INCENTIVES IN SUPPLY CHAIN MANAGEMENT

by

Mohammad M. Fazel-Zarandi

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Rotman School of Management
University of Toronto

© Copyright 2016 by Mohammad M. Fazel-Zarandi
Abstract

Incentives in Supply Chain Management

Mohammad M. Fazel-Zarandi
Doctor of Philosophy
Graduate Department of Rotman School of Management
University of Toronto
2016

My dissertation focuses on strategic issues in supply chain management. It consists of three studies. In the first chapter, I investigate the incentive issues in supply chain information sharing. Specifically, I provide an explanation for a long-standing question in supply chain management: why do supply chain firms often use simple contracts when our models typically show that more complex forms are needed to coordinate the supply chain? Using a stylized supply chain model consisting of a retailer who solicits forecast information from a manufacturer, I show that if there is sufficient operational flexibility in the supply chain, credible information sharing is indeed possible in a simple contracting environment.

The second chapter focuses on the structure of supply chains in a competitive environment. It is motivated by Ford’s support of the government bailout of its competitors General Motors and Chrysler. This is in sharp contrast with the general presumption that firms prefer fewer to more competitors. Using a stylized supply chain structure, I sort out conditions where firms have self-interest in promoting the sustainability of their competitors. To get some sense of whether the model can be a credible explanation for Ford’s support of the bailout, I calibrate it to pre- and post-recession data. The calibrated model suggests that the bailout of GM and Chrysler actually benefitted Ford.

In the third chapter, I present the Almost Robust Optimization model that is a simple and tractable methodology for incorporating data uncertainty into supply chain models. The model trades off the objective function value with robustness and finds optimal solutions that are almost robust. To efficiently solve the model, a novel decomposition approach is proposed that decomposes the model into a deterministic master problem and a subproblem which checks the master problem solution under different realizations. Finally, to demonstrate the efficiency of the methodology, I apply it to two important logistics problems.
Acknowledgements

I would like to express my sincere gratitude to my advisors, Professors Oded Berman, Dmitry Krass, and Opher Baron for their guidance, help and support throughout the course of my PhD studies. I have been extremely fortunate to have the opportunity to learn from and collaborate with them.

I would also like to thank my committee member, Professor Joseph Milner, and the external appraiser, Professor Gad Allon.

I am indebted to Professors Ig Horstmann and Frank Mathewson who have provided tremendous support and guidance, having great influence on my research. I am very fortunate to have worked with them over the years.

Last and most, I would like to thank my parents, Susan Bastani and Mohammad Hossein Fazel Zarandi, who have provided love, unfailing support and advice throughout the years, and my sisters, Maryam and Mahya, who are both a source of energy and encouragement.
Contents

1 Introduction 1

2 Can Supply Chain Flexibility Facilitate Information Sharing? 3
  2.1 Related Literature ................................................. 5
  2.2 The Model .......................................................... 7
    2.2.1 The Traditional System ...................................... 8
    2.2.2 Flexible System ............................................. 10
  2.3 The Cheap Talk Game ............................................ 15
    2.3.1 Communication in the Traditional System ................. 18
    2.3.2 Communication in the Flexible System .................... 19
  2.4 Conclusion ....................................................... 24
  2.5 Appendix .......................................................... 26

3 Can Supply Chain Firms Benefit From Competition? 32
  3.1 The Model .......................................................... 35
    3.1.1 Downstream Market .......................................... 35
    3.1.2 Upstream Market ............................................ 36
    3.1.3 Equilibrium ................................................ 37
  3.2 Firms Exit Both Market Segments: Competition and Network Effects ...................... 39
  3.3 Firms Exit One Market Segment: Competition and Network Effects ...................... 40
  3.4 Industry Equilibrium and Adjustment ................................................. 45
  3.5 A Calibration Experiment ......................................... 46
    3.5.1 Demand Data ................................................ 47
    3.5.2 Supply Data ................................................ 49
    3.5.3 Calibration Strategy ...................................... 49
Chapter 1

Introduction

In this dissertation, we study three problems in supply chain management. In the first study, we consider strategic information sharing within a supply chain. In this study, we attempt to provide an explanation for a long-standing observation in supply chain management: while our models typically show that simple contracts cannot induce credible forecast sharing between different parties within the supply chain, why do firms often use them in practice, and exchange information through unverifiable communication (“cheap talk”)? We address this question using a stylized model consisting of a retailer who solicits forecast information from a manufacturer launching a new product. Once information is shared, the retailer decides how much of its budget to invest in procuring inventory of an existing product, and how much to retain as a reserve for the anticipated launch of the manufacturer’s new product. To capture the effect of flexibility on credible forecast sharing, we examine two operational systems: a traditional system, where the retailer must make investment decisions at the beginning of the time horizon, and a flexible system, where the retailer has the option to postpone its procurement decision until learning whether the manufacturer’s product is released. We show that in the traditional system, the manufacturer has an incentive to inflate its forecast, thus, the retailer cannot trust the report. In the flexible system, however, procurement flexibility brings about a situation in which the manufacturer is uncertain about the retailer’s reaction to its report. This uncertainty may induce the manufacturer to report truthfully. Therefore, we show that under complete rationality, firms can employ simple contracts while, at the same time, induce truthful reporting.

The second study considers the structure of supply chains in a competitive environment. The motivation centers on the U.S. Federal Government bailout of the U.S. auto producers in 2008 - 2009. While Ford did not participate in the 2009 U.S. Federal Government bailout of General Motors and
Chrysler, Ford supported the bailout. This support by Ford stands in sharp contrast with the general presumption that firms prefer fewer to more competitors. Using a stylized supply chain structure, the goal is to sort out those conditions where downstream firms have a self-interest in promoting the sustainability of their downstream competitors. We show that, while Ford gains from reduced competition should Chrysler and/or GM exit, the resulting damage to the upstream supply industry harms Ford’s profitability. What matters is the relative size of mark-ups. If upstream parts industry mark-ups are large while auto mark-ups are small, the exit of competitors damages Ford’s profit.

To get some sense of whether our model can be a credible explanation for Ford’s support of the auto industry bailout, we calibrate it to pre- and post-recession data. In the pre-recession world, our calibration reveals that the equilibrium mark-up on variable cost in the downstream market is sufficiently large relative to the mark-up in the upstream market that exit would increase profits for the continuing firms. In the post-recession world, our calibration reveals just the opposite: the downstream mark-ups are small and the exit of GM and/or Chrysler would significantly lower Ford’s profits by dramatically raising the prices of parts. Thus, the bailout of GM and/or Chrysler benefited Ford.

The third study presents a simple and tractable methodology for incorporating data uncertainty into supply chain models, and to demonstrate the practicality of this methodology for logistics and manufacturing problems. We present the Almost Robust Optimization model that trades off the objective function value with robustness and finds optimal solutions that are almost robust (feasible under most realizations). The proposed model is attractive due to its simplicity, its ability to model dependence among uncertain parameters, and its ability to incorporate the decision maker’s attitude towards risk by controlling the degree of conservatism of the optimal solution. To solve the model efficiently, we decompose it into a deterministic master problem and a single subproblem that checks the master problem solution under different realizations and generates cuts if needed. We demonstrate our methodology on two important logistics problems: the capacitated facility location problem with uncertain demands, and the vehicle routing problem with uncertain demands and travel times. Computational experiments demonstrate the effectiveness of our approach.
Chapter 2

Can Supply Chain Flexibility Facilitate Information Sharing?

It is generally accepted that coordination between different levels of a supply chain is critical for the chain to operate effectively (see, for example, [26] and [65]). One of the main mechanisms to achieve such coordination is information sharing between various parties within the supply chain ([27] and [68]). Although the benefits of information sharing as a way to improve the performance of the supply chain is known, the issue of credibility is often a key obstacle in achieving such an objective: when the incentives of the multiple parties are not completely aligned, one party may have an incentive to distort its report to influence the other’s action.

To overcome the credibility issue in information sharing, operations management research has focused on eliciting truthful forecasts through carefully designed and often complex contracts ([28] and [78]), or through reputation mechanisms ([82]). While such contracts and mechanisms can be applied in some settings, empirical research suggest that many firms employ simple contracts and share information through nonbinding and unverifiable messages (see, for example, [55], [44], [11], [12], [56], [39], and [30]).

This discrepancy between theory and practice raises the question: given that our model typically claim that simple contracts cannot induce truthful information sharing, why do firms persist in using them and communicate through unverifiable reports? One possible explanation is that supply chain agents are not perfectly rational and are instead behavioral, caring about non-pecuniary factors such as trust and trustworthiness (see [79]). We provide an alternative and complementary explanation. We suggest that, under perfect rationality, if the supply chain entity sending the report is sufficiently uncertain about the receiving firm’s reaction to the report (due to both, the informational uncertainty
about the receiver and the supply chain structure being sufficiently flexible for the receiver to have alternative options), it may truthfully share its private information in equilibrium.

We develop a model in which the above issues are formally addressed by considering a supply chain that consists of a budget/capacity constrained retailer and a manufacturer. At the beginning of the time horizon, the manufacturer is in the development stage of producing a new product that will compete with a pre-existing product produced by a third party. Since a new product launch involves many sequential steps (including engineering development and testing, material sourcing, and manufacturing), the manufacturer is uncertain whether he can launch the product in time for the upcoming selling season. Such uncertainties are quite common in practice. For example, Microsoft delayed the release of Xbox from the fall of 2000 to the fall of 2001 since it was not able to acquire the right technology on time ([67]). Apple’s Lisa project ([38]) and Intel’s Broadwell processor\(^1\) are other examples where the launch was delayed by several quarters due to development and manufacturing issues. Thus, the manufacturer can only forecast the likelihood of his product being ready for the upcoming selling season. This forecast is the manufacturer’s private information, which he communicates (either truthfully or not) with the retailer. After the communication stage, the retailer faces an investment problem: how much of her budget (capacity) to invest in procuring inventory of an existing product in the market (manufactured by a third party) and how much of it to retain as a reserve for the anticipated launch of the manufacturer’s product (which has higher yield and/or demand). We focus on analyzing the settings under which the manufacturer has an incentive to truthfully share his private forecast with the retailer. In particular, we examine the role of “cheap talk” (costless, nonbinding, and unverifiable communication) in a sender (manufacturer)-receiver (retailer) game in two operational systems: a traditional system, in which the retailer must make her investment decision at the beginning of the time horizon; and a flexible system, where the retailer has the option to postpone (at a cost) her investment in the existing product until learning whether the manufacturer’s product is released. In both systems, to capture the effect of uncertainty on the manufacturer’s ability to share his forecast, we consider two information structures, one in which the manufacturer has perfect information about the retailer’s endowment, and the other in which the manufacturer is uncertain about it.

We show that in the traditional system, the manufacturer always (regardless of his knowledge of the retailer’s endowment) prefers to report a forecast that is excessively optimistic, so as to induce the retailer to retain a larger cash reserve for the possible release of his product (by investing less in the existing rival product in the market). Anticipating such behavior, the retailer discards the report. Therefore, in the traditional system, information cannot be shared credibly. The incentive of the informed party to

\[^1\]“Exclusive: Intel CEO promises Broadwell PCs on shelves for holidays”. Reuters, May, 2014.
provide an excessively optimistic forecast is also recognized in demand forecast sharing models as a key barrier to credible communication ([82] and [79]).

In the flexible system, on the contrary, the manufacturer may prefer to send different reports (inflate, deflate, or not alter his report) under different realizations of the retailer’s endowment. This mixture of preferred reports, coupled with the manufacturer’s uncertainty about the retailer’s endowment can, under fairly plausible conditions, result in truthful forecast sharing in equilibrium. Therefore, our model suggests that under complete rationality and in the absence of a complex contract or a reputation mechanism, uncertainty on the part of the reporting firm about the receiver’s reaction to the report, due to the combination of the sender’s informational uncertainty and supply chain flexibility, may result in truthful information sharing between the different tiers within the supply chain.

The chapter is organized as follows. In Section 2.1, we review the relevant literature. Section 2.2 sets out our model in the traditional and flexible systems, and provides benchmark analysis where information is symmetric and complete. Section 2.3 contains the cheap talk game. A final section concludes the chapter. Proofs of the main results appear in the Appendix.

2.1. Related Literature

Our analysis is related to the literature on strategic information sharing in supply chains, in general, and to the literature on “cheap talk” games in operations management, in particular. Many authors have emphasized the role of information sharing and its effect on improving the efficiency of the supply chain, for example, by better matching demand and supply and alleviating the bullwhip effect, see [33]. While earlier research considered nonstrategic interactions, a more recent stream of research has focused on strategic information sharing: when a party in the supply chain has superior information, it may have an incentive to be strategic in how the information is shared. To enforce credibility, the primary focus has been on designing complex contracts and mechanisms that ensure truthful information sharing, see, e.g., [28], [78], and [52]. In practice, however, simple contracts are often used, and the information is usually shared through costless, nonbinding, and unverifiable communication - cheap talk (see [44], [82], [79], and the references mentioned in the introduction). To address this inconsistency between theory and practice, [82] suggests the use of review or trigger policies with penalties to promote credible information sharing. [85] examine the role of public information sharing within a supply chain that is comprised of an incumbent and an entrant, and suggest that public information sharing by the incumbent is possible when it influences the entrant’s entry decision. Using controlled laboratory experiment, [79] demonstrate that behavioural factors such as trust and trustworthiness can play a role in sharing demand forecast
information.

Our forecast-sharing game also falls within the class of economic problems known as “cheap talk” models, which originate from Crawford and Sobel’s (1982) analysis of strategic information transmission between an informed sender and an uninformed receiver, where communication is costless, nonbinding, and unverifiable. Our work differs from the supply chain information sharing literature mentioned above along three broad dimensions. First, unlike the majority of the literature that assume one-sided asymmetric information, we consider a more general information structure: we allow both supply chain parties to possess private information, and permit their information to be correlated. To the best of our knowledge, our work is the first in this body of literature to consider such an information structure. Second, while many demand forecast sharing models assume that information flows from downstream to upstream, in practice, there are many cases where the upstream firm possess superior information, e.g., large manufacturers such as Apple, Boeing, and Samsung often possess superior information about their proprietary products. This is the case we consider. Note that while the direction of the flow of information may be different, the underlying incentive of the informed party to provide an excessively optimistic report is similar in both cases (to induce the receiver to reserve abundant cash/capacity). Third, and most importantly, we provide an alternative explanation on how in the absence of complex contracts, reputation mechanisms, or behavioral factors, firms within the supply chain may credibly share information.

In the service industry context, [4] and [3] investigate a service provider (call center) who uses cheap talk (delay announcement) to inform its customers. They show that such communication can increase the profit of the firm and at the same time improve the expected utility of the customers. In a retailer-customer setting, [2] show that when consumers are homogeneous, the retailer cannot influence customer behaviour through cheap talk, and provide conditions when the firm may be able to influence customer behavior in case of heterogenous customers. Our framework differs from these models in that we consider information sharing between different firms within a supply chain whereas they focus on settings in which a service provider transfers real-time information to its customers.

Our work is also related to the literature on operational flexibility, see [45], [43], [60], [89], [50], [5]. These works describes the benefit of reducing supply-demand mismatches by providing the firm the option to procure inventory after learning updated information. We add to this literature another important benefit of operational flexibility: supply chain flexibility can coordinate the chain by enabling information sharing.

Chapter 2. Can Supply Chain Flexibility Facilitate Information Sharing?

2.2. The Model

Players: We consider a stylized supply chain consisting of a manufacturer and a retailer. At the beginning of the decision horizon, the manufacturer is in the development stage of producing a new (seasonal) product, “Product N”, that will compete with an existing rival product in the market, “Product E”, manufactured by a third party. Since a new product launch involves many sequential steps (including engineering development and testing, material sourcing, and manufacturing), the manufacturer is uncertain whether he can launch the product in time for the upcoming selling season (see the Microsoft, Apple, and Boeing examples provided in the introduction). Thus, he can only forecast the likelihood $\theta$ of the product being ready for release at the start of the selling season, where we assume that $\theta < 1$. The value of $\theta$ is the manufacturer’s private information. Due to the seasonality of the product, we assume that it is not practical for the manufacturer to release a new product in the middle of the selling season.

The retailer is budget-constrained\(^3\) and has an endowment (budget) $b$, which can be fully or partially invested in inventory of the existing product, Product E, or retained in the firm as a reserve for the anticipated launch of Product N (which has higher yield and/or demand). We assume that the retailer does not have/want access to external sources of capital. This assumption is supported by the fact that, in practice, there are frictions in the capital market ([53]). The presence of financial constraints are also empirically well documented (see, for example, [91]).

The market is characterized by demand uncertainty: the total demand for Product E, $D_E$, is a random variable with positive support, strictly increasing cumulative distribution function $F_E(\cdot)$, and density $f_E(\cdot)$. If Product N is released, its demand, $D_N$, is a random variable with positive support, strictly increasing cumulative distribution function $F_N(\cdot)$, and density $f_N(\cdot)$\(^4\). The distributions of $D_E$ and $D_N$ are common knowledge.

Remark. The model can be extended to take into account product substitution, i.e., Product N, if released to the market, has a demand cannibalization effect on Product E. This can be incorporated into the model by assuming that the demand of Product E has a distribution function $F_{ERL}^E (F_{NRL}^E)$ if Product N is released (not released), where $F_{ERL}^E(x) \leq F_{NRL}^E(x), \forall x \geq 0$. For ease of exposition, we ignore such product substitutability. By removing this assumption, our main insights remain unchanged.

\(^3\)All our results go through if instead of a budget constraint, the retailer is capacity-constrained. This capacity can reflect the observation that shelf space is very scarce, see for example “Getting Your Product Onto Retail Shelves” The New York Times, Oct., 2010.

\(^4\)We ignore any superior demand information since both parties can be endowed with a good demand forecast for Product N: the manufacturer may be well informed because of his proprietary information about the product, whereas the retailer can be informed due to her proximity to the consumer market.
**Systems:** To capture the effect of procurement flexibility on credible forecast sharing, we analyze the interaction between the manufacturer and the retailer in two potential operational systems: (i) *Traditional System*, abbreviated *TS*, in which the retailer must decide on her order quantity at the beginning of the time horizon; and (ii) *Flexible System*, abbreviated *FS*, in which the retailer has the option to postpone (at a cost) her ordering decision of Product *E* until learning whether Product *N* is released.

### 2.2.1 The Traditional System

The timing of events in the traditional system, illustrated in Figure 2.1, is as follows: at the beginning of the time horizon, the retailer and the manufacturer share information regarding the forecasted likelihood of the release of Product *N*, θ. After the communication stage, the retailer decides on *Q_E*, the order quantity of Product *E*, and the amount of cash reserve *M* to retain in the firm for the anticipated launch of Product *N*. Then, Product *E* is procured at wholesale price *w_E*, and if Product *N* is released, the retailer places her order quantity *Q_N* at expected wholesale price *w_N* based on her remaining available cash reserve. Finally, during a finite selling season, the retailer sells Product *E* at unit price *p_E*, and if Product *N* is released, it is sold at expected unit price *p_N*. Without loss of generality, we ignore the possible salvage value of leftover inventory at the end of the selling season, which can be easily incorporated at the cost of extra parameters, but without any additional insights. Moreover, for ease of exposition, and without loss of generality, we normalize the risk-free interest rate to zero, i.e., the retailer does not make any profit by hoarding cash and not investing in either Product *E* or Product *N* (if released).

**Traditional System Benchmark (Full Information Sharing):** As stated above, given the manufacturer’s report of θ, the retailer seeks to choose the order quantity of Product *E*, *Q_E*, her cash reserve, *M* = *b* − *w_E* *Q_E*, and the order quantity of Product *N* (if released), *Q_N*, to maximize her overall expected profit. Assume that the manufacturer truthfully reports his forecast of θ (this assumption is relaxed in Section 2.3). Let *Q* = {*Q_E*, *Q_N*} denote a vector of order quantities. Then, the retailer’s problem can be formulated as follows (throughout the chapter, subscript *r* and *m* denote the retailer’s and manufacturer’s parameters, respectively):

\[
\max_{Q,M} \Pi^{TS}_{r}(Q, M, \theta, b) = \Pi^{TS}_{E}(Q_E) + \theta \Pi^{TS}_{N}(Q_N)
\]

s.t. \( w_E Q_E + M = b, \)

\( w_N Q_N \leq M, \)
where $\Pi^{TS}_E(Q_E) = p_E \mathbb{E}[\min(Q_E, D_E)] - w_E Q_E$ is the retailer’s expected profit when she places an order of $Q_E$, and $\Pi^{TS}_N(Q_N) = p_N \mathbb{E}[\min(Q_N, D_N)] - w_N Q_N$ is her expected profit from procuring inventory of Product $N$ (if released to the market). The first constraint is the budget constraint, while the second constraint ensures that the total investment in Product $N$ does not exceed the cash reserve.

Given the retailer’s cash reserve, and the expected per unit production cost of Product $N$, $c$, the manufacturer’s profit if he releases Product $N$ is:

$$\Pi_m(Q_N) = (w_N - c)Q_N.$$ 

It is obvious that if Product $N$ is not released, the manufacturer’s profit is 0.

As a benchmark, we next derive the retailer’s optimal decision in the traditional system, assuming that the manufacturer truthfully reveals his forecast. To avoid trivial solutions, throughout the chapter we assume that the retailer’s endowment is not sufficient to procure the optimal unconstrained order quantities of both products, but is sufficient to procure the optimal unconstrained order quantities of any single product (either Product $E$ or Product $N$). Proposition 1 characterizes the retailer’s optimal decision in the traditional system under truthful information exchange.

**Proposition 2.1.** Assume the manufacturer truthfully reports his forecast of $\theta$. In the traditional system:

---

5. Note that an equivalent optimization problem would maximize the retailer’s total expected cash holding: $p_E \mathbb{E}[\min(Q_E, D_E)] + \theta p_N \mathbb{E}[\min(Q_N, D_N)] + (1 - \theta)M$. 

---
(a) The optimal order quantity of Product \( E \), \( Q_{TS}^{*} \), is the unique solution to the following first-order condition:

\[
p_{E}(1 - F_{E}(Q_{TS}^{*}^{E})) - w_{E} - \theta \frac{w_{E}}{w_{N}} (p_{N}(1 - F_{N}(\frac{b-w_{E}Q_{TS}^{*}}{w_{E}})) - w_{N}) = 0.
\]

(b) The retailer’s optimal cash reserve is \( M_{TS}^{*} = [b - w_{E}Q_{TS}^{*}]^{+} \), and her optimal order quantity of Product \( N \) (if released) is \( Q_{TS}^{*} = \frac{M_{TS}^{*}}{w_{N}} \).

(c) Comparative statics: the table below indicates the changes to the values of columns as row parameters increases; “+” indicates an increase and “-” a decrease, e.g., a “-” in the \((\theta, Q_{TS}^{*})\) cell means \( \frac{\partial Q_{TS}^{*}}{\partial \theta} < 0 \).

<table>
<thead>
<tr>
<th></th>
<th>( Q_{TS}^{*} )</th>
<th>( M_{TS}^{*} )</th>
<th>( Q_{TS}^{*} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \uparrow \theta )</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( \uparrow b )</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Parts (a) and (b) of the proposition demonstrate that in the traditional system, there exist a unique optimal order quantity of Product \( E \) and a unique optimal cash reserve. The comparative statics presented in part (c) shows that the retailer’s optimal cash reserve is increasing in \( \theta \) (resulting in the optimal order quantity of Product \( N \) (if released) to be increasing in \( \theta \)), whereas the optimal order quantity of Product \( E \) is decreasing in \( \theta \). Moreover, as the retailer’s budget increases, both optimal order quantities (\( Q_{E}^{*}, Q_{N}^{*} \)) and the cash reserve (\( M^{*} \)) increase. From Proposition 2.1, we derive the following result.

Corollary 2.1. In the traditional system, the manufacturer’s profit (\( \Pi_{m} \)) is increasing in \( M_{TS}^{*}, \theta, \) and \( b \).

This corollary links the manufacturer’s profit with the retailer’s cash reserve and the likelihood of the release of his product. It shows that the manufacturer’s profit increases when either the retailer’s cash reserve or the likelihood of the release of his product increase. The intuition behind this result is that an increase in \( \theta \) increases the retailer’s cash reserve, thus, she has more cash on hand for the possible release of Product \( N \), which results in an increase in the manufacturer’s profit. This result will be used extensively in Section 2.3.

2.2.2 Flexible System

In the flexible system, the retailer has the option to postpone her procurement decision of Product \( E \) until after learning whether Product \( N \) is released. Specifically, she has the option to procure inventory
of Product $E$ at one of two times: *early* (where she is still uncertain whether Product $N$ will be released) or *late* (when her uncertainty about the release of Product $N$ is eliminated). If the retailer decides to postpone her procurement decision, she must purchase Product $E$ at a higher per unit cost $w^\tau_E > w_E$. The higher cost is due to expedited manufacturing and shipping expenses. Note that similar to [88], [7] and [89], we are assuming that the retailer is not able to place multiple orders of Product $E$, i.e., she must decide between early or late procurement, but cannot split her order between both. This assumption is reasonable when the fixed production/capacity/procurement cost of Product $E$ is high enough. The sequence of events in the flexible system, depicted in Figure 2.2, differs from that of the traditional system since after the information sharing stage, the retailer also faces an investment timing decision for Product $E$.

**Flexible System Benchmark (Full Information Sharing):** Assume that the manufacturer truthfully reports his forecast of $\theta$. If the retailer decides not to postpone her procurement decision, her optimal expected profit, $\Pi_{r}^{NPOS^*}$, is equal to that of the traditional system. Specifically, given $\theta$, her optimal expected profit for the early procurement case is:

$$\Pi_{r}^{NPOS^*} = \Pi_{r}^{TS^*} = \Pi_{r}^{TS}(Q_{TS}^*, M_{TS}^*, \theta, b)$$

If she decides to postpone, the following two outcomes must be considered:

(I) *Product $N$ is released (RL):* in this case, the retailer optimally allocates her endowment to procure both products (note that she must procure Product $E$ at cost per unit $w^\tau_E$). Therefore, her optimal
expected profit, $\Pi_{RL}^*, \text{ is the solution to:}$

$$
\max_{Q=\{Q_E, Q_N\}} \Pi_{RL}(Q, b) = p_E \mathbb{E}[\min(Q_E, D_E)] - w_E^* Q_E + p_N \mathbb{E}[\min(Q_N, D_N)] - w_N Q_N
$$

s.t. \quad w_E^* Q_E + w_N Q_N \leq b.

By comparing the above profit to the retailer’s profit in the traditional system we see that: the per unit cost of Product $E$ increases, while the uncertainty about the release of Product $N$ disappears (i.e., $\Pi_{RL}^* = \Pi_{TS}^*|_{w_E = w_E^*, \theta = 1}$).

\textbf{(I)} Product $N$ is released (RL): in this case, the retailer only procure inventory of Product $E$ at higher cost $w_E^*$. Therefore, her optimal expected profit, $\Pi_{NRL}^*$, is the solution to:

$$
\max_{Q_E} \Pi_{NRL}(Q_E, b) = p_E \mathbb{E}[\min(Q_E, D_E)] - w_E^* Q_E
$$

s.t. \quad w_E^* Q_E \leq b.

Note that $\Pi_{NRL}^* = \Pi_{TS}^*|_{w_E = w_E^*, \theta = 0}$.

From (I) and (II), and given the manufacturer’s report of $\theta$, the retailer’s optimal expected profit if she postpones her procurement decision is:

$$
\Pi_{POS}^* = \theta \Pi_{RL}^* + (1 - \theta) \Pi_{NRL}^*.
$$

Thus, when the retailer is deciding whether or not to postpone her procurement decision, she must compare the optimal expected profit from not postponing, $\Pi_{POS}^* = \Pi_{TS}^*$, with the optimal expected profit from postponing, $\Pi_{POS}^*$. Specifically, the retailer’s optimal expected profit in the flexible system (assuming a truthful report of $\theta$) is:

$$
\Pi_{FS}^* = \max\{\Pi_{NPOS}^*, \Pi_{POS}^*\}.
$$

The following proposition provides the details of the retailer’s optimal investment timing and procurement decisions in the flexible system, assuming that the manufacturer truthfully reports his forecast.

