 0.123, 0.134	0.943, 0.927, 0.911, 0.917	6, 3, 3, 3, 3	10.1	4, 4, 3, 3, 3	10.11	0.10	8, 3, 3, 3, 3	10.81	7.03
0.164, 0.220, 0.178, 0.248	0.932, 0.96 , 0.939, 0.974	6, 3, 4, 3, 4	11.25	6, 3, 4, 3, 4	11.25	0.00	7, 3, 4, 3, 5	12.62	12.18
0.213, 0.152, 0.125, 0.199	0.956, 0.926, 0.913, 0.949	4, 4, 3, 3, 4	10.74	4, 4, 3, 3, 4	10.74	0.00	8, 4, 3, 3, 4	12.43	15.74
0.174, 0.11 , 0.205, 0.176	0.937, 0.905, 0.952, 0.938	6, 3, 2, 4, 3	9.85	6, 3, 2, 4, 3	9.85	0.00	8, 3, 3, 4, 3	11.54	17.16
0.122, 0.242, 0.121, 0.236	0.911, 0.971, 0.911, 0.968	9, 2, 4, 2, 4	10.64	9, 2, 4, 2, 4	10.64	0.00	8, 3, 4, 3, 4	12.3	15.60
0.204, 0.145, 0.164, 0.111	0.952, 0.923, 0.932, 0.905	6, 4, 3, 3, 2	10	6, 4, 3, 3, 2	10	0.00	8, 4, 3, 3, 3	11.69	16.90
0.245, 0.202, 0.123, 0.232	0.972, 0.951, 0.911, 0.966	5, 4, 4, 3, 4	11.83	4, 5, 4, 3, 4	12.27	3.72	8, 4, 4, 3, 4	12.98	9.72
0.223, 0.187, 0.129, 0.127	0.962, 0.944, 0.914, 0.913	6, 4, 3, 3, 3	10.82	6, 4, 3, 3, 3	10.82	0.00	7, 4, 4, 3, 3	12.19	12.66
0.223, 0.171, 0.123, 0.176	0.961, 0.936, 0.912, 0.938	5, 4, 3, 3, 3	10.28	11, 4, 3, 2, 3	11.36	10.51	8, 4, 3, 3, 3	11.42	11.09
$\lambda_0 = 16$, $h_0 = 0.30$, $L_0 = 0.25$									
0.21 , 0.161, 0.142, 0.185	0.955, 0.93 , 0.921, 0.943	5, 4, 3, 3, 3	11.11	5, 4, 3, 3, 3	11.11	0.00	2, 5, 4, 4, 5	13.23	19.08
0.202, 0.213, 0.208, 0.171	0.951, 0.957, 0.954, 0.936	3, 4, 4, 4, 4	11.83	4, 4, 4, 4, 4	12.77	7.95	1, 5, 5, 5, 4	12.87	8.79
0.118, 0.155, 0.225, 0.105	0.909, 0.928, 0.963, 0.903	4, 3, 3, 4, 3	10.55	4, 3, 3, 4, 3	10.55	0.00	3, 4, 4, 5, 3	12.58	19.24
0.178, 0.199, 0.164, 0.116	0.939, 0.95 , 0.932, 0.908	5, 3, 4, 3, 3	11.27	4, 4, 4, 3, 3	11.33	0.53	2, 4, 5, 4, 4	12.4	10.03
0.242, 0.238, 0.182, 0.152	0.971, 0.969, 0.941, 0.926	5, 4, 4, 3, 3	11.64	5, 4, 4, 3, 3	11.64	0.00	2, 6, 5, 5, 4	14.75	26.72
0.171, 0.156, 0.227, 0.148	0.935, 0.928, 0.964, 0.924	4, 3, 3, 4, 3	10.17	4, 3, 3, 4, 3	10.17	0.00	4, 4, 4, 5, 4	14.13	38.94
0.168, 0.141, 0.247, 0.145	0.934, 0.92 , 0.974, 0.922	4, 3, 3, 5, 3	11.16	4, 3, 3, 5, 3	11.16	0.00	2, 4, 4, 6, 4	13.22	18.46
0.211, 0.185, 0.129, 0.214	0.955, 0.943, 0.915, 0.957	3, 4, 4, 3, 4	11.06	5, 4, 3, 3, 4	11.93	7.87	0, 5, 5, 4, 5	12.11	9.49
0.226, 0.16 , 0.175, 0.234	0.963, 0.93 , 0.938, 0.967	5, 4, 3, 3, 4	11.72	5, 4, 3, 3, 4	11.72	0.00	4, 5, 4, 4, 5	14.76	25.94
0.104, 0.249, 0.186, 0.108	0.902, 0.975, 0.943, 0.904	6, 2, 4, 3, 2	10.24	6, 2, 4, 3, 2	10.24	0.00	1, 3, 6, 5, 4	12.48	21.88
$\lambda_0 = 16$, $h_0 = 0.90$, $L_0 = 0.25$									
0.18 , 0.129, 0.226, 0.194	0.94 , 0.915, 0.963, 0.947	9, 7 , 5, 8, 7	16.04	9, 7, 5, 8, 7	16.04	0.00	9, 7, 5, 9, 8	18.01	12.28
