

MODELING AND SOLUTION OF SOME MULTI-PERIOD SUPPLY CHAIN OPTIMIZATION PROBLEMS

A Thesis
Presented to
The Academic Faculty

by

Weihong Hu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2016

Copyright © 2016 by Weihong Hu

MODELING AND SOLUTION OF SOME MULTI-PERIOD SUPPLY CHAIN OPTIMIZATION PROBLEMS

Approved by:

Professor Alejandro Toriello, Advisor
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Maged M. Dessouky
Department of Industrial and Systems
Engineering
University of Southern California

Professor Natasha L. Boland
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Alan L. Erera
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Chelsea C. White III
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Date Approved: May 4, 2016

To my family.

In memory of Lei Peng.

ACKNOWLEDGEMENTS

This thesis could not have been completed without the efforts of many people.

I would like to express my sincere gratitude to my advisor and the committee. Throughout my years in the PhD program, Professor Alejandro Toriello offered precious research opportunities, financial support as well as academic guidance. Professor Maged Dessouky closely participated in the majority of the studies, and constantly encouraged me even in the most difficult situations. Professor Natasha Boland, Professor Alan Erera and Professor Chelsea White provided timely and constructive comments from earlier proposal to the final defense. I have been privileged to be navigated by these excellent scholars on the path towards an effective researcher.

Numerous people kindly helped me adjust myself in the transitions first from China to the U.S. and then from the west coast to the southeast. Thank you for being there, Dr. James Moore II, Edward Wang, Julia Pan, Evelyn Felina, Pamela Morrison, Mark Danielson, and those I should not have forgotten.

I truly appreciate all the friends and acquaintances who made my stays joyful at the Georgia Institute of Technology and the University of Southern California. I am not social enough to know everyone here in Atlanta, but feel very fortunate to share bittersweet memories with Amelia Musselman, Minkyong Kang, Fangfang Xiao, Can Zhang, Jan Vlachy, Satya Malladi, Daniel Blado, Mathias Klapp and so many other knowledgeable but humble, smart but simple, serious but funny fellow students. I have always been impressed by the passionate folks in Los Angeles, including Christine Nguyen, Qian An, Joongkoo Cho, Michael Poremba, Huayu Xu, Xiaoqing Wang, Lunce Fu, Yihuan Shao, Han Zou, Wentao Zhang, Liang Liu, Zhanyuan Gao, Fangqi Liu, Zheng Zeng, etc. My graduate life might have been boring without their

inspiration.

My special thanks go to Yuening Yang and Lei Peng. Yuening, thanks for accompanying me like a sister. Your unwavering faith keeps me thinking in the pursuit of happiness. Lei, some people say you were not strong enough, but they don't know how much power you've passed during my long journey of self-therapy.

I owe the ultimate thanks to my family, whose love runs deep and never fades.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	ix
LIST OF FIGURES	x
SUMMARY	xii
I INTRODUCTION	1
1.1 Contribution	4
II STRATEGIC HEALTH WORKFORCE PLANNING	7
2.1 Introduction	7
2.1.1 Our Contribution	9
2.1.2 Literature Review	10
2.2 Model Formulation and Assumptions	11
2.3 Optimal System Behavior	17
2.4 Computational Study	23
2.4.1 Deterministic Experiments	23
2.4.2 Experiments with Stochastic Demand Growth	28
2.5 Conclusions	32
III INTEGRATED INVENTORY ROUTING AND FREIGHT CONSOLIDATION OF PERISHABLE GOODS	35
3.1 Introduction	35
3.2 Literature Review	38
3.3 Problem Statement	40
3.4 Solution Approach	46
3.4.1 Decomposition	47
3.4.2 Local Search	49
3.4.3 An Iterative Framework	57

3.5	Computational Study	59
3.5.1	Benchmark Approaches	59
3.5.2	Test Instances	60
3.5.3	Single Iteration Results	61
3.5.4	Multiple Iteration Results	64
3.5.5	Discussion	65
3.6	Conclusions	67
IV	THE “REVERSED” ONE-WAREHOUSE MULTI-RETAILER PROBLEM WITH MIXED TRUCKLOAD COSTS	69
4.1	Introduction	69
4.2	Literature Review	73
4.3	Preliminaries	78
4.4	Approximations for Mixed Truckload Costs	81
4.4.1	FTL	83
4.4.2	Short-haul FTL+stepwise LTL and long-haul FTL	85
4.4.3	FTL+linearized LTL or FTL+LTL+courier	89
4.5	Single-Echelon Subproblems	91
4.6	An Improved Approximation for FTL Costs	99
4.6.1	Algorithm	99
4.6.2	Approximation Ratio	111
4.6.3	Tightness	114
4.7	Asymptotic Analysis	117
4.8	Extensions	120
4.8.1	Regular Modified All-Unit Discounts	120
4.8.2	Perishable Products	122
4.9	Conclusions	124
V	CONCLUSION	127
APPENDIX A	— COMPLEX PROOFS	129

APPENDIX B	— EXTENDED WORKFORCE MODEL	150
APPENDIX C	— DETAILED DECOMPOSITION RESULTS . .	152
REFERENCES	159

LIST OF TABLES

1	Sample demand sequences and solutions with no hiring for $m = 3, 4, 5$ in Example 2	18
2	A sketch of three subproblems	47
3	Multiple iteration statistics	64
4	Comparison of our study to the OWMR literature	78
5	“Reversed” OWMR-FTL instances	80
6	Approximations for OWMR/JRP variants with FTL costs	84
7	Exact subproblem solutions	96
8	Worst-case example	115
9	Single iteration: small instances, nonpeak demand	153
10	Single iteration: small instances, peak demand	153
11	Single iteration: medium instances, nonpeak demand	153
12	Single iteration: medium instances, peak demand	154
13	Multiple iterations: small instances, nonpeak demand	154
14	Multiple iterations: small instances, peak demand	155
15	Multiple iterations: medium instances, nonpeak demand	156
16	Multiple iterations: medium instances, peak demand	157
17	Lower bounds and optimality gaps, medium instances	158

LIST OF FIGURES

1	Flow chart of model used in computational examples	25
2	Breakdown of the total number of admissions and hirings in baseline scenario and scenarios 1 through 5 over the course of 20 years	27
3	Breakdown of the total number of admissions and hirings in scenarios 6 through 8 over the course of 20 years	28
4	Gap between forecast factor δ lookahead policy and full information cost. All lookahead policy total costs are within 1.2%.	30
5	Breakdown of total cost under different policies: payroll costs dominate the other costs.	31
6	Breakdown of controllable cost under different policies: look-ahead policies stay within 40% from <i>full information</i> model with minimal gap of 22% at $\delta^* = 33\%$	31
7	Volume-dependent long-haul shipping costs	39
8	Integrated distribution network	41
9	MMD and reassignment strategies	52
10	Flowchart of the iterative framework	58
11	Demand distribution for two grower-retailer pairs	60
12	Average performance of a single iteration for small instances	62
13	Average performance of a single iteration for medium instances	63
14	Cost and volume breakdown for a small-sized sample	65
15	Per iteration cost and volume breakdown for a medium-sized sample	66
16	The distribution network: “reversed” OWMR	70
17	Mixed truckload costs	72
18	Relevant cost structures	77
19	Transformation of FTL costs	84
20	Transformation of FTL+stepwise LTL costs	86
21	Transformation of FTL+linearized LTL costs	90
22	Transformation of FTL+LTL+courier costs	91
23	Concave batch cost transformation of FTL+stepwise LTL costs	93

24	Concave batch cost transformation of FTL+LTL+courier costs	93
25	Recombination strategies for Case 1	103
26	Recombination strategies for Case 2	106
27	Patterns of holding cost changes	113
28	Transformations of regular modified all-unit discounts to concave batch costs	121
29	A network flow representation of subproblems for decentralized perishable C-OWMR	123

SUMMARY

This thesis consists of three parts, each of which contributes to an independent topic in the broad area of multi-period supply chain optimization, and provides modeling and solution approaches for the problem in concern.

Part I studies a strategic health workforce planning problem. Analysts predict impending shortages in the health care workforce, and wages for health care workers already account for over 50% of U.S. health expenditures. It is thus increasingly important to adequately plan to meet health workforce demand at reasonable cost. Using infinite linear programming (LP) methodology, we propose an infinite-horizon model for health workforce planning in a large health system for a single worker class, e.g. nurses. We give a series of common-sense conditions any system of this kind should satisfy, and use them to prove the optimality of a natural lookahead policy. We then use real-world data to examine how such policies perform in more complex systems; in particular, our experiments show that a natural extension of the lookahead policy performs well when incorporating stochastic demand growth.

Part II investigates an integrated inventory routing (IRP) and freight consolidation problem for perishable goods with a fixed lifetime. The problem is motivated by the status-quo of logistics in many U.S. markets, but also adapts to relevant two-echelon supply chain optimization problems e.g. combined production planning and distribution. We formulate the problem as a large-scale mixed-integer programming (MIP) model. We propose an iterative solution framework with a decomposition procedure and a local search scheme. In the decomposition, a freight consolidation subproblem is first solved to obtain crucial shipping decisions, and after fixing these a restrictive

model generates the other decisions for the integrated problem. The local search aims at fast identification of good neighborhoods by solving an assignment-style MIP which matches the consolidation decision with an IRP subproblem, and gradually strengthens the incumbent solution pool when executed in an iterative fashion. Experiments with empirical demand distributions based on real data demonstrate that 1) the integration can achieve remarkable efficiency compared to a sequential approach of the subproblems; 2) both the decomposition and the local search are effective in solving moderately-sized problem instances that are already challenging in practice.

Part III examines a two-echelon distribution problem which can be viewed as a one-warehouse multi-retailer (OWMR) problem reversed in time flows. Unlike the majority of the OWMR literature, where ordering is uncapacitated and the ordering cost is fixed in each period, we assume more realistic volume-dependent cost structures which can be interpreted as multiple transportation modes with batch capacities. The resulting transportation costs are piecewise linear non-convex functions of the shipping volume. Since this breaks classical optimality properties like zero-inventory-ordering, and the LP relaxation of the natural MIP formulation can be very weak, a straightforward application of previous OWMR methods may not be effective. We first introduce a technique that converts our problem to existing OWMR models by bounding the cost functions, and derive the corresponding worst-case approximation guarantees under each type of the transportation costs. We then treat or approximate the transportation costs as concave batch costs, and propose a polynomial-time 2-approximation algorithm by recombining single-echelon lot sizing subproblem solutions for the special case of full truckload costs, where trucks have an identical capacity and each incurs a fixed cost. This improves the best-known result (i.e. a 3.6-approximation) for the same cost structure. Finally, we utilize subproblem structural properties to prove the asymptotic optimality of a decentralized approach for pertinent two-echelon problems in a wider range of settings.

CHAPTER I

INTRODUCTION

Supply chain systems involve all kinds of industrial entities from raw material suppliers to end customers. As a consequence of the integral nature, planning, operation and control of a supply chain generally span multiple time epochs and geographical locations. Therefore, effective supply chain management necessitates trade-offs and coordination in the context of multi-period decision-making, which usually results in complicated optimization problems.

The broad area of multi-period supply chain optimization problems contains a vast literature with an enormous number of applications. There are several distinctions between these problems.

- Time horizon: This defines the length of time for which one must plan a decision, and can be categorized as *finite horizon* or *infinite horizon*. Finite horizon planning typically arises in two settings ([101]). First, some problems have a very specific horizon where the current decisions do not or rarely impact the future states. Second, some problems actually have an infinite horizon, but it matters most what to do right now given a particular situation. The cargo loading/discharging plan of a voyage, the scheduling of nurse shifts are such examples. Infinite horizon formulations are usually studied when long-term effects are crucial or the process eventually evolves into a steady state. Classical examples include strategic portfolio management, periodic inventory control, etc. An infinite horizon is *continuous-time* or *discrete-time* based as to whether a time interval can be divided into arbitrarily smaller granules, whereas a finite horizon is by name on a discrete-time basis. Both categories have been

extensively studied, but fundamental theories for finite horizon problems do not naturally extend to their infinite counterparts.

- Demand type: This characterizes the demand generation or evolution process. The demand is said to be *deterministic* if it is known at the time of decision-making, and *uncertain* (e.g. stochastic with some probabilistic distribution) otherwise. Rigorously speaking, the future demand is almost always uncertain; however, it is reasonable to assume deterministic demand in finite horizon planning where trustworthy data is available for the near future (e.g. with advanced information technology). Furthermore, deterministic theories can provide streamlined optimization models that capture various trade-offs in supply chain management ([89]); hence many infinite horizon problems also consider deterministic demand.

Another demand attribute is whether it is *stationary* or *dynamic* over time. Under deterministic demand, continuous-time infinite horizon problems usually assume constant demand arrival rates. Discrete-time stationary demand assumes an identical demand quantity in each time period, whereas discrete-time dynamic demand assumes fluctuating demand quantities from period to period. Dynamic demand is often (albeit not always) considered under a finite horizon, and the motivation again comes from knowledge of the demand only for a limited time interval. On the other hand, stationary demand is broadly studied with both finite horizon and infinite horizon models. The optimal solution of an infinite horizon problem is typically hard to compute even for stationary demand, whereas simplified solutions like periodic policies can be over-restrictive ([97]); therefore, a reasonable approach is to periodically repeat a non-periodic policy, which justifies the consideration of stationary demand for a finite horizon.

- Solution integrality: This specifies whether part of a solution must take integer

values. While decisions like staff size, shipping units, ordering quantities, etc. only make sense to be integral in practice, modeling them as integer variables can render a problem theoretically intractable. In fact, a decimal may not affect the solution quality much if the magnitude of a number is not too small; hence it is common to allow fractions for modeling convenience and then round the solution for the actual decision-making in such situations. On the other hand, if fractions are not negligible and integrality must be explicitly modeled, there are also cases where a model is specially structured so that it attains optimality at integral points even if the integrality constraints are relaxed or reformulated. These properties are what we would like to utilize when it comes to solution approaches.