\textbf{Proposition 2.2.} Assume the manufacturer truthfully reports his forecast of $\theta$. In the flexible system, the retailer prefers not to postpone if and only if: $\Pi_{NPOS}^* = \Pi_{TS}^* > \Pi_{POS}^*$, in which case the optimal procurement quantity of Product $E$, $Q_{E}^{TS^*}$, and optimal cash reserve, $M^{TS^*}$, are given by Proposition
2.1. Otherwise, the retailer uses the postponement option, yielding unique optimal quantities:

\[
\begin{align*}
Q_{NRL}^* &= F_E^{-1}(\frac{p_E - w^*_N}{w^*_E}), & \text{if Product } N \text{ is not released,} \\
p_E (1 - F_E(Q_{E}^{RL*}')) - w^*_E (p_N (1 - F_N(\frac{b - w^*_E Q_{E}^{RL*}}{w^*_N}) - w_N)) &= 0, & \text{if Product } N \text{ is released.} \\
Q_{NL}^{RL*} &= \frac{b - w^*_E Q_{E}^{RL*}}{w^*_N}.
\end{align*}
\]

From Proposition 2.2, the following corollary states the manufacturer’s preference in the flexible system.

**Corollary 2.2.** In the flexible system, the manufacturer prefers that the retailer postpones her procurement decision.

This result is quite intuitive and flows from two factors: first, if the retailer postpones, she retains her entire endowment until learning whether Product \( N \) is released, thus, she has more cash on hand in case it is released; second, postponement increases the input price of the rival product (Product \( E \)), making it less attractive, and consequently Product \( N \) more attractive.\(^6\)

To investigate the incremental value of having a postponement option (for the retailer), we define the postponement premium, \( V(\theta, b) \), as follows:

\[
V(\theta, b) = \Pi^*_{POS} - \Pi^*_N POS = \Pi^*_{POS} - \Pi^*_N S,
\]

where for a given \( \theta \) and \( b \), \( V(\theta, b) > 0 \) indicates that the retailer gains from having the postponement option, whereas \( V(\theta, b) \leq 0 \) implies that this option is of no value (and not used by the retailer). The following two lemmas provide further insight into the behaviour of the retailer in the flexible system.

**Lemma 2.1** (Postponement Interval). For any given endowment \( b \), if there exists a \( \tilde{\theta} \in [0, 1] \) for which \( V(\tilde{\theta}, b) \geq 0 \), then there exists a unique postponement interval \([\theta^b_{min}, \theta^b_{max}] \subset [0, 1]\), such that:

(i) for all \( \theta \in [\theta^b_{min}, \theta^b_{max}] \) the retailer postpones her procurement decision of Product \( E \),

(ii) for all \( \theta \notin [\theta^b_{min}, \theta^b_{max}] \) she does not postpone.

If \( V(\theta, b) < 0 \), \( \forall \theta \), the retailer never postpones (and we set \( \theta^b_{min} = \theta^b_{max} = 0 \)).

\(^6\)We should clarify that in some other settings, there may be situations where the upstream firm may prefer that the downstream party invests early. For example, Benetton was known to have a flexible capability in the form of the ability to dye sweaters after they have been sewn. However, since the cost of having this flexible capacity was high, the firm was trying to force the retailers to order most of the quantity before the beginning of the season, and only make minor adjustments during the main season.
As Lemma 2.1 demonstrates, the retailer prefers not to postpone for all \( \theta > \theta_{max}^b \) and \( \theta < \theta_{min}^b \). Intuitively, for higher (\( \theta > \theta_{max}^b \)) and lower (\( \theta < \theta_{min}^b \)) values of \( \theta \), the expected gain from postponing - as a result of the resolved uncertainty about the release of Product \( N \) and thus the ability to invest with complete information - is less than its expected cost - because of purchasing Product \( E \) at a higher wholesale price. Specifically, when \( \theta \) is close to 1, the retailer is almost certain that Product \( N \) will be released and thus can set aside the required reserve \( M^* \) at the early stage, benefiting from the lower procurement cost of Product \( E \). Indeed, if \( \theta \approx 1 \), then \( \Pi_{POS}^r \approx \Pi_{RL}^r = \Pi_{TS}^r \big|_{w_E = w_T} < \Pi_{TS}^r \). On the other hand, if \( \theta \) is close to 0, the likelihood that Product \( N \) will be released is too low to justify paying higher procurement cost for Product \( E \) under the postponement option. Thus, the postponement premium is, once again, negative.

The following result provides further precision on the relationship between the postponement premium and retailer’s endowment.

**Lemma 2.2.** For any given \( \theta \), the postponement premium decreases as \( b \) increases. Consequently, if the length of the postponement interval is greater than zero, it shrinks as \( b \) increases.

Lemma 2.2 implies that for sufficiently large endowments, the loss from postponing dominates its gain for all possible \( \theta \) values (i.e., the length of the postponement interval is equal to zero). Thus, for such endowments, the retailer’s incentive to postpone totally disappears. This is also quite intuitive: if the endowment is sufficiently large, the retailer can do both: procure Product \( E \) at the advantageous early price and have sufficient on-hand cash reserve should Product \( N \) be released. What is more interesting is that the dependence of the postponement interval \([\theta_{min}^b, \theta_{max}^b]\) on \( b \) creates “uncertainty” with respect to the “best” \( \theta \) value for the manufacturer. Recall that the manufacturer always wants the retailer to postpone, thus would like to have \( \theta \) inside the postponement interval. However, a value inside this interval for one level of \( b \) may be outside of the interval for another. As we will see in Section 2.3, this mechanism is crucial for the existence of the “truthful reporting” equilibrium.

**Example 1.** The key insights can be illustrated by a numerical example in which the demands for Products \( E \) and \( N \) are uniformly distributed on intervals \([0,50]\) and \([0,150]\), respectively. Assume, \( w_E = 4 \), \( w_T = 5 \), \( w_N = 5 \), \( p_E = 20 \), and \( p_N = 35 \). Figure 2.3 depicts the postponement premium, \( V(\theta, b) \), for different endowments. Several observations are noteworthy. First, as stated in Lemma 2.1, given endowment \( b \), there exists a unique postponement interval \([\theta_{min}^b, \theta_{max}^b]\) (the interval corresponding to the region under the curve and above \( V = 0 \)) where the retailer postpones if and only if \( \theta \in [\theta_{min}^b, \theta_{max}^b] \). Second, in line with Lemma 2.2, the postponement premium and thus the postponement interval shrinks
as $b$ increases. Moreover, for large enough endowments ($b = 500$), $V(\theta, b) < 0 \ \forall \theta$, thus, the retailer never postpones (the postponement cost outweighs the benefit).

### 2.3. The Cheap Talk Game

In this section, we explore the forecast communication game played between the manufacturer and the retailer in both the traditional and the flexible systems. In both settings, to analyze the effect of uncertainty on the manufacturer’s ability to share his forecast with the retailer, we consider two information structures: one in which the manufacturer has perfect information about the retailer’s endowment (which we label as one-sided asymmetric information structure as the only private information, the true value of $\theta$, is held by the manufacturer), and the other in which the manufacturer is uncertain about the retailer’s endowment (labeled as the two-sided asymmetric information structure). To avoid repetition, we set-up the communication game in the more general two-sided asymmetric information structure, and treat the one-sided structure as a special case.

In the two-sided asymmetric information structure, the manufacturer has a private forecast of the likelihood of the release of his product, $\theta$ (which we denote as his “type”), whereas the retailer has
perfect and private information about her endowment (her type) \( b \). To highlight our main insights, we make a standard simplifying assumption (see, for example, [28], [52], [6], and [63]) that allows the manufacturer and the retailer to be of one of two types: \( L \) or \( H \). That is, we assume \( \theta \in \Theta = \{ \theta^L, \theta^H \} \), with \( 0 \leq \theta^L < \theta^H < 1 \), and \( b \in B = \{ b^L, b^H \} \), where \( 0 < b^L < b^H \). Let \( \alpha \in \Delta(\Theta \times B) \) be a strictly positive joint prior probability distribution over the profile of the types, which is regularly assumed to be common knowledge. We make the information structure general by allowing the manufacturer’s type and the retailer’s type to be correlated. This could, for instance, reflect the observation that a manufacturer with a highly unpredictable product is more likely to interact with a smaller retailer (a retailer with low endowment), or be due to the fact that the manufacturer would exert more effort to release his product when he believes that the retailer has a high endowment (see, for example, [42], [71], [90], and [62] for articles in the economics literature that consider a correlated information structure). The above structure reduces to a one-sided asymmetric information structure by assuming that \( \alpha(b^i|\theta) \in \{0, 1\} \), \( i \in \{ L, H \} \) (i.e., conditional on \( \theta \), \( \alpha(\cdot) \) is a point mass on \( b^L \) or \( b^H \), implying that the retailer’s type is known to the manufacturer).

To formally define the communication game, we must first define the set of feasible reports that the manufacturer can use. As is common in the cheap talk literature, and without loss of generality ([77]), we assume that the set of feasible reports is equal to \( \Theta \). We define the manufacturer’s reporting strategy \( S : \Theta \rightarrow \Theta \), where \( S(\theta) = \hat{\theta} \) if the manufacturer reports \( \hat{\theta} \) when his private forecast is \( \theta \). (Note that we use symbol \( \hat{\theta} \) when we refer to a report, and \( \theta \) when we refer to the manufacturer’s private forecast.) Finally, let \( K : \Theta \times B \rightarrow \mathbb{R}^+ \) denote the cash holding strategy of the retailer, where \( K(\hat{\theta}, b) = M \) if the retailer retains the cash reserve \( M \) when her endowment is \( b \) and the manufacturer’s report is \( \hat{\theta} \).

Note that we only consider the retailer’s cash holding strategy since the procurement quantities can be automatically derived from the cash reserve (see Proposition 2.1 and 2.2).

The timing of the communication game is as follows: (i) at the beginning of the time horizon both the manufacturer and the retailer observe their type; (ii) having observed their types \( \theta \) and \( b \), the manufacturer has a (subjective) probability distribution \( \alpha_m(\cdot|\theta) \) over the retailer’s types, whereas the retailer has a (subjective) prior probability distribution \( \alpha_r(\cdot|b) \) over the manufacturer’s types. As in most incomplete information models, we assume that the manufacturer and the retailer share common prior beliefs expressed by conditional probabilities: \( \alpha_m(b^i|\theta) = \frac{\alpha(\theta^i, b^i)}{\sum_{y \in B} \alpha(\theta^i, y)} \) and \( \alpha_r(\theta^i|b) = \frac{\alpha(\theta^i, b)}{\sum_{y \in \Theta} \alpha(y, b)} \), \( i \in \{ L, H \} \); (iii) the manufacturer sends the report \( \hat{\theta} = S(\theta) \) to the retailer; (iv) the retailer observes the report \( \hat{\theta} \) and makes her cash holding decision \( M = K(\hat{\theta}, b) \); the rest of the events follows as described in Sections 2.2.1 and 2.2.2.

In this communication game, it is easy to see that truthful reporting by the manufacturer benefits
the retailer. However, it may not be beneficial for the manufacturer to truthfully reveal his forecast. Specifically, the manufacturer may want to distort his report of $\theta$ to ensure that the retailer holds abundant cash for the possible release of his product. Given such an incentive, the retailer may not find the report credible, and instead rely on her prior belief. The issues of incentives and credibility in such communication are due to two main factors: first, given the nature of the manufacturer’s private information, the retailer cannot verify (ex post) whether the manufacturer truthfully reported his private forecast; second, by distorting his report, the manufacturer does not incur a direct cost. This type of communication with costless, nonbinding, and unverifiable messages belongs to a class of games with incomplete information, referred to as “cheap talk” games ([41]).

As is common in cheap talk games, the equilibrium concept we employ to obtain a solution to our manufacturer-retailer communication game is Perfect Bayesian Nash Equilibrium (PBNE) ([46]). In our context, the PBNE requires that: (1) given the retailer’s cash holding strategy ($K(\hat{\theta}, b)$) and posterior belief ($\alpha_r(\theta|\hat{\theta}, b)$), the manufacturer will not deviate from his reporting strategy ($S(\theta)$), if it maximizes his total expected profit; and (2) given the manufacturer’s reporting strategy ($S(\theta)$) and her own endowment ($b$), the retailer chooses a cash reserve ($K(\hat{\theta}, b)$) that maximizes her total expected profit, updating her belief about the manufacturer’s private information using Bayes’ rule. To formalize this discussion, recall that the manufacturer’s profit is denoted by $\Pi_m(.)$, while the retailer’s expected profit in the traditional and flexible systems is represented by $\Pi_{TS}^r(.)$ and $\Pi_{FS}^r(.)$, respectively. The following definition formalizes the PBNE in our context.

**Definition 2.1.** (Perfect Bayesian Nash Equilibrium) The manufacturer’s reporting strategy $S(\cdot)$, the retailer’s cash holding strategy $K(\cdot)$, and the retailer’s posterior (updated) belief $\alpha_r(\theta|\hat{\theta}, b)$ constitute a perfect Bayesian Nash equilibrium if:

1. For each $\theta \in \Theta$, $S(\theta) = \arg\max_{\hat{\theta} \in \Theta} \sum_{b \in B} \alpha_m(b|\theta)\Pi_m(K(\hat{\theta}, b))$,

2. For each $\hat{\theta} \in \Theta$,

$$K(\hat{\theta}, b) = \begin{cases} \arg\max_{M} \sum_{\theta \in \Theta} \alpha_r(\theta|\hat{\theta}, b)\Pi_{TS}^r(M, \theta, b), & \text{in the traditional system,} \\ \arg\max_{M} \sum_{\theta \in \Theta} \alpha_r(\theta|\hat{\theta}, b)\Pi_{FS}^r(M, \theta, b), & \text{in the flexible system,} \end{cases}$$

3. The retailer’s belief is consistent, such that whenever possible, she updates her belief about the manufacturer’s private information using Bayes’ Rule, i.e., $\alpha_r(\theta|\hat{\theta}, b) = \frac{I(S(\theta) = \hat{\theta})\alpha_r(\theta|b)}{\sum_{\nu \in \Theta} I(S(\nu) = \hat{\theta})\alpha_r(\nu|b)}$, where $I\{\cdot\}$ is the indicator function.
In the above definition, the first condition implies that the manufacturer sends a report that maximizes his expected profit, taking the retailer’s cash holding strategy as given. The second and third conditions indicate that the retailer responds optimally (determines the optimal cash reserve that maximizes her total expected profit) to each possible report, using Bayes’ rule to update her prior belief (whenever possible).

Our focus is on identifying equilibrium conditions in which the manufacturer truthfully reports his forecast and the retailer believes him (it is without loss of generality that we restrict our analysis to such equilibrium (see [14])). We refer to such a PBNE as a “truthful revealing” equilibrium, which is formally defined as follows.

Definition 2.2. (Truthful Revealing Equilibrium) A truthful revealing equilibrium is a PBNE in which:

1. \( S(\theta) = \theta, \quad \forall \theta \in \Theta \),
2. \( \alpha_r(\theta|\theta,b) = 1, \quad \forall (\theta,b) \in (\Theta,B) \).

Definition 2.2 states that in a truthful revealing equilibrium, the manufacturer sends a truthful report (condition 1) and the retailer believes him (condition 2).

2.3.1 Communication in the Traditional System

We begin by examining the information exchange in the traditional system assuming the manufacturer has perfect information about the retailer’s endowment. Recall from Proposition 2.1 that in the traditional system, the retailer’s cash reserve is increasing in \( \theta \). Moreover, Corollary 2.1 indicates that in this system, the manufacturer’s profit is increasing in the retailer’s cash reserve \( M \). Consequently, to ensure that the retailer holds abundant cash for the possible release of his product, the manufacturer has an incentive to inflate his report of \( \theta \). Anticipating such an incentive, the retailer would not consider the manufacturer’s report credible, regardless of whether he is telling the truth or not. This discussion is formalized in the following lemma.

Lemma 2.3. In the traditional system with one-sided asymmetric information structure, the manufacturer always wants the retailer to believe that \( \theta = \theta^H \). Anticipating such an incentive, the retailer ignores the report in equilibrium, relying only on her prior belief. Thus, no truthful revealing equilibrium exists.

Driving the above result is the manufacturer’s incentive to inflate, and the retailer’s inability to verify ex post the truth of the report. Therefore, no matter what reporting strategy the manufacturer uses (note that the manufacturer’s equilibrium reporting strategy is not unique) the retailer would simply
ignore it and make her cash holding decision regardless of the report, relying only on her prior belief. Such an equilibrium, in which the manufacturer’s report is independent of his type and the retailer’s strategy is independent of the report, is referred to as a “babbling” equilibrium ([44]).

The result of Lemma 2.3 can be generalized to the case where the manufacturer is uncertain about the retailer’s endowment. Specifically, since for each possible endowment \( b \in B \), the manufacturer’s profit is increasing in the retailer’s cash reserve (Corollary 2.1), his expected profit (expectation taken over retailer types) is also increasing in it. Therefore, in order to induce the retailer to hold more cash, he has an incentive to inflate his report of \( \theta \), regardless of his knowledge of \( b \). Knowing this incentive, the retailer does not consider the report credible. The following theorem formalizes this argument.

**Theorem 2.1.** In the traditional system, the manufacturer always wants the retailer to believe that \( \theta = \theta^H \), regardless of his knowledge about \( b \). Knowing this incentive, the retailer ignores the report in equilibrium, determining her optimal strategy based on her prior belief. Therefore, in the traditional system, no truthful revealing equilibrium exists. In fact, the only PBNE is the babbling equilibrium.

The above discussion illustrates that in the traditional system, information cannot be shared credibly because of incentive misalignment between the manufacturer and the retailer. More precisely, since \( \theta \) only influences the retailer’s cash reserves, and since the reserve is increasing in \( \theta \), the manufacturer’s incentive is to always inflate \( \theta \). The question we next answer is whether procurement flexibility can align the incentives of the parties to allow credible forecast sharing.

Recall that the crucial difference between the traditional and the flexible system is that, in the latter, the value of \( \theta \) influences two decisions of the retailer: whether to postpone and how much cash reserve to hold for Product \( N \) in case the postponement option is not chosen. As we will show, in this case the manufacturer is no longer certain whether inflating or deflating \( \theta \) is beneficial.

### 2.3.2 Communication in the Flexible System

Before presenting our main result, to gain intuition, we first discuss the equilibrium outcome in the flexible system when the manufacturer has perfect knowledge about the retailer’s endowment. Recall from Lemma 2.1 that for each possible endowment \( b \in B \) there exists a unique postponement interval \( [\theta^b_{\text{min}}, \theta^b_{\text{max}}] \) (which can be empty), such that for all \( \theta \in [\theta^b_{\text{min}}, \theta^b_{\text{max}}] \) the retailer postpones her procurement decision. Moreover, Corollary 2.2 indicates that the manufacturer always prefers that the retailer postpones her procurement decision. Given these results, and assuming that the manufacturer has perfect knowledge of \( b \), we obtain the following cases:
**Completely Aligned Incentives:** if \( \Theta = \{ \theta^L, \theta^H \} \subset [\theta^b_{\min}, \theta^b_{\max}] \), the retailer postpones her procurement decision for both \( \theta^L \) and \( \theta^H \). In this case, the manufacturer has no incentive to misreport, thus, a truthful revealing equilibrium exists. While this shows that procurement flexibility can align the incentives of the two firms such that they achieve their mutually beneficial outcome even in the absence of communication (the mutually beneficial outcome can never be achieved in the traditional system), it also indicates that truthful reporting is supported in equilibrium because the shared information does not alter the retailer’s postponement decision (i.e., it is as if no communication has taken place). Such a truthful revealing equilibrium is referred to as a **noninfluential** equilibrium ([9] and [31]).

**Misaligned Incentives:** if \( \Theta \not\subset [\theta^b_{\min}, \theta^b_{\max}] \), there may be **partial** incentive misalignment, where only one manufacturer type falls within the postponement interval; or **full** incentive misalignment, where both types fall outside the postponement interval. In both cases, the manufacturer has an incentive to distort his report. Specifically, in case of partial misalignment, the manufacturer type (either \( \theta^L \) or \( \theta^H \)) that does not belong to the postponement interval \( [\theta^b_{\min}, \theta^b_{\max}] \) would alter his report - he inflates his report if \( \theta < \theta^b_{\min} \) and deflates it when \( \theta > \theta^b_{\max} \) - to induce the retailer to postpone her procurement decision. In the complete misalignment case, the retailer never postpones, thus, the manufacturer always wants to report \( \theta^H \) (as in the traditional system). In both cases, anticipating the manufacturer’s incentive to deceive, the retailer will not believe the report.

**Lemma 2.4.** In the flexible system with one-sided asymmetric information structure:

(i) If \( \Theta \subset [\theta^b_{\min}, \theta^b_{\max}] \), a truthful revealing (noninfluential) equilibrium exists.

(ii) If \( \Theta \not\subset [\theta^b_{\min}, \theta^b_{\max}] \), no truthful revealing equilibrium exists. In this case, the only PBNE is the babbling equilibrium.

While the above result seems somewhat encouraging (due to the emergence of truthful reporting in equilibrium in one case), it also shows that in the flexible system with one-sided asymmetric information, truthful reporting is supported in the equilibrium only when knowledge of \( \theta \) would not affect the retailer’s postponement decision. We next focus on our main result and show that procurement flexibility coupled with the manufacturer’s uncertainty about the retailer’s endowment can, under fairly plausible conditions, result in the manufacturer truthfully sharing his private forecast, i.e., a truthful revealing and **influential** equilibrium can emerge.\(^7\)

\(^7\)In an influential equilibrium ([9] and [31]), the manufacturer’s report influences the retailer’s postponement decision, i.e., the retailer’s postponement decision (in equilibrium) is not the same for all reports.
Main Result: As stated above, in the flexible system, the manufacturer always prefers that the retailer postpones her procurement decision (Corollary 2.2). Moreover, for each of the retailer’s possible endowment $b \in B$ there exists a unique postponement interval $[\theta_{min}^b, \theta_{max}^b]$ such that the retailer postpones her procurement decision if and only if $\theta \in [\theta_{min}^b, \theta_{max}^b]$ (Lemma 2.1), where the length of this interval shrinks as $b$ increases (Lemma 2.2). These results indicate that (unlike in the traditional system where the manufacturer always prefers $\theta^H$) in the flexible system the manufacturer’s preferred reports (of his type) may differ for various endowments (i.e., he may wish to report a particular type for one realization of $b$, while a different type may be reported for another realization). Specifically, if the manufacturer has perfect knowledge of $b$, we know from Lemma 2.4 that under partial incentive misalignment he may want to inflate, deflate, or not change his report to induce the retailer to postpone, i.e., given the postponement interval $[\theta_{min}^b, \theta_{max}^b]$, he would inflate his report (report $\theta^H$) if $\theta < \theta_{min}^b$, deflate (report $\theta^L$) if $\theta > \theta_{max}^b$, and truthfully report if $\theta \in [\theta_{min}^b, \theta_{max}^b]$. Furthermore, the manufacturer would inflate his report if he knows that the retailer’s endowment is sufficiently large (the reason for this is that the retailer never postpones for such endowments - see Lemma 2.2 and the complete misalignment case of Lemma 2.4). This mixture of preferred reports coupled with the endowment uncertainty results in the manufacturer not being able to perfectly predict the retailer’s reaction to each of his reports (because of the changing postponement interval as $b$ changes). This uncertainty may induce the manufacturer to truthfully reveal his private information.

The above discussion indicates that truthful reporting rests on the manufacturer’s mixture of preferred reports. Therefore, to formally characterize the equilibrium, we quantify this mixture of preferences.

**Definition 2.3.** (Preferred Report Mixture Measure) We define the preferred report mixture measure, $\delta$, as follows:

$$
\delta = \frac{\Delta \Pi(\theta, b^H)}{\Delta \Pi(\theta, b^L)} = \frac{\Pi_m(K(\theta^H, b^H)) - \Pi_m(K(\theta^L, b^H))}{\Pi_m(K(\theta^L, b^L)) - \Pi_m(K(\theta^H, b^L))}
$$

A negative $\delta$ implies that the manufacturer has a “preference reversal”, where he prefers to report a high $\theta$ for one value of $b$ and a low $\theta$ for the other. A positive $\delta$ means that the manufacturer prefers one value of $\theta$ over the other for all possible $b$ values, i.e., he prefers to report $\theta^L$ if $\Delta \Pi(\theta, b^i) < 0$, $i = L, H$, while preferring $\theta^H$ when $\Delta \Pi(\theta, b^i) > 0$, $i = L, H$. A special case is $\Delta \Pi(\theta, b^L) = \Delta \Pi(\theta, b^H) = 0$, where the manufacturer is indifferent between reporting $\theta^L$ and $\theta^H$ (because the retailer postpones her decision for all possible reports, i.e., $V(\theta, b) \geq 0$, $\forall (\theta, b) \in \{\Theta \times B\}$). This degenerate case generalizes the completely aligned incentives case of Lemma 2.4, thus, the implication of Lemma 2.4 is still applicable: a truthful revealing but noninfluential equilibrium exists.
The next theorem characterizes the necessary and sufficient conditions that guarantee the existence of a truthful revealing and influential equilibrium.

**Theorem 2.2.** *In the flexible system with two-sided asymmetric information structure, a truthful revealing (and influential) equilibrium exists if and only if:*

1. \( \delta < 0 \),
2. \( \frac{\alpha(\theta^L, b^L)}{\alpha(\theta^H, b^L)} \geq |\delta| \),
3. \( \frac{\alpha(\theta^H, b^H)}{\alpha(\theta^L, b^H)} \geq \frac{1}{|\delta|} \).

*In such an equilibrium, the manufacturer truthfully reports his forecast of \( \theta \), and the retailer believes him.*

As stated earlier, truthful reporting relies on the manufacturer being uncertain about the retailer’s reaction (or equivalently uncertain about the postponement interval). Using this fact, Theorem 2.2 establishes conditions on the prior probability distribution over the profile of the types for which credible information sharing is supported in equilibrium. The theorem implies that for a truthful revealing and influential equilibrium to exist, it must be that: (0) a postponement option must exist for the retailer, (1) the manufacturer is uncertain about the retailer’s endowment, (2) the manufacturer has preference reversal, where he prefers to send different reports for different endowments (condition (I)), and (3) there is relatively a greater probability that the manufacturer-retailer types match (conditions (II) and (III)).

The type matching conditions are plausible and are consistent with the observation that a manufacturer producing a highly unpredictable product (which can be denoted as a low type manufacturer) is more likely to interact with a smaller retailer.\(^8\) Theoretically, conditions (II) and (III) indicate that the prior distribution \( \alpha(\theta, b) \) satisfies the monotone likelihood ratio property ([72]), a standard assumption in mechanism design and contract theory (see, for example, [73], [54], [24], and [19]). In our context, an interpretation of this property is that high types are more likely to match because the manufacturer exerts more effort to release his product (thus is more likely to be a high type) when the retailer’s endowment is high. Collectively, the three conditions imply that, for each manufacturer type, his expected profit (the expectation is over \( b \)) is higher under truthful reporting.

It is noteworthy that when both conditions (II) and (III) are violated (i.e., \( \frac{\alpha(\theta^L, b^L)}{\alpha(\theta^H, b^L)} < |\delta| \) and \( \frac{\alpha(\theta^H, b^H)}{\alpha(\theta^L, b^H)} < \frac{1}{|\delta|} \)), a revealing and influential equilibrium still exists. However, in this equilibrium, the

---

\(^8\)E.g., for the first generation iPhone, a product with high uncertainty, Apple negotiated an exclusive deal with the smallest of the major wireless communications service providers (at the time), Cingular Wireless, later acquired by AT&T (see “The Untold Story: How the iPhone Blew Up the Wireless Industry”, Wired News, Jan., 2008.). Moreover, it is common to test innovative and new products through smaller outlets.
manufacturer is not truthful, i.e., his reporting strategy is $S(\theta^L) = \theta^H$ and $S(\theta^H) = \theta^L$, and the retailer reverses the report, i.e., $\alpha_r(\theta|\theta, b) = 0, \forall(\theta, b)$. Moreover, if only one of the two conditions (II) or (III) is violated, truthful reporting is not supported in the equilibrium. Specifically, when $\frac{\alpha(\theta^L, b^L)}{\alpha(\theta^H, b^H)} \geq |\delta|$ and $\frac{\alpha(\theta^L, b^L)}{\alpha(\theta^H, b^H)} < \frac{1}{|\delta|}$, the manufacturer is better off by reporting $\theta^L$, while if $\frac{\alpha(\theta^L, b^L)}{\alpha(\theta^H, b^H)} < |\delta|$ and $\frac{\alpha(\theta^H, b^H)}{\alpha(\theta^H, b^L)} > \frac{1}{|\delta|}$, he prefers $\theta^H$.