0.199, 0.13 , 0.226, 0.118	0.949, 0.915, 0.963, 0.909	9, 7 , 5, 8, 5	14.95	9, 7, 5, 8, 5	14.95	0.00	9, 8, 6, 9, 5	17.89	19.67
0.116, 0.211, 0.147, 0.241	0.908, 0.956, 0.924, 0.971	6, 5 , 8, 6, 9	15.69	6, 5, 8, 6, 9	15.69	0.00	9, 5, 8, 6, 9	17.22	9.75
0.143, 0.15 , 0.121, 0.21	0.921, 0.925, 0.911, 0.955	6, 6 , 6, 5, 8	14.18	6, 6, 6, 5, 8	14.18	0.00	9, 6, 6, 5, 8	15.7	10.72
0.225, 0.206, 0.19 , 0.212	0.963, 0.953, 0.945, 0.956	8, 8 , 8, 7, 8	17.92	8, 8, 8, 7, 8	17.92	0.00	9, 9, 8, 7, 8	19.34	7.92
0.138, 0.122, 0.1 , 0.109	0.919, 0.911, 0.9 , 0.905	9, 5 , 5, 4, 5	12.23	9, 5, 5, 4, 5	12.23	0.00	9, 6, 5, 5, 5	14.18	15.94
0.221, 0.228, 0.132, 0.117	0.96 , 0.964, 0.916, 0.909	9, 8 , 8, 5, 5	15.55	9, 8, 8, 5, 5	15.55	0.00	8, 9, 9, 6, 5	18.07	16.21
0.183, 0.102, 0.117, 0.168	0.941, 0.901, 0.909, 0.934	9, 7 , 4, 5, 6	13.6	9, 7, 4, 5, 6	13.6	0.00	8, 7, 5, 5, 7	15.12	11.18
0.213, 0.203, 0.182, 0.111	0.956, 0.951, 0.941, 0.906	6, 8 , 8, 7, 5	15.81	6, 8, 8, 7, 5	15.81	0.00	8, 8, 8, 7, 5	16.91	6.96
0.166, 0.13 , 0.204, 0.144	0.933, 0.915, 0.952, 0.922	8, 6 , 5, 8, 6	14.98	8, 6, 5, 8, 6	14.98	0.00	8, 7, 6, 8, 6	16.93	13.02
$\lambda_0 = 64$, $h_0 = 0.30$, $L_0 = 0.10$									
0.166, 0.135, 0.187, 0.18	0.933, 0.917, 0.943, 0.94	8, 6, 5, 7, 7	15.84	8, 6, 5, 7, 7	15.84	0.00	2, 8, 7, 9, 9	18.01	13.70
0.194, 0.124, 0.176, 0.244	0.947, 0.912, 0.938, 0.972	7, 7, 5, 7, 9	16.76	7, 7, 5, 7, 9	16.76	0.00	0, 9, 7, 9, 11	17.91	6.86
0.204, 0.239, 0.128, 0.15	0.952, 0.969, 0.914, 0.925	7, 8, 9, 5, 6	17.01	7, 8, 9, 5, 6	17.01	0.00	0, 1,0 1,1 7, 8	18.17	6.82
0.127, 0.249, 0.169, 0.25	0.913, 0.975, 0.934, 0.975	4, 6, 10, 7, 10	17.97	9, 5, 9, 6, 9	18.7	4.06	0, 7, 1,2 9, 12	20.98	16.75
0.115, 0.194, 0.114, 0.166	0.907, 0.947, 0.907, 0.933	5, 5, 8, 5, 7	14.29	8, 5, 7, 5, 6	15.09	5.60	1, 7, 9, 7, 8	16.3	14.07
0.24 , 0.107, 0.234, 0.144	0.97 , 0.904, 0.967, 0.922	6, 9, 5, 9, 6	17.02	6, 9, 5, 9, 6	17.02	0.00	1, 1,1 6, 1,1 8	19.11	12.28
0.134, 0.215, 0.162, 0.13	0.917, 0.958, 0.931, 0.915	8, 5, 8, 6, 5	15.26	8, 5, 8, 6, 5	15.26	0.00	1, 7, 1,0 8, 7	16.47	7.93
0.194, 0.191, 0.168, 0.17	0.947, 0.945, 0.934, 0.935	4, 8, 8, 7, 7	16.15	9, 7, 7, 6, 6	16.87	4.46	1, 9, 9, 8, 9	18.14	12.32
0.19 , 0.195, 0.228, 0.224	0.945, 0.948, 0.964, 0.962	4, 8, 8, 9, 9	18.28	10, 7, 7, 8, 8	19.91	8.92	1, 9, 9, 1,1 11	21.27	16.36
0.194, 0.208, 0.185, 0.156	0.947, 0.954, 0.942, 0.928	7, 7, 8, 7, 6	16.68	7, 7, 8, 7, 6	16.68	0.00	0, 9, 1,0 9, 8	17.84	6.95
$\lambda_0 = 64$, $h_0 = 0.90$, $L_0 = 0.10$									