Mathematically speaking, two modeling tools have been widely applied to multi-period decision-making, namely *dynamic programming* (DP) and *linear/mixed integer linear programming* (LP/MIP). DP, also known as a recursive solution technique, defines a problem with a set of states (e.g. inventory levels) and admissible actions (e.g. production lot sizes) from each state. As such, DP ideally captures the inherent system dynamics, and can be very powerful in optimal policy design when structural properties are established. However, the number of states in a DP tends to explode exponentially as the problem size increases, hence the notorious “curse of dimensionality” often limits its usage to moderately sized instances. On the other hand, LP defines a problem with a set of continuous variables, a linear objective function and a set of linear constraints, whereas MIP generalizes some of the variables to the discrete case (i.e. integers). In contrast to DP, LP/MIP are capable of explicitly modeling various requirements and flexible in optimal or near-optimal solution approaches, though the problem structure may suffer a lot to account for dynamics especially when data unavailability or uncertainty is an issue. Since both tools have advantages and disadvantages, there is no universal answer to which one fits better,

and the development of each should be case-by-case.

This thesis concerns three multi-period optimization problems that stem from the supply chain practice. The first problem represents decision-making over an infinite horizon and is amenable to LP techniques. The second problem ranges over a finite horizon and naturally leads to a large-scale MIP model. The third problem has variants in both finite and infinite horizon planning, but we focus on the finite horizon version and discuss combined methodologies of MIP, LP as well as DP. We consider deterministic dynamic demand and assume known input parameters for all the problems unless otherwise specified. We also conduct experiments based on real or empirical data for the first two problems, which can provide useful insights in much more complicated scenarios and justify the behavior of our approaches in a variety of stochastic settings.

1.1 Contribution

In Chapter 2, we take an initial step in understanding long-term implications of personnel staffing for large health care systems. While conventional health workforce management models have been focusing on operational/tactical time frames and the few relevant long-term models might over-simplify system dynamics, we analyze the interaction of diverse planning decisions like hiring, training and promotion in an infinite horizon framework, and derive easy-to-implement optimal policies implied by structural properties of our model. More specifically, we

- Propose an infinite-horizon LP model for long-term health workforce planning of a single worker type in large public health systems,
- Prove the optimality of a natural lookahead policy under mild assumptions for deterministic demand,
- Conduct experiments on stochastic demand based on real data using natural

policy extensions, and

- Provide managerial insights on health workforce management, e.g. forecast horizons, risk-aversion levels, cost control.

In Chapter 3, we investigate a novel MIP model that integrates inventory routing and freight consolidation of perishable goods. Our interest in this study is motivated by the challenges that many U.S. agricultural supply chains are facing, similar problems that prevail in production and distribution logistics, as well as the lack in pertinent literature and technical difficulties in attempts to address them. More specifically, we

- Formulate a MIP model with routing, consolidation, and inventory components,
- Propose a model-based decomposition consisting of three subproblems to balance computational time and solution quality,
- Build an ad hoc assignment-style MIP model to fix bad local optimality via iterative decomposition,
- Conduct experiments with empirical demand distributions based on real data to demonstrate the effectiveness of both solution approaches for small to medium size instances, and
- Deliver preliminary findings that shed light on possible extensions of the proposed approach for larger problem instances.

In Chapter 4, we study a two-echelon distribution problem with a bunch of non-convex transportation cost structures. The distribution network can be viewed as a variant of that in Chapter 3, where we do not allow routing and only consider one destination. When the network is reversed in time flows, it results in a conventional *one-warehouse multi-retailer* (OWMR) problem. Our cost assumptions, however, are

more realistic than the majority of the OWMR literature since we consider flexible options of transportation modes at both echelons. We analyze the system from two perspectives:

- By treating the transportation costs as special *piecewise linear* (PWL) non-convex functions on \mathbb{R}_+ , we introduce a technique to convert our problem to existing OWMR models, and derive the resulting worst-case approximation guarantees under each type of the transportation costs.
- By treating or approximate the transportation costs as concave batch costs, we propose a polynomial-time 2-approximation algorithm based on single-echelon lot sizing subproblems for the special case of full truckload costs. That is, the algorithm takes the optimal subproblem solutions as input, and yields a feasible solution for the two-echelon problem with a total cost no higher than 2 times the optimal objective value for any instance of the problem. This improves the best-known result (i.e. a 3.6-approximation) for the same cost structure. We also give an example that shows tightness of the approximation ratio in the worst case.
- Enlightened by the subproblem structural properties under the latter perspective, we prove the asymptotic optimality of a decentralized approach for relevant two-echelon problems in a wider range of settings.
- We extend some results to ordering cost discounts and perishable goods.

Chapter 2 is a joint work with Professor Mariel Lavieri (UMich), Professor Alejandro Toriello (GaTech), and Xiang Liu (UMich), who designed and conducted systematic experiments. Chapter 3 and Chapter 4 are joint works with Professor Alejandro Toriello and Professor Maged Dessouky (USC).

We provide concluding remarks in Chapter 5.

CHAPTER II

STRATEGIC HEALTH WORKFORCE PLANNING

2.1 Introduction

Health workforce planning plays a key role in the United States and worldwide. Analysts project that by 2020 the U.S. will experience a shortage of up to 100,000 physicians, up to one million nurses and up to 250,000 public health professionals [123]. Adequate staffing of medical units has been shown to have a direct impact in the quality of patient care [94], and also accounts for a considerable fraction of health care costs, with wages for health care workers representing 56% of the \$2.6 trillion spent on health care in the United States in 2010 [83].

As the U.S. population continues to age [121] and demand for health care continues to grow, different sectors of the population will compete for constrained and costly health care resources. It thus becomes increasingly important to understand how the health care needs of the population are linked to long-term workforce management plans of doctors, nurses and other medical personnel. The challenge is to ensure that sufficient resources are available in the future to meet the growing health care needs of the population, while accounting for the costs associated with meeting these needs. These workforce levels should meet the demand for resources in the present and be positioned to meet demand for the foreseeable future [109], an essentially infinite horizon. Furthermore, workforce plans should account for lags implied by training new members of the workforce, attrition stemming from retirements, firings and resignations, and also the adequate supervision of workers at different levels of the workforce hierarchy by their superiors.

Current practice has mostly focused on monitoring and evaluating health human resource systems [51], yet a systematic framework is needed to understand the long term implications of the sequential decisions made in those systems. Given the significant costs and the impact on health care outcomes associated with workforce decisions, it is essential for stakeholders in large health systems to understand the role of the planning horizon and the long-term consequences of health workforce plans.

We therefore propose to study the planning of workforce training, promotion and hiring within such systems, with the main goal of designing a natural policy for decision makers to implement, and concurrently determining common-sense conditions under which this policy is in fact optimal. Governments, regulatory bodies, professional associations, representatives from the private sector and senior health system executives may use the results presented in this chapter to gain a deeper understanding on where incentives should be placed to best meet the health workforce needs of the population. Our focus is on decisions at a health care policy or public policy level (i.e. not on individual hiring and firing decisions), and thus our model includes several stylized simplifications. The problem scale we are interested in has workforces numbering in the thousands or the tens of thousands, e.g. state or provincial health systems, large hospital conglomerates, or the U.S. Veterans Administration. We therefore model the workforce as a continuous flow and allow fractional quantities in our solutions.

We also assume centralized control of the system, which may only be realistic in some cases. Nevertheless, even for those systems in which this is not entirely the case, the conditions we list can help decision makers with limited control in monitoring the system's behavior and deriving policy recommendations; this is precisely the approach [110] take to study the U.S. pediatric nurse practitioner workforce.

Although uncertainty is present in any health system's dynamics, the model we propose is deterministic, and represents a first step in understanding how hiring,

training and promotion interact. The deterministic model allows for some preparation against uncertainty through sensitivity analysis. In addition, the structure of solutions suggested by our analysis can be successfully extended to models with uncertainty; we include computational experiments on a model with stochastic demand growth to demonstrate this.

2.1.1 Our Contribution

We propose a discounted, infinite linear programming model for strategic workforce planning, which includes training, promotion and hiring decision for a class of health workers within a hierarchical system. The model takes as input a demand forecast, workforce payroll, training and hiring costs, workforce hierarchy parameters and a discount factor. Though similar finite models have appeared in earlier work [84, 85, 86], our focus here is to derive structural results and study workforce management policies that are provably optimal under reasonable assumptions. Specifically, we consider the following to be our main contributions:

- i)* We give a series of common-sense conditions any system of this kind should satisfy under our assumptions, demonstrate the pathological behavior that can occur when they are not satisfied and derive useful structural properties of the optimal solutions from the conditions. Though based on our assumptions, these conditions may help guide decision making in more complex systems.
- ii)* We prove that a natural lookahead policy is optimal for our model. In addition to optimizing this model in particular, the result is useful because lookahead policies mimic how more complex models may be managed in practice.
- iii)* We provide a two-part computational study based on real-world nursing workforce data. The first component demonstrates the effectiveness of the lookahead policy in a more complex deterministic system with additional detail, such as

worker age. The second component shows that lookahead policies remain near-optimal in a setting with stochastic demand growth, arguably the most important source of uncertainty in our model.

The remainder of the chapter is organized as follows: This section closes with a literature review. Section 2.2 formulates our model and states the conditions we assume. Section 2.3 uses the conditions to show some structural properties of optimal solutions, proves the optimality of our proposed policy and discusses duality and sensitivity. Section 2.4 discusses experiments that test our policy on more complex models, and Section 2.5 concludes outlining future research avenues. The Appendix contains technical proofs and some additional modeling information.

2.1.2 Literature Review

Workforce planning models are not new to the industrial engineering and operations research literature, with work stretching back several decades, such as [1, 22, 76, 91, 103, 111]. Workforce management models have been developed to manage workforce in call centers [65], military personnel [67], medical school budgets [38, 87], as well as to address cross-training and flexibility of the workforce [99, 124]. [27, 56, 120] provide overviews of workforce/manpower planning models, while [36] discuss the need of a greater interface between operations and human resource management models and the complexities associated with those models. Recent work continues to address workforce issues in operational or tactical time frames, e.g. [26]; this focus on shorter horizons extends also to health care and emergency workforce planning [32, 54, 63, 129]. The long-term workforce capacity planning models [10, 66, 117] are related to our work, yet they concentrate only on the recruitment and retention of personnel without incorporating some of the other decisions required to manage health care personnel. On the other hand, models such as [20, 128] concentrate on skill acquisition and on-the job learning, focusing on a shorter time scale. The results in [125, 126] and

the recent survey [120] particularly highlight the need to research long-term health workforce planning, among other areas.

Infinite-horizon optimization has been widely applied to various operational problems; the dynamic programming paradigm in particular is extensively used [102, 130]. The last two or three decades have also seen the direct study of infinite mathematical programming models and specifically infinite linear programs for operations management applications. Some problems studied in the literature include inventory routing [3, 4], joint replenishment [5, 6], production planning [49, 105], and equipment replacement [80]. However, workforce management possesses differences with other resource management problems that deal mostly with products [66]. A general reference for infinite linear programming is [12]. Our models operate in countable dimensions; recent results for this topic include [49, 70, 71, 104, 105, 106, 113]. To our knowledge, although dynamic programming has been applied to model strategic workforce management, e.g. [66, 103], infinite linear programming has not yet been considered in the literature to address this topic.

For a recent overview of optimization in health care, we refer the reader to [28].

2.2 Model Formulation and Assumptions

We consider an infinite-horizon, discounted workforce planning model with the following characteristics. There is a deterministic demand forecast for each period, and the population of workers at the lowest level of the system, e.g. junior nurses, must be at least equal to that period's demand. The system has a fixed number of levels above this first level; worker population at each higher level must be at least a fixed fraction of the same period's population one level below, to ensure adequate supervision. Between one period and the next, a fixed fraction of each level's population leaves the system, accounting for retirements, firings and resignations. New workers may be added to any level directly via hiring, or indirectly through student admission

and training at the first level, and promotion at higher levels; there is no down-sizing, i.e. mass firing to reduce workforce levels. Student populations take one period to train before entering at the first workforce level; similarly, only workers who have been in a level for at least one period may be promoted. We discuss how to extend our results to models with longer training in the subsequent sections.

The model is defined by the following parameters.

- $n \geq 2$: Number of workforce hierarchy levels.
- $h_k > 0$: Per-period variable payroll costs for level $k = 1, \dots, n$.
- $c_k > 0$: Variable training ($k = 0$) or hiring ($k \geq 1$) costs for level $k = 0, \dots, n$.
- $c_{k,k+1} \geq 0$: Variable promotion cost from level $k = 1, \dots, n-1$ to $k+1$. Workers may only be promoted once they have worked at a particular level for at least one period.
- $\gamma \in (0, 1)$: Discount factor, adjusted to account for cost increases. That is, if $\bar{\gamma}$ is the nominal discount rate and $\alpha > 1$ is the cost growth rate, then $\gamma = \alpha\bar{\gamma}$; this is the reciprocal of the “health care inflation.”
- $d_t > 0$: Forecasted level-1 workforce demand for period $t = 1, \dots$.
- $q_{k,k+1} \in (0, 1)$: Minimum fraction of level- k workers needed at level $k + 1$, for $k = 1, \dots, n - 1$.
- $p_k \in (0, 1)$: Per-period *retention* rate of workers that stay in the system at level $k = 0, \dots, n$ from one period to the next. The *attrition* rate $1 - p_k$ is the fraction of workers at level k expected to leave the system from one period to the next; this includes firing, retirement and quitting.
- s_k^0 : Students ($k = 0$) or workers in level $k = 1, \dots, n$ at the start of the current period, before attrition.