We next construct a simple example to better demonstrate the intuition behind Theorem 2.

**Example 2.** Suppose $\Theta = \{\theta^L = 0.5, \theta^H = 0.85\}$ and $B = \{b^L = 350, b^H = 450\}$, and assume that the other parameters are identical to those in Example 1. The resulting postponement premiums for each endowment are shown in Figure 2.4. The top (bottom) plot shows the postponement premium when the retailer’s endowment is low (high). The figure illustrates the postponement interval for the two endowments. As can be seen, when the retailer’s endowment is low ($b^L$), she gains from having the postponement option (she postpones when $\theta$ falls within the postponement interval $[\theta_{min}^{b^L}, \theta_{max}^{b^L}]$), whereas for a high endowment ($b^H$), the postponement option is useless (she never postpones). The retailer’s best response, assuming that she trusts the manufacturer and makes her cash holding decision according to the manufacturer’s report is:

$$\begin{align*}
    \text{if } b = b^L: & \quad \begin{cases}
    \text{if } S(\theta) = \theta^L \rightarrow \text{postpone and hold } K(\theta^L, b^L) = b^L \\
    \text{if } S(\theta) = \theta^H \rightarrow \text{not postpone and hold } K(\theta^H, b^L) < b^L
    \end{cases}
\end{align*}$$

Figure 2.4: Example 2
From Section 2.2 we can infer: 

\[ b_L = K(\theta^L, b_L) > K(\theta^H, b_L) \text{ and } K(\theta^L, b^H) < K(\theta^H, b^H). \]

Thus, since the manufacturer’s profit is increasing in the retailer’s cash reserve (see Corollaries 2.1 and 2.2), 

\[ \Pi_m(K(\theta^L, b_L)) > \Pi_m(K(\theta^H, b_L)) \text{ and } \Pi_m(K(\theta^L, b^H)) < \Pi_m(K(\theta^H, b^H)). \]

Therefore, if the manufacturer had perfect information about the retailer’s endowment, he would send the report \( S(\theta) = \theta^L \) if \( b = b_L \), while reporting \( S(\theta) = \theta^H \) if \( b = b^H \). This implies that the manufacturer has a preference reversal, hence a negative \( \delta \). In this example, the preferred report mixture measure is \( \delta = \frac{\Pi_m(K(\theta^H, b^H)) - \Pi_m(K(\theta^L, b^L))}{\Pi_m(K(\theta^H, b^L)) - \Pi_m(K(\theta^L, b^L))} = \frac{2850 - 2770}{2360 - 2440} = -1. \) Therefore, from Theorem 2.2, a truthful revealing equilibrium exists if and only if:

\[
\begin{align*}
\frac{\alpha(\theta^L, b^L)}{\alpha(\theta^L, b^H)} &\geq 1, \\
\frac{\alpha(\theta^H, b^H)}{\alpha(\theta^H, b^L)} &\geq 1;
\end{align*}
\]

that is, there is a greater probability that the manufacturer-retailer types match, i.e., \( \alpha(\theta^L, b^L) \geq \alpha(\theta^L, b^H) \text{ and } \alpha(\theta^H, b^H) \geq \alpha(\theta^H, b^L) \), or equivalently, the prior distribution \( \alpha(\theta, b) \) satisfies the monotone likelihood ratio property. Therefore, under these conditions (see the discussion above on the plausibility of the conditions) the manufacturer is on average better off by truthfully sharing his forecast.

### 2.4. Conclusion

Using a stylized supply chain model, consisting of a retailer who solicits forecast information from a manufacturer, our goal is to examine conditions where the manufacturer would have an incentive to truthfully share his forecast with the retailer. The motivation for this analysis is the observation that while theoretical supply chain models often predict the need for complicated contracts or reputation mechanism to achieve credible information sharing, simple contracts are common in practice, and firms often use nonbinding and unverifiable reports (cheap talk) to communicate.

In our model, we find that if the manufacturer can anticipate the retailer’s reaction to his report (as in the traditional system), he has an incentive to be strategic on how he shares his forecast. Predicting such a behaviour, the retailer cannot find the report credible. Thus, in such a setting, credible forecast sharing is not supported in equilibrium.

Our main result, however, suggests that when the retailer has alternative options (e.g., a procurement
postponement option), the manufacturer is no longer able to perfectly predict the retailer’s reaction to his report. This uncertainty can, under fairly plausible conditions, induce the manufacturer to truthfully share his forecast with the retailer. Therefore, we show that, under perfect rationality, and in the absence of a complex contract or a reputation mechanism, if the supply chain entity providing the report is uncertain about the receiving firm’s reaction to its report, it may, in equilibrium, credibly share its private forecast. This suggests that nonbinding and unverifiable forecast sharing can indeed improve supply chain coordination.

There are several interesting avenues for future research. A natural extension is to consider the communication game to include more than two states for $\theta$ and $b$, i.e., $\theta \in \{\theta_E, ..., \theta_N\}$ and $b \in \{b_E, ..., b_N\}$. While most of our qualitative insights continue to hold (i.e., equilibrium information sharing is possible in the flexible system, while information sharing is blocked in equilibrium in the traditional system), such an extension would make the analyses more complex since it requires the introduction of partially revealing (also known as partial-pooling) equilibria, in which the manufacturer partially reveals his information. A second possible direction is to extend the model to allow the retailer to place multiple orders of the existing product. In particular, by considering a low enough fixed procurement/production/capacity cost of Product $E$, we can endogenize the number times the retailer procures this product. In this case, a revealing equilibrium may still exist, but the analysis is more cumbersome.
2.5. Appendix

Proof of Proposition 2.1. Due to no slackness, we can substitute $Q_N$ by $b - \frac{w_E}{w_N}Q_E$ in the objective function. To prove part (a), we show that the total expected profit, $\Pi_T^{TS}(Q, M, \theta, b)$, is concave in $Q_E$: it is sufficient to show that both $\Pi_T^{ES}(Q_E)$ and $\Pi_T^{SN}(Q_N) = \Pi_T^{SN}(b - \frac{w_E}{w_N}Q_E)$ are concave in $Q_E$ (positive weighted sum of concave functions is concave). The first and second derivatives of these two functions are (using the Leibniz integral rule):

\[
\begin{align*}
\frac{\partial \Pi_T^{ES}(Q_E)}{\partial Q_E} &= p_E (1 - F_E(Q_E)) - w_E, \\
\frac{\partial^2 \Pi_T^{ES}(Q_E)}{\partial Q_E^2} &= -p_E f_E(Q_E) < 0, \\
\frac{\partial \Pi_T^{SN}(Q_N)}{\partial Q_N} &= -\frac{w_E}{w_N} (p_N (1 - F_N(b - \frac{w_E}{w_N}Q_E)) - w_N), \\
\frac{\partial^2 \Pi_T^{SN}(Q_N)}{\partial Q_N^2} &= -p_N \left(\frac{w_E}{w_N}\right)^2 f_N(b - \frac{w_E}{w_N}Q_E) < 0.
\end{align*}
\]

The second derivative of both functions are negative, thus, both functions are concave. Since $\theta > 0$, we can conclude that $\Pi_T^{TS}(Q, M, \theta, b)$ is also concave in $Q_E$. This implies that the optimal order quantity of Product $E$, $Q_T^{E*}$, is the unique solution of the first-order condition (FOC):

\[p_E (1 - F_E(Q_T^{E*})) - w_E - \theta \frac{w_E}{w_N} (p_N (1 - F_N(b - \frac{w_E}{w_N}Q_T^{E*})) - w_N) = 0.\]

The result of part (b) follows from no slackness and by the definitions of $M$. Finally, the comparative statics of part (c) is derived by employing the implicit function theorem.

Proof of Corollary 2.1. The proof follows from $\frac{\partial \Pi_m}{\partial \theta} = (w_N - c) \frac{\partial Q_N}{\partial \theta}$, and from Proposition 2.1(c) $\frac{\partial Q_T^{S*}}{\partial \theta} > 0$ and $\frac{\partial Q_T^{S*}}{\partial b} > 0$.

Proof of Proposition 2.2. Since $\Pi_T^{RL}(Q, b)$ and $\Pi_T^{NR}(Q_E, b)$ are special cases of $\Pi_T^{TS}(Q, M, \theta, b)$, i.e., when $w_E = w^*_E$ and $\theta = 1$ or $\theta = 0$, they are both concave in $Q_E$ (see proof of Proposition 2.1), and thus have a unique maximum. Therefore, as proven in Proposition 2.1, the optimal order quantities can be derived by setting the FOC equal to zero.

Proof of Corollary 2.2. Since $\frac{\partial \Pi_m}{\partial \theta} > 0$, it is sufficient to show that $Q_T^{RL*} \geq Q_T^{SN*}$. From Proposition 2.1 we know $\frac{\partial Q_T^{S*}}{\partial \theta} > 0$, thus, $Q_T^{S*}\big|_{\theta=1} \geq Q_T^{SN*}$. Moreover, since $\frac{\partial Q_T^{SN*}}{\partial w_E} > 0$, thus $Q_T^{SN*}\big|_{w_E=w^*_E} \geq Q_T^{SN*}\big|_{w_E<w^*_E}$. Therefore:

\[Q_T^{RL*} = Q_T^{SN*}\big|_{\theta=1, w_E=w^*_E} > Q_T^{SN*}.\]
Proof of Lemma 2.1. The proof will follow by showing the for any given $b$: (i) the postponement premium at the two extreme values of $\theta$ is negative, and (ii) the postponement premium is concave in $\theta$. We first show (i), that is $V(\theta = 0, b) < 0$ and $V(\theta = 1, b) < 0$.

- $\theta = 0$: Since $\frac{\partial \Pi^T_S|_{\theta=0}}{\partial w_E} = \frac{\partial \Pi^T_S|_{\theta=0}}{\partial w_E} < 0$, and given the assumption $w_E < w'_E$ we have

$$\Pi^T_S|_{\theta=0} = \Pi^N_{RL} + \Pi^T_S|_{\theta=0} < \Pi^T_S|_{\theta=0}.$$

Therefore, $V(\theta = 0, b) = \Pi^N_{RL} - \Pi^T_S|_{\theta=0} < 0$.

- $\theta = 1$: Since $\frac{\partial \Pi^T_S|_{\theta=1}}{\partial w_E} < 0$, and given $w_E < w'_E$ we have

$$\Pi^T_S|_{\theta=1} = \Pi^N_{RL} + \Pi^T_S|_{\theta=1} < \Pi^T_S|_{\theta=1}.$$

Therefore, $V(\theta = 1, b) = \Pi^N_{RL} - \Pi^T_S|_{\theta=1} < 0$.

To conclude the proof, we now prove (ii), that is $V(\theta, b)$ is concave in $\theta$. Since $\frac{\partial^2 \Pi^{POS*}}{\partial \theta^2} = 0$, we have $-\frac{\partial \Pi^{T_S*}}{\partial \theta} = -\frac{\partial \Pi^{T_S*}}{\partial \theta}$. Therefore, we must show that $\Pi^{T_S*}$ is convex in $\theta$. The derivative of $\Pi^{T_S*}$ with respect to $\theta$ yields:

$$\frac{\partial \Pi^{T_S*}}{\partial \theta} = \Pi^{T_S*} + \frac{\partial \Pi^{T_S*}}{\partial \theta} = (p_E(1 - F_E(Q^T_{E})) - w_E - \frac{\theta w_E}{w_N}p_N(1 - F_N(b - w_E Q^T_{E})) - w_N) = \Pi^{T_S*},$$

where the second equality holds because (from Proposition 2.1) $p_E(1 - F_E(Q^T_{E})) - w_E - \frac{\theta w_E}{w_N}p_N(1 - F_N(b - w_E Q^T_{E})) = 0$. Thus, $\frac{\partial^2 \Pi^{T_S*}}{\partial \theta^2} = \frac{\partial \Pi^{T_S*}}{\partial \theta} < 0$, which implies $\frac{\partial^2 V}{\partial \theta^2} < 0$. 

Proof of Lemma 2.2. We must show that $V(\theta, b)$ is monotonically decreasing in the endowment. Since we assume that the retailer’s endowment is not sufficient to procure the optimal unconstrained order quantity of both products, but is sufficient to procure the optimal unconstrained order quantity of any single product, the derivative of $V(\theta, b)$ with respect to $b$ yields

$$\frac{\partial V(\theta, b)}{\partial b} = \theta \frac{\partial \Pi^{N_{RL}}}{\partial b} - \frac{\partial \Pi^{T_S*}}{\partial b},$$

which after simplifying is

$$\frac{\partial V(\theta, b)}{\partial b} = \frac{\theta p_E}{w_E} (F_N(Q^T_{N}) - F_N(Q^N_{RL})).$$

From the proof of Corollary 2.2 we know $Q^N_{RL} > Q^T_{N}$, which implies $F_N(Q^N_{RL}) > F_N(Q^T_{N})$. Thus, $\frac{\partial V(\theta, b)}{\partial b} < 0$. 

**Proof of Lemma 2.3.** Assume that the retailer trusts the manufacturer and makes her cash holding decision according to the manufacturer’s report. For any given $b$, let $K^*(\hat{\theta} = \theta^i, b) = \arg\max_M \Pi_T^S(M, \theta^i, b)$, $i \in \{L, H\}$ denote her optimal cash reserve given the received report $\hat{\theta}$. We use contradiction to prove that truthful reporting is not supported in the equilibrium; thus assume a truthful revealing equilibrium exit. From Definition 2.2, in such an equilibrium, $S(\theta_L) = \theta_L$ and $S(\theta_H) = \theta_H$. This implies that for any given $b$ the following conditions must hold:

$$\Pi_m(K^*(\theta^i, b)) \geq \Pi_m(K^*(\theta^j, b)), \quad \forall i, \forall j \in \{L, H\}, \ j \neq i \quad (A1)$$

Moreover, from Proposition 2.1, $Q^TS^*_L$ and $Q^TS^*_H$ are the unique solution to the following two FOCs:

$$p_E(1 - F_E(Q^TS^*_L|\theta_L)) - w_E - \theta_L w_E (p_N(1 - F_N(\frac{b - w_E Q^TS^*_L|\theta_L}{w_N})) - w_N) = 0,$$

$$p_E(1 - F_E(Q^TS^*_H|\theta_H)) - w_E - \theta_H w_E (p_N(1 - F_N(\frac{b - w_E Q^TS^*_H|\theta_H}{w_N})) - w_N) = 0.$$  

From Proposition 2.1 we also know that $\frac{\partial Q^TS^*_L}{\partial \theta_L} < 0$. Therefore, $Q^TS^*_L > Q^TS^*_H$. Since $K^*(\theta^L, b) = b - w_E Q^TS^*_L|\theta_L$ and $K^*(\theta^H, b) = b - w_E Q^TS^*_H|\theta_H$, then $K^*(\theta^L, b) < K^*(\theta^H, b)$. Finally, since $\frac{\partial \Pi_m}{\partial \theta} > 0$ (Corollary 2.1), we can conclude that $\Pi_m(K^*(\theta^H, b)) > \Pi_m(K^*(\theta^i, b)) \quad \forall j \in \{L, H\}$, which contradicts (A1). Therefore, truthful reporting is not supported in equilibrium. \hfill \Box

**Proof of Theorem 2.1.** Assume that the retailer trusts the manufacturer and makes her cash holding decision according to the manufacturer’s report. Let $K^*(\hat{\theta} = \theta^i, b) = \arg\max_M \Pi_T^S(M, \theta^i, b)$, $i \in \{L, H\}$ denote her optimal cash holding decision given the report. As in Lemma 2.3, we use contradiction and thus assume that truthful revealing equilibrium exists. From Definitions 2.1 and 2.2, in such an equilibrium, the following conditions hold:

$$\sum_{b \in B} \alpha_m(b|\theta^i) \Pi_m(K^*(\theta^i, b)) \geq \sum_{b \in B} \alpha_m(b|\theta^j) \Pi_m(K^*(\theta^j, b)), \quad \forall i, j \in \{L, H\}, \ j \neq i \quad (A2)$$

Using the same steps as in Lemma 2.3 we observe that $K^*(\theta^L, b) < K^*(\theta^H, b), \forall b \in B$, and thus $\Pi_m(K^*(\theta^H, b)) > \Pi_m(K^*(\theta^L, b)), \forall b \in B$. This implies:

$$\sum_{b \in B} \alpha_m(b|\theta^i) \Pi_m(K^*(\theta^H, b)) \geq \sum_{b \in B} \alpha_m(b|\theta^j) \Pi_m(K^*(\theta^j, b)), \quad \forall i, j \in \{L, H\}$$

which contradicts (A2). Therefore, truthful reporting is not supported in equilibrium. Since every cheap-talk game has a babbling equilibrium ([35]), we can conclude that the only PBNE in the traditional system is the babbling equilibrium. \hfill \Box

**Proof of Lemma 2.4.** We must look at the following two possible outcomes:
Completely Aligned Incentives ($\Theta \subseteq [\theta_{b_{\min}}, \theta_{b_{\max}}]$) Under this condition, the retailer will postpone her procurement decision for all possible reports. Moreover, from Lemma 2.2 we know that the manufacturer always prefers that the retailer postpones her procurement decision. From these two facts we observe that the manufacturer has no incentive to distort his report, since the retailer’s optimal strategy will always maximize his profit. Knowing that the manufacturer has no incentive to misreport, the retailer does not have any incentive not to trust the manufacturer.

Misaligned Incentives ($\Theta \not\subseteq [\theta_{b_{\min}}, \theta_{b_{\max}}]$) Under this condition we must look at two cases: (i) full incentive misalignment $\Theta \cap [\theta_{b_{\min}}, \theta_{b_{\max}}] = \emptyset$ and (ii) partial incentive misalignment $\Theta \cap [\theta_{b_{\min}}, \theta_{b_{\max}}] \neq \emptyset$.

We will show that in both cases no revealing PBNE exists:

Case (i): When $\Theta \cap [\theta_{b_{\min}}, \theta_{b_{\max}}] = \emptyset$, regardless of the manufacturer’s type, the retailer will not postpone her procurement decision. Furthermore, from Section 2.2.2 we know that in the flexible system, when the retailer does not postpone, her optimization problem is the same as that of the traditional system. Thus, under this condition, the manufacturer’s incentive to inflate his report still remains. Therefore, the proof of Lemma 2.3 holds in this case; thus, we can conclude that in the flexible system when $(\Theta \cap [\theta_{b_{\min}}, \theta_{b_{\max}}] = \emptyset)$, no revealing PBNE exists.

Case (ii): We next show that no revealing PBNE exists in the partial incentive misalignment case.

As in case (i), for $\theta \not\in [\theta_{b_{\min}}, \theta_{b_{\max}}]$, the manufacturer’s profit in the flexible system is equal to his profit in the traditional system. Moreover, for $\theta \in [\theta_{b_{\min}}, \theta_{b_{\max}}]$, the manufacturer’s profit in the flexible system is equivalent to that of the traditional system with $\theta = 1$ and $w_e = w_e^r$, i.e., $\Pi_m(MFS^*) = \Pi_m(MTS^*|\theta = 1 & w = w_e^r)$. It follows that we can construct a payoff equivalent type space for the manufacturer: (1) let $\theta^{H'}$ denote an artificial manufacturer type with prior probability equal to the prior probability of the manufacturer type that falls within the postponement interval, and with the payoff $\Pi_m(MTS^*|\theta = 1 & w = w_e^r)$; (2) define the sorted set $\Theta' = \{(\Theta - \{\theta \in \Theta | \theta \in [\theta_{b_{\min}}, \theta_{b_{\max}}]\}) + \theta^{H'}\}$, which is comprised of the non-postponement element of $\Theta$ plus $\theta^{H'}$. We can see that for the manufacturer, the sets $\Theta$ and $\Theta'$ are payoff equivalent, i.e., the manufacturer profit for the non-postponement type is the same in both sets, while his profit for the postponement type is equal to his profit for $\theta^{H'}$. Therefore, for truthful reporting to be supported in equilibrium for the types in $\Theta$, it must also be supported for the types in $\Theta'$. However, following the proof of Lemma 2.3 we know that no such equilibrium exists for $\Theta'$; thus, no revealing PBNE exists in the partial incentive misalignment case.

Proof of Theorem 2.2. From Definitions 2.1 and 2.2, necessary conditions for truthful reporting to be supported in equilibrium are:
\[
\sum_b \alpha_m(b|\theta^L)\Pi_m(K(\theta^L, b)) \geq \sum_b \alpha_m(b|\theta^L)\Pi_m(K(\theta^H, b)) \\
\sum_b \alpha_m(b|\theta^H)\Pi_m(K(\theta^H, b)) \geq \sum_b \alpha_m(b|\theta^H)\Pi_m(K(\theta^L, b))
\]

i.e., the left hand side is the manufacturer's expected profit from truthfully revealing his private information \((S(\theta) = \theta)\) while the right side is his expected profit from misreporting \((S(\theta) \neq \theta)\). The above two conditions can be re-written as:

\[
\alpha_m(b^L|\theta^L)[\Pi_m(K(\theta^L, b^L)) - \Pi_m(K(\theta^H, b^L))] \geq \alpha_m(b^H|\theta^L)[\Pi_m(K(\theta^H, b^L)) - \Pi_m(K(\theta^L, b^L))], \quad (A3) \\
\alpha_m(b^H|\theta^H)[\Pi_m(K(\theta^H, b^H)) - \Pi_m(K(\theta^L, b^H))] \geq \alpha_m(b^L|\theta^H)[\Pi_m(K(\theta^L, b^L)) - \Pi_m(K(\theta^H, b^L))]. \quad (A4)
\]

To proceed, we need the following lemma.

**Lemma 2.5.** In the Flexible System and under non-degenerate postponement strategies,

\[
\Pi_m(K(\theta^L, b^L)) < \Pi_m(K(\theta^H, b^L)) \Rightarrow \Pi_m(K(\theta^H, b^H)) > \Pi_m(K(\theta^L, b^H)).
\]

**Proof.** Under non-degenerate postponement strategies (by non-degenerate strategies we mean that one retailer type’s optimal postponement strategy is to postpone whereas the other type’s strategy is not to postpone), \(\Pi_m(K(\theta^L, b^L)) < \Pi_m(K(\theta^H, b^L))\) implies that when \(b = b^L\), only \(\theta^H\) belongs to the postponement interval. Moreover, from Lemma 2.2 we know that the postponement interval shrinks as \(b\) increases. Therefore, when \(b = b^H\), the retailer will never postpone if \(\theta = \theta^L\), while it may or may not postpone if \(\theta = \theta^H\). If it does postpone, then, from Corollary 2.2, \(\Pi_m(K(\theta^H, b^H)) > \Pi_m(K(\theta^L, b^H))\). If she does not postpone, then from Corollary 2.1, \(\Pi_m(K(\theta^H, b^H)) > \Pi_m(K(\theta^L, b^H))\).

From the above lemma and from conditions (A3) and (A4), truthful reporting can be supported in equilibrium if:

\[
\Pi_m(K(\theta^L, b^L)) - \Pi_m(K(\theta^H, b^L)) > 0, \\
\Pi_m(K(\theta^H, b^H)) - \Pi_m(K(\theta^L, b^H)) > 0.
\]

Or, equivalently, the preferred report mixture measure must be negative:

\[
\delta = \frac{\Pi_m(K(\theta^H, b^H)) - \Pi_m(K(\theta^L, b^H))}{\Pi_m(K(\theta^H, b^L)) - \Pi_m(K(\theta^L, b^H))} < 0.
\]

By the definition of \(\delta\), and since \(\alpha_m(b|\theta) = \frac{\alpha(\theta, b)}{\sum_{\theta \in \Theta} \alpha(\theta, y)}\), conditions (A3) and (A4) reduce to:

\[
\frac{\alpha(\theta^L, b^L)}{\alpha(\theta^L, b^H)} \geq |\delta|, \\
\frac{\alpha(\theta^H, b^L)}{\alpha(\theta^H, b^H)} \geq |\frac{1}{\delta}|.
\]
Finally, the conditions of the theorem are sufficient since they incorporate the retailer’s best-response into the manufacturer’s expected payoff so that truth-telling for the manufacturer is evaluated conditional on the retailer taking her best-reply in the fully revealing (separating) equilibrium in which $\alpha_r(\theta|\hat{\theta} = \theta, b) = 1$. 

\[ \square \]
Chapter 3

Can Supply Chain Firms Benefit From Competition?

For firms in imperfectly competitive markets, it is generally presumed that fewer competitors are preferred to more competitors. At first blush, then, it seems strange that Ford should have supported the 2009 U.S. Federal Government bailout of General Motors and Chrysler. If fewer competitors are preferred to more, why not push to let some of its competitors fail so that Ford can reap the rewards of reduced competition? We argue in this chapter that the initial presumption that fewer competitors are preferred to more need not be correct. Indeed, we show that, if the prices that producers face for inputs vary with the number of competitors, the opposite may be the case. This linkage between (upstream) input prices and downstream competition can explain Ford’s preference for the bailout.¹

We explore these issues in a supply chain model in which, initially, there are a fixed number, \( n \), of final goods producers - the downstream market. To allow for changes in competition at both the extensive and intensive margins, we assume that each producer provides a single product to each of two related market segments. For example, in the case of vehicles, one can think of this as each producer selling automobiles in one segment and vans and sports utility vehicles (vans/SUVs) in the other. In this world, competition can change because a producer exits the market entirely (exits both market segments) or because a producer rationalizes production, exiting one of the two market segments.

Without loss of generality we assume that, to manufacture any product, each final goods producer requires an (aggregate) input that is provided by an imperfectly competitive, upstream market. For

¹Ford was concerned that the disappearance of GM and/or Chrysler would be to Ford’s detriment because it would significantly reduce the demand for upstream auto parts with the consequent effect of reducing parts suppliers and the general competitive vigor in the upstream parts market which could potentially sink the entire auto industry(see [49] and “Ford Would Have Shut Without Auto Bailouts”, Bloomberg, Oct. 9, 2012).
simplicity, we assume that manufacturers in the upstream market produce a homogeneous product under increasing returns to scale.\(^2\) The price of each input is determined via a free-entry, Cournot-Nash equilibrium (CN). The combination of scale economies and free-entry in the input market imply that the input price varies with the number of competitors in the downstream market. Together, these results portray a situation that is consistent with the one described by those involved in the auto bailout.

Within this framework, we show that any reduction in competition in the downstream market – either exit of a producer from both market segments or from just one – results in an increase in the price of the input provided by the upstream market. All else equal, this rise in the input price increases costs and reduces profits of each of the remaining producers. As is standard, complete exit of a producer from the market increases the profits of all remaining producers at constant input prices. Overall, a reduction in competition due to the exit of a competitor from the market is more likely to reduce the profits of the remaining firms the larger are either the fixed or variable costs of production in the upstream market and the smaller is the markup over input costs in the downstream market. The reason is that (i) with significant scale economies in the upstream market, any reduction in input demand due to the exit of a downstream firm results in a large increase in the input price and (ii) with a small markup, any reduction in competition leads to only a small increase in profits. The overall impact of exit from the downstream market in this case is that profits of the remaining producers decline. This effect is enhanced if the number of downstream producers is small.

Rather than exiting completely from the market, some competitors may rationalize production by exiting one of the two market segments.\(^3\) In this situation, the two forces at play above continue to be at play here: (i) the reduction in demand for the input due to the exit of producers from one of the market segments causes the input price to rise and (ii) competition declines in the sub-market in which exit occurs and so profits rise for the remaining producers in that market segment. The input price rise will not be as large in this case since (rationalizing) producers exit only a single market segment rather than exiting entirely. At the same time, the positive impact on profits is not as large either. One reason for this is obvious: the (rationalizing) producers exit only one market segment rather than two and so the remaining producers obtain a profit increase only in that market segment with fewer competitors. Another, less obvious reason is that downstream producers with only a single product increase output and now compete more vigorously in their remaining market segment. The reason is that the pecuniary

\(^2\)In [76], the claim is that beginning in 2009 to 2011, the auto-parts sector faced overcapacity and, for survival, continuing suppliers reduced break-even output numbers. Further, it states that there was pressure from vehicle manufacturers for wholesale price reductions. While our simple characterization abstracts from many of the detailed features of this auto supply sector, open-entry (and exit) with economies of scale are critical features.