Table 4.5: Instances for Non-identical Retailer with $J = 4$.

J	c^*	c_{NA}	c_{NVA}	c_{DFA}	L_0	c^*	c_{NA}	c_{NVA}	c_{DFA}
2	8.02	6.67	7.96	18.00	0.10	9.09	7.62	9.55	18.78
3	9.64	7.90	9.75	20.70	0.20	8.88	7.44	9.16	18.97
4	11.64	8.94	11.26	22.97	0.30	8.69	7.22	8.77	18.85
6	14.89	10.88	13.79	26.78	0.40	8.49	6.97	8.37	18.51
8	16.41	12.13	15.92	30.00	0.50	8.02	6.67	7.96	18.00
10	20.12	14.16	17.80	32.83	0.60	7.45	6.34	7.54	17.31
12	23.38	17.14	19.50	35.39	0.70	7.26	5.94	7.12	16.39
14	27.42	19.74	21.06	37.75	0.80	7.07	5.47	6.68	15.18
16	25.10	16.70	22.51	39.94	0.90	6.11	4.88	6.23	13.42
	$R^2 =$	98.92%	96.69%	96.68%		$R^2 =$	97.88%	96.32%	92.48%
	$\beta_j = 0.90,\ h_0 = 0.50,\ L_0 = 0.50$					$\beta_j = 0.90,\ h_0 = 0.50,\ J = 2$			
β_j	c^*	c_{NA}	c_{NVA}	c_{DFA}	h0	c^*	c_{NA}	c_{NVA}	c_{DFA}
0.860	6.94	5.88	7.27	14.87	0.1	7.60	6.01	6.36	14.68
0.875	7.64	6.15	7.51	15.87	0.2	7.77	6.18	6.78	15.79
0.900	8.02	6.67	7.96	18.00	0.3	7.95	6.35	7.19	16.65
0.915	8.29	7.04	8.27	19.69	0.4	7.99	6.52	7.58	17.37
0.930	8.90	7.45	8.63	21.87	0.5	8.02	6.67	7.96	18.00
0.945	9.53	7.94	9.06	24.87	0.6	8.06	6.80	8.33	18.57
0.960	10.21	8.61	9.61	29.39	0.7	8.09	6.91	8.68	19.10
0.975	11.97	9.45	10.36	37.47	0.8	8.13	7.00	9.02	19.59
0.990	13.95	11.02	11.71	59.70	0.9	8.16	7.08	9.36	20.05
	$R^2 =$	98.94%	99.01%	95.37%		R^2	90.79%	87.27%	93.09%
	$L0 = Lj = 0.5,\ h_0 = 0.50,\ J = 2$					$\beta_j = 0.90,\ L_0 = L_j = 0.5,\ J = 2$			