The model's decision variables are:

- s_k^t : Students ($k = 0$) or workers in level $k = 1, \dots, n$ at end of period $t = 1, \dots$
- x_k^t : Students admitted ($k = 0$) or workers hired at level $k = 1, \dots, n$ in period $t = 1, \dots$
- $x_{k,k+1}^t$: Workers promoted from level $k = 1, \dots, n-1$ to $k+1$ in period $t = 1, \dots$

Our strategic workforce planning problem then has the following formulation.

$$\inf C(s, x) = \sum_{t=1}^{\infty} \gamma^{t-1} \left(\sum_{k=0}^n c_k x_k^t + \sum_{k=1}^{n-1} c_{k,k+1} x_{k,k+1}^t + \sum_{k=1}^n h_k s_k^t \right) \quad (1a)$$

$$\text{s.t. } s_1^t \geq d_t, \quad \forall t = 1, \dots \quad (1b)$$

$$s_{k+1}^t - q_{k,k+1} s_k^t \geq 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (1c)$$

$$p_1 s_1^{t-1} - s_1^t + p_0 x_0^{t-1} - x_{12}^t + x_1^t = 0, \quad \forall t = 1, \dots \quad (1d)$$

$$p_k s_k^{t-1} - s_k^t + x_{k-1,k}^t - x_{k,k+1}^t + x_k^t = 0, \quad \forall k = 2, \dots, n-1, \quad \forall t = 1, \dots \quad (1e)$$

$$p_n s_n^{t-1} - s_n^t + x_{n-1,n}^t + x_n^t = 0, \quad \forall t = 1, \dots \quad (1f)$$

$$p_k s_k^{t-1} - x_{k,k+1}^t \geq 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (1g)$$

$$s^t, x^t \geq 0, \quad \forall t = 1, \dots, \quad (1h)$$

where $s_0^t = x_0^t$ for $t = 0, \dots$. We take the feasible region to be the subset of solutions for which the objective is well defined and finite [105]. In the model, the objective (1a) minimizes discounted cost over the infinite horizon. The demand satisfaction constraint (1b) ensures enough level-1 workers are present to satisfy projected demand each period, while (1c) ensures the minimum required fraction of level- $(k+1)$ workers are present to supervise level- k workers. The flow balance constraints (1d–1f) track workers present at each level from one period to the next, and (1g) limits the promoted workers from level k to $k+1$ to those present in level k for at least one period. The domain constraints (1h) ensure non-negativity of worker levels, hires, promotions and student admissions.

Whereas most of the model’s parameters are stationary and can thus be explicitly given or recorded, the demand forecast is an infinite sequence that cannot be explicitly given. In practical terms this forecast can only be modeled implicitly, for instance by giving a first-period demand and a per-period growth rate. While our results hold for an arbitrary sequence satisfying our assumptions, our policy requires explicit knowledge of only the first few values of the sequence (two in the model as currently stated, but see Corollary 2.3.4 below for an extension). For a discussion of related issues with non-stationary data in infinite-horizon optimization, see e.g. [69].

We next list several conditions the model should satisfy. These conditions are common in many real world settings or are reasonable approximations, and are necessary for most of our subsequent results. Many are also necessary to avoid pathological behavior. We begin with technical assumptions.

Assumption 2.2.1 (Technical assumptions).

- i)* Finite total demand: Total discounted demand converges.

$$\sum_{t=1}^{\infty} \gamma^{t-1} d_t < \infty \tag{2a}$$

- ii)* Linear costs: The variables’ costs are linear, and described by c and h .

The former assumption is necessary to have a finite objective and thus a feasible problem. The latter is required to apply linear programming techniques. Though large changes in a system’s workforce could render some costs non-linear (e.g. a large increase in hiring leading to an increase in hiring and payroll costs because of the labor market’s supply), our results suggest that in the long run moderate decisions predominate, and thus the assumption of linearity is reasonable.

Assumption 2.2.2 (Growing demand). The sequence (d_t) is non-decreasing.

$$d_t \leq d_{t+1}, \quad \forall t = 1, \dots \tag{2b}$$

This assumption reflects most contemporary health care systems in which demand is expected to grow for the foreseeable future, and ensures that training and promotion will be perpetually necessary within the system. As discussed by [52], given changes in the demographics of the population, as well as expanded coverage under the Affordable Care Act, demand for primary care services in the United States is expected to grow by 14% by 2025. This expected growth in demand for health care workers is not unique to the United States; it is estimated that an additional 1.9 billion people will seek access to health care by 2035 [40]. In more general cases, even if demand is only expected to be eventually non-decreasing, our conclusions can be applied starting at the period where non-decreasing growth begins, with a finite model accounting for the system in preceding periods.

The first non-technical assumption concerns the relative costs of payroll, promotion and hiring.

Assumption 2.2.3 (Promotion is preferable). Even when factoring attrition, payroll costs and discounting, promotion is cheaper than hiring.

$$\frac{c_0}{\gamma p_0} \leq c_1, \quad \frac{c_k + h_k}{\gamma p_k} + c_{k,k+1} \leq c_{k+1}, \quad \forall k = 1, \dots, n-1 \quad (2c)$$

If this assumption does not hold at some level in the hierarchy, there is no incentive to train and promote from within beyond that point. This condition should be satisfied by many workforce systems, both in health care and in other industries.

The next assumption is slightly more specific to the health care industry, but still common in other industries.

Assumption 2.2.4 (Non-increasing retention). The hierarchy does not tend to become top-heavy:

$$p_k \geq p_{k+1}, \quad \forall k = 1, \dots, n-1 \quad (2d)$$

This assumption is natural in health care hierarchies such as nursing, where higher-level workers are usually older, since older workers tend to retire or leave the system for other reasons at a higher rate. The assumption is more problematic, for example, in industries where tenure guarantees at an intermediate level imply an unnaturally high attrition at lower levels.

For some of our results, it is necessary to further strengthen the previous assumption.

Assumption 2.2.4' (Equal retention). Retention and attrition are equal at all hierarchy levels:

$$p_k = p_{k+1}, \quad \forall k = 1, \dots, n-1 \quad (2d')$$

Though it appears restrictive, in many real-world systems the top and bottom retention rates in fact only differ by a few percentage points [84, 85, 86].

Assumption 2.2.5 (Non-decreasing payroll). Salaries increase within the hierarchy, even when accounting for attrition:

$$\frac{h_k}{1 - \gamma p_k} \leq \frac{h_{k+1}}{1 - \gamma p_{k+1}}, \quad \forall k = 1, \dots, n-1 \quad (2e)$$

As the next example shows, this condition prevents undesirable behavior.

Example 1 (Down-sizing by promotion). Consider a two-level system which is drastically over-staffed. Let $d_t = \varepsilon$ for all t , where $\varepsilon > 0$ is a small positive number, and let $s_1^0 \gg \varepsilon$. If (2e) is not satisfied, it may be optimal because of (2d) to promote all but ε workers to level 2, effectively down-sizing the workforce by promoting most of it, and achieving lower costs in the process. Such behavior could lead to detrimental side effects, such as poor morale in the remaining workforce.

Assumption 2.2.6 (Moderate demand growth). Demand does not grow too quickly:

$$\frac{d_{t+1}}{d_t} \leq \frac{p_{\min}}{q_{\max}}, \quad \forall t = 1, \dots, \quad (2f)$$

where $p_{\min} = \min_k p_k$ and $q_{\max} = \max_k q_{k,k+1}$.

Intuitively, the assumption ensures enough worker population at each level to promote to the next level as demand grows; it is easily satisfied in most systems. For example, if $n = 2$, $p_1 = p_2 = 0.8$ and $q_{12} = 0.25$, (2f) requires the demand growth to be no more than 320% per period, a condition met in virtually any system. Furthermore, as the next example shows, when this assumption is not met, the planning horizon necessary to compute an optimal solution may be arbitrarily long.

Example 2 (Excessive demand growth). We consider a two-level system that experiences excessive demand growth for a given number of periods, and constant demand thereafter. To simplify the numbers in the example, we set $p_1 = p_2 = q_{12} = 1$. For a fixed $m \geq 2$ let

$$d_1^m = 1, \quad d_t^m = \begin{cases} 2^{t-1} - 1, & t = 2, \dots, m \\ 2^{m-1} - 1, & t = m + 1, \dots \end{cases},$$

and $s_0^0 = 0$, $s_1^0 = s_2^0 = 1$; note that $d_{t+1}/d_t > p_1/q_{12} = 1$ for $t = 2, \dots, m - 1$. Table 1 details the first demand values in the sequence, for $m = 3, 4, 5$. The table also lists a solution that satisfies demand without any hiring, which can be made optimal by choosing large enough hiring costs. Although projected demand for the first three periods is identical in all cases, the optimal number of students admitted in the first period changes with m ; for general m , we get $x_0^1 = (2^{m-2} - 1)/2^{m-3}$. In other words, the current period's decision may depend on a horizon of arbitrary length m .

As Example 2 suggests, the condition (2f) can be relaxed; we include the best possible condition of this kind we could derive in the Appendix (see the proof of Claim A.1.3). However, (2f) is much simpler to state and suffices for any practical situation.

2.3 *Optimal System Behavior*

We begin our characterization of optimal solutions of (1) by outlining structural properties satisfied in models that meet our assumptions. We include only simple

Table 1: Sample demand sequences and solutions with no hiring for $m = 3, 4, 5$ in Example 2

t	1	2	3	4	5	...
d_t^3	1	1	3	3	3	...
x_0^t	1	3	0	0	0	...
s_1^t	1	3/2	3	3	3	...
s_2^t	1	3/2	3	3	3	...
d_t^4	1	1	3	7	7	...
x_0^t	3/2	7/2	7	0	0	...
s_1^t	1	3/2	7/2	7	7	...
s_2^t	1	3/2	7/2	7	7	...
d_t^5	1	1	3	7	15	...
x_0^t	7/4	15/4	15/2	15	0	...
s_1^t	1	15/8	15/4	15/2	15	...
s_2^t	1	15/8	15/4	15/2	15	...

proofs here and relegate any complex proof to the Appendix.

Lemma 2.3.1 (No unnecessary hiring). *Suppose the model parameters satisfy Assumptions 2.2.1 through 2.2.3. There is an optimal solution of (1) in which no hiring takes place when promotion is possible:*

$$x_1^t = 0, \quad \forall t = 2, \dots \quad (3a)$$

$$(p_k s_k^{t-1} - x_{k,k+1}^t) x_{k+1}^t = 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (3b)$$

Proof. If a solution does not satisfy either condition, a simple substitution produces another solution with equal or lesser objective that does satisfy the conditions. \square

Lemma 2.3.2 (No excess training or promotion). *Suppose Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.5 and 2.2.6 hold. Furthermore, suppose either Assumption 2.2.4 holds and $n = 2$, or Assumption 2.2.4' holds. Then there is an optimal solution of (1) in which no excess promotion or student admittance occurs:*

$$(s_1^t - d_t) x_0^{t-1} = 0, \quad \forall t = 2, \dots \quad (4a)$$

$$(s_{k+1}^t - q_{k,k+1} s_k^t) x_{k,k+1}^t = 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (4b)$$

Like the preceding lemma, Lemma 2.3.2 follows from applying a substitution or perturbation to any solution that does not satisfy it. However, unlike in the hiring case, a perturbation in promotion has ripple effects in higher levels of the hierarchy and in later periods that render it much more complex.

With these two structural properties in place, we are able to characterize optimal solutions of (1). Consider the two-period restriction of (1) given by

$$\min \sum_{k=0}^n c_k(x_k^1 + \gamma x_k^2) + \sum_{k=1}^{n-1} c_{k,k+1}(x_{k,k+1}^1 + \gamma x_{k,k+1}^2) + \sum_{k=1}^n h_k(s_k^1 + \gamma s_k^2) \quad (5a)$$

$$\text{s.t. } s_1^t \geq d_t, \quad \forall t = 1, 2 \quad (5b)$$

$$s_{k+1}^t - q_{k,k+1}s_k^t \geq 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, 2 \quad (5c)$$

$$p_1 s_1^{t-1} - s_1^t + p_0 x_0^{t-1} - x_{12}^t + x_1^t = 0, \quad \forall t = 1, 2 \quad (5d)$$

$$p_k s_k^{t-1} - s_k^t + x_{k-1,k}^t - x_{k,k+1}^t + x_k^t = 0, \quad \forall k = 2, \dots, n-1, \quad \forall t = 1, 2 \quad (5e)$$

$$p_n s_n^{t-1} - s_n^t + x_{n-1,n}^t + x_n^t = 0, \quad \forall t = 1, 2 \quad (5f)$$

$$p_k s_k^{t-1} - x_{k,k+1}^t \geq 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, 2 \quad (5g)$$

$$s^t, x^t \geq 0, \quad \forall t = 1, 2, \quad (5h)$$

A *one-period lookahead policy* constructs a solution to (1) by iteratively solving (5), fixing the variables for $t = 1$, stepping one period forward by relabeling $t \leftarrow t + 1$ for all variables and parameters, and repeating the process. In practice, this corresponds to a decision maker planning the current period's promotion, training and hiring based on current demand and the next period's forecasted demand, while ignoring demand for subsequent periods.

Theorem 2.3.3 (Optimality of one-period lookahead policy). *Suppose Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.5 and 2.2.6 are satisfied. Suppose either Assumption 2.2.4 holds and $n = 2$, or Assumption 2.2.4' holds. Then one-period lookahead policies are optimal.*

Corollary 2.3.4 (Increased training time). *Suppose students require $L \geq 1$ periods to train instead of one, with all other system characteristics remaining the same. Under the conditions of Theorem 2.3.3, L -period lookahead policies are optimal, where an L -period lookahead is defined analogously to a one-period lookahead but with L additional periods instead of one.*

Proof. The proof of Theorem 2.3.3 still applies; we are simply relabeling level-0 variables. □

These results indicate that good workforce planning decisions can be made using a minimal amount of forecasted information, which strengthens the robustness of the resulting solution since forecasts of more distant demand naturally tend to be less reliable. This also places our result within the context of *solution* and *forecast horizons*; see, e.g., [45] for formal definitions and discussion. Moreover, lookahead policies mimic how such large workforce systems might be managed in practice, suggesting that these policies could be useful in more complex settings. We explore this idea experimentally in Section 2.4.

Another important question related to (1) is duality. A dual satisfying the typical complementary relationships can shed additional light on the structure of optimal solutions to (1). Furthermore, optimal dual prices may also be useful as indicators of the model's sensitivity to parameters such as demand. However, the infinite horizon implies significant technical complications and gives rise to pathologies not encountered in the finite case.