\(^3\)For Chrysler and GM to restructure, the two firms where obligated to cut capacity and eliminate weaker brands ([49]), e.g., GM phase out its Hummer vehicle in 2010, while Chrysler intends to replace its Dodge Grand Caravan minivan and the Dodge Avenge with one single crossover vehicle.
externality that encouraged these firms to mute their price competition when they competed in both market segments has been eliminated. Faced with this enhanced competition in the market segment, the profits of the other producers decline. We show in this case that rationalization leads to an increase in all downstream prices but that prices increase by more in the market segments in which rationalization occurs than in the other segment. We also show that (i) the larger the extent of rationalization and (ii) the greater the scale economies in the upstream market, the more likely that rationalization leads to a decline in the profits of the continuing, multi-product producers. This profit decline is smaller, however, than the decline that would be observed if the rationalizing firms exited the market entirely.

To get some sense of whether our model can be a credible explanation for Ford’s support of the auto industry bailout, we calibrate the model to pre- and post-recession data. In the pre-recession world, our calibration reveals that the equilibrium mark-up on variable cost in the downstream market is sufficiently large relative to the mark-up in the upstream market that exit would increase profits for the continuing firms. In the post-recession world, our calibration reveals just the opposite. Specifically, at post-recession parameter values, downstream mark-ups are small and the exit of GM and/or Chrysler would significantly lower Ford’s profits by dramatically raising parts prices. A bailout of one or both companies, with or without product rationalization, increases Ford’s profits relative to the no bailout case.

These issues are not just relevant for the automotive sector. A report from RTI International and National Institute of Standards and Technology ([59]) outlines vertical supply chain issues for a sample of U.S. industries. This sample includes not only the U.S. automotive sector but industries in the electronics sector such as the production of computers and peripheral equipment. Computer manufacturers, for instance, face trade-offs between the intensity of competition in their downstream market and a sufficiently large market to support a vigorously competitive upstream supply market. In this or any other case, our analysis can be applied. Further, since the condition that determines whether profits of the downstream producers increase or decrease with exit requires only information on mark-ups in the upstream and downstream markets, the analysis is easily implemented empirically.

This is not the first work to be concerned with the structure of supply chains in a competitive environment. Other papers study the impact of entry and exit on competition and market performance. The focus of these papers, however, is different from ours, and are mainly directed towards issues of vertical integration. [69] present an early discussion of successive oligopoly and vertical integration.

---

4 For example, Sony is selling its Vaio PC business in order to focus on smartphones and tablets. Sony’s claim is that the way in which tablets are replacing PCs for many users played a part in its decision (see “Bye Bye-O to Vaio: Sony Is The Canary In Two Coal Mines”, Forbes, June, 2014). Similarly, HP is also shifting resources from PCs to tablets (see “HP Shifting Focus from PCs to Tablets”, dailytech, Feb, 2013).
Chapter 3. Can Supply Chain Firms Benefit From Competition?

[51] show that vertical integration of successive oligopolists leads to higher output and prices. [36] and [37] study vertical integration and possible disintegration and whether exclusive dealing contracts can exclude equally efficient upstream producers and cartelize the downstream industry.

The competition of supply chain firms has also been studied in the operations management literature. [29] study multi-tier supply chains with an assembly structure. [80] study the efficiency of price-only contracts in three tier supply chains under demand uncertainty. [70] study competition between separate supply chains and investigate the effects of contract leadership on the efficiency of the supply chain. [1] study the effects of product and retailer differentiation, stochastic demand, and retailer risk aversion on the decentralized supply chain equilibrium and its efficiency. In contrast to the papers described above, our goal is to examine those conditions where downstream firms would have a self-interest in promoting the sustainability of their downstream competitors. Moreover, while the above papers consider single-product firms, we consider multi-product firms, which allows us to capture the changes in competition at both extensive and intensive margins.

The chapter is organized as follows. In Section 3.1, we set out the downstream and upstream equilibrium in a two tier supply chain model. Section 3.2 analyses the setting where downstream producers completely exit the market. Section 3.3 contains the analysis where downstream producers rationalize production by exiting only one of the two market segments. Section 3.4 develops results consistent with Ford’s revealed preference to support that U.S. Government bailout of Ford’s two U.S. auto rivals. Section 3.5 reports on our calibration results. A final section contains concluding remarks. Proofs of the principal results appear in a Mathematical Appendix and detail of the inputs to our calibration exercise appear in a Calibration Appendix.

3.1. The Model

In what follows, we present a stylized model of supply chains. The model contains the salient features of the vehicle and electronic markets while at the same time being simple enough to derive analytical results. More importantly, it allows us to provide intuition on the elements of supply chain environments that drive our results.

3.1.1 Downstream Market

The downstream market consists of two market segments, $A$ and $B$. For simplicity, we assume that any product in segment $A$ is a perfect substitute for any other product in segment $A$; the same is assumed to be true for segment $B$. Products in segment $A$ are assumed to be imperfect substitutes
Chapter 3. Can Supply Chain Firms Benefit From Competition?

for products in segment B (and vice versa). For example, this assumption captures the idea that one variety of SUV is very similar to any other variety of SUV (and similarly for automobiles) while SUVs and automobiles are much less substitutable. Initially, we assume that there are \( n \) symmetric downstream multi-product firms that produce products for both market segments. For now, we imagine that the value of \( n \) is given exogenously and defer the determination of the value of \( n \) to Section 3.4. We define \( q^A_{i,(n,n)}, q^B_{i,(n,n)} \) as the output of firm \( i \) in market segments A and B respectively, with the subscript \((n,n)\) capturing the initial configuration of \( n \) producers providing a product in each market segment. The variable \( Q^A_{(n,n)} = \sum_i q^A_{i,(n,n)} \) (\( Q^B_{(n,n)} = \sum_i q^B_{i,(n,n)} \)) gives total output in segment A (B).

In both market segments A and B, we assume that demand is linear and symmetric between markets, with the (inverse) demand functions given by:

\[
\begin{align*}
p^A_{(n,n)} &= \alpha - \beta Q^A_{(n,n)} - \gamma Q^B_{(n,n)} \\
p^B_{(n,n)} &= \alpha - \beta Q^B_{(n,n)} - \gamma Q^A_{(n,n)}.
\end{align*}
\]

Note that \( \beta > \gamma \), capturing our assumption that products in segment A are imperfect substitutes for products in segment B (and vice versa). We assume that each firm, \( i \), chooses output levels \( q^A_{i,(n,n)}, q^B_{i,(n,n)} \) to maximize total profit from selling to the two segments. Firms produce at constant (and identical) marginal cost. Market quantities and prices are given by the Nash equilibrium quantities and corresponding market clearing prices.

3.1.2 Upstream Market

We assume that production in the downstream market requires a single input and that this input is provided via an imperfectly competitive upstream market. The products of all upstream manufacturers are assumed identical and to sell at a market-determined wholesale price, \( w_{(n,n)} \), which the downstream firms take as given (i.e., there is no bargaining over the per unit exchange price). We further normalize output so that one unit of output produced in the downstream industry requires one unit of the (homogeneous) input produced in the upstream market. Together, these assumptions imply that the (constant) marginal cost of every downstream producer is \( w_{(n,n)} \). The upstream industry is an open-entry industry. Upstream firms are symmetric and produce with a fixed cost \((F)\) and a constant marginal cost \((c)\) technology described by the cost function \( C_i = cq^u_i + F \), where \( q^u_i \) is the output of upstream firm \( i \), common to each upstream firm. Each firm chooses output level \( q^u_i \) to maximize profits from sale to

---

\(^5\)These demand functions are derived from quadratic utility functions (see [86] and [58]).

\(^6\)If production is Hicks-Leontief, then the assumption of a single input is without loss of generality. See the Appendix for more details on this issue.
the downstream industry. Since upstream firms simultaneously produce their input prior to downstream firms producing their output, the upstream equilibrium is subgame perfect Nash.

3.1.3 Equilibrium

**Downstream Equilibrium**

Given any input price \( w(n,n) < \alpha \) set by the upstream market, the profit of any downstream firm \( i \) is:

\[
\pi_{i,(n,n)} = \pi_{i,(n,n)}^A + \pi_{i,(n,n)}^B = (p_{(n,n)}^A - w(n,n))q_{i,(n,n)}^A + (p_{(n,n)}^B - w(n,n))q_{i,(n,n)}^B \quad i \in 1, ..., n,
\]

where \( p_{(n,n)}^A \) and \( p_{(n,n)}^B \) are given by the system (3.1) above. The following lemma characterizes the Nash equilibrium quantity, price and profit, as a function of \( w(n,n) \).

**Lemma 3.1.** The Nash equilibrium quantity, price and profit, as a function of \( w(n,n) \) are given by:

\[
q_{i,(n,n)}^A(w(n,n), n) = q_{i,(n,n)}^B(w(n,n), n) = \frac{\alpha - w(n,n)}{(\beta + \gamma)(n + 1)}, \quad (3.2)
\]

\[
p_{(n,n)}^A(w(n,n), n) = p_{(n,n)}^B(w(n,n), n) = p^* = \frac{\alpha + nw(n,n)}{n + 1} \quad (3.3)
\]

\[
\pi_{i,(n,n)}^*(w(n,n), n) = 2\pi_{i,(n,n)}^A(w(n,n), n) = 2\pi_{i,(n,n)}^B(w(n,n), n) = \frac{2(\alpha - w(n,n))^2}{(\beta + \gamma)(n + 1)^2} \quad (3.4)
\]

As expected, for an exogenous \( w(n,n) \), each downstream firm’s profit increases with a decrease in the number of downstream producers (\( -\frac{\partial \pi_{i,(n,n)}^*}{\partial n} > 0 \)) and decreases in the input price (\( -\frac{\partial \pi_{i,(n,n)}^*}{\partial w(n,n)} < 0 \)). That is, any disturbance upstream (with economies of scale) that increases wholesale input prices will negatively impact the output and profits of firms in both segments of the downstream market.

**Upstream Equilibrium**

Since upstream firms move sequentially prior to downstream firms, each upstream manufacturer considers the impact of its quantity decision on wholesale price, \( w(n,n) \), when choosing its profit maximizing output.

From equation (3.2) above, then, market demand for the upstream manufacturers is given by the inverse

---

7 The restriction that \( w(n,n) < \alpha \) is required to guarantee that the downstream firms make positive profits.

8 For a given \( w(n,n) \), the own-market price elasticity and cross-market price elasticity at the equilibrium are:

\[
\varepsilon_{AA} = -\frac{\beta(\alpha + nw(n,n))}{(\beta - \gamma)n(\alpha - w(n,n))},
\]

\[
\varepsilon_{AB} = -\frac{\gamma(\alpha + nw(n,n))}{(\beta - \gamma)n(\alpha - w(n,n))} = \frac{-\gamma}{\beta} \varepsilon_{AA}
\]

For \( \beta > \gamma \), own price elasticity exceeds cross-price effects, as expected.
upstream demand function:
\[ w_{(n,n)} = \alpha - \delta (q_{u1}^n + ... + q_{u_n}^n), \]  \( (3.5) \)

where \( \delta \equiv (\beta + \gamma)(n + 1)/2n \), \( n_u \) is the total number of firms that the upstream market can support, and \( q_i^u \) \( (i = 1, ..., n_u) \) is the output of each of the \( n_u \) upstream firms.

Under open-entry, each active manufacturer must operate such that its average cost at its equilibrium output level equals the wholesale price given by (3.5) evaluated at the equilibrium market output level.

**Lemma 3.2.** The equilibrium wholesale input price is:
\[ w^*(n,n) = c + \sqrt{\frac{F(\beta + \gamma)(n + 1)}{2n}} \]  \( (3.6) \)

Note that the wholesale price is a mark-up on cost, where the mark-up depends on the number of downstream firms \( (n) \), firm-specific demand \( (\beta + \gamma) \) and the fixed cost \( (F) \). It is straightforward to check that the equilibrium wholesale price is decreasing in the number of downstream firms \( (\frac{\partial w^*(n,n)}{\partial n} < 0) \) and increasing in upstream firm costs \( (\frac{\partial w^*(n,n)}{\partial F} > 0, \frac{\partial w^*(n,n)}{\partial c} > 0) \). Based on these facts, one can also show that:

**Proposition 3.1.** The equilibrium number of upstream manufacturers is increasing in the number of downstream manufacturers.

To understand this result, imagine that the number of downstream producers falls. From (3.5) above, the market demand for the input rotates inward and becomes steeper. All active upstream manufacturers now earn negative profits and so exit occurs. Because the market demand curve is steeper, the new equilibrium is established at a point where the average cost curve is steeper, implying a smaller output level for each remaining active manufacturer and a higher wholesale price. In terms of the vehicle example, for Ford, departure of domestic competitors in the downstream segments of the U.S. automobile market has the obvious downstream effect of increasing Ford’s outputs and profits at constant wholesale prices. However, the exit of competitors drives wholesale prices higher so that the departure has the potential to harm Ford’s profits overall. An identical result prevails for computer manufacturers: the departure of one downstream computer manufacturer has an obvious competitive benefit to the remaining producers but there is the impact on the upstream supply market that harms the remaining downstream producers. This proposition also highlights Ford’s concern regarding the reduction of parts suppliers should GM and/or Chrysler disappear and the impact on computer manufacturers from any reduction of upstream suppliers/assemblers as downstream producers leave the market. We analyze these issues in the section that follows.
3.2. Firms Exit Both Market Segments: Competition and Network Effects

With the equilibrium results, we can determine the impact that exit of competitors in the downstream market has on profits of continuing downstream producers. We show that exit has two effects: (i) downstream firm profits increase as a result of the reduction in the degree of downstream market competition and (ii) downstream firm profits decrease as a result of an increase in the wholesale price. Specifically, from above, we have that:

\[
\frac{d\pi^*_i, (w_{(n,n)}, n)}{dn} = 2 \left[ \frac{\partial \pi^*_i, (w_{(n,n)}, n)}{\partial n} \right] + 2 \left[ \frac{\partial \pi^*_i, (w_{(n,n)}, n)}{\partial w_{(n,n)}} \right] \frac{dw^*_{(n,n)}}{dn}
\]

The first term is the Competition Effect (CE); it captures the change in firm \(i\)'s profit due to both a reduction in the number of competitors that \(i\) faces and changes in the outputs of its remaining competitors. We saw above that the sign of this term is negative, so that a reduction in the number of competitors increases \(i\)'s profit. The second term is the Network Effect (NE); it captures the change in firm \(i\)'s profit due to changes in the wholesale price of the input generated by reduced downstream demand. From above, we have that this term is positive, so that exit from the downstream market increases the value of \(w^*_{(n,n)}\) and thus reduces \(i\)'s profit. The overall impact on \(i\)'s profit of exit from the downstream market, then, depends on whether or not the positive competition effect more than offsets the negative network effect.

From \(\pi^*_i, (w_{(n,n)}, n)\) (equation (3.4)) the condition under which exit reduces the profits of continuing downstream firms is:

\[
\frac{\partial \pi_{i, (n,n)}(w^*_{(n,n)}, n)}{dn} > 0 \text{ if and only if } 2(n + 1)(p^* - w^*_{(n,n)}) < \frac{1}{n}(w^*_{(n,n)} - c)
\]

Finally, by utilizing the expression for \(w^*_{(n,n)}\) from equation (3.6) above, we have the following result:

**Proposition 3.2.** \(\frac{\partial \pi_{i, (n,n)}(w^*_{(n,n)}, n)}{dn} > 0 \text{ if and only if }\)

\[
2(n + 1)(p^* - w^*_{(n,n)}) < \frac{1}{n}(w^*_{(n,n)} - c)
\]
cost in the downstream market. The term \((w^*_{(n,n)} - c)\) gives the equilibrium mark-up of the wholesale price over the marginal cost of the upstream product. Exit from the downstream market reduces the profit of continuing firms when the mark-up over cost of the input is large relative to the mark-up on the downstream product. This situation arises when \(w^*_{(n,n)}\) is large relative to \(p^*\). (From equation (3.3), as \(w^*_{(n,n)}\) increases, \(p^*\) increases less than proportionately.) In this case, the reduction in downstream competition due to exit results in only a small profit increase (\(p^* - w^*_{(n,n)}\) is small) while it results in even higher wholesale prices, creating large profit reductions (\(w^*_{(n,n)} - c\) is large).

We expect the value of \(w^*_{(n,n)}\) to be large when \(F\) and \(c\) are both large. In this situation, average cost for input producers is high and the average cost curve is steeply sloped: there are significant scale economies in the upstream market. As a result, the reduction in input demand due to the exit of downstream firms results in a large increase in the input price. Combining this with a situation in which profits in the downstream market are small (the mark-up is small) leads to an overall effect where exit from the downstream market causes the profit of continuing producers to decline. This effect is enhanced if the number of downstream producers is small.

### 3.3. Firms Exit One Market Segment: Competition and Network Effects

An alternative to full exit from the downstream market is that some firms rationalize their product portfolio and exit only one of the two downstream markets. As we note above, the qualitative difference between total and partial exit is that with partial exit, the firm no longer faces any internal cross-substitution effect between its own products in the two market segments. The absence of this pecuniary externality causes a producer to increase its output in the remaining sector where it continues to operate. This rationalization is, in fact, what occurred as a result of the bailout of GM and Chrysler. Again, we can examine the impact that such a rationalization has on the profits of other firms through both direct competition and upstream effects. In the analysis that follows, we (i) treat product rationalization as an exogenous event and (ii) assume that market demand is stable and unchanging. In this section, we make no attempt to account for any change in demand or new products that might motivate a firm to consolidate its brand offerings in one market segment. In the calibration section, however, we will take demand changes into account.

Without loss of generality, we consider a setting where \(k\) of the multi-product firms rationalize production and arbitrarily exit market segment \(B\) (we label the firms making this adjustment as the
rationalizing firms). In this case, there are \((n - k)\) symmetric multi-product firms who are present in both market segments \(A\) and \(B\) (firms \(i = 1, \ldots, n - k\)), and \(k\) firms who exit market \(B\) and specialize in market \(A\) (firms \(j = n - k + 1, \ldots, n\)). The subscript \((n, n - k)\) denotes this market configuration (i.e., \(n\) refers to the number of firms in market \(A\) and \(n - k\) refers to the number of firms in market \(B\)).

In this market configuration, the inverse demand functions (given by equation (3.1)) become:

\[
p^A_{(n, n-k)} = \alpha - \beta \left( \sum_{i=1}^{n-k} q^A_i(w_{(n, n-k)}) + \sum_{j=n-k+1}^{n} q^B_j(w_{(n, n-k)}) - \gamma \sum_{i=1}^{n-k} q^A_i(w_{(n, n-k)}) \right)
\]

\[
p^B_{(n, n-k)} = \alpha - \beta \left( \sum_{i=1}^{n-k} q^B_i(w_{(n, n-k)}) - \gamma \sum_{i=1}^{n-k} q^A_i(w_{(n, n-k)}) + \sum_{j=n-k+1}^{n} q^A_j(w_{(n, n-k)}) \right)
\]

Now, for any given value of \(k\) and fixed value of \(w_{(n, n-k)}\), the Nash equilibrium outputs are as follows.

**Lemma 3.3.** For a given \(w_{(n, n-k)}\), the Nash equilibrium outputs of each multi-product firm \(i \in 1, \ldots, n - k\) are:

\[
q^A_{i,(n, n-k)}(w_{(n, n-k)}, n, k) = \frac{(\beta(n - k + 1) - \gamma k)(\alpha - w_{(n, n-k)})}{\beta(n + 1 - k)(\beta + \gamma)(n + 1)}
\]

\[
q^B_{i,(n, n-k)}(w_{(n, n-k)}, n, k) = \frac{\alpha - w_{(n, n-k)}}{(\beta + \gamma)(n + 1 - k)}
\]

The output of each rationalizing firm \(j \in n - k + 1, \ldots, n\) in market \(A\) is:

\[
q^A_{j,(n, n-k)}(w_{(n, n-k)}, n, k) = \frac{\alpha - w_{(n, n-k)}}{\beta(n + 1)}
\]

One can check that, for a given value of \(w_{(n, n-k)}\), the output in market segment \(A\) of each multi-product firm is decreasing and concave in \(k\), while multi-product firm output in market segment \(B\) is increasing and convex in \(k\). An immediate implication of these facts is that, for any fixed value of \(w\) and any value of \(k\), \(q^A_{i,(n, n-k)}(w, n, k) < q^A_{i,(n,n)}(w, n)\) while \(q^B_{i,(n, n-k)}(w, n, k) > q^B_{i,(n,n)}(w, n)\). For the rationalizing firms that now only produce in segment \(A\), \(q^A_{j,(n, n-k)}(w, n, k) > q^A_{j,(n,n)}(w, n)\). The reason that the rationalizing firms expand output in segment \(A\) is that, while, prior to rationalization, such an expansion damaged the profits of these firms in segment \(B\), this no longer happens given they have withdrawn from this market segment. It can also be shown that, for any fixed value of \(w\) and any value of \(k\), total output in segment \(A\) under rationalization is greater than the pre-rationalization output for segment \(A\), i.e., \(Q^A_{(n, n-k)}(w, n, k) > Q^A_{(n,n)}(w, n)\). For segment \(B\), the opposite holds: \(Q^B_{(n, n-k)}(w, n, k) > Q^B_{(n,n)}(w, n)\).}

\[9\] That is, \(\partial q^A_{i,(n, n-k)}(w_{(n, n-k)})/\partial k < 0\) and \(\partial^2 q^A_{i,(n, n-k)}(w_{(n, n-k)})/\partial k^2 < 0\) while \(\partial q^B_{i,(n, n-k)}(w_{(n, n-k)})/\partial k > 0\) and \(\partial^2 q^B_{i,(n, n-k)}(w_{(n, n-k)})/\partial k^2 > 0\).

\[10\] See the Appendix for a proof of these market output results.
Given the above lemma, we next characterize the Nash equilibrium profits.

**Lemma 3.4.** The Nash equilibrium profits of multi-product firms $i \in 1, \ldots, n - k$ and rationalizing firms $j \in n - k + 1, \ldots, n$ are:

$$
\begin{align*}
\pi^A_{i,(n,n-k)}(w_{(n,n-k)}) &= \frac{(\beta(n-k+1) - \gamma k)(\alpha - w_{(n,n-k)})^2}{\beta(\beta + \gamma)(n-k+1)(n+1)^2} \\
\pi^B_{i,(n,n-k)}(w_{(n,n-k)}) &= \frac{(\beta(n+1) - \gamma k)(\alpha - w_{(n,n-k)})^2}{\beta(\beta + \gamma)(n-k+1)^2(n+1)} \\
\pi^A_{j,(n,n-k)}(w_{(n,n-k)}) &= \frac{(\alpha - w_{(n,n-k)})^2}{\beta(n+1)^2}
\end{align*}
$$

An implication from above is that, for any given $w_{(n,n-k)}$, an increase in $k$ decreases the profits from segment $A$ and increases the profits from segment $B$ for any multi-product firm.

Finally, the subgame perfect Nash equilibrium input price is:

**Lemma 3.5.** The subgame perfect Nash equilibrium wholesale input price under rationalization is

$$
w^∗_{(n,n-k)} = c + \sqrt{\frac{F(\beta + \gamma)\beta(n-k+1)(1+n)}{\beta(2(n-k+1) - k) + \gamma k}}
$$

One can check that $\frac{\partial w^∗_{(n,n-k)}}{\partial k} > 0$; that is, the upstream wholesale prices increase as more firms rationalize their production. As before, the reason is that, with $k$ firms exiting segment $B$, total demand decreases for the upstream product. In turn, because of economies of scale in upstream production, this reduction in demand increases the wholesale price. This fact, combined with the results above, imply that, post-rationalization equilibrium output in market segment $B$ falls as does the total output of multi-product firms in market segment $A$. In the Appendix, we show that the post-rationalization equilibrium prices rise in both market segments ($p^A_{(n,n-k)}$ and $p^B_{(n,n-k)}$ are increasing in $k$) and that, for small values of $k$ at least, the relative price $p^B_{(n,n-k)}/p^A_{(n,n-k)}$, is also increasing in $k$. We summarize these results below:

**Proposition 3.3.** Rationalization leads to an increase in the equilibrium input price and the equilibrium prices in both downstream market segments. Initial rationalization ($k$ in a neighborhood of $k = 0$) results in an increase in relative downstream prices, $p^B_{(n,n-k)}/p^A_{(n,n-k)}$.

To determine the impact on multi-product firm profits from a rationalization by $k$ competitors, we need only compare the pre- and post-rationalization profits. Letting $\Delta \pi^A = \pi^A_{i,(n,n-k)} - \pi^A_{i,(n,n)}$ and $\Delta \pi^B = \pi^B_{i,(n,n-k)} - \pi^B_{i,(n,n)}$, we have that:
\[
\Delta \pi^A = \frac{1}{R} \left[ (\alpha - w^*_{(n,n-k)})^2 (1 - \frac{\gamma k}{\beta (n-k+1)}) - (\alpha - w^*_{(n,n)})^2 \right] < 0
\]
\[
\Delta \pi^B = \frac{1}{R} \left[ (\alpha - w^*_{(n,n-k)})^2 \frac{(n+1)}{(n-k+1)^2} (n+1 - \gamma k/\beta) - (\alpha - w^*_{(n,n)})^2 \right] \leq 0
\]

where \( R \equiv (\beta + \gamma)(n+1)^2 \). From above, we know that \( \frac{\partial w^*_{(n,n-k)}}{\partial k} > 0 \), which means that the network effect is always negative. In market segment \( A \), the competition effect is also negative. This result stands in sharp contrast to the previous situation in which firms exited the market completely. Now, the rationalizing firms expand output in market segment \( A \) since doing so no longer impacts profits in market segment \( B \). The result is that profits of the multi-product firms unambiguously fall in market segment \( A \): \( \Delta \pi^A < 0 \).

In market segment \( B \), the exit of the rationalizing firms from that segment reduces competition and so enhances the profitability of the remaining, multi-product firms (the competition effect in market segment \( B \) is positive). Rationalization is more likely to have a negative overall profit effect for the multi-product firms (\( \Delta \pi^B \) is more likely to be negative) the larger is \( k \) (the more firms that rationalize) and the greater are the scale economies in the upstream market. In this case, the large rationalization causes a significant reduction in upstream demand. If scale economies are significant in the sense that the average cost curve is steeply sloped, then the demand reduction leads to significant exit from the upstream market, large increases in the wholesale price, \( w^*_{(n,n-k)} \) and lowering of profits for the multi-product firms.

Comparing the marginal impact of a rationalization to the marginal impact of a full exit from the market yields further insight into the impact of rationalization on the profits of the multi-product firms. From the previous section, we have that

\[
- \frac{d\pi^*_i(n,n)}{dn} = \left[ \frac{2}{n+1} \pi^*_i(n,n) \right]_{CE(+)} + \left[ \frac{4(\alpha - w^*_i(n,n))}{(\beta + \gamma)(n+1)^2} \times \frac{\partial w^*_i(n,n)}{\partial n} \right]_{NE(-)}
\]

The marginal impact at \( k = 0 \) on profits from rationalization is given by

---

11Formally, we know that \( (\alpha - w^*_{(n,n-k)}) < (\alpha - w^*_{(n,n)}) \) since \( w^*_{(n,n-k)} \) is increasing in \( k \).

12In market segment \( B \), we again have that the negative network effect implies that \( (\alpha - w^*_{(n,n-k)}) < (\alpha - w^*_{(n,n)}) \). However, one can check that \( \frac{(n+1)}{(n-k+1)^2} (n+1 - \gamma k/\beta) > 1 \) for all \( k \geq 1 \) so that the sign of \( \Delta \pi^B \) is ambiguous.

13Formally, the large increase in \( w^*_{(n,n-k)} \) means that \( (\alpha - w^*_{(n,n-k)})^2 \) is small relative to \( (\alpha - w^*_{(n,n)})^2 \), making \( \Delta \pi^B \) negative.
Chapter 3. Can Supply Chain Firms Benefit From Competition?

∂π∗_i,(n,n−k) / ∂k |k=0 = \left[ \frac{2β - γ}{2β} \pi∗_i,(n,n) \right] + \frac{2β - γ}{β} \frac{\partial w∗_i,(n,n)}{\partial n} \right)

We can see, first, that the network effect is smaller (less negative) under rationalization than under full exit. There are two reasons. First, the rationalizing firm exits only the one market segment and so upstream demand falls by less than if it exits both segments. Additionally, however, the rationalizing firm expands output in its remaining market segment (segment A) and does so by enough that total output in segment A, and so total demand for the input from segment A, increases. This latter effect is captured by the term \((β - γ)/β\).