Table 4.6: Regression Analysis for Approximations ($\lambda_0 = 16$).

		Fill-Rate Constrained Model		PONI Constrained Model	
	# Experiments	mean	STDEV	mean	STDEV
$J = 2$	80	19.07%	19.27%	54.63%	21.39%
$J = 32$	80	-30.35%	9.47%	51.23%	13.52%
$\lambda_0 = 16$	200	-14.98%	21.17%	54.50%	16.41%
$\lambda_0 = 64$	200	-0.41%	16.42%	57.72%	29.06%
$L_0 = 0.10$	200	-14.69%	14.95%	47.04%	20.07%
$L_0 = 0.25$	200	-0.69%	22.42%	65.19%	23.47%
$h_0 = 0.30$	200	-10.89%	17.71%	52.74%	20.68%
$h_0 = 0.90$	200	-4.49%	22.14%	59.48%	25.86%

Table 4.7: Summary Statistics for Backorder-cost Method: Non-identical Retailers.

Bibliography

[1] Agrawal, V., Seshadri., S. (2000) "Distribution Free Bounds for Service Constrained (Q, r) Inventory Systems," *Naval Research Logistics*, Vol. 47, pp. 635–656.

[2] Aubin, J.-P., Ekeland, I. (1984) *Applied Nonlinear Analysis*, Wiley, New York, NY.

[3] Axsäter, S. (1990) "Simple Solution Procedures for a Class of Two-Echelon Inventory Problems," *Operations Research*, Vol. 38, pp. 64–69.

[4] Axsäter, S. (1993) "Continuous Review Policies for Multi-Level Inventory Systems with Stochastic Demand," Chapter 4 in: Graves, A. H. G. Rinnooy Kan, Zipkin, P. H. (Eds.), *Handbook in Operations Research and Management Science, Vol. 4: Logistics of Production and Inventory*, Elsevier Science/North-Holland, Amsterdam, NL, pp. 175–198.

[5] Axsäter, S. (2003) "Note: Optimal Policies for Serial Inventory Systems Under Fill-Rate Constraints," *Management Science*. Vol 49, pp. 247–253.

[6] Bashyam, S., Fu., M. (1998) "Optimization of (s,S) Inventory Systems with Random Lead Times and a Service Level Constraint," *Management Science*, Vol. 44, pp. 243–256.

[7] Berge, C. (1963) *Topological Spaces*, Oliver and Boyd, Edinburgh, UK. Reprinted by Dover Publications, Mineola, NY, in 1997.

[8] Bernheim, B.D., Whinston, M.D. (1986) "Menu Auctions, Resource Allocations, and Economic Influence," *Quarterly Journal of Economics*, Vol. 101, No. 1, pp. 1–31.

[9] Bernheim, B.D., Peleg, B., Whinston, M.D. (1987) "Coalition Proof Nash Equilibria, I: Concepts," *Journal of Economic Theory*, Vol. 42, No. 1, pp. 1–12.

[10] Bernstein, F., Federgruen, A. (2005) "Decentralized Supply Chains with Competing Retailers Under Demand Uncertainty," *Management Science*, Vol. 51, No. 1, pp. 18–29.

[11] Bertsekas, D.P. (1995) *Nonlinear Programming*, Athena Scientific, Belmont, MA.

[12] Boyaci, T., Gallego, G. (2001) "Serial Production/Distribution Systems Under Service Constraints," *Manufacturing & Service Operations Management*, Vol. 3, pp. 3–50.