Extending the typical LP dual construction to (1) yields

$$\sup D(\mu, \lambda, \eta) = \sum_{t=1}^{\infty} d_t \mu_1^t - p_0 s_0^0 \lambda_1^1 - \sum_{k=1}^{n-1} p_k s_k^0 (\lambda_k^1 + \eta_{k,k+1}^1) - p_n s_n^0 \lambda_n^1 \quad (6a)$$

$$\text{s.t. } \mu_k^t - q_{k,k+1} \mu_{k+1}^t - \lambda_k^t + p_k \lambda_k^{t+1} + p_k \eta_{k,k+1}^{t+1} \leq \gamma^{t-1} h_k, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (6b)$$

$$\mu_n^t - \lambda_n^t + p_n \lambda_n^{t+1} \leq \gamma^{t-1} h_n, \quad \forall t = 1, \dots \quad (6c)$$

$$p_0 \lambda_1^t \leq \gamma^{t-2} c_0, \quad \forall t = 2, \dots \quad (6d)$$

$$\lambda_k^t \leq \gamma^{t-1} c_k, \quad \forall k = 1, \dots, n, \quad \forall t = 1, \dots \quad (6e)$$

$$-\lambda_k^t + \lambda_{k+1}^t - \eta_{k,k+1}^t \leq \gamma^{t-1} c_{k,k+1}, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (6f)$$

$$\mu^t, \eta^t \geq 0, \quad \lambda^t \text{ unrestricted}, \quad \forall t = 1, \dots, \quad (6g)$$

where we similarly define the feasible region as a subset of the points for which the objective is well defined and finite. However, this model does not satisfy strong or even weak duality with (1).

Example 3 (No weak duality; adapted from [113]). Suppose $s_n^0 > 0$ and let $M > 0$. Define $\hat{\lambda}_n^t = -M/p_n^t$, $\forall t = 1, \dots$, and set all other variables to zero. The solution is feasible for (6), and its objective function value is positive and goes to infinity as $M \rightarrow \infty$. However, (1) is clearly feasible and bounded below by zero.

The following result addresses this problem.

Theorem 2.3.5. *Suppose we can change the equality constraints (1d–1f) to greater-than-or-equal constraints (and thus impose $\lambda^t \geq 0$) for all but a finite number of indices t without affecting optimality in (1). Let (\hat{s}, \hat{x}) and $(\hat{\mu}, \hat{\lambda}, \hat{\eta})$ be feasible for (1) and (6) respectively.*

i) *Weak duality: $D(\hat{\mu}, \hat{\lambda}, \hat{\eta}) \leq C(\hat{s}, \hat{x})$.*

ii) *Strong duality: Both solutions are optimal and $D(\hat{\mu}, \hat{\lambda}, \hat{\eta}) = C(\hat{s}, \hat{x})$ if and only if complementary slackness holds (in the usual sense) and transversality [105, 106, 113] holds:*

$$\liminf_{t \rightarrow \infty} p_0 \hat{\lambda}_1^{t+1} \hat{x}_0^t + \sum_{k=1}^{n-1} p_k (\hat{\lambda}_k^{t+1} + \hat{\eta}_k^{t+1}) \hat{s}_k^t + p_n \hat{\lambda}_n^{t+1} \hat{s}_n^t = 0. \quad (7)$$

Proof. If all constraints eventually become greater-than-or-equal, then the off-diagonal constraint matrix of (1) in inequality form is eventually non-negative, implying that [105, Assumption 3.1] holds, and thus the results follow from [105, Theorems 3.3 and 3.7]. \square

Corollary 2.3.6. *The conditions of Theorem 2.3.5 apply, and therefore weak and strong duality hold, if demand is eventually non-decreasing.*

The results in [105] imply we can use optimal solutions of (6) as shadow prices to perform sensitivity analysis on (1).

Example 4 (Sensitivity analysis). Consider a two-level system in which the incoming worker populations in period 1 require some promotion from level 1 to level 2, with enough level-1 workers remaining after promotion to meet demand in period 1 but not later. Based on these initial conditions and Assumptions 2.2.1 through 2.2.6, Lemmas 2.3.1 and 2.3.2 imply the following structure to the optimal solution:

$$\begin{aligned} x_1^t = x_2^t = 0, \quad x_0^t, x_{12}^t > 0, \quad x_{12}^t < p_1 s_1^{t-1}, \quad s_2^t = q_{12} s_1^t, \quad \forall t = 1, \dots \\ s_1^1 > d_1; \quad s_1^t = d_t, \quad \forall t = 2, \dots \end{aligned}$$

The solution for (6) that satisfies complementary slackness and transversality is:

$$\begin{aligned} \mu_1^1 &= 0 \\ \mu_1^t &= \frac{\gamma^{t-2} c_0}{p_0} (1 - \gamma p_1 + q_{12} (1 - \gamma p_2)) + \gamma^{t-1} q_{12} c_{12} (1 - \gamma p_2) \\ &\quad + \gamma^{t-1} (h_1 + q_{12} h_2), \quad \forall t = 2, \dots \\ \mu_2^1 &= \frac{1}{1 + q_{12}} \left(\frac{c_0}{p_0} (p_1 - p_2) + c_{12} (1 - \gamma p_2) + (h_2 - h_1) \right) \\ \mu_2^t &= \gamma^{t-2} (1 - \gamma p_2) \left(\frac{c_0}{p_0} + \gamma c_{12} \right) + \gamma^{t-1} h_2, \quad \forall t = 2, \dots \\ \lambda_1^1 &= \frac{1}{1 + q_{12}} \left(\frac{c_0}{p_0} (p_1 + q_{12} p_2) - q_{12} c_{12} (1 - \gamma p_2) - (h_1 + q_{12} h_2) \right) \\ \lambda_1^t &= \frac{\gamma^{t-2} c_0}{p_0}, \quad \forall t = 2, \dots \end{aligned}$$

$$\lambda_2^1 = \frac{1}{1 + q_{12}} \left(\frac{c_0}{p_0} (p_1 + q_{12}p_2) + c_{12}(1 + \gamma q_{12}p_2) - (h_1 + q_{12}h_2) \right)$$

$$\lambda_2^t = \gamma^{t-2} \left(\frac{c_0}{p_0} + \gamma c_{12} \right), \quad \forall t = 2, \dots$$

$$\eta_{12}^t = 0, \quad \forall t = 1, \dots$$

It can be verified that this solution is dual feasible provided the assumptions hold. Suppose in particular that demand grows based on a rate $1 < \beta < 1/\gamma$, so that $d_t = \beta^{t-1}d_1$. It follows that

$$\sum_{t=1}^{\infty} d_t \mu_1^t = d_1 \left[\frac{c_0}{\gamma p_0} (1 - \gamma p_1 + q_{12}(1 - \gamma p_2)) + q_{12} c_{12} (1 - \gamma p_2) + h_1 + q_{12} h_2 \right] \frac{\beta \gamma}{1 - \beta \gamma}.$$

This expression indicates how the optimal cost would change if either d_1 or β vary slightly from their given values.

2.4 Computational Study

To evaluate the efficacy of our proposed models and policies, we performed computational experiments based on the British Columbia nursing workforce described in [84, 85, 86]. Health care human resource data is more readily available from Canadian provinces because of their centralized control of health care. However, similar data from U.S. systems can be used within a model such as ours to derive policy recommendations, e.g. [110], even though U.S. health systems are usually de-centralized.

We first discuss the performance of lookahead policies applied to more complex, albeit deterministic, settings. We then develop a simulation model that considers uncertainty in demand growth and evaluate the performance of our lookahead policies in this setting.

2.4.1 Deterministic Experiments

While Model 1 provides useful insights into the behavior of strategic workforce planning models, possible extensions include the differentiation of workers by age (as it affects attrition rates), and the extension of the length of student training (to four

years). In order to evaluate the performance of our lookahead policies, we began by solving the problem over a 25-year planning horizon (a *full information* model) and used the solution to the first 20 years as our benchmark. We then compared the results to a solution obtained by implementing a four-year lookahead policy of this extended model.

Figure 1 outlines the structure of the extended model and its parameters (see further model description in the Appendix). Students are admitted into the training program, where they take four years to train before entering the workforce. The probabilities of students continuing their education depend on the school year of the student (with greater attrition in the first year of the program). After graduation, students enter the first workforce level as direct care nurses. In level 1, the number of workers has to meet current demand. This demand is met by workers that have not retired or been promoted, graduates from the training program and workers hired externally. Level 2 consists of nurse managers, a supervisory position to the first level; nurse managers are either hired externally or promoted from the first workforce level. To account for transition shock and adaptation to the profession [55], we assume that Level-1 workers must have worked for at least one year before being promoted into the second workforce level. In both levels, retention rates depend on the age of the workers. The average retention in level 2 is slightly higher than the average retention in level 1, which would violate Assumption 2.2.4 if the averages applied to all age groups. Furthermore, since the parameter is age-dependent in this model, the actual retention in each level depends on the age distribution of the worker population. This difference in attrition rates did not impact our results, further supporting the robustness of our findings. We set the discount factor to $\gamma = 0.95$.

We tested the model in nine scenarios. Among the nine scenarios, the baseline scenario represents the estimated demand in British Columbia, Canada, starting in 2007; we calculated demand by extrapolating the population growth between 1996

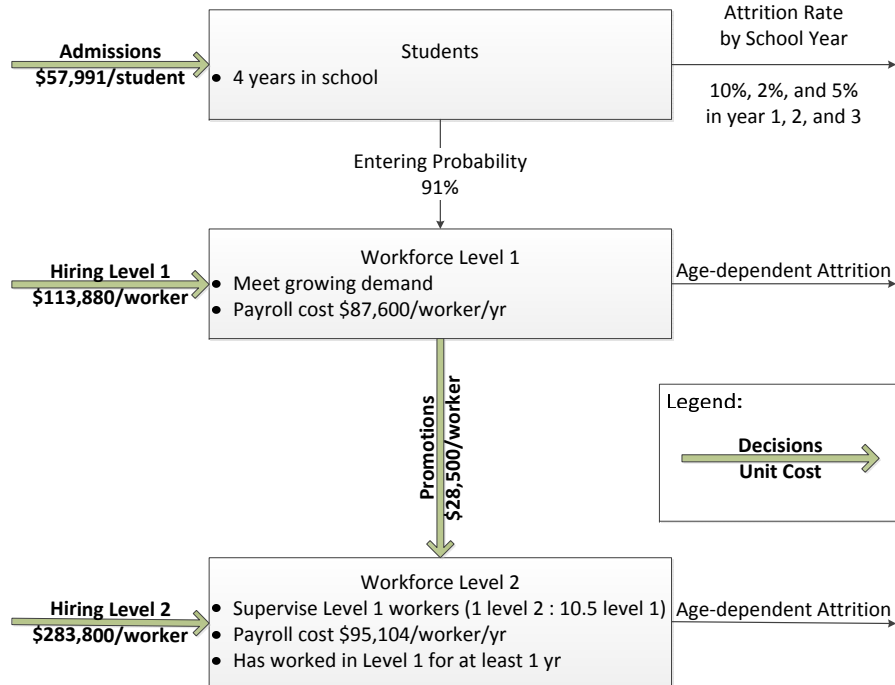


Figure 1: Flow chart of model used in computational examples

and 2006 [29]. Scenarios 1 through 4 evaluate the impact of different demand growth rates. Scenario 5 evaluates the impact of limiting the growth of the training program. Scenarios 6, 7 and 8 evaluate the performance of the lookahead policy in extreme conditions where demand has a peak, hiring growth is limited, and costs are varied. The parameter characteristics and descriptions of the scenarios are summarized in the following list.

Baseline Scenario Fixed demand growth rate of 1.25% per year. Projected demand growth in British Columbia, Canada.

Scenario 1 Fixed demand growth rate of 0.01% per year. Very low demand growth.

Scenario 2 Fixed demand growth rate of 2.5% per year. High demand growth.

Scenario 3 Linearly accelerating demand growth from 0% per year to 2.5% per year over 25 years.

Scenario 4 Linearly decelerating demand growth from 2.5% per year to 0% per year over 25 years.

Scenario 5 Fixed demand growth rate of 1.25% per year and student population growth limited to no more than 1% per year. Major restrictions in training growth.

Scenario 6 Fixed demand growth rate of 1.25% per year in years 1 through 9 and 11 through 25, demand doubled in year 10. Level-1 hiring growth limited to no more than 50% per year. This scenario simulated a sudden jump of demand, which might be due to a drastic change in roles and scope of practice of the workforce. We assumed that drastic changes in the number of workers hired could not be made without incurring very large recruitment costs.

Scenario 7 Fixed demand growth rate of 1.25% per year in years 1 through 9 and 11 through 25, demand doubled in year 10. Level-1 hiring growth limited to no more than 50% per year, and zero student admission cost. In addition to the jump in demand and limited hiring growth, we eliminated the admission cost to increase the incentive to admit students in advance and thus potentially undermine the four-year lookahead model.

Scenario 8 Fixed demand growth rate of 1.25% per year in years 1 through 9 and 11 through 25, demand doubled in year 10. Level-1 hiring growth limited to no more than 50% per year, and zero level-1 payroll cost. In addition to the jump in demand and limited hiring growth, we eliminated the level-1 payroll cost to increase the incentive to admit students in advance and thus potentially undermine the four-year lookahead model.