Next, note that the competition effect is also smaller. Again the reason is twofold: (i) the rationalizing firm only exits market segment B and (ii) it expands output in segment A. This latter effect not only reduces the multi-product firms’ profits in segment A but also lowers profits in segment B since the products in segments A and B are substitutes. This latter difference is captured in the competition effect by the term \((2β - γ)/2β\).

Finally, substitution for the profit expressions in equation (3.10) reveals that \(\frac{∂π∗_i,(n,n−k)}{∂k} |k=0 = -(β - γ) \frac{dπ∗_i,(n,n)}{dn}\) so that, at the margin at \(k = 0\), the condition under which \(\frac{∂π∗_i,(n,n−k)}{∂k} |k=0 < 0\) is the same as the case of exit from the market: from equation (3.7), this occurs when scale economies in the upstream market are large and the potential mark-up in the downstream market is small. This means that, in the neighborhood of \(k = 0\) at least, profits for the multi-product firms can be ranked when network effects dominate; specifically, \(π∗_i,(n,n) > π∗_i,(n,n−k) > π∗_i,(n−k,n−k)\). For our vehicle example, the implication is that Ford prefers a bailout with rationalization to no bailout at all. We summarize these results below.

**Proposition 3.4.** Rationalization results in (i) a decrease in multi-product firm profits from market segment A and (ii) a decrease in overall profits if \(k\) is large and scale economies in the upstream market are significant. Initial rationalization (in a neighborhood of \(k = 0\)) results in a decrease in overall profits for multi-product firms iff \(2(n+1)(p∗ − w∗_i,(n,n)) < \frac{1}{n}(w∗_i,(n,n) − c)^{14}\). In this case, profits are ranked such that \(π∗_i,(n,n) > π∗_i,(n,n−k) > π∗_i,(n−k,n−k)\).

Note that the condition for rationalization to reduce profits of the multi-product firms is the same as the condition for exit to reduce profits of the remaining competitors. This condition is relatively easy to

---

14This condition is identical to the mark-up condition in Proposition 3.2 when firms exit both markets.
implement empirically since it only requires information on the number of downstream competitors and the mark-ups in the upstream and downstream markets.

3.4. Industry Equilibrium and Adjustment

So far, we have assumed that the number of downstream firms is exogenous and asked whether an exogenous exit from the market damages the remaining downstream firms. As is clear from the analysis, the answer to this question depends, in part, on the value of \( n \), the initial number of downstream producers. An obvious concern is whether a situation in which the remaining producers would benefit from keeping competitors active can be consistent with an initial industry equilibrium in the downstream market. To address this concern, we consider a setting where the initial number of firms \( n \) is determined endogenously but exit remains an exogenous action. Specifically, suppose that each downstream firm has a fixed cost of operation, \( F^d_i \), determined by an initial draw from some distribution \( G(F^d) \).\textsuperscript{15} Each firm pays a sunk entry cost to obtain its draw from \( G(\cdot) \) and there is open-entry in the market. Then, in the initial equilibrium, there will be \( n \) active firms with fixed costs \( \{ F^d_1 < F^d_2 < \ldots < F^d_n \} \) such that

\[
\pi^*_i(n,n)(w^*_i(n,n)) - F^d_i \geq 0 \quad \text{for all } i.
\]

Further, as long as fixed costs are not too heterogeneous, it must also be the case:

\[
\frac{\partial \pi^*_i(n,n)}{dn} = \frac{2(\alpha - w^*_i(n,n))}{(\beta + \gamma)n(n + 1)^2} \left[ -2n(n + 1)(p^* - w^*_i(n,n)) + (w^*_i(n,n) - c) \right] < 0; \quad (3.11)
\]

that is, the industry equilibrium is such that additional entry reduces the profits of all active firms. This equilibrium outcome is illustrated in 3.1 for the example of the 3 U.S. auto manufacturers. In the figure, the dashed curve gives each firm’s initial equilibrium variable profits as a function of \( n \). Given fixed cost draws \( F_{Ford}, F_{GM} \) and \( F_{Chrysler} \), the initial value of \( n \) occurs at point \( R \) on the dashed curve where all three make positive profits in equilibrium. These profits are given by \( RS \) for Chrysler, \( RT \) for GM, and \( RY_1 \) for Ford.

The fact that the initial equilibrium occurs on the downward-sloping part of the profit function raises the obvious question: How can our condition for a bailout to be desirable – exit lowers the profits of remaining firms – be fulfilled in equilibrium? The answer can be found by imagining, at this initial equilibrium value of \( n \), that a negative demand shock occurs (the Great Recession) that lowers the profits of all producers. This outcome is illustrated in Figure 3.1 by the solid curve which gives each firm’s post-shock equilibrium variable profits as a function of \( n \). Suppose also that this shock is sufficiently large, as in the figure, that some of the downstream firms – GM and Chrysler in our illustration – can

\[\text{\textsuperscript{15}[49] point out that many significant costs were fixed and not variable. Examples include pension and health costs for retirees, a wage settlement where workers when laid off received 95\% of their salary and interest payments to bond holders.}\]
no longer cover their fixed costs of operation \((\pi_i^{\ast}(n,n)(w_{n,n}^{\ast}) - F_m^d < 0 \text{ for some } m < n)\). If exit of these firms results in a sufficient reduction in the number of competitors, the remaining firms could find themselves on the upward sloping segment of their profit curve. This outcome is illustrated by the point \(Z\) below. In this event, the remaining firm(s) – Ford in our illustration – can benefit from the bailout of some or all of its competitors. In our illustration, a bailout of Chrysler and GM results in a profit increase for Ford from \(ZY_2\) to \(XY_1\). This illustrates precisely the incentive for Ford to endorse the bailout of both Chrysler and GM.

### 3.5. A Calibration Experiment

In order to examine the plausibility of our results and to investigate further the issues highlighted in our analysis, we study a calibrated version of the model in this section. Our calibration utilizes information on the U.S. automotive and auto-parts industries. The parameters to be calibrated are demand parameters, \(\{\beta, \gamma\}\), and technology parameters, \(\{c, F\}\). To calibrate the demand parameters \(\{\beta, \gamma\}\), we use information from the U.S. Consumer Expenditure Survey (CES). Conducted by the Bureau of Labor Statistics, the CES provides information on the buying habits of American consumers, including data on their expenditures, income, and consumer unit (families and single consumers) characteristics.

\(^{16}\) To calibrate the cost parameters \(\{c, F\}\), we use data from the U.S. Annual Survey of Manufacturers
(ASM). The ASM provides sample estimates of statistics such as payroll and materials costs for all manufacturing establishments with one or more paid employee. In the previous sections, for analytical ease, our operative assumption was that the downstream marginal production cost was simply the aggregate parts costs, $w$. To match the model to the data, however, we introduce an additional cost parameter, $v$, that represents all other downstream marginal costs incurred in producing a unit of output in the downstream market (e.g., production labour cost). We calibrate $v$ using our data. Finally, as part of our price calibration, we estimate vehicle salvage values from the Canadian Red Book Vehicle Valuation Guide (Red Book) converted to U.S. prices by a purchasing power parity exchange rate. We calibrate the model using pre-recession data from 2002-2005 and then update the demand parameters using recession data from 2009. This choice reflects our interest in explaining Ford’s revealed preference for the bailout precipitated by the shift in auto demand and prices realized during the recession of 2008 - 2009.

3.5.1 Demand Data

The U.S. Consumer Expenditure Survey is designed to be representative of the U.S. civilian, non-institutionalized population from both urban and rural areas. In each quarter, approximately 7,000 households are interviewed about large purchases, including vehicle purchases and leases. Approximately half of the households are re-interviewed in subsequent quarters. Table 2.1 in the Calibration Appendix provides the number of the non-repeat households interviewed and the total number of households in the U.S. for pre-recession and recession years. For those households interviewed, the CES contains detailed information on vehicle characteristics such as the purchase price, the vehicle make and model purchased, the model year, and information on whether the vehicle is purchased or leased. The purchase data contains information on whether the vehicle is financed, including net prices, down payments and monthly payments. The lease data contains information of both the down payment and the monthly payments. Tables 2.2 and 2.3 in the Calibration Appendix provide a summary description of the percentage of households buying/leasing new vehicles each year.\footnote{In the data, a vehicle with a new model year can be purchased in the prior year and the next year beyond the model year. In Table 3.3, we calculate these ‘early’ and ‘late’ purchases and attribute them to their respective model year.}

The CES data divide vehicles into two groups: (A) trucks, minivans, vans, and SUVs (vans/SUVs); and (B) automobiles. Using this classification, on average between the years 2002 and 2010, we observe that 54% of vehicles produced by the three U.S. manufacturers are in category (A) and 46% are in category (B). This division corresponds reasonably well to the assumption in our model of two market segments with similar market shares. Even though vehicles within each group are not identical, our model assumes within group homogeneity and between group heterogeneity. For example, the operative
assumption is that vans/SUVs in group (A) are homogeneous and automobile models in group (B) are homogeneous but vans/SUVs and autos are not perfect substitutes. This aggregation assumption serves analytical ease; we maintain this assumption in our calibration.

The price data record purchase prices (included both financed and not financed vehicle) and leased prices. The purchase data contain information on the trade-in allowance received (for the buyers’ old vehicles), net price after discount/rebate/trade-in allowance, down payment and monthly payments (if the vehicle is financed). We take account of any manufacturer discounts through lower finance rates during the recession. Financed sales require adjustments to arrive at an estimate of the purchase price of the vehicle. For these sales, we start with the final prices (cash purchases) which are equal to the net price plus any trade-in allowance received for the old vehicle. To calculate the net present value on the day of purchase, we discount future payments to the day of the purchase and add any down payment and trade-in allowance to derive the final price of the vehicle. That is, the final price (of the financed vehicle) is equal to the net present value of all future payments plus any down payment plus any trade-in allowance received for the old vehicle. To discount future payments, we use the relevant U.S. prime rate, which is a conservative discount rate as auto loans will be above prime, conditional on risk.

The lease data record trade-in allowance received (for the old vehicle), down payments, monthly payments and the number of months in the contract. There is a missing observation required to calculate the final price of the vehicle on the purchase date. What is unobserved is the price specified in the lease that gives the buyer the option to purchase the vehicle at the end of the lease. To overcome this missing observation, using the Canadian Red Book Vehicle Valuation Guide (Red Book), we collect the used-vehicle prices of leased vehicles on the month of the lease expiry. This Guide is published monthly and specifies the price of different used vehicles based on their make and model year. For example, if a Chevy Malibu is leased in August 2003 for 36 months, we look at the 2006/volume 8 version of the Red Book to see the used price for a 2003 Malibu. We use the Purchasing Power Parity exchange rate to convert values from Canadian dollars to U.S. Dollars. We then discount these values to the date of purchase using the U.S. Consumer Price Index. These values permit an approximation of the final price of the leased vehicle on the purchase date. The final price is equal to the net present value of all monthly payments plus any down payment plus any trade-in allowance (for the old vehicle) plus the present value of the salvage price. Finally, since prices are in nominal terms, we use the Consumer Price Index to calculate real values adjusted to 2006. Table 2.4 in the Calibration Appendix reports the average adjusted real final U.S. auto prices. Also, following [20], we set the average vehicle price elasticity to 5.18

18[40] estimates a much lower elasticity of 1.7. We tested our model for this elasticity and observed that our results are
3.5.2 Supply Data

On the manufacturing and parts supplier side, we need to calibrate technology parameters to arrive at the upstream and downstream costs of vehicle manufacturing. ASM provides data on production-workers payroll and material cost of both the automobile industry and the auto-parts manufacturing industry. Table 2.5 in the Calibration Appendix provides a summary description of the variable cost (production workers wage plus material cost) as a percentage of the value of the shipments for both the downstream and the upstream. The downstream percentages are consistent with the average ratio of markups of retail price (the Lerner Index) of 0.239 reported in [20]. Further examination of the data reveals that auto parts, on average, account for 65% of the value of the vehicle.

3.5.3 Calibration Strategy

To proceed, we need to calibrate the model’s parameters. This task uses the pre-recession data from 2002 to 2005. The assumption is that the recession has two effects: (i) it reduces the number of households in the vehicle market and (ii) it reduces vehicle prices. To capture these effects, we assume that market demand for autos and vans/SUVs is derived from a collection of identical individual demand functions of the form:

\[ q^A = a - bp^A + cp^B \]
\[ q^B = a - bp^B + cp^A \]

Then, if \( m_{pre} \) denotes the number of households purchasing/leasing a new vehicle pre-recession, the market inverse demand functions are:

\[ p^A = \frac{a}{b - c} - \frac{b}{m_{pre}(b^2 - c^2)} Q^A - \frac{c}{m_{pre}(b^2 - c^2)} Q^B \]
\[ p^B = \frac{a}{b - c} - \frac{b}{m_{pre}(b^2 - c^2)} Q^B - \frac{c}{m_{pre}(b^2 - c^2)} Q^A \]

so that, corresponding to our model

\[ \beta_{pre} \equiv \frac{b}{m_{pre}(b^2 - c^2)} \text{ and } \gamma_{pre} \equiv \frac{c}{m_{pre}(b^2 - c^2)} \]

Next, we turn to some data. From Tables 2.1, 2.2 and 2.3 in the Calibration Appendix, the average number of U.S. households buying a new vehicle between 2002 and 2005 is \( m_{pre} = 5,004,940 \). The demand-weighted average price in both market sectors between 2002 and 2005, in real 2006 values, is robust.
$p_{pre}^A = p_{pre}^B = p_{pre} = $29,825. Using the variable cost percentages in Table 2.5 in the Calibration Appendix, on average, the variable production cost accounts for 76% of the price of the vehicle. Furthermore, as stated above, 65% of the value of the vehicle is produced by independent auto suppliers. These two observations permit a calculation of the pre-recession average marginal cost of production as well as the pre-recession average wholesale input price:

$$\begin{align*}
MC_{pre} &= 0.76 \times p_{pre} = $22,667 \\
w_{pre} &= $19,386 \\
v &= $3,281
\end{align*}$$

These estimates give a per unit variable cost of $22,667 for an auto/van product that sells for $29,825.

The above estimates (with the assumption that the U.S. auto manufacturers operated in equilibrium in 2002 to 2005) and the Nash equilibrium price and quantity formulas permit a calibration of the pre-recession demand parameters \( \{\beta_{pre}, \gamma_{pre}\} \). From equations (3.2) and (3.3), the Nash equilibrium total output in each market segment is given by:

$$Q_{pre}^* = \frac{m_{pre}}{2} = \frac{n \times (n + 1) \times (p_{pre}^* - MC_{pre})}{(\beta_{pre} + \gamma_{pre})(n + 1)}, i \in \{A, B\}$$

As our focus is on the U.S. auto manufacturers, we set \( n = 3 \), and substitute our estimates of the price, marginal cost, and household demand into the above equations to obtain:

$$\begin{align*}
(\beta_{pre} + \gamma_{pre}) &= \frac{n \times (n + 1) \times (p_{pre}^* - MC_{pre})}{(\frac{m_{pre}}{2})(n + 1)} = 0.0086 \\
\text{(3.12)}
\end{align*}$$

The large number of households purchasing vehicles (which appears in the denominator of the above expression) means that the value of \( \beta_{pre} + \gamma_{pre} \) is small.

From [20], the average vehicle price elasticity is 5. This assumption can be restated in terms of the parameters in our model as:

$$|\varepsilon_{AA}| = 5 = \frac{\beta_{pre} \times p_{pre}^*}{(\beta_{pre} - \gamma_{pre})n(p_{pre}^* - MC_{pre})} \tag{3.13}$$

By substituting the estimated and calibrated parameters into equations (3.13), we obtain:

$$\gamma_{pre} = 0.7 \times \beta_{pre}$$
This completes the calibration of the demand parameters.

The data in Table 2.5 in the Calibration Appendix permit a calculation of the technology parameters. From this table, the variable costs for the auto-parts manufacturing sector, on average, account for 70% of the total value of shipments. In our model, the upstream is an open-entry industry so that, for each upstream firm \( i \), the total value of shipments must equal the total cost. Letting \( q_{u}^{*} \) give the equilibrium output of upstream firm \( i \), we have that:

\[
q_{u}^{*} \times w_{pre} = q_{u}^{*} \times c + F
\]

\[
\Rightarrow c = 0.7 \times w_{pre} = \$13,570 = \frac{2n(w_{pre} - c)^2}{(\beta_{pre} + \gamma_{pre})(n + 1)} = \$5,899,858,605
\]

At first glance, the value of \( F \) may seem high. However, since the calibration aggregates the 6000+ auto-parts manufacturers into 6 upstream firms, such a fixed cost does not seem out of order. All of this establishes the pre-recession state.

These pre-recession parameters applied to the profit-margin analogue of equation (3.7) indicate that:

\[
(p_{i}^{pre} - w_{pre} - v) = 7158 > (w_{pre} - c) = 5816
\]

The interpretation is that between 2002-2005, the change in per firm profits as \( n \) changes, (given by equation (3.11)), is negative. That is, in the pre-recession period, the industry equilibrium is such that additional entry reduces the profits of all active firms, i.e., the pre-recession equilibrium occurs on the downward-sloping part of the profit function illustrated by the dashed curve in Figure 3.1.

What is the impact of the recession on this equilibrium? The recession produces a demand shift. Data from 2009 permit an update of the demand parameter values to account for this shift. The same calibration procedure for this recession period yields the following demand parameters (where \( m_{rec} \) denotes the number of households purchasing a new car during the recession):

\[
\beta_{rec} = \frac{b}{m_{rec}(b^2 - c^2)}, \gamma_{rec} = \frac{c}{m_{rec}(b^2 - c^2)}
\]

From Tables 2.1, 2.2, and 2.3 in the Calibration Appendix, \( m_{rec} = 1,849,302 < m_{pre} \). As expected, during the recession, the demand curve becomes steeper, i.e., \( \beta_{rec} = \beta_{pre} \times \frac{m_{pre}}{m_{rec}} > \beta_{pre} \) and \( \gamma_{rec} = \gamma_{pre} \times \frac{m_{pre}}{m_{rec}} > \gamma_{pre} \). Thus,

\[
(\beta_{rec} + \gamma_{rec}) = \frac{m_{pre}}{m_{rec}} \times (\beta_{pre} + \gamma_{pre}) = 0.023
\]
With the updated recession slope parameters, the input price, $w_{rec}$, becomes:

$$w_{rec} - w_{pre} = c + \sqrt{\frac{F(\beta_{rec} + \gamma_{rec})(n+1)}{2n}} - c - \sqrt{\frac{F(\beta_{pre} + \gamma_{pre})(n+1)}{2n}}$$

$$= \sqrt{\frac{F(\beta_{pre} + \gamma_{pre})(n+1)}{2n}} \left( \frac{m_{pre}}{m_{rec}} - 1 \right)$$

yielding a value of

$$w_{rec} = $23,138$$

Using $w_{rec}$ and the average real product price in 2009 of $26,854$ permits a calculation of the price markup over the marginal cost during the recession:

$$p_{rec}^* - w_{rec} - v = $26,854 - $23,138 - $3,281 = $435$$

This seemingly small markup of price over marginal cost is consistent with the observation of a depressed demand for vehicles and losses accruing to the U.S. auto manufacturers during the period from 2006 to 2009.19

### 3.5.4 Experiment

In the setting described by these calibrated parameters, what is the plausibility for a bailout to be desirable? Recall from the analogue to equation (3.7), the necessary condition for Ford to support the bailout of at least one of GM or Chrysler is that:

$$4(p_{rec}^i - w_{rec} - v) < (w_{rec} - c)$$

Substitution of the calibrated parameters values yields $1,740 < 9,568$ so that the profit mark-up condition to support the bailout is satisfied. Moreover, by substituting the calibrated parameters into the equilibrium profit formula, we have the following profit inequalities:

$$\pi_{i,(3,3)}^* = 1.01 \times \pi_{i,(3,1)}^* > \pi_{i,(3,1)}^* \gg \pi_{i,(1,1)}^* \approx 0$$

This set of inequalities implies that the departure of GM and Chrysler from both market segments would have increased the input price sufficiently that the downstream price and marginal cost of production

19http://fortune.com/fortune500/
20Profit is gross of any fixed cost.
Chapter 3. Can Supply Chain Firms Benefit From Competition?

Figure 3.2: Difference in Profits as a Function of $k$ under Product Rationalization (in 1000s).

for Ford would be virtually equal – the value of $\pi^*_i(1,1)$ is close to zero. This is, in fact, the claim by the U.S. auto producers and the rationale for Ford to support the bailout of GM and Chrysler. If Chrysler and GM were to exit one market segment (segment $B$ in our model), $\pi^*_i(3,1) \gg \pi^*_i(1,1) \approx 0$. Again, the implication is that Ford prefers a bailout with rationalization to no bailout at all, as was revealed to be the case in the actual bailout deal. Finally, a bailout without any product rationalization is only marginally better for Ford than a bailout with product rationalization. That is, Ford only very weakly prefers to have it and its two competitors in both market segments as opposed to all competing in market segment $A$ with Ford as the sole supplier in market segment $B$. Thus the values of our calibration are consistent with industry facts.

Further insight is gained by examining the changes in profits as firms depart market sectors. In Figure 3.2, we show, for our calibrated model, the difference in post-recession profits of each multi-product firm in market segments $A$ and $B$ and the total difference in profits as a function of $k$ (the number firms that exit market $B$ and specialize in market $A$). The figure has three plots. With the departure of firms from market $B$ (measured on the horizontal axis), the top plot shows the change in profits for each firm remaining in market $B$, the bottom plot shows the change in profits for multiproduct firms in market $A$, and the middle plot shows the sum of the change in the profits for the multiproduct firm(s).

This figure reveals that, for market segment $B$, whether one or two firms depart segment $B$, the profit for the remaining multiproduct firms or firm increase.22 Put differently, for segment $B$, at the

---

21 Steven Rattner who headed the US auto task force claimed that Ford would have closed because it would not have been able to get parts, as the upstream parts industry was in arguably worse shape than the assemblers (see “Ford Would Have Shut Without Auto Bailouts”, Bloomberg, Oct. 9, 2012)

22 Obviously, if two firms (GM and Chrysler) leave market $B$, the remaining firm (Ford) becomes a monopolist.
calibrated parameters, the competition effect dominates the network effect and so profits rise due to rationalization in segment $B$. For segment $A$, the rationalizing firms (GM or GM and Chrysler) now focus on their product(s) in market $A$. As discussed in Section 3.3, when a firm departs segment $B$, there is a decline in the profits of the multiproduct firms in segment $A$. At the calibrated parameter values, the bottom plot in Figure 3.2 shows that, as $k$ increases, there is a further decline in the profits of the one remaining multiproduct firm in market $A$. Finally, the net effect of rationalization is shown in the middle plot of Figure 3.2. It shows that, at our calibrated parameter values, if one active firm departs segment $B$, each of the remaining multiproduct firms net profit increases; however, if two active firms (GM and Chrysler) depart market $B$, the net profits for the remaining multiproduct firm (Ford) fall. That is, under our calibration, contrary to naive intuition, Ford does not want to become a monopolist in segment $B$.

Finally, recall that Proposition 3.3 indicates that, as some firms rationalize their production, the equilibrium input price as well as the equilibrium retail price increase. Do these price changes hold under our calibration? Figure 3.3 (below) illustrates the resulting input price changes as rationalization occurs: under our calibration, the input price increases as more firms rationalize their production. In addition, the rate of this increase is increasing in $k$. Proposition 3.3 also highlights that the initial product rationalization ($k$ in a neighborhood of $k = 0$) yields an increase in relative downstream prices. Under our calibration, Figure 3.4 (below) shows that this result is robust for all $k$. As the figure illustrates, as more firms rationalize their production, the price in segment $B$ increases more than the price in segment $A$. Moreover, the magnitude of the increase in relative prices increases with $k$.

These results provide a check of our calibration experiment. In turn, this experiment reinforces the interpretive value of our original specification.
3.6. Conclusions

Using a stylized supply chain structure, our goal was to examine those conditions where downstream firms would have a self-interest in promoting the sustainability of their downstream competitors. The motivation for this analysis was the U.S. Federal Government bailout of the U.S. auto producers, specifically GM and Chrysler. Ford supported the bailout, which may seem counterintuitive. This is not the case once the interaction between downstream producers and upstream suppliers is endogenized.

As part of the bailout, GM and Chrysler were obliged to rationalize their product offerings. In our structure, this rationalization meant that active producers were going to leave one segment to focus exclusively on the other market segment. We analyze the impact of this market segment refocus on the profits in both market segments for both types of continuing producers, multi-product and single-product firms. Again, this analysis endogenizes the impact of any product rationalization on upstream supply as well as downstream competition.