[13] Cachon, G. (2003) "Supply Chain Coordination with Contracts," in: Graves, S., de Kok, T. (Eds.), *Handbooks in Operations Research and Management Science: Supply Chain Management*, North-Holland, Amsterdam, NL, pp. 229–339.

[14] Cachon, G., Lariviere, M. (2005) "Supply Chain Coordination with Revenue-Sharing Contracts: Strengths and Limitations," *Management Science*, Vol. 51, No. 1, pp. 30–44.

[15] Carr, S.M., Karmarkar, U.S. (2005) "Competition in Multi-Echelon Assembly Supply Chains", *Management Science*, Vol. 51, No. 1, pp. 45–59.

[16] Chain Store Age. (2002) "Inventory Management 2002: Data, Detail And Discipline: Supermarkets," *December Issue*, New York, NY 17A.

[17] Chen, F., Zheng, Y. S. (1994) "Lower Bounds for Multi-Echelon Stochastic Inventory Systems," *Management Science*. Vol. 40, pp. 1426–1443.

[18] Cheng, F., Ettl, M., Lin, G., Yao, D. (2002) "Inventory-Service Optimization in Configure-To-Order Systems," *Manufacturing & Service Operations Management.* Vol. 4, pp. 114–132.

[19] Clark, A. J., Scarf, H. (1960) "Optimal Policies for a Multi-Echelon Inventory Problem," *Management Science*, Vol. 6, pp. 475–490.

[20] Cohen, M. A., Deshpande, V., Wang, Y. (2003) "Managing Supply Chains with Differentiated Service Requirements–Models and Applications," Chapter 11 in: Harrison, T. P., Lee, H. L., Neale, J. J. (Eds.), *The Practice of Supply Chain Management: Where Theory And Application Converge,* Kluwer's International Series in Operations Research & Management Science. Kluwer Academic Publishing, Boston, MA, pp. 175–195.

[21] Cohen, M., Kleindorfer, P., Lee, H. L. (1988a) "Service Constrained (s,S) Inventory Systems with Priority Demand Classes and Lost Sales," *Management Science*, Vol. 34, pp. 482–499.

[22] Cohen, M., Kleindorfer, P., Lee, H. L. (1988b) "Near-Optimal Service Constrained Stocking Policies for Spare Parts," *Opearations Research.* Vol. 37, pp. 104–117.

[23] Diks, E. B., de Kok, A. G., Lagodimos, A. G. (1996) "Multi-Echelon Systems: a Service Measure Perspective," *European Journal of Operations Research* Vol. 95, pp. 241–263.

[24] Gallego, G., Özer, Ö. (2005) "A New Algorithm and a New Heuristic for Serial Supply Systems," *Operations Research Letters*, Vol. 33, pp. 349–362.

[25] Gallego, G. , Özer, Ö., Zipkin, P. (2003) "Bounds, Heuristics, and Approximations For Distribution Systems," *To appear in Operations Research.* Stanford University.

[26] Gallego, G., Zipkin, P. (1999) "Stock Positioning and Performance Estimation for Multi-Stage Production-Transportation Systems," *Manufacturing and Service Operations Management.* Vol. 1, pp. 77–88.

[27] Grossman, S.J., Hart, O.D. (1986) "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy,* Vol. 94, No. 4, pp. 691–719.

[28] Federgruen, A. (1993) "Centralized Planning Models for Multi-Echelon Inventory Systems Under Uncertainty," Chapter 3 in: Graves, A. H. G. Rinnooy Kan, Zipkin, P. H. (Eds.), *Handbook in Operations Research and Management Science, Vol. 4: Logistics of Production and Inventory,* North-Holland, Amsterdam, NL, pp. 133–173.

[29] Jensen, M.C., Meckling, W.H. (1976) "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics,* Vol. 3, No. 4, pp. 305–360.

[30] Jeuland, A.P., Shugan, S.M. (1983) "Managing Channel Profits," *Marketing Science,* Vol. 2, No. 3, pp. 239–272.

[31] Katz, M.L. (1989) "Vertical Contractual Relations," in: Schmalensee, R., Willig, R.D. (Eds.), *Handbook of Industrial Organization, Vol. 1,* North Holland, Amsterdam, NL, pp. 655–721.