We compared the solutions obtained using the full information model and the lookahead model. Figure 2 shows results for the baseline scenario and scenarios 1

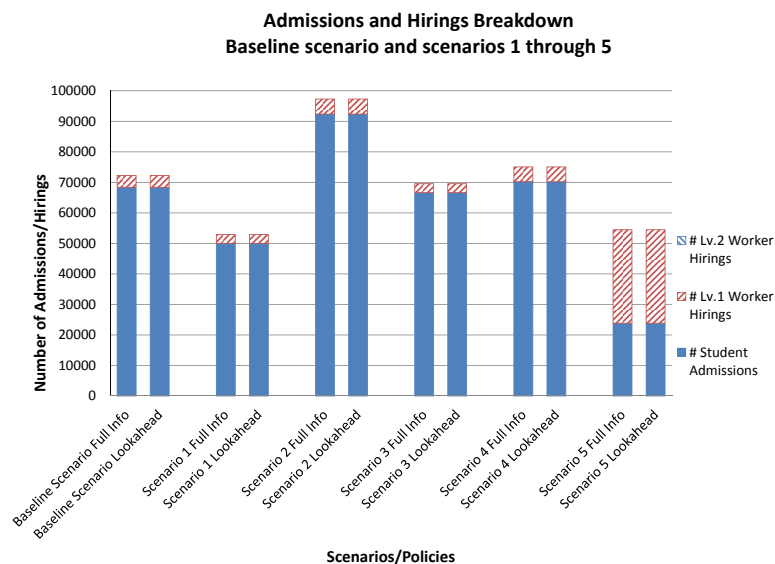


Figure 2: Breakdown of the total number of admissions and hirings in baseline scenario and scenarios 1 through 5 over the course of 20 years

through 5. In these scenarios, we obtained the same solutions using the full information and the lookahead models. The lookahead model was robust in these scenarios, even if Assumption 2.2.4 was slightly violated by our system’s parameters. Even in Scenario 5, where education growth was drastically limited, the full information model did not differ from the lookahead policy because training students a year in advance incurred extra payroll costs, making early training more expensive than hiring. Though education growth was limited, hires served as a back-up action in Scenario 5 and made the lookahead and the full information methods operate identically.

Figure 3 shows results for scenarios 6, 7, and 8; in this case, the lookahead policy resulted in slightly higher total costs. Compared to the full information solution, the percentage differences in total cost were only 0.026%, 0.129%, and 0.014% respectively. The lookahead model resulted in more admissions, more level-2 hirings, and fewer level-1 hirings than the full information model. Since level-1 hiring was limited, fewer level-1 workers were hired and more students were trained as an alternative.

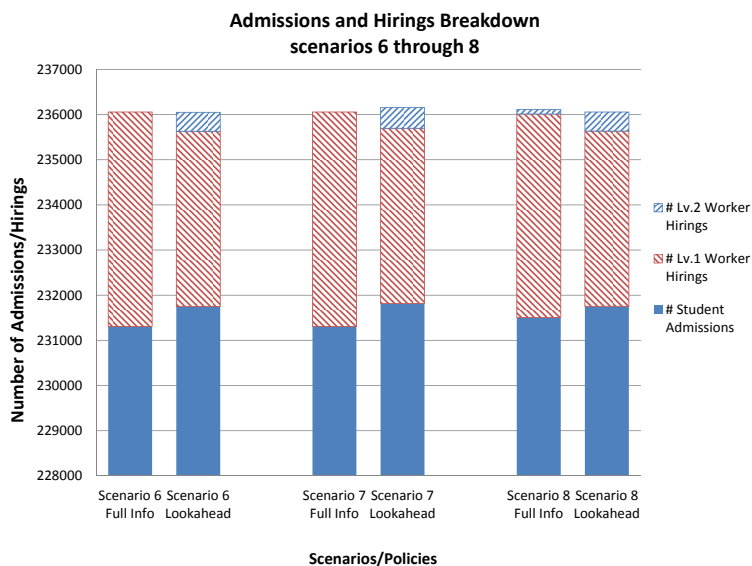


Figure 3: Breakdown of the total number of admissions and hirings in scenarios 6 through 8 over the course of 20 years

Level-2 workers were hired when the model reached a point where promotions could not meet the level-2 workforce demand due to the insufficient number of level-1 workers. The lookahead model failed to anticipate future changes in demand, not training sufficient students nor hiring sufficient level-1 workers in advance.

Overall, the lookahead policy showed robustness in the nine scenarios modeled. In the most extreme scenarios, where demand had a sudden jump and hirings or admissions were limited, the lookahead policy and the full information policy still showed very little difference, particularly in total cost.

2.4.2 Experiments with Stochastic Demand Growth

To further evaluate the lookahead policy, we examined the performance of our model in a stochastic setting where the demand growth rate in each year (denoted ρ) is an i.i.d. random variable uniformly distributed between 0% and 2.5% (the mean growth rate is thus kept at 1.25%, as in the deterministic baseline scenario [86]).

We applied the lookahead policy sequentially. After the simulated demand d_t is realized in year t , we project year $(t + 1)$'s demand to be $\hat{d}_{t+1} = (1 + (1 + \delta)\mathbf{E}[\rho])d_t$, where δ is a *forecast factor* used to represent the planner's level of risk-aversion. When $\delta > 0$, the planner assumes demand grows faster than the mean; for $\delta < 0$, the planner assumes the demand grows slower than the mean; for $\delta = 0$, the planner plans for the expected growth. After solving the lookahead model for years t and $t + 1$, the process steps forward one year, true demand in $t + 1$ is observed, and hiring decisions are made if the workforce is insufficient to meet the demand. The algorithm proceeds to the next period and the look-ahead policy is sequentially applied. This procedure iterates until period 20. In our simulation, each policy was solved with 2000 replications.

We benchmarked the lookahead policy described above with the full information model; as in the deterministic experiments, the full information solution solves a single LP with full (deterministic) access to the uncertain parameters. In the stochastic case, this implies solving one full information LP for every simulated replication and averaging the resulting costs. Because this solution has earlier access to the uncertain data, it provides a lower bound on any policy's cost.

As shown in Figure 4, by varying the forecast factor δ over 1% increments between -100% and 100%, the lookahead policy achieves lowest cost at $\delta^* = -33\%$ (the *best delta* policy). All lookahead policies tested were within 1.2% of the full information cost. The percentage gap for the mean-growth policy ($\delta = 0$) is 0.72%, the gap of the no-growth policy ($\delta = -100\%$) is 0.95%, and the gap of the highest-growth policy ($\delta = 100\%$) is 1.2%.

In our simulation, δ^* is less than 0. This implies that it is more favorable to adopt a policy that plans for demand growth smaller than the mean. To explain the rationale behind this behavior, Figure 5 shows the breakdown of the total cost as a proportion of the full information cost. The model assumes the workforce cannot

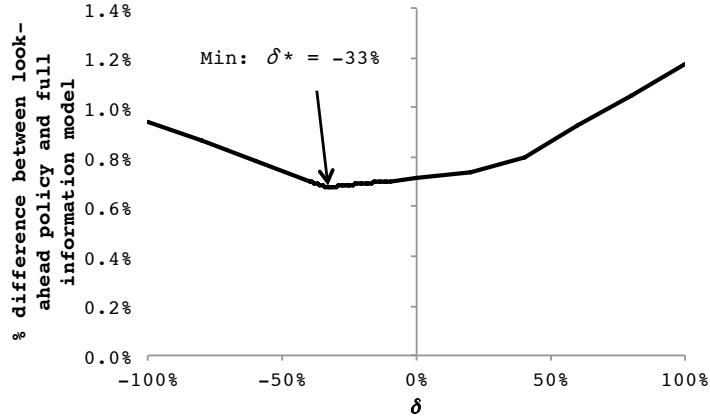


Figure 4: Gap between forecast factor δ lookahead policy and full information cost. All lookahead policy total costs are within 1.2%.

be downsized, and payroll cost makes up more than 90% of the total. Therefore, an oversized workforce will remain in the system many years and will thus increase the cost dramatically. To further explore this idea, we also simulated a policy with no training that directly hires 100% of its workforce. However, the no-training policy performed far worse, with a gap of 4.7%.

Payroll cost makes up more than 90% of the total. Specifically, there is a fixed amount of unavoidable payroll cost needed to satisfy demand, regardless of any decisions. By subtracting the unavoidable payroll cost from the total cost, we are left with the controllable costs, i.e. promotion cost, hiring cost, admission cost, and payroll cost in excess of the unavoidable. Figure 6 shows the controllable cost breakdown. The no-training policy exceeds the full information model with respect to the controllable cost by over 1.5 times, whereas the gaps between the lookahead policies with forecast factor and the full information model are within 40% with respect to the controllable cost, with minimal gap of 22% at $\delta^* = -33\%$.

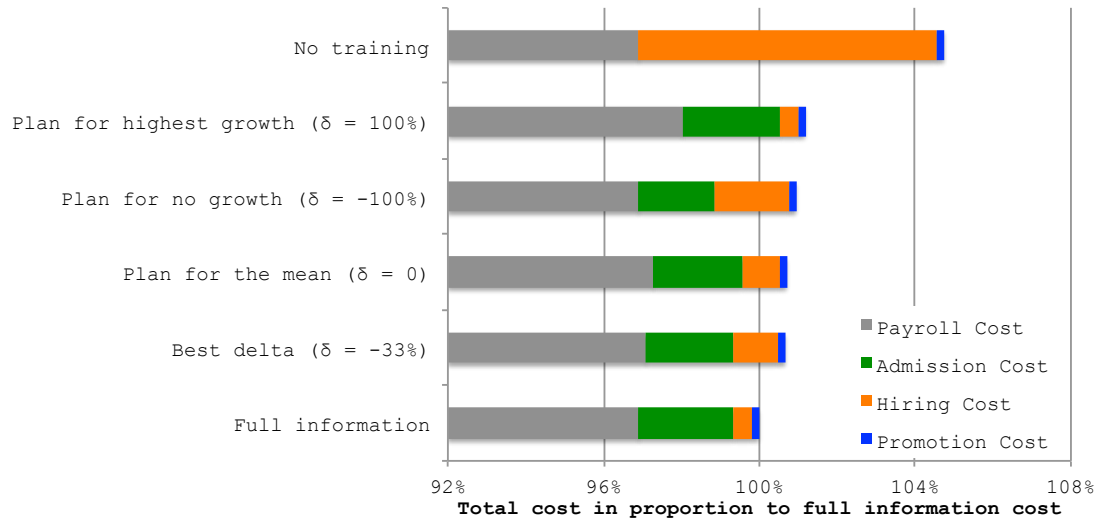


Figure 5: Breakdown of total cost under different policies: payroll costs dominate the other costs.

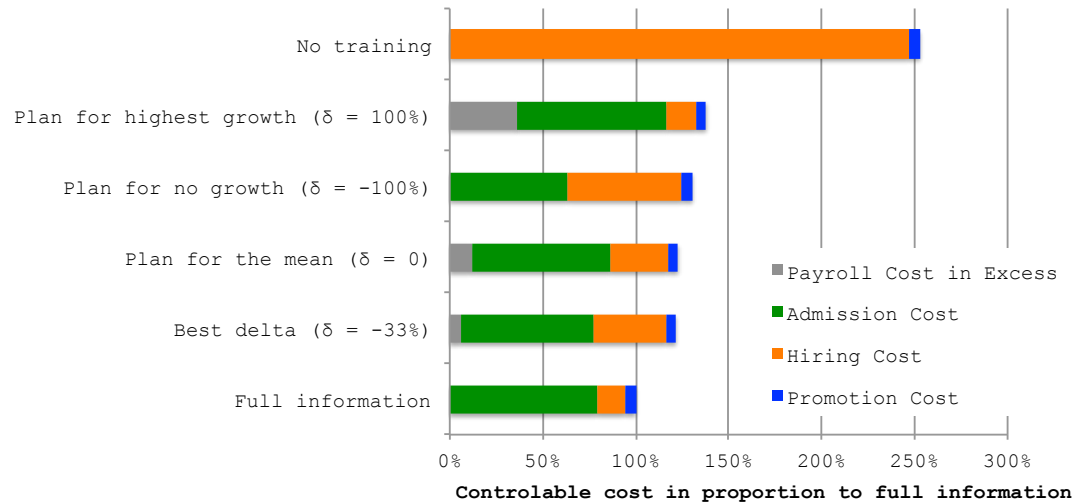


Figure 6: Breakdown of controllable cost under different policies: look-ahead policies stay within 40% from *full information* model with minimal gap of 22% at $\delta^* = 33\%$.

2.5 *Conclusions*

This chapter contributes a new modeling framework for strategic health workforce planning. Through infinite-horizon optimization, we are able to model the long-term implications of training, hiring and promotion decisions made within a health care system. Our approach enables us to understand the planning horizon length necessary to obtain optimal decisions. We derive common-sense system conditions that should hold in any situation and also imply the optimality of a simple lookahead workforce management policy. Using real-world data from British Columbia, we further demonstrate how lookahead policies perform well in a variety of scenarios, particularly with uncertain demand growth. These results are particularly useful, as the lookahead solution mirrors workforce management policies implemented in practice.

Given that long-term workforce planning should be an important component of a well-functioning health care system, this kind of model can be used to obtain qualitative checks on whether a particular health workforce system is behaving optimally, or what conditions it must meet to do so. For example, in [110] the authors apply a similar model to derive policy recommendations for the U.S. pediatric nurse practitioner workforce.

A next step in our work is to directly model and optimize the system's uncertainty, specifically in demand growth or retention rates [92]. It is important to understand whether the conditions we develop in this work and their structural consequences (or appropriate modifications) still hold in more general settings. For example, it is possible that under uncertain demand growth condition (4a) of Lemma 2.3.2 does not hold in level 0 – we may need to train in excess of forecasted demand – but it may be that a similar property holds which accounts for the risks of under-training and over-training. The more nuanced analysis required in this case may give insight into the impact of uncertainty on health workforce costs and management decisions;

for example, [53, 75] investigate similar questions for short-term nurse staffing. From a theoretical perspective, the infinite linear programming tools we use still apply in the presence of uncertain demand growth or retention, provided these can be modeled as finite-support random variables. In more general cases, such as the uniformly distributed demand growth used in Section 2.4.2, the model is no longer a countably infinite linear program. Nevertheless, recent results in non-stationary infinite optimization, e.g. [72], may suggest alternative approaches.