A calibrated application of our upstream and downstream model reinforces our specification and our analytical results. While vehicle production and manufacturer bailout are the principal drivers of our analysis, other industries face similar issues. For example, the PC/laptop segment of the computer industry is also going through such a rationalization as producers leave the PC/laptop segment to focus on the table/smart phone segment. This is also an industry with significant links between downstream product development and assembly and upstream supply of both parts and even final product.
3.7. Appendices

3.7.1 Mathematical Appendix

Deriving Equilibrium Solutions  

Recall that the profit function of each downstream firm is:

\[ \pi_i(n, n) = \pi_i^A + \pi_i^B, \]

\[ \pi_i^A(n, n) = (p_i(n) - w(n))q_i^A(n, n) = (\alpha - \beta(q_i^A(n, n) + \ldots + q_i^A(n, n)) - \gamma(q_i^B(n, n) + \ldots + q_i^B(n, n))q_i^A(n, n) \]

\[ = (\alpha - \beta \sum_{j \neq i} q_j^A(n, n) - \beta q_i^A(n, n)) - \gamma \sum_{j \neq i} q_j^B(n, n) - \gamma q_i^B(n, n) - w(n)nq_i^A(n, n), \]

\[ \pi_i^B(w(n, n)) = (p_i^B(n) - w(n, n))q_i^B(n, n) = (\alpha - \beta(q_i^B(n, n) + \ldots + q_i^B(n, n)) - \gamma(q_i^A(n, n) + \ldots + q_i^A(n, n))q_i^B(n, n) \]

\[ = (\alpha - \beta \sum_{j \neq i} q_j^B(n, n) - \beta q_i^B(n, n)) - \gamma \sum_{j \neq i} q_j^A(n, n) - \gamma q_i^A(n, n) - w(n)nq_i^B(n, n). \]

Since \( \pi_i(n, n) \) is concave in \( q_i^A(n, n) \) and \( q_i^B(n, n) \), First-order Conditions yield,

\[ q_i^A(n, n) = \frac{\alpha - \beta \sum_{j \neq i} q_j^A(n, n) - 2\beta q_i^A(n, n) - \gamma \sum_{j \neq i} q_j^B(n, n) - w}{2\beta}, \]

\[ q_i^B(n, n) = \frac{\alpha - \beta \sum_{j \neq i} q_j^B(n, n) - 2\beta q_i^B(n, n) - \gamma \sum_{j \neq i} q_j^A(n, n) - w}{2\beta}. \]

Solving the above equations with respect to \( q_i^A(n, n) \) and \( q_i^B(n, n) \) yields:

\[ q_i^A(n, n) = \frac{\alpha - (\beta + \gamma) \sum_{j \neq i} q_j^A(n, n) - w}{2(\beta + \gamma)}, \]

\[ q_i^B(n, n) = \frac{\alpha - (\beta + \gamma) \sum_{j \neq i} q_j^B(n, n) - w}{2(\beta + \gamma)}. \]

Since the firms in each market are symmetric, \((n - 1)q_i^A(n, n) = \sum_{j \neq i} q_j(n, n)\) and \((n - 1)q_i^B(n, n) = \sum_{j \neq i} q_j(n, n)\). By replacing these in the above equations and solving we derive the optimal quantity of firm \( i \) given \( w \):

\[ q_i^A(n, n)w(n, n) = q_i^A(n, n)w(n, n) = \frac{(n - 1)(\alpha - w(n, n))}{(\beta + \gamma)(n + 1)}. \]

By substituting the optimal quantities into the profit function we have:

\[ \pi_i(n, n)w(n, n) = \frac{2(\alpha - w(n, n))^2}{(\beta + \gamma)(n + 1)^2}. \]

Using the optimal downstream quantities we can derive the total upstream demand:

\[ Q_u = nq_i^A(n, n)w(n, n) + nBq_i^B(n, n)w(n, n) = \frac{2n(\alpha - w(n, n))}{(\beta + \gamma)(n + 1)}. \]
Therefore, the inverse wholesale price is

\[ w = \alpha - \frac{(\beta + \gamma)(n+1)}{2n} (q_1^u + \ldots + q_n^u) \]

Given the fixed cost plus constant marginal cost assumption of each upstream firm, the profit function of each upstream firm is:

\[ \pi_i^u = (w - c)q_i^u - F = (\alpha - c - \frac{(\beta + \gamma)(n+1)}{2n} (\sum_{j \neq i} q_j^u + q_i^u))q_i^u - F \]

Since \( \pi_i^u \) is concave in \( q_i^u \), first-order conditions yield:

\[ q_i^u^* = \frac{(\alpha - c)n}{(\beta + \gamma)(n + 1)} - \frac{\sum_{j \neq i} q_j^u}{2} \]

Since the upstream firms are symmetric \( \sum_{j \neq i} q_j^u = (n_u - 1)q_i^u^* \). By replacing this into the above equation, we obtain

\[ q_i^u^*(n_u) = \frac{2(\alpha - c)n}{(\beta + \gamma)(n + 1)(n_u + 1)} \]

Thus, the optimal wholesale price as a function of \( n_u \) is

\[ w(n_u)^* = \frac{\alpha + cn_u}{n_u + 1} \]

Since the upstream is a free-entry industry, we set the equilibrium CN wholesale price \( (w(n_u)^*) \) equal to the equilibrium average cost \( (c + F/q_i^{u*}) \) to determine the number of firms the upstream market can support (the equation has two roots; we consider the one the greater one). The number of firms that the upstream can support is:

\[ n_u^* = (\alpha - c) \sqrt{\frac{2n}{F(\beta + \gamma)(n + 1)}} - 1 \]

And thus the optimal wholesale price is:

\[ w^* = c + \sqrt{\frac{F(\beta + \gamma)(n + 1)}{2n}} \]
Proof of Proposition 3.2. Recall that

\[
\frac{\partial \pi_{i,(n,n)}(w^*_{(n,n)})}{dn} = -\frac{4(\alpha - w^*_{(n,n)})^2}{(\beta + \gamma)(n+1)^3} - \frac{4(\alpha - w^*_{(n,n)})}{(\beta + \gamma)(n+1)^2} \frac{dw^*_{(n,n)}}{dn}
\]

Therefore,

\[
\frac{\partial \pi_{i,(n,n)}(w^*_{(n,n)})}{dn} > 0 \quad \text{iff} \quad \frac{dw^*_{(n,n)}}{dn} < -\frac{(\alpha - w^*_{(n,n)})}{(n+1)}
\]

where

\[
\frac{dw^*_{(n,n)}}{dn} = -\frac{1}{2n^2} \sqrt{\frac{F(\beta + \gamma)2n}{n+1}}
\]

Let

\[
r \equiv \sqrt{\frac{F(\beta + \gamma)(n+1)}{2n}}
\]

Therefore, \(w^*_{(n,n)}\) and \(dw^*_{(n,n)}/dn\) are:

\[
w^*_{(n,n)} = c + r
\]

\[
\frac{dw^*_{(n,n)}}{dn} = -\frac{r}{2n(n+1)}.
\]

Thus,

\[
\frac{dw^*_{(n,n)}}{dn} < -\frac{(\alpha - w^*_{(n,n)})}{(n+1)}
\]

can be simplified to

\[
2n(\alpha - w^*_{(n,n)}) < (w^*_{(n,n)} - c)
\]

Finally, since \(p^* = \frac{\alpha + nw^*_{(n,n)}}{n+1}\), we can rewrite the above inequality as:

\[
2(n+1)(p^* - w^*_{(n,n)}) < \frac{1}{n}(w^*_{(n,n)} - c)
\]

Metal Exercise: To illustrate that our assumption of a single input is without loss of generality, assume that instead of a single input, the production in the downstream market requires multiple inputs from multiple upstream markets (e.g., auto-parts markets), i.e., each unit of output produced in the downstream industry requires \(L\) inputs, each produced in an upstream industry (e.g., in our auto example, there is an upstream engine manufacturing market and a transmission manufacturer market). The downstream, purchases each input at a per-unit market-determined wholesale price \(w^*_{h,(n,n)}\) \(h = 1, \ldots, L\). Moreover, assume each upstream industry is an open-entry industry and that firms in each
market are symmetric. The technology in the upstream market $h$ is characterized by a fixed cost ($f_h$) and a constant marginal cost ($c_h$). Thus, the total cost and average cost of upstream firm $i$ are $c_h q_{i,h}^u + f_h$ and $c_h + \frac{f_h}{q_{i,h}^u}$. Let $h = 1, \ldots, L$, where $q_{i,h}^u$ is the output of firm $i$ in upstream market $h$. By using the same procedure as before, the equilibrium wholesale input prices are:

$$w_{h,(n,n)}^* = c_h + \frac{f_h(\beta + \gamma)(n+1)}{2n}, \quad h = 1, \ldots, L$$

We can now aggregate the upstream markets and use the aggregate equilibrium input price

$$W_{(n,n)}^* = \sum_h w_{h,(n,n)}^* = C + \sqrt{F(\beta + \gamma)(n+1)}$$

where

$$C = \sum_h c_h$$

$$F = \left( \sum_h \sqrt{f_h} \right)^2$$

This aggregate input is what we consider in the body of the chapter. Therefore, our assumption of a single (aggregate) input is without loss of generality.

**Proof of pre/post rationalization Market Output.** We want to show that for a given $w$:

$$Q_{(n,n-k)}^A(w, n, k) \geq Q_{(n,n)}^A(w, n)$$

$$Q_{(n,n-k)}^B(w, n, k) < Q_{(n,n)}^A(w, n)$$

From equation (2) we can derive:

$$Q_{(n,n)}^A(w) = n q_{i,(n,n)}^A(w) = n \frac{\alpha - w}{(\beta + \gamma)(n+1)}$$

Moreover, from (8) and (9) we have:

$$Q_{(n,n-k)}^A(w) = \left( (n-k)q_{i,(n,n-k)}^A(w) + k q_{j,(n,n-k)}^A(w) \right), i \in \{1, \ldots, n-k\}, j \in \{n-k+1, \ldots, n\}$$

$$= \left( n + \frac{\gamma}{\beta (n-k+1)} \right) \frac{\alpha - w}{(\beta + \gamma)(n+1)}$$

$$Q_{(n,n-k)}^B(w) = \left( (n-k)q_{i,(n,n-k)}^B(w) \right) = \left( \frac{n-k}{n-k+1} \right) \frac{\alpha - w}{\beta + \gamma}$$

From the above results, it is easy to see $Q_{(n,n)}^A(w(n,n)) \leq Q_{(n,n-k)}^A(w(n,n))$. Moreover, since $\frac{n}{n+1} > \frac{n-k}{n-k+1}$, we can conclude that $Q_{(n,n)}^B(w) > Q_{(n,n-k)}^A(w)$. \qed
Post-Rationalization Equilibrium Prices. The post-rationalization equilibrium prices in market segment A and B are:

\[
p_{(n,n-k)}^A = \frac{\alpha + nw_{(n,n-k)}^*}{n + 1}
\]

\[
p_{(n,n-k)}^B = \frac{\alpha(\beta(n + 1) - \gamma k) + (\beta(n + 1)n - (\beta - \gamma)k)w_{(n,n-k)}^*}{\beta(n-k+1)(n+1)}
\]

From the above prices, it is easy to check that:

\[
\frac{\partial p_{(n,n-k)}^A}{\partial k} = \frac{n}{(n+1)} \frac{\partial w_{(n,n-k)}^*}{\partial k} > 0
\]

\[
\frac{\partial p_{(n,n-k)}^B}{\partial k} = \frac{(\beta - \gamma)(n+1)(\alpha - w_{(n,n-k)}^*) + (n - k - 1)(\beta(n-k)(1+n) + \gamma k)\partial w_{(n,n-k)}^*/\partial k}{\beta(1+n)(n-k+1)^2} > 0
\]

Therefore, post-rationalization equilibrium prices in both market segments increase as some firms rationalize their production. Moreover, the derivative of relative price \(p_{(n,n-k)}^B / p_{(n,n-k)}^A\) evaluated at \(k = 0\) (initial rationalization) is given by:

\[
\left. \frac{\partial (p_{(n,n-k)}^B / p_{(n,n-k)}^A)}{\partial k} \right|_{k=0} = \frac{((\beta - \gamma)(\alpha - w_{(n,n)}^*))}{\beta(n+1)(\alpha + nw_{(n,n)}^*)} > 0
\]

This implies that initial rationalization results in an increase in relative downstream prices.

3.7.2 Calibration Appendix

<table>
<thead>
<tr>
<th>Year</th>
<th>Number Interviewed</th>
<th>Number of Households</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>15650</td>
<td>112108</td>
</tr>
<tr>
<td>2003</td>
<td>16125</td>
<td>115536</td>
</tr>
<tr>
<td>2004</td>
<td>15632</td>
<td>116282</td>
</tr>
<tr>
<td>2005</td>
<td>14734</td>
<td>117356</td>
</tr>
<tr>
<td>2008</td>
<td>13788</td>
<td>120171</td>
</tr>
<tr>
<td>2009</td>
<td>13975</td>
<td>120847</td>
</tr>
<tr>
<td>2010</td>
<td>14231</td>
<td>121107</td>
</tr>
</tbody>
</table>

Table 3.1: The Number of Non-Repeat Households Interviews and The Number of Households in the U.S. (in 1000s).
### Chapter 3. Can Supply Chain Firms Benefit From Competition?

#### Table 3.2: Percentage of Households Purchasing New Vehicles in Each Year.

<table>
<thead>
<tr>
<th>Model Year</th>
<th>% New Vehicle Purchases</th>
<th>U.S.</th>
<th>Foreign</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>6.81%</td>
<td>4.12%</td>
<td>2.69%</td>
</tr>
<tr>
<td>2003</td>
<td>7.15%</td>
<td>4.11%</td>
<td>3.04%</td>
</tr>
<tr>
<td>2004</td>
<td>6.47%</td>
<td>3.58%</td>
<td>2.89%</td>
</tr>
<tr>
<td>2005</td>
<td>6.99%</td>
<td>3.45%</td>
<td>3.54%</td>
</tr>
<tr>
<td>2008</td>
<td>4.97%</td>
<td>2.18%</td>
<td>2.79%</td>
</tr>
<tr>
<td>2009</td>
<td>3.78%</td>
<td>1.42%</td>
<td>2.36%</td>
</tr>
<tr>
<td>2010</td>
<td>4.06%</td>
<td>1.61%</td>
<td>2.45%</td>
</tr>
</tbody>
</table>

#### Table 3.3: Percentage of Households Leasing New Vehicles in Each Year.

<table>
<thead>
<tr>
<th>Model Year</th>
<th>% New Vehicle Leases</th>
<th>U.S.</th>
<th>Foreign</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>1.12%</td>
<td>0.57%</td>
<td>0.55%</td>
</tr>
<tr>
<td>2003</td>
<td>0.87%</td>
<td>0.42%</td>
<td>0.45%</td>
</tr>
<tr>
<td>2004</td>
<td>0.90%</td>
<td>0.45%</td>
<td>0.45%</td>
</tr>
<tr>
<td>2005</td>
<td>1.25%</td>
<td>0.70%</td>
<td>0.55%</td>
</tr>
<tr>
<td>2008</td>
<td>1.08%</td>
<td>0.48%</td>
<td>0.60%</td>
</tr>
<tr>
<td>2009</td>
<td>0.51%</td>
<td>0.17%</td>
<td>0.34%</td>
</tr>
<tr>
<td>2010</td>
<td>0.69%</td>
<td>0.13%</td>
<td>0.56%</td>
</tr>
</tbody>
</table>

#### Table 3.4: Average Final Prices (Adjusted to 2006).

<table>
<thead>
<tr>
<th>Model Year</th>
<th>Average Prices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>29,617</td>
</tr>
<tr>
<td>2003</td>
<td>30,376</td>
</tr>
<tr>
<td>2004</td>
<td>30,280</td>
</tr>
<tr>
<td>2005</td>
<td>29,025</td>
</tr>
<tr>
<td>2008</td>
<td>26,315</td>
</tr>
<tr>
<td>2009</td>
<td>26,854</td>
</tr>
<tr>
<td>2010</td>
<td>27,662</td>
</tr>
</tbody>
</table>
Table 3.5: Variable Cost Percentage of the Total Value of Shipments for Both the Upstream and Downstream Auto Industry.
Chapter 4

Almost Robust Optimization and Its Applications in Logistics

Binary requirements on decision variables are essential in many operational and planning models. A variety of applications with binary requirements include: logistics problems such as facility location and vehicle routing, and manufacturing problems such as production planning and scheduling. Traditional optimization models for these applications usually assume that the parameters and data are known and certain. In practice, however, such assumptions are often unrealistic, e.g., the data may be noisy or incomplete.

Two main approaches used to incorporate uncertainty are stochastic programming (see [83] and references therein) and chance constrained programming ([32]). In stochastic programming (which are usually modeled in two stages) the decision maker optimizes an expected value of an objective function that involves random parameters. In Chance Constrained Programming, the decision maker wants to ensure that the constraints hold with at least a specified probability. In the presence of discrete decision variables, both approaches are typically very challenging to solve, and are thus less effective.

Another method to incorporate uncertainty into optimization problems is robust optimization (RO). Unlike the two methods mentioned above that use probabilistic information, in RO the uncertainty is set-based, requiring no specific information about the distribution functions. In such models, the decision-maker seeks to find solutions that are feasible for any realization of the uncertainty in a given set (see [17, 23]). The majority of the RO literature, however, focus on convex optimization and limited work is done on discrete optimization problems; see [64, 10, 22, 8], and [13] for exceptions.

In this chapter, we present the Almost Robust Discrete Optimization (ARDO) model that combines
ideas from stochastic programming, chance constrained programming, and robust optimization. ARDO, trades off the objective function value with robustness of the solution (measured by the amount of violation of the uncertain constraints under different realizations) to find optimal solutions that are almost robust (feasible for most data realizations). ARDO is attractive for five reasons: (i) it has a simple structure; (ii) it permits dependence among uncertain parameters; (iii) it permits an incorporation of the decision maker’s attitude towards risk by controlling the degree of conservatism of the optimal solution; (iv) unlike exiting approaches that either assume that the decision maker has complete probabilistic information (e.g., stochastic programming) or assume that she has no probabilistic information (e.g., RO), ARDO can be setup to suit a wide spectrum of available probabilistic information, e.g., full, partial, or no probabilistic information at all; and (v) compared to the existing approaches, it is computationally easier to solve. In Section 4.1.4, we provide a comparison of our approach with stochastic programming, chance constrained programming, and RO. Note that while the focus of this paper is on discrete optimization problems, the framework is also applicable for problems with continues decision variables (and as discussed in Section 4.1.4, unlike other methods that lose tractability even with continues variables, the almost robust framework is much more tractable).

To solve ARDO efficiently, we develop a decomposition approach, which decomposes ARDO into a deterministic master problem and a subproblem which checks the master problem solution under different realizations and generates cuts if needed. An important upside of our approach is that since the master problem is a deterministic discrete optimization problem, existing approaches for solving such problems can be incorporated into the decomposition model to speed up the solution procedure. The decomposition approach also overcomes the main drawbacks of the integer L-shape method ([66]), the main tool used to solve stochastic integer programs. Specifically, unlike the L-shaped method, in which the subproblems are integer programs and are hard to solve, the subproblems in our model are very simple to solve; furthermore, unlike the L-shape method where the cuts are generated from the relaxed LP dual of the subproblems (which can drastically decrease its efficiency), our cuts are uniquely designed for integer programs.

We demonstrate our methodology, both the modeling and the decomposition approaches, on two important logistics problems: the capacitated facility location problem with uncertain demands and the vehicle routing problem with uncertain demands and travel times. Our methodology can be applied to other problems in manufacturing and logistics, where binary requirements on decision variables are required. We show that our approach is very effective in solving such problems, i.e., we find an optimal solution up to three orders-of-magnitude faster than the original IP formulation. Furthermore, when the data uncertainty is represented by a set of scenarios (see Section 4.1), the number of scenarios does not
affect the computability of the decomposition approach, unlike other existing approaches.

In summary, the main contributions of this chapter are threefold. First, we introduce the Almost Robust Optimization framework that is simple, intuitive, flexible, and easy to communicate to managers. This formulation allows uncertainty in the decision making process and bridges the gap between stochastic programming and robust optimization. Second, we develop a decomposition approach to efficiently solve ARDO and show that our methodology is both practical and computationally tractable. Third, by defining the Robustness Index we demonstrate the potential advantages of being almost robust as opposed to being fully robust.

The chapter is organized as follows. Section 4.1 presents the general framework and formulation of ARDO. The details of the decomposition approach are described in Section 4.2. Two applications are provided in Section 4.3. Computational results are presented in Section 4.4. Section 4.5 extends the model by considering a general uncertainty structure. Section 4.6 concludes the paper. Proofs and extensions are included in the Appendices.

4.1. Almost Robust Discrete Optimization

We present models that incorporate data uncertainty in the decision making process using a scenario-based description of the data (in Section 4.5, we extend our approach to cases where the uncertainty has other structures). Scenarios are used since they are very powerful, convenient and practical in representing uncertainty, and they allow the decision maker flexibility in the specification of uncertainty (e.g., uncertain data may be highly correlated, sampled from simulation, and/or obtained from historical data).

We next present prior formulations in which scenario-based data are incrementally incorporated into a deterministic optimization model, and discuss their advantages and disadvantages. Then, we present our Almost Robust Discrete Optimization model.

4.1.1 Prior Optimization Models

**Deterministic Model:** Consider the deterministic discrete optimization problem:

\[
\begin{align*}
\min \quad & z = c^T x \\
\text{s.t.} \quad & Ax \leq b \\
& x \in X \subseteq \{0,1\}^N
\end{align*}
\]

(Deterministic Model)
where \( x, c, \) and \( b \) are \( N \times 1 \) vectors and \( A \) is a \( (M + J) \times N \) matrix.

The classical paradigm in mathematical programming assumes that the input data \( c, A \) and \( b \) is known. In practice, however, portions of this data may be uncertain.

Without loss of generality, we assume that the data uncertainty affects only a partition of \( A \). As is standard (see [17]), both \( c \) and \( b \) are assumed to be deterministic. (These can be achieved by incorporating them into the uncertainty partition of \( A \), by adding the constraint \( z - c^T x \leq 0 \), introducing the variable \( 1 \leq x_{n+1} \leq 1 \), and transforming the constraints to \( Ax - bx_{n+1} \leq 0 \).) A treatment on uncertain objectives in a more general manner is discussed in Section 4.1.2. We partition \( A \) into a deterministic \( M \times N \) matrix, \( D \), and an uncertain \( J \times N \) matrix \( U \). Similarly, \( b \) is partitioned into a \( M \times 1 \) vector \( b^D \) corresponding to \( D \), and a \( J \times 1 \) vector \( b^U \) corresponding to \( U \).

In the body of the chapter (except in Section 4.5) we assume that the uncertain data is represented by a set of scenarios \( S = \{1, 2, ..., |S|\} \), with probabilities \( p_s \) for each \( s \in S \). Each scenario is associated with a particular realization of the uncertainty matrix, \( U^s \), with corresponding entries \( u^s_{ij} \). Note that later in this section, we show that ARDO can also be setup to suit a wide spectrum of available probabilistic information such as: full probabilistic information, no probabilistic information, and partial probabilistic information.

We next introduce two existing models that incorporate scenarios into the deterministic model above and discuss their advantages and disadvantages.

**Robust Model:** Consider the following robust formulation

\[
\begin{align*}
\text{min} \quad & c^T x \\
\text{s.t.} \quad & Dx \leq b^D \\
& U^s x \leq b^U \quad \forall s \in S \\
& x \in X
\end{align*}
\]

The robust model follows the formulation of RO (see [17]). This formulation ensures that the optimal solution does not violate any of the constraints under any possible scenario realization.

This formulation has three main drawbacks. First, due to the scenarios, it may no longer have a feasible solution. Second, even when feasible solutions exist, they may be too conservative, i.e.,
the optimal solution may be very costly. The conservatism is magnified when a scenario with a low probability has a large effect on the optimal outcome. Specifically, a major flaw of the robust model is that it completely ignores available information regarding the probabilities of the different scenarios. Indeed, RO was developed to be effective when no specific probabilistic knowledge on the different realizations of $U$ exists. Finally, the size of the robust formulation grows with the number of scenarios and therefore this formulation becomes less tractable as the number of scenarios increases. Note that the deterministic problem is a discrete optimization problem, thus, it may by itself be difficult to solve, and the introduction of scenarios drastically increases this difficulty.

**Worst-Case Model:** An even more conservative problem is to optimize the worst-case. Let each element of the matrix $U'$ be equal to the maximum value of the corresponding element over all scenarios, i.e., $u'_{ij} = \max_s u^s_{ij}$. The worst-case formulation is

$$
\begin{align*}
\min & \quad c^T x \\
\text{s.t.} & \quad Dx \leq b^D \\
& \quad U'x \leq b^U \\
& \quad x \in X
\end{align*}
$$

(Worst-Case Model)

As can be seen, the worst-case model, which is a special case of [87] in which the uncertainty matrix belong to a convex set of scenarios, has all the drawbacks of the robust model (its solution is even more conservative). The main upside of this model is its smaller size that makes it more tractable (its complexity is the same as that of the deterministic model independent of the number of scenarios).

### 4.1.2 The Almost Robust Discrete Optimization Model

We next introduce a model that overcomes the issues of infeasibility and overconservatism, and which integrates the decision maker’s risk attitude into the optimization process. This model, which we refer to as the *Almost Robust Discrete Optimization* (ARDO) model trades off the solution value with robustness to find solutions that are almost robust (feasible under most realizations).

In ARDO, infeasibility of the uncertain constraints under some of the scenarios may be allowed at a penalty. Consider the $j$th uncertain constraint, $U^s_j x \leq b^U_j$. For every $j \in 1, \ldots, J$, the amount of infeasibility of solution $x$ in scenario $s$ can be measured by
\[ Q^*_j(x) = \max \{0, (U^*_j x - b^*_j)\} = (U^*_j x - b^*_j)^+ . \] (4.1)

Our approach allows other measures of infeasibility such as the absolute value of infeasibility, \( Q^j(x) = |U^*_j x - b^*_j| \), which is applicable for equality constraints. For ease of exposition, we focus on the infeasibility measure (4.1), and defer the explanation of the incorporation of the absolute value measure to Appendix C.

We now introduce penalty functions to penalize uncertain constraint violations. Three alternative penalty functions (based on the available probabilistic information) are:

- **Max Penalty Function**: this penalty is applicable when the decision maker does not have knowledge of the probabilities (a main assumption in RO):

\[
\overline{Q}_j(x) = \max_{s \in S} \alpha_s Q^s_j(x),
\] (4.2)

where \( \alpha_s \) is the per unit constrain violation penalty under scenarios \( s \). This penalty function determines the maximum (worst-case) penalty over all scenarios.

- **Expected Penalty Function**: this penalty is suitable when the decision maker has full probabilistic knowledge of the scenarios:

\[
\overline{Q}_j(x) = \sum_{s \in S} p_s \alpha_s Q^s_j(x),
\] (4.3)

- **Distributionally-Robust Penalty Function**: this penalty is applicable when the decision maker has some but limited information regarding the probability distribution of the scenarios, i.e., she knows that the probability distribution of the scenarios belongs to a known set of distributions, \( F \). Let \( p^f_s \) be the probability of scenario \( s \) under distribution \( f \in F \), and \( \overline{Q}^f_j(x) = \sum_{s \in S} p^f_s \alpha_s Q^s_j(x) \) be the expected penalty of uncertain constraint \( j \) under distribution \( f \). This penalty function has the form:

\[
\overline{Q}_j(x) = \max_{f \in F} \overline{Q}^f_j(x).
\] (4.4)

This penalty function determines the worst-case expected penalty over distributions in \( F \), thus, they enforce robustness with respect to a set of possible distributions.
For ease of exposition, we focus on the expected penalty function as in (4.3) and assume $\alpha_s = 1$. The inclusion of different $\alpha_s$ can be easily incorporated into the solution methodology and the proofs. In Appendix B, we extend the results to the max and distributionally-robust penalty functions. Furthermore, in Section 4.5, we extend the summation used in (4.3) to represent an expectation with respect to any well defined random variable.

**Remark.** Note that chance-constraints ([32]) and integrated chance constraints ([61]) are special case of ARDO, e.g., by modifying the penalty function to $\overline{Q}_j(x) = \sum_{s \in S} p_s I\{Q^*_s > 0\}$ where $I\{Q^*_j(x) > 0\} = 1$ ($I\{Q^*_j > 0\} = 0$) if $Q^*_j > 0$ ($Q^*_j \leq 0$), the model can be converted into a chance-constrained model.

We further define $\overline{Q}(x) = [\overline{Q}_1(x), \overline{Q}_2(x), ..., \overline{Q}_J(x)]^T$. Let $l_j$ denote the maximum penalty accepted by the decision maker for violating constraint $j$, so that the $J \times 1$ vector $l$ denotes the penalty limits. Specifically, the vector $l$ incorporates the decision maker attitude towards risk, i.e., a decrease in the values of $l$ implies the decision maker is less tolerant and is more conservative. Later in this section, we introduce the Robustness Index that the decision maker can use to determine an appropriate vector $l$ based on her risk attitude.

Using the above notation, the ARDO formulation is

$$
\begin{align*}
\min & \quad c^T x \\
\text{s.t.} & \quad Dx \leq b^D \\
& \quad \overline{Q}(x) \leq l \\
& \quad x \in X \subseteq \{0, 1\}^N
\end{align*}
$$

(ARD0)

As can be seen, this model allows the decision maker to control the magnitude of the penalties by changing $l$. For example, the robust model is a special instance of ARDO where $l_j = 0$, $\forall j$, i.e., when infeasibility of any constraint under any scenario is not allowed. Note that the model with $l_j = 0$, $\forall j$ is equivalent to the robust counterpart ([17]) in which the uncertainty set is a convex hull of the scenarios.

As stated earlier, when the coefficients of the objective function are uncertain, we augment them into the uncertain constraints by using the objective function $\min z$ and adding constraint $z - c^T x \leq 0$. In general, however, other objective functions can also be considered, e.g., the mean value objective, $\min \sum_{s \in S} p_x c^T x \equiv \min \overline{c}^T x$, where $\overline{c} = \sum_{s \in S} p_x c$, and the distributionally-robust objective: $\min \max_{f \in F} \sum_{s \in S} p_x f^T c^T x$. 

4.1.3 Robustness Index

As stated earlier, in order to aid the decision maker in determining the appropriate penalty limits, $l$, and to show a connection between the decision maker’s risk preference and the penalty limit, we introduce the Robustness Index. The Robustness Index can be used to investigate the trade-off between the changes in the objective function and in the penalty as the penalty limit changes. To formally introduce the Robustness Index, let $x^*_l$ denote the optimal solution of ARDO given the penalty limits $l$, and let $x^*_0$ be the optimal solution when violation is not allowed. Recall that the objective function value of the problem where infeasibility is not allowed is equal to the objective function of the (fully) robust model. Let $I_{1 \times J}$ denote a $J \times 1$ vector of 1s. We introduce the following Robustness Index:

$$\text{Robustness Index} = \frac{\text{improvement (decrease) in objective function value}}{\text{increase in total penalty}} = \frac{c^T x^*_0 - c^T x^*_l}{Q(x^*_l)^T I_{1 \times J}}. \quad (4.5)$$

The Robustness Index indicates the relative improvement in the objective function value, $c^T x^*_0 - c^T x^*_l$, (the difference between the objective function with penalty limits, $l$, and the objective function values where infeasibility is not allowed, i.e., $l = 0$) over the increase in the total penalty $(Q(x^*_l)^T I_{1 \times J})$.

Similar to “the price of robustness” ([23]) we expect that allowing some infeasibility while paying a penalty should improve the objective function. Moreover, it is reasonable to expect that the benefit from allowing a higher deviation alongside an increase in the corresponding penalty will be less effective as the allowed penalty increases. That is we expect some decreasing returns for allowing a higher penalty. As the Robustness Index in (4.5) measures the benefit of allowing a higher expected penalty, we expect that it will be a convex decreasing function of the penalty, $l$. Indeed our computational results in Section 4.4.2 agree with this intuition. Moreover, our results suggest that often a little increase in the penalty limit is all that is required in order to substantially improve the objective function.

4.1.4 Comparison with Existing Approaches

Here, we compare the exiting approaches that are used to incorporate uncertainty with the ARDO. Table 4.1 summarizes this discussion.