[32] Lee, H. L., Billington, C., Carter, B. (1993) "Hewlett-Packard Gains Control of Inventory for Service Through Design for Localization," *Interfaces,* Vol. 23, pp. 1–11.

[33] Magaril-Il'yaev, G.G., Tikhomirov, V.M. (2003) *Convex Analysis: Theory and Applications,* American Mathematical Society, Providence, RI.

[34] Majumder, P., Srinivasan, A. (2003) "Asset-Specific Supply Tree Competition and Leadership," Working Paper, Fuqua School of Business, Duke University, Durham, NC.

[35] Martimort, D. (2006) "Multi-Contracting Mechanism Design," in: Blundell, R., Newey, W.K., Persson, T. (Eds.), *Advances in Economics and Econometrics, Vol. I*, Cambridge University Press, Cambridge, UK, pp. 57–101.

[36] Maskin, E., Tirole, J. (1990) "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica*, Vol. 58, No. 2, pp. 379–409.

[37] Moorthy, K.S. (1987) "Managing Channel Profits: Comment," *Marketing Science*, Vol. 6, No. 4, pp. 375–379.

[38] Oren, S.S., Smith, S.A., Wilson, R.B. (1982) "Nonlinear Pricing in Markets with Interdependent Demand," *Marketing Science*, Vol. 1, No. 3, pp. 287–313.

[39] Pasternack, B.A. (1985) "Optimal Pricing and Return Policies For Perishable Commodities," *Marketing Science*, Vol. 4, No. 2, pp. 166–176.

[40] Prat, A., Rustichini, A. (2003) "Games Played Through Agents," *Econometrica*, Vol. 71, No. 4, pp. 989–1026.

[41] Rosenbaum, B. A. (1981) "Service Level Relationships in a Multi-Echelon Inventory System," *Management Science*, Vol. 27, pp. 926–945.

[42] Schwarz, L. B., Deuermeyer, B., Badinelli, R. (1985) "Fill Rate Optimization in a One-Warehouse n-Identical Retailer Distribution System," *Management Science*, Vol. 31, No. 4, pp. 488–498.

[43] Segal, I. (1999) "Contracting with Externalities," *Quarterly Journal of Economics*, Vol. 104, No. 2, pp. 337–388.

[44] Segal, I., Whinston, M.D. (2003) "Robust Predictions for Bilateral Contracting with Externalities," *Econometrica*, Vol. 71, No. 3, pp. 757–791.

[45] Shang, K., Song, J. (2003) "Newsvendor Bounds and Heuristic for Optimal Policies in Serial Supply Chains," *Management Science*, Vol. 49, pp. 618–638.

[46] Sobel, M. (2003) "Fill Rates of Single-State and Multi-Stage Supply Systems," *Manufacturing & Service Operations Management*, Vol. 6, pp 41–52.

[47] Spengler, J.J. (1950) "Vertical Integration and Antitrust Policy," *Journal of Political Economy*, Vol. 58, No. 4, pp. 347–352.

[48] Strulovici, B.H., Weber, T.A. (2004) "Constructing Efficient Equilibria in Games Played Through Agents," Working Paper, Department of Management Science and Engineering, Stanford University, Stanford, CA.

[49] Taylor, T. (2002) "Coordination Under Channel Rebates with Sales Effort Effects," *Management Science*, Vol. 48, No. 8, pp. 992–1007.

[50] Tsay, A.A. (1999) "The Quantity Flexibility Contract and Supplier-Customer Incentives," *Management Science*, Vol. 45, No. 10, pp. 1339–1358.

[51] van Houtum, G. J., Zijm, W.H.M. (2000) "On the Relation Between Cost and Service Models for General Inventory Systems," *Statistica Neerlandica*, Vol. 54, No. 2, pp. 127–147.

[52] Von Neumann, J., Morgenstern, O. (1944) *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

[53] Weber, T.A., Xiong, H. (2006) "Efficient Contract Design in Multi-Principal Multi-Agent Supply Chains," Working Paper, Department of Management Science and Engineering, Stanford University, Stanford, CA.

[54] Zipkin, P. (2000) *Foundation of Inventory Management,* Irwin McGraw-Hill, New York, NY.