Because this work is applied to guide strategic health workforce decisions, we can formulate more realistic models by incorporating other elements. For instance, (1) could be expanded to include a variety of health care providers and changes in scopes of practice. As a first step, the impact of multiple worker types can be modeled indirectly in (1) through scaling or modification of the demand forecast. This approach has the advantage of allowing for non-linear interactions between multiple health care providers and demand, if, for example, different worker types cannot serve patient demand in the same fashion. Assuming that the interaction of all worker types with demand is linear, multiple worker types can be incorporated in models similar to (1), by differentiating across both type and level, where each worker type includes its own hierarchy with its own supervision constraints (1c) and dynamics, but the multiple types serve patient demand jointly.

Furthermore, clinical inactivity has been a well documented phenomenon among health care providers [41, 61, 62, 78], and therefore policy makers may be interested in understanding the role of such inactivity in workforce planning. As before, one possibility is to incorporate expected inactivity in the demand calculations. A more complex option is to incorporate additional states representing the number of health care providers that are inactive each period. While this second option entails an expansion of the model, by following this option it could be possible to study the impact of adding incentives to bring inactive health care professionals back to the

workforce.

By providing an initial understanding of this infinite-horizon model, our goal is to move a step forward in the field of strategic health workforce planning, and to motivate others to continue doing research in this important application.

CHAPTER III

INTEGRATED INVENTORY ROUTING AND FREIGHT CONSOLIDATION OF PERISHABLE GOODS

3.1 Introduction

Transportation costs, the single largest cost element which accounts for 50-65% of the total logistics cost, has long been a focus of supply chain strategies ([74]). In practice, a supplier or carrier often achieves efficient use of transportation assets by intelligently routing a fleet to serve multiple customers. When the transportation decision is coupled with inventory trade-offs, it leads to the *inventory routing problem* (IRP) in well-known stories e.g. Walmart's *vendor-managed inventory* (VMI) initiative ([13]), ExxonMobil's liquefied natural gas projects ([98]).

Nowadays, the ever-increasing global competition also encourages transportation cost savings through horizontal cooperation within a supply chain echelon. This is the case, e.g., when a warehouse sorts different suppliers' orders to schedule outbound shipments at minimized per-unit freight cost of common products. A real lesson is taught by the California cut flower industry, where local growers ship the product individually to national sellers, and because of small disaggregated volume, usually suffer from expensive *less-than-truckload* (LTL) or courier (e.g. FedEx, UPS) services. During the last two decades, this disadvantage kept playing a major role in the loss of over 40% U.S. market share to the South American counterparts, who enjoy favorable *full truckload* (FTL) rates by aggregating all shipments prior to domestic distribution. On the other hand, [96] estimated that consolidation could reduce annual transportation costs by \$6-17 million with 20-50 California cut flower growers' participation.

Competitive transportation is further attainable by vertical coordination across supply chain echelons. Although consolidation ameliorates outbound transportation with less frequent deliveries and higher shipping volume, sole emphasis on it may restrict the pickup plans of target orders to be recombined at a terminal. Since trucks are the only form of agriculture transportation in at least 80% of U.S. cities and communities ([122]), flexible inbound transportation also actively contributes to lower total freight costs. Moreover, it is hard to evaluate the impact of consolidation alone on inventory control at each facility without considering both echelons together. Therefore, many agricultural sectors are realizing the collaboration imperative for growers, consolidation terminals, sellers as well as third-party carriers.

Generally speaking, cooperation and coordination can benefit almost every area from procurement to last mile delivery. Besides the aforementioned truckload costs, economies of scale prevail when a supplier offers batch ordering discounts, a factory processes identical jobs on heterogeneous machines, a liner company ships packed cargoes with container vessels, etc. Given complex entities involved and distinctive features of each, the ideal supply chain performance usually necessitates joint management of cross-functional activities such as production, transportation, consolidation and inventory. Abundant research opportunities thus arise in the contexts of synchronized sourcing and fulfillment, combined production and distribution, hub-and-spoke transshipments, etc.

This chapter integrates two classic problems that are typically solved independently in these and similar supply chains. As an interpretation in agriculture logistics, the first decision is the shipments and routes from local growers to a consolidation center, which is the short-haul problem and can be modeled as an IRP. The second decision is the direct shipments from the consolidation center to individual retailers or wholesalers, which is the long-haul freight consolidation problem. The separate decisions are difficult themselves with real-world complications like demand

dynamics, perishability, truckload costs, routing capacities and duration limits. We propose a *mixed-integer programming* (MIP) formulation of the integrated problem, and develop an iterative framework with a decomposition and a search scheme to obtain solutions of high quality in reasonable time. In the decomposition, a freight consolidation subproblem is first solved to obtain long-haul FTL schedules, and after fixing these a restrictive model generates the other decisions for the integrated MIP. The local search aims at fast identification of good neighborhoods by solving an assignment-style MIP which matches the consolidation decision with an IRP subproblem, and gradually strengthens the incumbent solution pool when executed in an iterative fashion.

A notable fact of the agricultural industry is that long-haul transportation costs usually dominate the total distribution cost, which intuitively suggests consolidation-based strategies to control the main cost elements. Hence, the standard approach is to first solve the long-haul freight consolidation decision, and then use the delivery quantities as the demand for the short-haul IRP. However, our results show that the overall costs can be reduced by utilizing a system-wide optimization approach to solving the integrated problem. We also demonstrate the potential of the iterative framework in balancing solution efficacy and computational efficiency as the problem size increases. The standard approach tends to exhaust 5 hours in solving a moderate-sized IRP subproblem with CPLEX, whereas the proposed method can significantly improve the global solution by exploring more neighborhoods in less average CPU time per iteration.

In the remainder of the chapter, we provide a brief literature review in Section 3.2, describe problem details in Section 3.3, illustrate the iterative heuristic in Section 3.4, present experimental results in Section 3.5, and discuss potential future research in Section 3.6.

3.2 Literature Review

Both inventory routing and freight consolidation problems have drawn extensive attention in the operations research community. The IRP simultaneously decides 1) when a central facility dispatches vehicles; 2) which customers and in which order to visit for a trip; and 3) how much demand to fulfill or inventory to maintain at each facility. Over the past thirty years, numerous IRP models have been studied with quite specific characteristics, among which the closest variants to our short-haul problem are the class of finite horizon multi-period single-vehicle one-to-many IRP. The solution techniques widely range from exact methods to metaheuristics to optimization-based heuristics. We refer the reader to [47] and [13] for comprehensive surveys on state-of-the-art methodologies and industrial applications, respectively.

Freight consolidation addresses the question of how much volume or how many time periods to accumulate before releasing a shipment that leverages economies of scale in transportation costs. Since the pioneering work of [33, 77], quantity-based, time-based as well as hybrid policies have been investigated on which [42] gave an excellent summary. In particular, [95, 96] proposed near-optimal heuristics to solve the version of our long-haul problem without inventory aspects. When inventory is taken into account, the long-haul decision can be viewed as a unique *lot-sizing problem* (LSP) where the ordering cost function is *piecewise linear* (PWL) as depicted in Figure 7. Pertinent research includes LSP with multiple set-up costs ([11, 60, 90]) and volume discounts ([18, 43]), but usually assumes concave or monotonic properties which do not generalize to our case.

The literature is relatively rare on integration of inventory, routing, and consolidation. Representative problems in this venue are the *production routing problem*

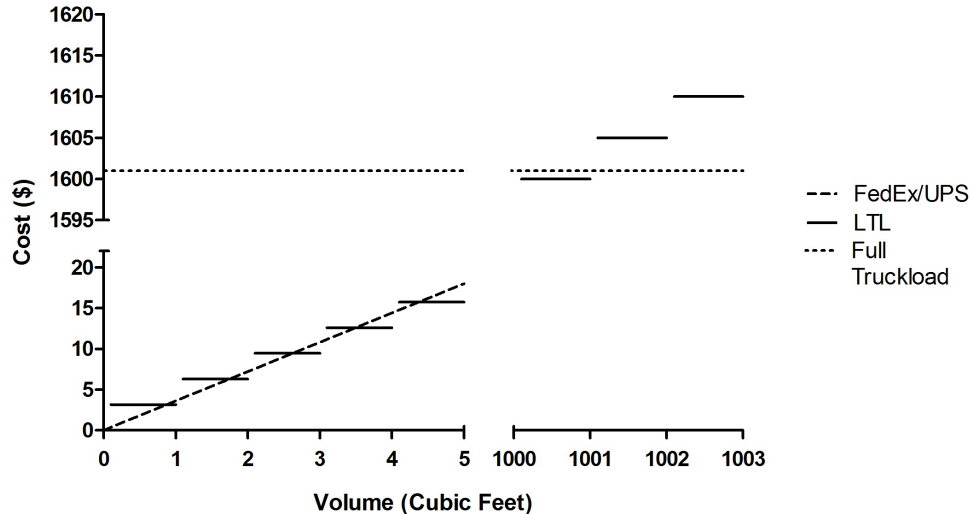


Figure 7: Volume-dependent long-haul shipping costs

(PRP, [9]) and the *maritime inventory routing problem* (MIRP, [98]). PRP coordinates two core supply chain functions, namely production planning and distribution, which are the origins of LSP and IRP thus resemble our long-haul and short-haul decisions, respectively. However, the few publications on this subject concern much simpler settings e.g. specified inventory replenishment rules ([2, 34]), uncapacitated production and/or routing ([108]), fixed or PWL concave production costs ([8, 21, 23, 24, 25, 34, 35]). Furthermore, existing PRP models largely consider a homogeneous product from a single plant to customers that do not require delivery time windows. In contrast, we treat the demand for each grower-seller pair as an individual commodity, and impose hard constraints on the time in transit or storage to accommodate the perishable nature of the agricultural product. On the other hand, MIRP typically involves multiple commodities, loading/discharging time slots and intricate cost elements, but fundamentally differs in network topology and does not share the cost structure of our kind. Another related yet less-studied problem is the integration of various transportation modes in the conventional IRP, e.g., when a main route serves a few major customers who then transship the goods to minor customers via direct shipping ([46]).

From the perspective of solution approaches, many exact and heuristic IRP algorithms have been successfully extended to MIRP and IRP with transshipment, whereas the PRP literature centers more on heuristics. Early attempts are primarily metaheuristics especially those very powerful in vehicle routing e.g. GRASP ([34]), memetic algorithm ([35]), tabu search ([21, 24]). Metaheuristics are capable of tackling PRP instances up to 200 customers and 20 time periods ([34]), but the implementation is difficult because of increased complexity in combined decisions and constraints. Exact methods have also been developed such as branch-and-price algorithms ([25]), branch-and-cut algorithms ([7, 17, 46, 108]), and Lagrangian relaxations ([64, 116]), which unfortunately are effective only in more basic variants and smaller problem instances. A seemingly promising alternative is thus optimization-based heuristics. In this spirit, [8] put forward a hybrid *adaptive large neighborhood search* (ALNS) scheme where upper-level search operators handle binary setup and routing decisions, and lower-level network flow problems yield the corresponding production, inventory and shipping quantities. Recently, [2] introduced a new two-phase scheme which considers a lot-sizing problem with approximated routing costs first and a routing problem subsequently. Both approaches exploit diversification mechanisms to prevent fast convergence to local optima in an iterative fashion, and outperform previous methods for most benchmark instances with 14-200 customers and 6-20 time periods.

3.3 Problem Statement

We consider the distribution of a perishable product with a fixed lifetime and known deterministic demand over a finite discretized horizon (Figure 16). The product is moved from local growers to a consolidation center via short-haul routing, and then from the center to geographically dispersed retailers/wholesalers via long-haul direct shipping. We assume one vehicle available per period for local pickup, but allow more

expensive direct shipping alternatives for possible excess demand. We differentiate long-haul delivery options by volume-dependent services, including fixed FTL rates, fixed LTL rates, and linear courier rates. The growers and the center may keep inventory to delay pickup or delivery with facility-specific unit holding cost rates, but all commodities must leave the system before spoilage.

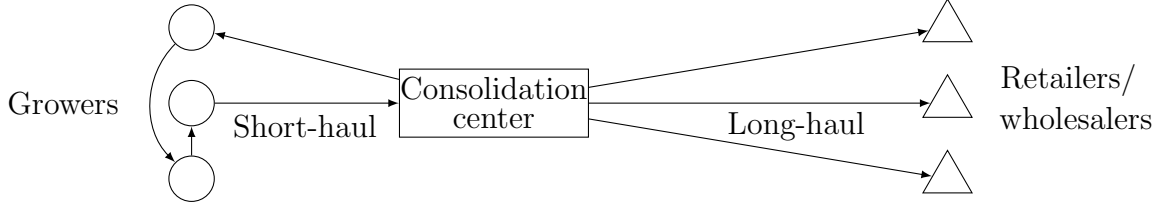


Figure 8: Integrated distribution network

Parameters

$\mathcal{T} = \{1, \dots, T\}$: set of periods in the planning horizon.

θ : product lifetime.

$G, 0, D$: set of growers, the consolidation center, and set of sellers, respectively.

$\{m\} = M_{ik}^s$: a commodity tuple of product that is ready for pickup at the beginning of period s and will be moved from grower i to seller k , $\forall i \in G, k \in D, 1 \leq s < T$. M denotes the set of all commodity tuples, and \cdot indicates all elements of the respective index. E.g., M_i is the set of commodity tuples that originate from grower i .

d_m : demand for commodity m .

Q : local vehicle capacity.

l_{ij} : travel time from facility i to facility j , $\forall i, j \in G \cup \{0\}, i \neq j$.

c_{ij} : mileage cost from facility i to facility j , $\forall i, j \in G \cup \{0\}, i \neq j$.

B : alternative short-haul direct shipping cost per shipment.

$\{F, L, U\}$: set of long-haul direct shipping modes, where F, L, U represent FTL, LTL and courier services, respectively.

$\mathcal{K}_F, \mathcal{K}_L$: maximum capacities in cubic feet for a long-haul FTL and LTL, respectively.

c_{kF}, c_{kL} : transportation costs for an FTL and LTL, respectively, from the consolidation center to seller k , $\forall k \in D$.

α : conversion factor (lbs per cubic foot).

c_{kU} : transportation cost (per pound) for a courier shipment to seller k , $\forall k \in D$.

h_i : inventory holding cost per unit product per period at facility i , $\forall i \in G \cup \{0\}$.