**RO vs. ARDO:** Both RO and ARDO seek to find solutions that are robust given an uncertainty set and can incorporate the decision makers’ risk attitude ([23], [75], and [15]); there are two main differences between them. First, unlike RO, ARDO can directly use probabilistic knowledge of the uncertain data. As stated above, ARDO can be setup to suit a wide spectrum of available probabilistic
information. Second, in RO the constraints are usually hard, i.e., the constraints must be satisfied for all possible realizations, (exceptions include [16] and [15]) while ARDO accommodates both hard and soft constraints, i.e., the decision maker has the choice to decide whether to allow infeasibility under some realizations. Note that even the framework of generalized RO that allows large data variations, which can be thought of as rare scenarios, treats constraints as hard ([17]).

Stochastic programming and Chance constrained programming vs. ARDO: The main similarity between these three models is that they use probabilistic knowledge of the uncertain parameters; but ARDO also applies to situations with limited probabilistic information. The main difference between stochastic programming and ARDO is that the former seek to minimize the expected cost (dis-utility) of a decision made in the first stage, while the latter seeks to find solutions that are almost robust (feasible under most realizations).

As stated in the Remark above, chance constraints are a special case of ARDO. Conceptually, however, there are two main differences between these models: first, ARDO is more suitable for cases where the decision maker faces quantitative risk (where measuring the amount of constraint violation is important), whereas chance-constraints are more suited for cases with qualitative risk (where instead of the amount of violation, any violation and its probability are important); second, while chance constraints require probabilistic knowledge of the uncertain parameters, ARDO can be applied to setting with limited (or no) probabilistic information.

Most importantly, while discrete variables can make stochastic programming and chance constrained programming computationally intractable, ARDO is much more applicable, e.g., unlike chance constraint, ARDO preserves the convexity of the constraints and as we show below it is much more tractable.

Robust mathematical programming vs. ARDO: Robust mathematical programming ([74]) is an approach that integrates goal programming with scenario-based description of the problem data. It seeks to generate solutions that are less sensitive to the specific realization of the data. ARDO retains the advantages of robust mathematical programming while offering the following additional advantages: first, ARDO allows control of the degree of conservatism for every constraint (i.e., gives the decision maker control over the amount of violation allowed for each uncertain constraint), thus, fully integrating the decision makers’ risk attitude in the optimization model. Second, while robust mathematical programming assumes that the probability distribution of the scenarios are known, ARDO can also be used for setting in which the probabilistic information of the uncertain data is limited. Finally, while robust mathematical programming focuses on continues optimization, we focus on discrete
Table 4.1: Comparison of ARDO with Other Approaches.

optimization models.

4.2. The Decomposition Approach

Due to the integrality of the decision variables and the need to linearize the penalty constraints (by introducing new variables and adding new inequality constraints that significantly increases the problem’s size) ARDO generally becomes computationally intractable as the number of scenarios increases. To overcome this issue and to solve ARDO efficiently, we decompose it into a deterministic master problem and a simple subproblem. As in most decomposition methods (see, for example, [18], [25], and [57]), we start by solving a relaxation of the original problem (the master problem) and iteratively adding cuts (derived from solving the subproblem) until a feasible (optimal) solution is found. Specifically, in the master problem, the uncertain parameters are replaced with their expected values (the master problem is similar to the deterministic problem). The single subproblem calculates the penalty of the different scenarios given the master problem solution and generates cuts when this solution is infeasible for the ARDO (i.e., when the master problem solution violates at least one penalty constraint). The cuts, when added to the master problem, eliminate at least the current infeasible master solution. The decomposition approach iterates between solving the master problem and the subproblem until the optimal solution to ARDO is found.

In order for the decomposition model to generate an optimal solution of ARDO, it must satisfy the following two conditions:

- Condition 1 - Lower Bound: In any iteration, the master problem solution must be a lower bound on that of ARDO.

- Condition 2 - Validity of the Cuts: A cut is valid if it (i) excludes the current master problem solution, which is not feasible to ARDO, and (ii) does not remove any feasible solutions of ARDO.

Condition 1 ensures that any feasible solution of ARDO is also feasible for the master problem. Condition 2(i) ensures that the decomposition converges in a finite number of steps, i.e., this condition
guarantees that the decomposition does not get stuck in a loop. Condition 2(ii) guarantees optimality, i.e., because the cuts never remove feasible solutions of ARDO, while removing infeasible solutions of ARDO (condition 2(i)), and since all feasible solutions of ARDO are also feasible for the master problem (condition 1), the decomposition is guaranteed to find an optimal solution to ARDO, if one exists.

4.2.1 Master Problem

Let $\mathbf{U} = E_s[U^s] = [\sum_{s \in S} p_s U^s_1, \ldots, \sum_{s \in S} p_s U^s_J]^T$ be a $J \times N$ matrix where each element is equal to the expectation of the corresponding element over all possible realization, and let the $J \times 1$ vector $l' = l + b^U$.

The deterministic master problem formulation is:

$$\begin{align*}
\min & \quad \mathbf{c}^T \mathbf{x} \\
\text{s.t.} & \quad \mathbf{D} \mathbf{x} \leq b^D \\
& \quad \mathbf{U} \mathbf{x} \leq l' \\
& \quad \text{cuts} \\
& \quad \mathbf{x} \in \{0, 1\}
\end{align*}$$

(Master Problem)

As stated earlier, cuts are constraints that are added to the master problem when its current solution is not feasible for ARDO. As can be seen, the master problem formulation is equivalent to a deterministic formulation with an extra set of cuts. Note that the cuts that we introduce later are both linear and deterministic.

4.2.2 Subproblem

The subproblem focuses on checking the master problem solution under different realizations and creating cuts for the master problem when the current solution is not feasible for ARDO. These cuts prevent the current infeasible master problem solution from being considered in future iterations without removing any feasible solution of the original ARDO.

The subproblem is comprised of three steps: (i) penalty calculation, (ii) feasibility check, and (iii) cuts generation. The pseudo code of the implementation of the decomposition approach is given in Algorithm 1. To discuss these steps let $x^t$ denote the optimal master problem solution in iteration $t$ of the decomposition, $\circ$ denote the entrywise product of two vectors (i.e., this binary operation takes two vectors of the same dimension and produces another vector where each element $i$ is the product of elements $i$ of the original two vectors), and $I$ be a vector of 1s with the same dimension as $x$ (i.e., $I$ is a
Algorithm 1 The Decomposition Approach for Solving ARDO

1. $cuts \leftarrow \emptyset$, $t \leftarrow 1$, $terminate \leftarrow \text{false}$;

2. while $terminate \neq \text{true}$ do

3.     $terminate \leftarrow \text{true}$;

4.     Solve master problem to obtain $x^t$;

5.     Solve subproblem given $x^t$:

6.     for each uncertain constraint $j$ do

6.1.     Run penalty calculation to obtain $\overline{Q}_j(x^t)$;

6.2.     Run feasibility check for $\overline{Q}_j(x^t) \leq l_j$;

6.3.     if $\overline{Q}_j(x^t) > l_j$ then

6.4.         Generate new cuts;

6.5.         $cuts \leftarrow cuts + \text{new cuts}$;

6.6.         $terminate \leftarrow \text{false}$;

6.7.     return $cuts$;

6.8.     $t \leftarrow t + 1$;

N $\times$ 1 vector of 1s).

Step 1 - Penalty Calculation: For any scenario $s \in S$ and uncertain constraint $j = 1, \ldots, J$ calculate $Q^s_j(x^t)$ . Then, calculate the penalty of each uncertain constraint given the master solution, $\overline{Q}_j(x^t) = \sum_{s \in S} p_s Q^s_j(x^t)$, $j = 1, \ldots, J$.

Step 2 - Feasibility Check: Check if $\overline{Q}(x^t) \leq l$. The feasibility check has two possible outcomes:

- No Violation: If the optimal solution of the master problem does not violate any penalty constraints, the decomposition approach is terminated with the current solution, which is optimal.

- Violation: If the optimal master solution violates at least one of the penalty constraints, determine the set of penalty constraints that are violated in iteration $t$, denoted by $V^t$, and the set of scenarios where uncertain constraint $j \in V^t$ has a non-zero penalty, denoted by $S^t_j$, where $S^t_j = \{ s | Q^s_j(x^t) > 0 \}$, and continue to step 3.

Step 3 - Cuts Generation: Generate and add to the master problem the cuts:

$$\overline{Q}_j(x^t) - \sum_{s \in S^t_j} p_s ((U^s_j \circ x^T)(I - x \circ x^t) - U^s_j((I - x^t) \circ x)) \leq l_j, \quad \forall j \in V^t. \quad (4.6)$$

As can be seen, the cuts are linear and deterministic.

The cuts ensure that the changes in the current infeasible solution are sufficient to reduce the current penalty to a level below the allowable limit. That is, the cuts remove all supersets and some infeasible
subsets of the current infeasible solution of the ARDO from the feasible region of the master problem. Specifically, the first term in the left hand side of the cuts (4.6) is the penalty of the master problem solution at iteration $t$ as calculated in step 1. The term $(U_s^s \circ x^T_s)(I - x \circ x^t)$ measures the change in the penalty of constraint $j$ in scenario $s$ when some elements of $x^t$ are switched from 1 to 0. That is, by removing elements (switching their solution value from 1 to 0) with positive coefficients, the penalty decreases, while removing elements with negative coefficients increases the penalty. Similarly, the term $U_s^j((I - x^t) \circ x)$ measures the change in the penalty of constraint $j$ in scenario $s$ when some elements of $x^t$ are switched from 0 to 1. That is, by adding elements (switching their solution value from 0 to 1) with negative coefficients, the penalty decreases, while adding elements with positive coefficients increases the penalty. Therefore, for any future master problem solution, $x$, the summation measures the maximum change in the penalty of constraints $j$. The cuts ensure that this change is sufficient, i.e., that the corresponding uncertain constraint is no longer violated.

Theorem 4.1, whose proof appears in Appendix A, states that the decomposition model generates an optimal solution of ARDO because it satisfies conditions 1 and 2 from above.

**Theorem 4.1 (Convergence of the decomposition approach).** If ARDO has an optimal solution, the decomposition approach finds this solution in a finite number of iterations.

In theory, the number of iterations of the decomposition may be large. However, as confirmed by our computational experiments, in practice, the number of iterations of the decomposition is small; in our experiments the average number of iterations was about 3.

Note that the cuts in (4.6) are the most general cuts, however, other problem specific cuts can also be used. For example, if all coefficients of $U$ are non-negative (e.g., units of time), an alternative simpler valid cut would exclude the term $U_s^j((I - x^t) \circ x)$ in (4.6).

### 4.3. Applications

We next illustrate our methodology on two important logistics problems: the capacitated facility location problem with uncertain demands and the vehicle routing problem with uncertain demands and travel times.

#### 4.3.1 The Capacitated Facility Location Problem with Uncertain Demand

Consider the single-source capacitated facility location problem. There is a set of potential facilities, $H$, and a set of customers, $G$. For each facility $h$ there is an opening cost, $f_h$, and a capacity on the total
demand that can be served, \( b_h \). Furthermore, each customer has demand, \( d_g \), and providing service to customer \( g \) from facility \( h \) has an associated cost, \( c_{gh} \). The goal is to select the set of facilities to open and assign customers to facilities in order to minimize the total cost, under the conditions that facility capacities are not exceeded.

Let the binary variable \( y_h \) indicate whether facility \( h \) is selected and \( x_{gh} \) indicate whether customer \( g \) is assigned to facility \( h \). The integer programming formulation of the single-source capacitated facility location problem is:

\[
\begin{align*}
\min & \quad \sum_{h \in H} f_h y_h + \sum_{g \in G} \sum_{h \in H} c_{gh} x_{gh} \\
\text{s.t} & \quad \sum_{h \in H} x_{gh} = 1 \quad g \in G \\
& \quad x_{gh} \leq y_h \quad g \in G, h \in H \\
& \quad \sum_{g \in G} d_g x_{gh} \leq b_h y_h \quad h \in H \\
& \quad x_{gh}, y_h \in \{0, 1\} \quad g \in G, h \in H
\end{align*}
\] (4.7)

The above formulation assumes that the customer demand data is perfectly known. In practice, however, this assumption may not hold. We use ARDO to consider such uncertainties. Assume that the customer demands are uncertain, and that different demand conditions are considered as different scenarios, \( d_{gs} \), with known probability \( p_s \). Due to this uncertainty, the capacity constrains in (4.9) may no longer hold under all scenarios. We will allow violations of constraints \( h \in H \) under some scenarios, \( Q^*_h(x) = (\sum_{g} d_{gs} x_{gh} - b_h y_h)^+ \), at a penalty, \( Q_h(x) = \sum_{s} p_s Q^*_h(x) \). The objective is to minimize the total cost such that the penalty of each facility, \( Q_h(x) \), does not exceed a given threshold value, \( l_h \). Recall that the threshold value for each facility reflects the decision makers willingness to have less than the required capacity at this potential facility. Note that if there is no facility located at location \( h \), we have \( y_h = x_{gh} = 0 \ \forall g \), and therefore \( Q_h(x) = 0 \).

In this problem, the deterministic part of the formulation \( (Dx \leq b^D) \) is captured by constraints (4.7) and (4.8), and the uncertain part \( (Ux \leq b^U) \) is represented by constraint (4.9).

The ARDO formulation of the single-source capacitated facility location problem with uncertain demand is:
\[
\min \sum_{h \in H} f_h y_h + \sum_{g \in G} \sum_{h \in H} c_{gh} x_{gh} \\
\text{s.t. } (4.7), (4.8), (4.10) \\
\quad \overline{Q}_h(x) \leq l_h, \quad h \in H
\]

**Decomposition Approach:** We solve this problem using the decomposition.

**Master Problem:** Let \( \overline{d}_g = \sum_s p_s d^s_g \). The formulation of the master problem is:

\[
\min \sum_{h \in H} f_h y_h + \sum_{g \in G} \sum_{h \in H} c_{gh} x_{gh} \\
\text{s.t. } (4.7), (4.8), (4.10) \\
\quad \sum_{g \in G} \overline{d}_g x_{gh} \leq l_h + b_h y_h \quad h \in H
\]

**Subproblem:**

- **Step 1 - Penalty Calculation:** Given the set of facilities opened in the solution of the master problem in iteration \( t \), \( H^t = \{ h | y^t_h = 1 \} \), and the set of customers assigned to facility \( h \in H^t \), \( G^t_h = \{ g | x^t_{gh} = 1 \} \), calculate the penalty cost of all facilities \( h \in H^t \) in scenario \( s \), \( Q^s_h(x^t) = ( \sum_{g \in G^t_h} d^s_g - b_h)^+ \), and the expected penalty cost of each facility over all the scenarios, \( \overline{Q}_h(x^t) = \sum_s p_s Q^s_h(x^t) \), \( h \in H^t \).

- **Step 2 - Feasibility Check:** If \( \overline{Q}_h(x^t) \leq l_h \), \( \forall h \in H \), terminate. Otherwise, determine the set of facilities that violate the penalty constraints, \( V^t \), and go to step 3

- **Step 3 - Cut Generation:** Generate and add the cuts:

\[
\overline{Q}_h(x^t) - \sum_{s \mid Q^s_h(x^t) > 0} p_s (\sum_{g \in G^t_h} d^s_g (1 - x_{gh}) - \sum_{g \in G - G^t_h} d^s_g x_{gh}) \leq l_h, \quad h \in V^t.
\]

The inner summation term is the decrease/increase in the penalty cost of the corresponding facility, given some of the customers are reassigned. As can be seen, the cut prevents the customers in \( G^t_h \) or a super-set of them from being assigned to facility \( h \) in future iterations. As discussed in Section 4.2, since demand is non-negative, by removing \( \sum_{g \in G - G^t_h} d^s_g x_{gh} \), the cut would still be valid.
4.3.2 Vehicle Routing Problem with Uncertain Demands and Travel-Times

Let $G = (N, A)$ be an undirected graph, where $N$ is the set of nodes and $A$ is the set of arcs. Node 0 is a depot of a fleet of $K$ identical vehicles each with capacity $q$, and all routes must start and end at the depot. All other nodes represent customers with an associate demand $d_u$, $u \in N \\setminus \{0\}$. Each arc $(u, v)$ has a distance or cost $c_{uv}$, and using arc $(u, v)$ generates travel time $z_{uv}$ for the vehicle performing the service. The objective is to design $K$ vehicle routes of minimum total cost/distance, such that (i) the total demand of all customers in any route does not exceed the vehicle capacity, $q$, and (ii) the total travel time of any vehicle $k$ does not exceed its maximum travel time limit, $r_k$ (e.g., due to union work time regulation).

Let the binary variable $x_{uvk}$ indicate if arc $(u, v)$ appears on the route of vehicle $k$, and let $y_{uk}$ indicate whether customer $u$ is served by vehicle $k$. The integer programming formulation of the vehicle routing problem is:

$$\min \sum_k \sum_{(u,v) \in A} c_{uv}x_{uvk}$$

s.t. $\sum_{v \in N \\setminus \{0\}} x_{uvk} = y_{uk}$ \hspace{1cm} u \in N, \forall k$ \hspace{1cm} (4.11)

$\sum_k y_{uk} \leq 1$ \hspace{1cm} u \in N \hspace{1cm} (4.12)

$\sum_{u \in N \\setminus \{0\}} x_{iok} - \sum_{v \in N \\setminus \{0\}} x_{ovk} = 0$ \hspace{1cm} o \in N, \forall k \hspace{1cm} (4.13)

$\sum_{v \in N \\setminus \{0\}} x_{0vk} = 1$ \hspace{1cm} \forall k \hspace{1cm} (4.14)

$\sum_{u \in N \\setminus \{0\}} x_{u0k} = 1$ \hspace{1cm} \forall k \hspace{1cm} (4.15)

$\sum_{u \in N \\setminus \{0\}} d_u y_{uk} \leq q$ \hspace{1cm} \forall k \hspace{1cm} (4.16)

$\sum_{(u,v) \in A} z_{uv}x_{uvk} \leq r_k$ \hspace{1cm} \forall k \hspace{1cm} (4.17)

$x_{uvk}, y_{uk} \in \{0,1\}$ \hspace{1cm} (u, v) \in A, u \in N, \forall k \hspace{1cm} (4.18)

Unlike the previous application in which only one aspect was uncertain, we now assume that the decision maker faces uncertainty in both the demands and the travel times. Assume different demands and travel times are represented by a set of demand scenarios ($S^{DE}$) and a set of travel time scenarios ($S^{TT}$), i.e., serving customer $u$ in scenario $s \in S^{DE}$ generates demand $d^s_u$ and using arc $(u, v)$ in scenario $s' \in S^{TT}$ generates a travel time $z^s_{uv}$. Furthermore, each demand scenario $s \in S^{DE}$ occurs
with probability \( p_s^{DE} \), while each travel time scenario \( s' \in S^{TT} \) has a known probability \( p_s^{TT} \). Since demands and travel times are uncertain, constraints (4.16) and (4.17) may be violated under some scenarios. ARDO will allow such infeasibility at a penalty. Since two aspects of the problem are uncertain, we need to determine the infeasibility for both demands, \( Q_s^{DE} k(x) = \left( \sum_{u \in N \setminus \{0\}} d_s^u y_{uk} - q \right)^+ \forall s \), and travel times, \( Q_s^{TT} k(x) = \left( \sum_{(u,v) \in A} z_{uv}^s x_{uvk} - r_k \right)^+ \forall s' \). The objective of ARDO is to minimize total cost/distance, such that the demand penalty, \( Q_s^{DE} k(x) = \sum_{s \in S^{DE}} p_s^{DE} Q_s^{DE} k(x) \), and travel time penalty, \( Q_s^{TT} k(x) = \sum_{s' \in S^{TT}} p_s^{TT} Q_s^{TT} k(x) \), of each vehicle do not exceed threshold values \( l_k^{DE} \) and \( l_k^{TT} \), respectively.

In this problem, the deterministic part of the formulation \((Dx \leq b^D)\) is comprised of constraints (4.11)-(4.15), while the uncertain part \((Ux \leq b^U)\) is represented by constraints (4.16) and (4.17). The ARDO formulation of the vehicle routing problem with uncertain travel times is:

\[
\begin{align*}
\min & \quad \sum_{k \in K} \sum_{(u,v) \in A} c_{uv} x_{uvk} \\
\text{s.t} & \quad (4.11) - (4.15), (4.18) \\
& \quad \overline{Q}_k^{DE} (x) \leq l_k^{DE} \quad \forall k \\
& \quad \overline{Q}_k^{TT} (x) \leq l_k^{TT} \quad \forall k 
\end{align*}
\]

The above formulation produces solutions that are almost robust with respect to the different realizations of demands and travel times.

**Decomposition Approach:** We apply the decomposition approach to solve this problem:

**Master Problem:** Let \( \overline{d}_u = \sum_{s \in S^{DE}} p_s d_u^s \) be the mean demand of customer \( u \) and \( \overline{z}_{uv} = \sum_{s' \in S^{TT}} p_s^{TT} z_{uv}^{s'} \) be the mean travel time of arc \((u,v)\). Furthermore, let \( l_k^{DE'} = l_k^{DE} + q \) and \( l_k^{TT'} = l_k^{TT} + r_k \). The master problem is:

\[
\begin{align*}
\min & \quad \sum_{k \in K} \sum_{(u,v) \in A} c_{uv} x_{uvk} \\
\text{s.t} & \quad (4.11) - (4.15), (4.18) \\
& \quad \sum_{u \in N \setminus \{0\}} \overline{d}_u y_{uk} \leq l_k^{DE'} \quad \forall k \\
& \quad \sum_{(u,v) \in A} \overline{z}_{uv} x_{uvk} \leq l_k^{TT'} \quad \forall k \\
& \quad \text{cuts}
\end{align*}
\]

**Subproblem:**
Step 1 - Penalty Calculation: Given the set of nodes ($N^t_k = \{u | y^t_{uk} = 1\}$) and the set of arcs ($A^t_k = \{(u, v) | x^t_{uvk} = 1\}$) assigned to vehicle $k$ in iteration $t$, calculate the demand penalties, $Q^DE_{k}(x^t) = \sum_{s \in S^DE} p^DE_s (\sum_{u \in N^t_k} d^s_u - q) \forall k$, and the travel time penalties, $Q^{TT}_{k}(x^t) = \sum_{s' \in S^{TT}} p^{TT}_{s'} (\sum_{(u,v) \in A^t_k} z^s_{uv} - r)^+ \forall k$.

Step 2 - Feasibility Check: If $Q^DE_{k}(x^t) \leq l^DE_k$ and $Q^{TT}_{k}(x^t) \leq l^TT_k \forall k$, terminate. Otherwise, determine the set of vehicles that violate the demand penalty constraints, $V^DE_t$, and the travel time penalty constraints, $V^{TT}_t$. Go to step 3.

Step 3 - Cut Generation: Generate and add the following two sets of cuts (one for the demands and one for the travel times):

$$Q^DE_{k}(x^t) - \sum_{\{s | Q^DE_{k}(x^t) > 0\}} p^DE_s (\sum_{u \in N^t_k} d^s_u (1 - y^t_{uk}) - \sum_{u \in N^t_k \setminus N^t_k} d^s_u y^t_{uk}) \leq l^DE_k, \forall k \in V^DE_t,$$

$$Q^{TT}_{k}(x^t) - \sum_{\{s' | Q^{TT}_{k}(x^t) > 0\}} p^{TT}_{s'} (\sum_{(u,v) \in A^t_k} z^{s'}_{uv} (1 - x^t_{uvk}) - \sum_{(u,v) \in A^t_k \setminus A^t_k} z^{s'}_{uv} x^t_{uvk}) \leq l^{TT}_k, \forall k \in V^{TT}_t.$$

The inner summation terms of the two cuts are the decrease/increase in the demand and travel time penalties of the corresponding vehicle, given that the nodes/arcs are reassigned. Note that in iterations where only one type of uncertain penalty constraint (either the demand or the travel time penalty constraint) is violated, the decomposition approach only generates cuts for the violated type, e.g., when only demand penalty constraints are violated, the decomposition approach solely generates demand cuts.

4.4. Computational Experiments

In previous sections we claimed that our decomposition approach is efficient in solving the ARDO. Here, we present experimental results for the capacitated facility location problem with uncertain demand presented in the previous section to support our claim.

We conducted two experiments to address the following:

1. Efficiency: Our initial experiment addresses the computational efficiency of the decomposition approach by comparing its performance with the original IP formulation of ARDO (note that, in what follows, the abbreviation ARDO denotes the original IP formulation of ARDO).
2. **Value of Almost Robustness:** In the second experiment, we use the Robustness Index (defined in Section 4.1.3) to show the connection between the decision-maker’s risk preference and the penalty limit. Furthermore, by using this index, we highlight the potential benefits of being almost robust as opposed to being fully robust.

**Problem Instances:** The problem instances were randomly generated as follows: the facility capacities are uniformly drawn from an integer interval $b_h \in [50, 200]$; the fixed facility opening costs, $f_h$, are randomly generated based on the capacities $b_h$ and are calculated as: $f_h = b_h(10 + \text{rand}[1; 15])$, where 10 is the per unit capacity cost. We have added a random integer multiplier from $[1, 15]$ to take into account differences in property costs. The assignment costs are uniformly generated from an interval $c_{gh} \in [1, 100]$. The demands for each scenario is uniformly generated from the interval $d_{sg} \in [1, 30]$. To generate the probabilities of the different scenarios we assigned each one of them a random number in the range $[1, 100]$ in a uniform manner. We then normalized these numbers to probabilities by dividing by the sum of the numbers assigned to all scenarios.

Overall, the problem instances consist of three sizes: $30 \times 15$ (i.e., 30 customers, 15 possible facility sites), $40 \times 20$, and $50 \times 25$; four number of scenarios: 5, 15, 30, and 60; four different penalty limits ($l$) 10, 20, 30, and 40. Finally, six instances of each size-scenario-limit combination are generated for a total of $3 \times 4 \times 4 \times 6 = 288$ instances.

The tests were performed on a Dual Core AMD 270 CPU with 1 MB cache, 4 GB of main memory, running Red Hat Enterprise Linux 4. The models are implemented in ILOG CPLEX. A time limit of three hours was used, and for unsolved instances (instance that where timed-out) the time-limit is used in the calculations (we also report the percentage of unsolved instances).

### 4.4.1 Efficiency of the Decomposition Approach

Figures 4.1 and 4.2 illustrate the effects of the number of scenarios on the run-time of ARDO and the decomposition approach. As can be seen, the mean and median run time of ARDO drastically increases with the number of scenarios. However, the number of scenarios has only a minor effect if any on the run time of the decomposition approach.
Table 4.2 gives the mean, median, and the 10th and 90th percentiles of the CPU time in seconds required to solve each problem instance for each scenario size for both ARDO and the decomposition approach. The column labeled “% Uns.” indicates the percentage of the problems for which the models were terminated because of the time limit (of 3 hours). For the decomposition approach, the column labeled “Iter.” indicates the average number of iterations needed to converge to optimality. The “Time ratio” for a given instance is calculated as the ARDO runtime divided by the decomposition run-time.
Chapter 4. Almost Robust Optimization and Its Applications in Logistics

Table 4.2: The mean, median, and the 10th and 90th percentiles of the CPU time (in Seconds) and percentage of unsolved problem instances ("% Uns.") for the ARDO and decomposition approaches, and the mean number of iterations ("Iter.") for the decomposition approach. "Overall" indicates the mean results over all problem instances.

<table>
<thead>
<tr>
<th>Number of Scenarios</th>
<th>ARDO Mean</th>
<th>ARDO Median</th>
<th>ARDO Percentile 10</th>
<th>ARDO Percentile 90</th>
<th>ARDO % Uns.</th>
<th>Decomposition Mean</th>
<th>Decomposition Median</th>
<th>Decomposition Percentile 10</th>
<th>Decomposition Percentile 90</th>
<th>Decomposition % Uns.</th>
<th>Decomposition % Iter.</th>
<th>Time Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>655</td>
<td>13.2</td>
<td>1.13</td>
<td>309.41</td>
<td>4.2</td>
<td>258</td>
<td>2.1</td>
<td>0.24</td>
<td>61.17</td>
<td>1.3</td>
<td>3.2</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>1188</td>
<td>98.9</td>
<td>4.12</td>
<td>1187.88</td>
<td>8.9</td>
<td>133</td>
<td>4.9</td>
<td>0.28</td>
<td>72.99</td>
<td>0.2</td>
<td>2.4</td>
<td>91</td>
</tr>
<tr>
<td>30</td>
<td>3238</td>
<td>557.4</td>
<td>18.39</td>
<td>10800</td>
<td>23.6</td>
<td>248</td>
<td>3.2</td>
<td>0.33</td>
<td>65.63</td>
<td>1.3</td>
<td>3.4</td>
<td>240</td>
</tr>
<tr>
<td>60</td>
<td>5524</td>
<td>5883.7</td>
<td>70.67</td>
<td>10800</td>
<td>43.1</td>
<td>286</td>
<td>5.1</td>
<td>0.28</td>
<td>77.78</td>
<td>1.3</td>
<td>3.1</td>
<td>566</td>
</tr>
<tr>
<td>Overall</td>
<td>2651</td>
<td>222.8</td>
<td>5.03</td>
<td>10800</td>
<td>19.5</td>
<td>231</td>
<td>3.7</td>
<td>0.27</td>
<td>76.73</td>
<td>1.0</td>
<td>3.0</td>
<td>229</td>
</tr>
</tbody>
</table>

The mean over each instance in each subset was then calculated.

As Table 4.2 indicates the decomposition model is very efficient. Specifically, the Decomposition approach is unable to find the optimal solution within the time limit in only 1.0% of instances, while ARDO is unable to find the optimal solution in 19.5% of the instances. For ARDO, we see an increasing trend in the number of unsolved instance as the number of scenarios increases (up to 43.1%). Furthermore, as the time ratios indicate, on average, the decomposition approach is 229 times faster than ARDO. We again see an increasing trend of the time ratios as the number scenarios increase (up to 566). Finally, the table demonstrates that the decomposition approach converges to optimality on average in 3 iterations.

Figure 4.3 shows a scatterplot of the run-times for both ARDO and the decomposition approach. Both axes are log-scaled, and the points below the $x = y$ line indicate a lower runtime for the decomposition approach. On all but 8 of the 288 instances (over 97%), the decomposition achieves equivalent or better run-time (Note that 6 out of the 8 instances that are dominated are for the problem instances with 5 scenarios). We summarize:

**Observation 1. (Computational Efficiency)** The decomposition is much more efficient than the original IP formulation of ARDO, i.e., it is up to three orders-of-magnitude faster, and can find the optimal solution for many more problems within the time limit.

4.4.2 Value of Almost Robustness

In this section, we show how the Robustness Index in (4.5) can be used by the decision makers to choose an appropriate penalty limit to better suit her risk preferences. We also highlight the benefits of being almost robust as opposed to fully robust.

To study the Robustness Index, we calculated it for all the problem instances discussed earlier for the capacitated facility location problem with uncertain demand. We summarized the results in Table 4.3 and Figure 4.4. In Table 4.3, we show the mean, median, max, and the 10th and 90th percentiles of the Robustness Index over all problems instances with a particular $l$ value. Figure 4.4 depicts the
changes in the average Robustness Index (the average is taken over all the instances with a given $l$) as a function the penalty limit. As expected from the discussion in Section 4.1.3, as the penalty limit increases the average incremental improvement in the objective function value increases slower than the average incremental increases in the penalty. Therefore, it may not be necessary for the decision maker to choose a very high penalty limit since her relative gain (in terms of improvement in the solution value) may not grow as fast as the loss (in terms of the penalty).

This experiment also demonstrates that allowing a small amount of infeasibility can substantially improve the solution value. Specifically, from Table 4.3 we can see that as we go from $l = 0$ to $l = 5$, the Robustness Index can be as high as 1397, which indicates that the improvement in the objective function value can be up to 1397 times more than the expected penalty. This suggests that the solution of ARDO can be (substantially) less conservative than that of the robust model, and therefore, highlights the potential benefits of being almost robust (instead of being fully robust). Moreover, Table 4.3 also

<table>
<thead>
<tr>
<th>Penalty Limit ($l$)</th>
<th>Robustness Index Mean</th>
<th>Median</th>
<th>Max</th>
<th>Percentile 10</th>
<th>Percentile 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>105.4</td>
<td>100.2</td>
<td>1397.8</td>
<td>6.4</td>
<td>306.6</td>
</tr>
<tr>
<td>10</td>
<td>43.3</td>
<td>40.8</td>
<td>146.3</td>
<td>5.3</td>
<td>71.3</td>
</tr>
<tr>
<td>20</td>
<td>30.6</td>
<td>27.7</td>
<td>103.1</td>
<td>4.5</td>
<td>50.3</td>
</tr>
<tr>
<td>30</td>
<td>23.1</td>
<td>22.8</td>
<td>103.1</td>
<td>4.1</td>
<td>31.5</td>
</tr>
<tr>
<td>40</td>
<td>21.7</td>
<td>19.3</td>
<td>103.1</td>
<td>3.8</td>
<td>26.9</td>
</tr>
</tbody>
</table>

Table 4.3: The mean, median, max, and the 10th and 90th percentiles of the Robustness Index as a function of the penalty limit, $l$.  

Figure 4.3: Runtime of the ARDO (x Axis, Log-Scale) vs. decomposition model (y Axis, Log-Scale)
emphasizes that a little increase in the expected penalty is all that is required in order to substantially improve the objective function. To summarize:

**Observation 2. (The Robustness Index)** The decision maker can use the Robustness Index to determine the appropriate penalty limit based on her risk attitude. The index demonstrates that a (possibly) small amount of penalty can substantially improve the objective function.

### 4.5. General Uncertainty Structure

In this section, we extend ARDO to allow a more general uncertainty structure. Assume that the uncertainty matrix $U$ has a multivariate cumulative distribution, $f$ (representing the joint distribution of $u_{ij}$, for each $i, j$), with mean $\mu$, and covariance matrix $\Sigma$, and support $\Omega(U)$. Based on the available probabilistic information, penalty function (3) can be extended to $Q_j(x) = E[(U_jx - b_j^U)^+] = \int_{\Omega(U)}(U_jx - b_j^U)^+dF$, while penalty function (4) can be extended to $Q_j(x) = \max_{f \in F}E_f[(U_jx - b_j^U)^+]$, where $F$ denotes the of set of probability distributions with mean $\mu$ and covariance matrix $\Sigma$ that is assumed to include the true distribution $f$. Given the appropriate penalty function, the almost robust model with a general uncertainty structure is:
\[
\begin{align*}
\min & \quad c^T x \\
\text{s.t.} & \quad Dx \leq b^D \\
& \quad Q_j(x) \leq l_j, \quad j = 1, \ldots, J \quad (\text{Generalized ARDO}) \\
& \quad x \in X
\end{align*}
\]

As an example, when \( l = 0 \) (when no violation is allowed) and \( \Omega(U) \) is bounded, the generalized model is equivalent to the robust counterpart ([17]), where the uncertain parameter belongs to the uncertainty set \( \Omega(U) \).

Due to the multidimensional integrals, which are computationally prohibitive when the dimension exceeds four (see [34]), the Generalized ARDO (even with continues variables) may be computationally intractable. Therefore, it may be essential to find tractable approximation for the penalty function. For each constraint \( j \), let vector \( U_j \) denote the mean of \( U_j \) and \( \Sigma_j \) denotes the corresponding covariance matrix. The next lemma provides a distribution-free bound, \( B_j(x) \), on the penalty function.

**Lemma 4.1 (Distribution free bound on the penalty function).** For any distribution, the following upper bound always holds:

\[
E_f[\{(U_j x - b_j U_j)^+\}] \leq \frac{1}{2}((U_j x - b_j U_j) + \sqrt{\Sigma_j xx^T + (U_j x - b_j U_j)^2}) = B_j(x), \quad \forall f \in F. \tag{4.19}
\]

In the Generalized ARDO, by replacing constraints \( Q_j(x) \leq l_j, j = 1, \ldots, J \) with constrains \( B_j(x) \leq l_j, j = 1, \ldots, J \), the formulation is converted into a second-order cone programming model, which is much more tractable. Since, from Lemma 4.1, \( \overline{Q}_j(x) \leq B_j(x) \), the optimal solution of this model is a feasible (and conservative) solution of the Generalized ARDO.

**Decomposition Approach** The decomposition approach presented in Section 4.2 can be modified to solve the Generalized ARDO. As before, we use the first moment matrix, \( \overline{U} \), in the master problem. Furthermore, the subproblem follows the same steps as in Section 4.2: (i) in the penalty calculation step of the subproblem, we calculate \( \overline{Q}_j(x^t) = E[(U_j x^t - b_j U_j)^+] \) (if we have full probabilistic knowledge) or \( \overline{Q}_j(x^t) = \max_f E_f[(U_j x^t - b_j U_j)^+] \) (if we have partial probabilistic knowledge). Note that in circumstances where calculating \( \overline{Q}_j(x^t) \) is not computationally feasible, we can instead calculate \( B_j(x^t) \). Doing so however, can result in finding more conservative solutions; (ii) in the feasibility check step we examine if any penalty constraints is violated; (iii) in the cut generation step, the following modified cuts are
generated and added to the master problem:

\[ \mathcal{Q}_j(x^t) - ((U_j \circ x^t)(I - x \circ x^t) - U_j((I - x^t) \circ x)) \leq l_j, \quad \forall j \in V^t. \]

**Corollary 4.1 (Optimality of the decomposition approach with general uncertainty).** *If the Generalized ARDO has an optimal solution, the decomposition approach produces an optimal solution for it in a finite number of iterations.*

### 4.6. Conclusions

We presented the Almost Robust Discrete Optimization model that is simple, intuitive, flexible, and easy to communicate to managers. This model allows uncertainty in the decision making process and bridges stochastic programming and robust optimization. We also developed a decomposition approach to efficiently solve ARDO. Furthermore, we defined and studied the Robustness Index that can be used by decision makers to adjust ARDO to better suit her risk preferences. By using this index we, demonstrated the potential advantages of being almost robust as opposed to being fully robust. Finally, we demonstrated our methodology to two important logistics problems.

We believe that the approach of Almost Robust Discrete Optimization is applicable in many settings and for many problems. We hope that this approach will be used in many future studies. We also intend to establish additional theoretical results on the solvability of ARDO problems and the Robustness Index for different families of optimization models.
Proof of Theorem 4.1. As explained in Section 4.2, conditions 1 and 2 are sufficient to establish Theorem 4.1. We first prove condition 2, that the cuts (4.6) are valid. We start by establishing 2(i), i.e., that the cut excludes the current infeasible master problem solution from all subsequent solutions. Suppose the optimal master problem solution is infeasible in iteration $t$. Then there exist at least one penalty constraint, $i$, such that:

$$Q_i(x^t) > l_i.$$  \hspace{1cm} (4.20)

Then, the generated cut for this constraint, from (4.6), is

$$Q_i(x^t) - \sum_{s \in S^t_i} p_s ((U^s_i \circ x^{t^T}) (I - x \circ x^t) - U^s_i ((I - x^t) \circ x)) \leq l_i.$$ \hspace{1cm} (4.21)

If the same solution $x^t$ is found in future iterations, $(U^s_i \circ x^{t^T}) (I - x \circ x^t) - U^s_i ((I - x^t) \circ x) = 0$, so that the left hand side of (4.21) equal to $Q_i(x^t)$ that from (4.20) is greater than $l_i$, creating a contradiction. Thus, $x^t$ is infeasible in future iterations of the master problem.

We now prove 2(ii), that is, the proposed cuts do not remove any feasible solution of ARDO: let $x^w$ be a feasible solution of ARDO found in iteration $w > t$,

$$Q_j(x^w) \leq l_j, \hspace{1cm} j = 1, ..., J.$$ \hspace{1cm} (4.22)

We define $x^{10} = x^t \circ (I - x^w)$, $x^{11} = x^t \circ x^w$, and $x^{01} = (I - x^t) \circ x^w$. Thus, $x^{10}$ is a vector with elements equal to 1 if these elements are 1 in $x^t$ and 0 in $x^w$, 0 otherwise. Similarly, $x^{11}$ is a vector with elements equal to 1 if these elements are 1 in both $x^t$ and $x^w$, 0 otherwise. Finally, $x^{01}$ is a vector with elements equal to 1 if these elements are 0 in $x^t$ and 1 in $x^w$, 0 otherwise.

We present a proof by contradiction. Assume that $x^w$ does not satisfy a cut formed in iteration $t$ for constraint $i$. Since the elements that are 1 in $x^{11}$ are not included in the cut, they do not contribute to the LHS of (4.21). Thus, from the LHS of (4.21) when $x = x^w$,

$$Q_i(x^t) - \sum_{s \in S^t_i} p_s (U^s_i x^{10} - U^s_i x^{01}) > l_i.$$ \hspace{1cm} (4.23)

Furthermore, since $U^s_i x^t = U^s_i x^{10} + U^s_i x^{11}$, $Q_i(x^t)$ can be rewritten as,
where the last equality follows by only summing over the scenarios with non-negative penalties, \(S'_i\).

Therefore, by replacing \(\mathcal{Q}_i(x^*)\) in the LHS of (4.23) by (4.24) we have

\[
\sum_{s \in S'_i} p_s(U_i^s x^{11} + U_i^s x^{01} - b_i^U) > l_i. \tag{4.25}
\]

Since \(U_i^s x^w - b_i^U = U_i^s x^{11} + U_i^s x^{01} - b_i^U\), thus, \(\mathcal{Q}_i(x^w) = \sum_{s \in S} p_s(U_i^s x^w - b_i^U) = \sum_{s \in S} p_s(U_i^s x^{11} + U_i^s x^{01} - b_i^U)^+\). Therefore,

\[
\mathcal{Q}_i(x^w) = \sum_{s \in S'_i} p_s(U_i^s x^{11} + U_i^s x^{01} - b_i^U)^+ \geq \sum_{s \in S} p_s(U_i^s x^{11} + U_i^s x^{01} - b_i^U)^+ \\
\geq \sum_{s \in S'_i} p_s(U_i^s x^{11} + U_i^s x^{01} - b_i^U) > l_i, \tag{4.26}
\]

where the first inequality is due to \(\sum_{s \in S-S'_i} p_s(U_i^s x^{11} + U_i^s x^{01} + b_i^U) \geq 0\), the second due to \((U_i^s x^{11} + U_i^s x^{01} - b_i^U) \leq \max\{0, (U_i^s x^{11} + U_i^s x^{01} - b_i^U)\}\), and the last from (4.25). Therefore, from (4.26) we have \(\mathcal{Q}_i(x^w) > l_i\), which contradicts (4.22) and thus, our assumption that the feasible solution \(x^w\) does not satisfy the cut. This establishes that the cut does not remove any feasible solution of ARDO.

Since both conditions 2(i) and 2(ii) are satisfied, we conclude that the cut is valid.

We now prove condition 1, that the optimal objective value of the master problem is a lower bound on that of ARDO. We first note that because the cuts are valid, they do not remove any feasible solution of ARDO at any iterations of the decomposition. Moreover, since the objective function and the deterministic constraints of ARDO also appear in the master problem, we must only take into consideration the penalty constraints of ARDO, \(\mathcal{Q}_j(x) = \sum_{s} p_s(U_j^s x - b_j^U)^+ \leq l_j \ \forall j\), and the expected-value constraints of the master problem, \(\bar{U}_j x \leq l_j' = \sum_{s} p_s(U_j^s x - b_j^U) \leq l_j \ \forall j\). In particular, we must show that any feasible solution of ARDO is also feasible to the master problem, i.e., if \(\mathcal{Q}_j(x) \leq l_j\) then \(\bar{U}_j x \leq l_j' = l_j + b_j^U\). Indeed,

\[
\bar{U}_j x - b_j^U = \sum_{s} p_s(U_j^s x - b_j^U) \leq \sum_{s} p_s(U_j^s x - b_j^U)^+ = \mathcal{Q}_j(x), \quad \forall j. \tag{4.27}
\]

Thus, we conclude that any feasible solution of ARDO is also feasible for the master problem. Therefore,
the optimal objective value of the master problem is a lower bound on that of \( \text{ARDO} \).

Since both conditions are satisfied, and since the decision variables have a finite domain, the decomposition model produces an optimal solution to \( \text{ARDO} \) in a finite number of iterations.

**Proof of Lemma 4.1.** It is easy to see:

\[
(U_j x - b_j^U)^+ = \frac{|U_j x - b_j^U| + (U_j x - b_j^U)}{2}.
\]

By taking the expectation of both sides of the above equation, and applying the Cauchy-Schwarz inequality to \( E_f [U_j x - b_j^U] \), we obtain the bound.

**4.7.2 Appendix B - Extensions**

The exposition of the paper focuses on the expected penalty function. In this section, we modify the decomposition approach to capture both the worst-case and the distributionally-robust penalty functions.

**Worst-Case Penalty Function:** The following modifications in the decomposition approach are required: (i) in the master problem, replace the expected value matrix \( U = E_s[U^s] \) by a minimum matrix \( U' \) with elements equal to the minimum of the corresponding element over all scenarios; (ii) in Step 1 of the subproblem calculate the maximum penalty \( (\bar{Q}_j(x^t) = \max_s Q_s^j) \) instead of the expected penalty; (iii) in Step 2 of the subproblem let \( s^j_t \) denote the scenario with the maximum penalty; and (iv) in Step 3 use the modified cut:

\[
\bar{Q}_j(x^t) - (U_j^s \circ x^T)(I - x \circ x^t) - U_j^s((I - x^t) \circ x) \leq l_j, \quad \forall j \in V^t.
\] (4.28)

**Corollary 4.2.** Cuts of the form (4.28) are valid, and the modified decomposition approach produces an optimal solution to \( \text{ARDO} \) with the worst-case penalty function in a finite number of iterations.

The proof of Theorem 4.1 directly applies here, we thus omit the proof.

**Distributionally-Robust Penalty Function:** This penalty function is applicable when the decision maker has limited probabilistic information. As is common in the literature (see, for example, [84, 21] and [81]), we assume that the decision maker knows that the probability distribution of the scenarios belongs to a known set of probability distributions, \( F \), and has an estimate of the first moment of the uncertainty matrix, \( \bar{U} \).

The following modifications in the decomposition approach are required: (i) as in the decomposition approach use the first moment matrix, \( \bar{U} \), in the master problem (if no information on first moment is
available, use \( \mathcal{U} \), as described for the worst-case penalty function; (ii) in Step 1 of the subproblem, for each uncertain constraint \( j = 1, \ldots, J \) calculate the expected penalty under distribution \( f \), \( \mathcal{Q}_j^f(x^f) \) \( j = 1, \ldots, J \), \( \forall f \in F \), determine the distribution with the maximum expected penalty \( f_t^j = \arg \max_{f \in F} \mathcal{Q}_j^f(x^f) \) and let \( \mathcal{Q}_j = \mathcal{Q}_j^{f_t}(x^t) \); (iii) in Step 3 use the modified cuts:

\[
\mathcal{Q}_j(x^t) - \sum_{s \in S^t_j} p_{s}^{f_t} [(U^s_j \circ x^T)(I - x \circ x^t) - U^s_j((I - x^t) \circ x)] \leq l_j, \quad \forall j \in V^t. \tag{4.29}
\]

Corollary 4.3. Cuts of the form (4.29) are valid, and the modified decomposition approach produces an optimal solution to ARDO with the distributionally-robust penalty function in a finite number of iterations.

\textbf{Proof.} The proof of condition 2(i) is similar to that in Theorem 4.1 and is omitted. Similar to Theorem 4.1, we use contradiction to prove condition 2(ii): we assume \( x^w \) is a feasible solution of the distributionally-robust model, i.e.,

\[
\max_{f \in F} \mathcal{Q}_j^f(x^w) \leq l_j, \quad j = 1, \ldots, J, \tag{4.30}
\]

that does not satisfy the cut (4.29) of constraint \( i \). Therefore, given \( \mathcal{Q}_j(x^t) = \mathcal{Q}_j^{f_t}(x^t) \) the cut when \( x = x^w \) is

\[
\mathcal{Q}_j^{f_t}(x^t) - \sum_{s \in S^t_j} p_{s}^{f_t} [U^s_j x^T + U^s_j x^1 - b_j^U] > l_i. \tag{4.31}
\]

Since \( \mathcal{Q}_j^{f_t}(x^t) = \sum_{s \in S^t_j} p_{s}^{f_t} [U^s_j x^T + U^s_j x^1 - b_j^U] \) (as in Theorem 4.1 we removed superscript \( ^+ \) by summing over \( S^t_j \)), (A19) is reduced to

\[
\sum_{s \in S^t_j} p_{s}^{f_t} [U^s_j x^1 + U^s_j x^0 - b_j^U] > l_i. \tag{4.32}
\]

We now have two possible outcomes:

(i) \( f^w_i = \arg \max_{f \in F} \mathcal{Q}_i^f(x^w) \equiv f^t_i \): In this case, the distribution in iteration \( w \) with the maximum penalty is the same as that of iteration \( t \). Under this condition, \( \max_{f \in F} \mathcal{Q}_i^f(x^w) = \sum_{s \in S} p_{s}^{f^w_i} [U^s_i x^1 + U^s_i x^0 - b_i^U]^+ \) \( = \sum_{s \in S} p_{s}^{f^t_i} [U^s_i x^1 + U^s_i x^0 - b_i^U]^+ \) \( \). From this and following the argument in (4.26) we get a contradiction, thus under this condition, the cut is valid.

(ii) \( f^w_i \neq f^t_i \): In this case:
\[
\max_{f \in F} Q_f^T(x^w) = \sum_{s \in S} p_s^w [U_s^x x^{11} + U_s^x x^{01} - b^l_j]^+ \\
\geq \sum_{s \in S} p_s^w [U_s^x x^{11} + U_s^x x^{01} - b^l_j]^+, \quad (4.33)
\]

where the inequality is due to that \( f^w \) has the maximum expected penalty in iteration \( w \). From (4.33) and following the argument in (4.26) we again get a contradiction, thus, under this condition, the cut is valid.

From the above two conditions we conclude that the cut is always valid. Furthermore,

\[
E_f[U_s^x x - b^l_j] = \bar{U}_j x - b^l_j \leq E_f[U_s^x x - b^l_j]^+, \quad \forall f \in F, \ j = 1, \ldots, J. \quad (4.34)
\]

From (4.34) and using the same argument as the proof of condition 1 of Theorem 4.1, we can conclude that the master problem solution is a lower bound on that of the distributionally-robust model. \( \square \)

### 4.7.3 Appendix C - Absolute Value of Infeasibility Measure

An alternative measure of uncertainty is the absolute value of infeasibility, \(|U_s^x x - b^l_j|\), which is suitable for equality constraints where neither upward nor downward infeasibility is desired. For such measure, we consider the expected penalty function \( \overline{Q}_j(x) = \sum_{s \in S} p_s[U_s^x x - b^l_j]|. \) Note that the max or the distributionally-robust penalty functions introduced in Section 4.1 could also be used if full probabilistic information is not available.

The decomposition approach now requires two major modification. First, in step 1 of the subproblem, we calculate the absolute infeasibility, \( Q_j^T(x^t) = |U_j^s x^t - b^U_j|, \ j = 1, \ldots, J. \) Second, in step 3, the cuts are modified to take into account both downward and upward deviations. In order to introduce the modified cuts, let \( n^{s,t} = 0 \) if \( U_j^s x - b^U_j > 0 \) and \( n^{s,t} = 1 \) if \( U_j^s x - b^U_j < 0 \). The cuts are:

\[
\overline{Q}_j(x^t) - \sum_{s \in S_j^t} p_s(-1)^{n^{s,t}} ((U_s^x \circ x^t^T)(I - x \circ x^t) - U_j^s((I - x^t) \circ x)) \leq l_j, \quad \forall j \in V^t. \quad (4.35)
\]

As can be seen, the cuts are identical to the cuts presented in (4.6) with an addition term \((-1)^{n^{s,t}}\). \( n^{s,t} \) is defined such that in scenarios with downward deviation \( (n^{s,t} = 1) \), removing an element with a positive (negative) coefficients from \( x^t \) increases (decreases) the penalty of the scenario (reflected by \( U_j^s((I - x^t) \circ x) \)), while adding an element with negative (positive) coefficient decreases (increases) the penalty (enforced through \( (U_j^s \circ x^t^T)(I - x \circ x^t) \)). For scenarios with upward deviation the effects are
reversed.

**Corollary 4.4** (Convergence of the decomposition approach with the absolute value measure). Cuts of the form (4.35) are valid, and the modified decomposition approach produces an optimal solution to ARDO with the absolute value measure in a finite number of iterations.

**Proof.** Since the proof of condition 1 and condition 2(i) (see Section 4.2) is similar to the proof presented for Theorem 4.1, we omit it.

We again use contradiction to prove condition 2(ii), thus assume \( x^w \) is a feasible solution of ARDO with the absolute value measure, i.e.,

\[
\overline{Q}_j(x^w) \leq l_j \quad j \in 1, ..., J,
\]

and it does not satisfy the cut of constraint \( i \) (of the form (4.35)). Let \( x^{10}, x^{11}, \) and \( x^{01} \) be as defined in Appendix A. From (4.35) when \( x = x^w \):

\[
\overline{Q}_i(x^t) - \sum_{s \in S^t_i} p_s(-1)^{n^{s,t}}(U^s_i x^{10} - U^s_i x^{01}) > l_i.
\]

Furthermore, since \( U^s_i x^t = U^s_i x^{10} + U^s_i x^{11} \) and \( \overline{Q}_i(x^t) = \sum_{s \in S^t_i} p_s |U^s_i x^t - b^U_i| \), (4.37) can be rewritten as,

\[
\sum_{s \in S^t_i} p_s(|U^s_i x^{10} + U^s_i x^{11} - b^U_i| - (-1)^{n^{s,t}}(U^s_i x^{10} - U^s_i x^{01})) > l_i.
\]

We define \( S^t_i^+ = \{ s \in S^t_i | n^{s,t} = 0 \} \) to be the set of scenarios with upward deviation, i.e., \( U^s_i x^t - b^U_i > 0 \) \( \forall s \in S^t_i^+ \), and \( S^t_i^- = \{ s \in S^t_i | n^{s,t} = 1 \} \) to be the set of scenarios with downward deviation, i.e., \( U^s_i x^t - b^U_i < 0 \) \( \forall s \in S^t_i^- \). Since \( U^s_i x^w = U^s_i x^{11} + U^s_i x^{01} \), thus \( \overline{Q}_i(x^w) = \sum_{s \in S} p_s |U^s_i x^w - b^U_i| = \sum_{s \in S} p_s |U^s_i x^{11} + U^s_i x^{01} - b^U_i| \). Therefore,

\[
\overline{Q}_i(x^w) = \sum_{s \in S} p_s |U^s_i x^{11} + U^s_i x^{01} - b^U_i| \\
= \sum_{s \in S^t_i^+} p_s |U^s_i x^{11} + U^s_i x^{01} - b^U_i| + \sum_{s \in S-S^t_i^+} p_s |U^s_i x^{11} + U^s_i x^{01} - b^U_i| \\
\geq \sum_{s \in S^t_i^+} p_s |U^s_i x^{11} + U^s_i x^{01} - b^U_i| \\
= \sum_{s \in S^t_i^+} p_s (U^s_i x^{11} + U^s_i x^{01} - b^U_i) + \sum_{s \in S^t_i^-} p_s (-U^s_i x^{11} + U^s_i x^{01} - b^U_i) \\
\geq \sum_{s \in S^t_i^+} p_s (U^s_i x^{10} + U^s_i x^{11} - b^U_i) + \sum_{s \in S^t_i^-} p_s (-U^s_i x^{11} + U^s_i x^{01} - b^U_i) \\
= \sum_{s \in S^t_i^+} p_s (U^s_i x^{10} + U^s_i x^{11} - b^U_i) - (-1)^{n^{s,t}}(U^s_i x^{10} - U^s_i x^{01}) \\
> l_i.
\]
where the first inequality is due to \( \sum_{s \in S - S_i^+} p_s |U_i^s x^{11} + U_i^s x^{01} - b_i^U| \geq 0 \), the second due to \( |U_i^s x^{11} + U_i^s x^{01} - b_i^U| \geq (-(U_i^s x^{11} + U_i^s x^{01} - b_i^U)) \), the last equality is by the definition of \( n^{s,t} \) and \( S_i^t^+ \) and \( S_i^t^- \), and the last inequality is from (4.38).

From (4.39) we have \( \bar{Q}_i(x^w) > t_i \), which contradicts (4.36). Therefore, the cut does not remove any feasible solution of ARDO with the absolute value measure (condition 2(ii)). \( \square \)
Bibliography