The goal is to minimize total transportation and inventory costs while satisfying the demands, local vehicle capacity and routing duration limits, as well as long-haul shipping capacities for each mode. We propose the MIP formulation below.

Decision variables

$$x_{ij}^t = \begin{cases} 1, & \text{arc } (i, j) \text{ is traversed in period } t \\ 0, & \text{otherwise} \end{cases}, \quad \forall i, j \in G \cup \{0\}, i \neq j, 1 \leq t < T.$$

$$y^t = \begin{cases} 1, & \text{a trip occurs in period } t \\ 0, & \text{otherwise} \end{cases}, \quad 1 \leq t < T.$$

u_i^t : number of alternative local vehicles used by grower i in period t , $\forall i \in G, 1 \leq t < T$.

q_m^t : pick-up volume of commodity m in period t , $\forall m \in M^s, \max\{0, t - \theta\} < s \leq t, 1 \leq t < T$.

v_m^t : volume of commodity m picked up in period t via alternative direct shipping, $\forall m \in M^s, \max\{0, t - \theta\} < s \leq t, 1 \leq t < T$.

f_{ijm}^t : flow volume of commodity m on arc (i, j) in period t , $\forall i, j \in G \cup \{0\}, i \neq j, m \in M^s, \max\{0, t - \theta\} < s \leq t, 1 \leq t < T$.

z_{mp}^t : delivery volume of commodity m with mode p in period t , $\forall p \in \{F, L, U\}, m \in M^s, \max\{1, t - \theta\} \leq s < t, 1 < t \leq T$.

r_{kp}^t : FTL numbers or LTL units sent to seller k in period t , $\forall k \in D, p \in \{F, L\}, 1 < t \leq T$.

I_{im}^t : grower inventory of commodity m at the end of period t , $\forall i \in G, m \in M_i^s, \max\{0, t - \theta\} < s \leq t, 1 \leq t < T$.

I_{0m}^t : central inventory of commodity m at the end of period t , $\forall m \in M^s, \max\{1, t - \theta\} \leq s < t, 1 < t \leq T$.

Objective function

The total distribution cost includes three parts: 1) the short-haul transportation cost, which equals regular arc routing costs plus alternative costs; 2) the long-haul transportation cost, which is the sum of FTL/LTL numbers and courier volume multiplied by the respective dispatch cost rates; 3) the inventory cost, which covers the linear costs for keeping commodities in the system, and is evaluated with facility-wise holding cost rates.

$$\begin{aligned}
\min \quad & \sum_{t=1}^{T-1} \sum_{i,j \in G \cup \{0\}, i \neq j} c_{ij} x_{ij}^t + \sum_{t=2}^T \sum_{k \in D} \sum_{p \in \{F, L\}} c_{kp} r_{kp}^t + \alpha \sum_{t=2}^T \sum_{s=\max\{1, t-\theta\}}^{t-1} \sum_{k \in D} \sum_{m \in M_{k,s}^s} c_{kU} z_{mU}^t \\
& + \sum_{t=1}^{T-1} \sum_{i \in G} h_i \sum_{s=\max\{1, t-\theta+1\}}^t \sum_{m \in M_i^s} I_{im}^t + h_0 \sum_{t=2}^T \sum_{s=\max\{1, t-\theta\}}^{t-1} \sum_{m \in M^s} I_{0m}^t + \sum_{t=1}^{T-1} \sum_{i \in G} B u_i^t
\end{aligned} \tag{8}$$

Route definition

Arc degree constraints: (9a) and (9b) relate the node degrees in the short-haul network to whether a trip occurs in each period. Specifically, the outdegree of the center equals 1 if there is a trip, and 0 otherwise; the outdegree of a grower is no more than that of the center. (9c) balance each facility's indegree and outdegree.

Commodity flow constraints (9d)-(9e): for a commodity originated at a grower, the total outflow equals the total inflow plus the pickup volume in any period; for a commodity originated elsewhere, the total outflow equals the total inflow. These constraints also eliminate subtours in a trip.

Vehicle capacity constraints (9f)-(9g): the total volume carried by a regular vehicle does not exceed its capacity on any arc traversed in each period; the total local direct shipping volume does not exceed the total capacity of alternative vehicles dispatched from a grower in any period.

Duration constraints (9h): the total duration of a tour is within one period.

$$\sum_{i \in G} x_{0i}^t = y^t, \quad 1 \leq t < T \quad (9a)$$

$$\sum_{j \in G \cup \{0\}} x_{ij}^t \leq y^t, \quad \forall i \in G, 1 \leq t < T \quad (9b)$$

$$\sum_{j \in G \cup \{0\}, j \neq i} x_{ij}^t - \sum_{j \in G \cup \{0\}, j \neq i} x_{ji}^t = 0, \quad \forall i \in G \cup \{0\}, 1 \leq t < T \quad (9c)$$

$$\sum_{j \in G \cup \{0\}, j \neq i} f_{ijm}^t - \sum_{j \in G, j \neq i} f_{jim}^t = q_m^t, \quad \forall i \in G, m \in M_i^s, \max\{0, t - \theta\} < s \leq t, 1 \leq t < T \quad (9d)$$

$$\sum_{j \in G \cup \{0\}, j \neq i} f_{ijm}^t - \sum_{j \in G, j \neq i} f_{jim}^t = 0, \quad \forall i \in G, m \notin M_i^s, \max\{0, t - \theta\} < s \leq t, 1 \leq t < T \quad (9e)$$

$$\sum_{s=\max\{1, t-\theta+1\}}^t \sum_{m \in M^s} f_{ijm}^t \leq Qx_{ij}^t, \quad \forall i, j \in G \cup \{0\}, i \neq j, 1 \leq t < T \quad (9f)$$

$$\sum_{s=\max\{1,t-\theta+1\}}^t \sum_{m \in M_i^s} v_m^t \leq Qu_i^t, \quad \forall i \in G, 1 \leq t < T \quad (9g)$$

$$\sum_{i,j \in G \cup \{0\}, i \neq j} l_{ij} x_{ij}^t \leq 1, \quad 1 \leq t < T \quad (9h)$$

Demand satisfaction

(10a): For the short-haul echelon, the total pickup quantity of each commodity, including regular routing and alternative direct shipping volume, equals its demand.

(10b): For the long-haul echelon, the total delivery quantity of each commodity, including FTL, LTL and courier volume, equals its demand.

$$\sum_{t=s}^{\min\{s+\theta, T\}-1} (q_m^t + v_m^t) = d_m, \quad \forall m \in M^s, 1 \leq s < T \quad (10a)$$

$$\sum_{t=s+1}^{\min\{s+\theta, T\}} \sum_{p \in \{F, L, U\}} z_{mp}^t = d_m, \quad \forall m \in M^s, 1 \leq s < T \quad (10b)$$

Long-haul direct shipping capacities

In each period, the total FTL and LTL delivery quantity to a seller does not exceed the total capacity of the dispatched FTLs and LTLs.

$$\sum_{s=\max\{1,t-\theta\}}^{t-1} \sum_{m \in M_k^s} z_{mp}^t \leq \mathcal{K}_p r_{kp}^t, \quad \forall k \in D, p \in \{F, L\}, 1 < t \leq T \quad (11)$$

Inventory conservation

(12a)-(12b): grower inventory conservation, i.e., the ending inventory of a commodity equals the demand in its ready period or the beginning inventory in other periods, minus the corresponding pickup volume.

(12c): central inventory conservation, i.e., the ending inventory of a commodity equals the beginning inventory plus the pickup volume and minus the delivery volume in each period.

(12d): boundary condition.

$$I_{im}^t + q_m^t + v_m^t = d_m, \quad \forall i \in G, m \in M_i^s, 1 \leq t = s < T \quad (12a)$$

$$I_{im}^t - I_{im}^{t-1} + q_m^t + v_m^t = 0, \quad \forall i \in G, m \in M_i^s, 1 \leq s < t < \min\{s + \theta, T\} \quad (12b)$$

$$I_{0m}^t - I_{0m}^{t-1} - q_m^{t-1} - v_m^{t-1} + \sum_{p \in \{F, L, U\}} z_{mp}^t = 0, \quad \forall m \in M^s, 1 \leq s < t \leq \min\{s + \theta, T\} \quad (12c)$$

$$I_{0m}^t = 0, \quad \forall m \in M^s, 1 \leq t < s + 1 \leq T \quad (12d)$$

Domain

$$x \in \{0, 1\}, y \in \{0, 1\}, u \in \mathbb{Z}_+, r \in \mathbb{Z}_+, q \geq 0, v \geq 0, z \geq 0, f \geq 0, I \geq 0 \quad (13)$$

We will refer to (8)-(16g) as the *full MIP* model. The problem incorporates the conventional IRP and so is \mathcal{NP} -hard. In preliminary tests, we observed that instances with 15 periods, 10 growers and 5 sellers contain over 1,500 binary variables and 100 general integer variables, which usually took 5 hours to solve with CPLEX. In addition, the multi-commodity coefficient matrix is high-dimensional and could run out of memory on a CONDOR system when the number of growers exceeds 15. Hence the formulation easily becomes intractable as the problem size increases.

3.4 Solution Approach

At the heart of our solution approach are a decomposition procedure and a local search scheme. For illustrative purposes, we interpret the consolidation center as a common customer of the growers and a common supplier of the sellers. Accordingly, we define *central demand* as the quantity of each commodity that is required to arrive at the center by the end of a period, and let *central supply* be the quantity of each commodity that is expected to be ready at the center at the beginning of a period.

3.4.1 Decomposition

The basic idea of decomposition is to reduce computational burden by solving a series of subproblem decisions with partial information fixed in the full MIP. As Table 2 shows, we identify three subproblems to be used in different phases of the approach. The *direct shipping* (DS) subproblem assumes known central supply and determines the delivery quantities as well as FTL numbers and LTL units in each period, which satisfy demand, truck capacity and central inventory balance constraints in a way that minimizes the corresponding long-haul transportation and central inventory costs. The IRP subproblem assumes known central demand and determines the pickup quantities, commodity flows as well as local routes in each period, which satisfy demand, regular and alternative vehicle capacities, node degree and commodity flow constraints in a way that minimize the corresponding short-haul transportation, grower and central inventory costs. The *restricted full MIP* subproblem assumes fixed FTL numbers in each period, and determines all the other decisions that minimize the corresponding total distribution cost.

Table 2: A sketch of three subproblems

The DS subproblem	
Fixed information	q
Decision variables	z, r, I_0
Goal (to minimize)	Long-haul costs
Constraints	(10b)-(11), (12c)-(16g)
The IRP subproblem	
Fixed information	z
Decision variables	x, y, q, v, u, f, I
Goal (to minimize)	Short-haul costs
Constraints	(9a)-(10a), (12a)-(12b), (16g)
The restricted full MIP subproblem	
Fixed information	$r.F$
Decision variables	$x, y, q, v, u, f, z, I, r.L$
Goal (to minimize)	Total distribution cost
Constraints	(9a)-(16g)

The subproblems are interrelated in that the output of each problem naturally defines an input neighborhood for the others. Depending on how we combine them, we may obtain various decompositions to solve the full MIP. For instance, a standard *DS-guided* approach sequentially solves the DS subproblem assuming earliest possible central supply and the IRP subproblem with central demand derived from the long-haul schedules. In each iteration of our approach, we recommend an alternative *DS-restricted full MIP* decomposition where the FTL numbers instead of the entire long-haul schedules are fixed as the DS phase ends. Since constraints (10b)-(11), (12c)-(12d) are irrelative to the IRP objective, the restricted full MIP is in fact a relaxation of the IRP subproblem. The reason for not following the standard decomposition is to balance flexibility and tractability of the subproblems. As Example 5 indicates, an over-restrictive fixing scheme could severely affect the overall solution quality.

Example 5. Consider a network with two growers and one retailer where the planning horizon length $T = 3$ days, the product lifetime $\theta = 2$ days, and the local vehicles and the long-haul FTLs have identical capacities i.e. $Q = \mathcal{K}_F$. Suppose two orders are ready at the beginning of the planning horizon, $0.5Q$ for one grower and $0.5Q + \epsilon$ ($\epsilon > 0$) for the other. The total shipping volume exceeds the FTL capacity by a small volume of ϵ units which will be shipped via LTL or courier services. Since $h_0 > 0$ and the long-haul transportation cost depends on the volume rather than the shipping period, the optimal DS solution ships all $\mathcal{K}_F + \epsilon$ units on the second day to minimize central inventory costs. In the consequent IRP subproblem, this results in a violation of the local vehicle capacity on the first day and thus induces a penalty cost of B . On the other hand, the global optimum would postpone ϵ units to the third day as a compromise between the two echelons. The DS-restricted full MIP decomposition attains this optimum since the final LTL and courier schedules are subject to short-haul decisions. The excess total cost of the DS-guided decomposition is then $B - h_2\epsilon$, which can be very high for small holding cost rate at the second grower.

The DS subproblem can be solved with CPLEX for moderate instances and the heuristics in [95, 96] for large instances assuming just-in-time central supply (i.e. $I_0 = 0$). The IRP subproblem and the restricted full MIP can be solved with CPLEX for small instances. Note that in an iterative approach, we do not need to solve each subproblem to optimality since the decisions interact with each other and vary from iteration to iteration. Also, existing heuristics for IRP variants with time windows, though harder to implement, may be employed to solve our subproblems.

3.4.2 Local Search

The motivation for developing a local search scheme is to offset bad local optimality in a single iteration of the decomposition. Despite a theoretical improvement by substituting the IRP subproblem with the restricted full MIP, there is no quality guarantee for the resulting global solution. Also, the desired improvement is valid only if the subproblems are solved to optimality. However, as we already mentioned, efficient subproblem solutions can be an issue for large problem instances, and the restricted full MIP is conceptually more complicated than the IRP subproblem. A subroutine is thus necessary to explore more and potentially better neighborhoods in the solution space.

In this section, we propose an MIP-based local search which takes advantage of the inherent “incompatibility” between DS and IRP. We introduce two terms before elaborations.

Definition 3.4.1 (*Mismatched demand, MMD*). Assume both IRP and DS subproblems are solved simultaneously under a given central demand/supply assignment, a subcommodity (i, k, s, ι, τ) is said to be mismatched if the IRP pickup time ι and the DS delivery time τ are such that $s < \tau \leq \iota < \min\{s + \theta, T\}$.

Definition 3.4.2 (*Pickup/delivery time windows*). A time interval $[\tau, \iota]$ is called a pickup window when a commodity (i, k, s) is allowed for shipments from the grower,

$s \leq \tau \leq \iota < \min\{s + \theta, T\}$. A time interval $[\tau, \iota]$ is called a delivery window when a commodity (i, k, s) is allowed for shipments to the seller, $s < \tau \leq \iota \leq \min\{s + \theta, T\}$.

Since later central demand benefits short-haul routing whereas earlier central supply facilitates long-haul consolidation, MMD can arise when the IRP subproblem and the DS subproblem are solved separately. According to Definition 3.4.2, we can eliminate MMD by adjusting the pickup/delivery time windows for mismatched commodities. We explain the basic idea with the example below.

Example 6. Consider the time-space network in Figure 9. There are three planning periods, two growers and two retailers, i.e. $T = 3$ days, $G = \{i, k\}$, $D = \{j, \ell\}$. The product has a lifetime $\theta = 2$ days, and the demands are $d_{ij}^1 = 5$, $d_{i\ell}^1 = 5$, $d_{i\ell}^2 = 5$, $d_{kj}^1 = 10$, $d_{k\ell}^2 = 5$. Suppose local mileage costs are symmetric, holding costs are such that $h_i > h_0 > h_k$, and vehicle capacities $Q = \mathcal{K}_F = 15$. The consolidation center is split into two copies representing the IRP customer and the DS supplier, respectively. Assume the central supply is ready for long-haul delivery the day after a demand is ready for pickup at the grower, whereas the central demand is not due until end of the horizon. This encourages each subproblem to best utilize the transportation capacity. The total demand is $\sum_{s=1}^2 (d_{ij}^s + d_{i\ell}^s + d_{kj}^s + d_{k\ell}^s) = 30 = 2Q$, and the portion ready on the first day is $d_{ij}^1 + d_{i\ell}^1 + d_{kj}^1 + d_{k\ell}^1 = 20 = Q + 5$. To avoid expensive alternative direct shipping, the IRP subproblem tries to fully utilize the local vehicle capacity, which means a volume of 5 units will be held in inventory on day 1. Since $h_i > h_k$, grower i 's demands are prioritized whereas half of commodity $(k, j, 1)$ is delayed until day 2 for local pickup. On the other hand, the total demands for retailer j and retailer k both equal K_F , and are expected to be ready for long-haul delivery on day 2 and day 3, respectively; hence the optimal DS subproblem solution sends out an FTL on each day. These are given in Figure 9 (a), where the central flow imbalance from day 2 to day 3 indicates that subcommodity $(k, j, 1, 2, 2)$ induces an MMD of 5 units. Therefore, the corresponding full MIP solution would be infeasible if we piece the

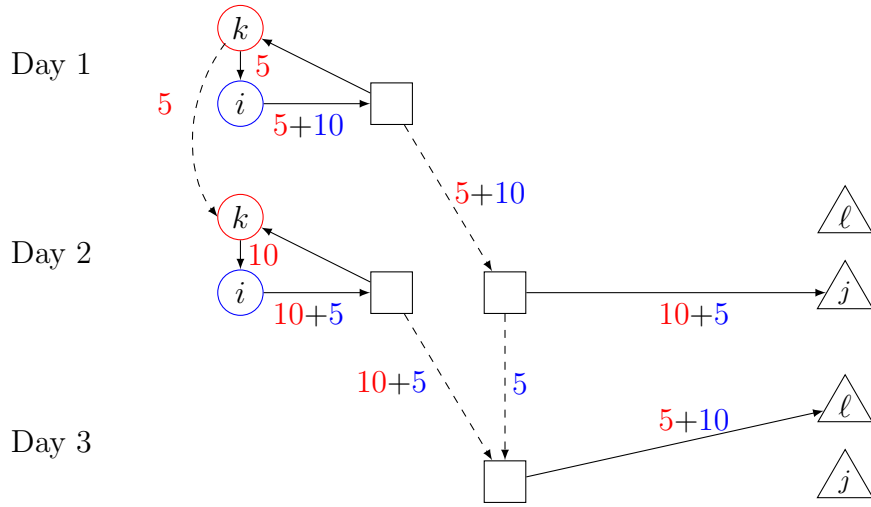
subproblem solutions together. To remove the MMD, we may narrow the delivery window for the entire commodity $(k, j, 1)$ to day 3 in the DS subproblem, or narrow the pickup window to day 1 in the IRP subproblem, or take a mixture given by Figure 9 (b), where the pickup window for 2 units of subcommodity $(k, j, 1, 2, 2)$ is narrowed to day 1, and the delivery window for the remaining 3 units is narrowed to day 3, respectively.

In essence, MMD provides a guide to reassign central demand/supply so that the new neighborhoods are more compatible when we revisit the subproblems. Since reassignment rules impact routing and consolidation decisions as well, we propose to find a candidate strategy by solving a *demand reassignment problem*.

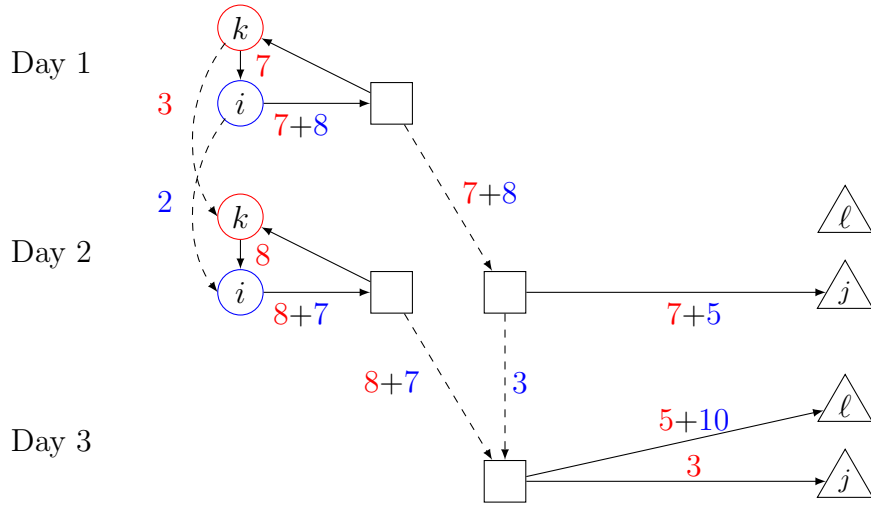
The input of the reassignment problem includes relevant full MIP parameters as well as extra data from the IRP and DS subproblem solutions. Specifically, we calculate mismatched subproblem demand, residual short-haul and long-haul transportation capacities as well as remaining time allowed for local routes. We also estimate the routing cost and duration changes for each pair of grower and regular route. If grower i is visited by a regular vehicle in period t (abbr. *route t*), we consider no insertion cost for (i, t) , and approximate the routing cost savings with the amount obtained by joining grower i 's predecessor and successor when it is removed from route t . Meanwhile, if grower i is not visited by route t , we consider no removal savings for (i, t) , and approximate the insertion cost with the amount obtained by the cheapest insertion heuristic for the *traveling salesman problem* (TSP). The duration changes are determined consequently.

Additional input

$\{m\} = \mathcal{M}_{ik}^{s\iota\tau} \subseteq M_{ik}^s$: a tuple of MMD where a portion of commodity (i, k, s) is picked up in period ι for the IRP subproblem and delivered in period τ for the DS subproblem, $s < \tau \leq \iota < \min\{s + \theta, T\}$. Let d_m be the associated subcommodity demand.



(a) Full MIP infeasibility due to subproblem MMD



(b) A reassignment strategy by narrowing pickup and delivery windows

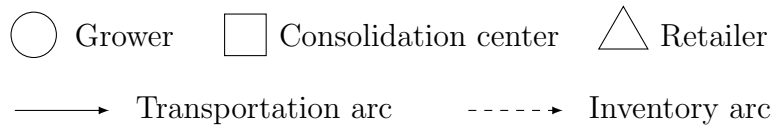


Figure 9: MMD and reassignment strategies

σ_i^t : indicator, equals 1 if grower i is visited by route t , 0 otherwise.

η_{i+}^t : insertion cost for grower-route pair (i, t) , equals 0 if $\sigma_i^t = 1$.

l_{i+}^t : duration increase for route t after inserting grower i , equals 0 if $\sigma_i^t = 1$.

η_{i-}^t : cost savings by removing i from t , equals $\sigma_i^t = 0$.

l_{i-}^t : duration reduction for route t after removing grower i , equals 0 if $\sigma_i^t = 0$.

\mathcal{L}^t : remaining time allowed for route t , equals 1 minus the duration of route t if it occurs, 1 otherwise.

\mathcal{Q}^t : residual capacity for route t , equals Q minus the total pickup volume for route t if it occurs, Q otherwise.

\mathcal{V}_k^t : total FTL and LTL volume sent to seller k in period t in the DS subproblem solution.

\mathcal{N}_{kp}^t : number of FTL trucks or LTL units sent to seller k in period t in the DS subproblem solution, $p \in \{F, L\}$.

The decision variables of the reassignment problem include insertion or removal for each grower-route pair (i, t) , pickup and delivery quantities for each mismatched subcommodity, extra or saved long-haul FTL and LTL numbers, as well as courier volume and inventory changes in each period.

Decision variables

$$\pi_i^t = \begin{cases} 1, & \text{if grower } i \text{ is inserted to route } t \\ 0, & \text{otherwise} \end{cases}, i \in G, 1 \leq t < T.$$

$$\rho_i^t = \begin{cases} 1, & \text{if grower } i \text{ is removed from route } t \\ 0, & \text{otherwise} \end{cases}, i \in G, 1 \leq t < T.$$

$$\nu^t = \begin{cases} 1, & \text{if the new route } t \text{ exceeds capacity or duration limit} \\ 0, & \text{otherwise} \end{cases}, \quad 1 \leq t < T.$$

$\Delta_m^t \in \mathbb{R}_+$: reassigned pickup volume of subcommodity m to route t , $m \in \mathcal{M}^s$, $1 \leq s \leq t < \min\{s + \theta, T\}$.

$\delta_m^t \in \mathbb{R}_+$: reassigned delivery volume of subcommodity m in period t , $m \in \mathcal{M}^s$, $1 \leq s < t \leq \min\{s + \theta, T\}$.

$r_{kp+}^t, r_{kp-}^t \in \mathbb{Z}_+$: extra or saved FTL or LTL numbers dispatched to seller k in period t , $k \in D, p \in \{F, L\}$, $1 < t \leq T$.

$z_{k+}^t, z_{k-}^t \in \mathbb{R}_+$: extra or saved courier volume shipped to seller k in period t , $k \in D$, $1 < t \leq T$.

$I_{im+}^t, I_{im-}^t \in \mathbb{R}_+$: inventory increase or reduction of subcommodity m at grower i in period t , $m \in \mathcal{M}_i^s$, $i \in G$, $1 \leq s \leq t < \min\{s + \theta, T\}$.

$I_{0m}^t \in \mathbb{R}_+$: central inventory of subcommodity m in period t , $m \in \mathcal{M}^s$, $1 \leq s < t \leq \min\{s + \theta, T\}$.

We model the reassignment problem as a MIP. To allow some flexibility in local direct shipping, we assume that if a new route exceeds the local vehicle capacity or duration limit, one alternative vehicle is sufficient to make the solution feasible.

Reassignment MIP

$$\begin{aligned} \min \quad & \sum_{t=1}^{T-1} \sum_{i \in G} (\eta_{i+}^t \pi_i^t - \eta_{i-}^t \rho_i^t) + \sum_{t=1}^{T-1} B\nu^t + \sum_{t=2}^T \sum_{k \in D} \sum_{p \in \{F, L\}} c_{kp} (r_{kp+}^t - r_{kp-}^t) + \alpha \sum_{t=2}^T \sum_{k \in D} c_{kU} (z_{k+}^t \\ & - z_{k-}^t) + \sum_{t=1}^{T-1} \sum_{i \in G} \sum_{s=\max\{1, t-\theta+1\}}^t \sum_{m \in \mathcal{M}_i^s} h_i (I_{im+}^t - I_{im-}^t) + \sum_{t=2}^T \sum_{s=\max\{1, t-\theta\}}^{t-1} \sum_{m \in \mathcal{M}^s} h_0 I_{0m}^t \end{aligned} \quad (14a)$$

$$\text{s.t. } \Delta_m^t \leq d_m(\sigma_i^t + \pi_i^t - \rho_i^t), \quad \forall m \in \mathcal{M}_i^{s^*}, i \in G, 1 \leq s \leq t < \min\{s + \theta, T\} \quad (14b)$$

$$\pi_i^t \leq 1 - \sigma_i^t, \quad \forall i \in G, 1 \leq t < T \quad (14c)$$

$$\rho_i^t \leq \sigma_i^t, \quad \forall i \in G, 1 \leq t < T \quad (14d)$$

$$\sum_{s=\max\{1, t-\theta+1\}}^t \left(\sum_{m \in \mathcal{M}^{s^*}: \tau \neq t} \Delta_m^t - \sum_{m \in \mathcal{M}^{s^*}} (d_m - \Delta_m^t) \right) \leq \mathcal{Q}^t + Q\nu^t, \quad \forall 1 \leq t < T \quad (14e)$$

$$\sum_i l_{i+}^t \pi_i^t - \sum_i l_{i-}^t \rho_i^t \leq \mathcal{L}^t + \nu^t, \quad \forall 1 \leq t < T \quad (14f)$$

$$\sum_{t=s}^{\min\{s+\theta, T\}-1} \Delta_m^t = \sum_{t=s+1}^{\min\{s+\theta, T\}} \delta_m^t = d_m, \quad \forall m \in \mathcal{M}^{s^*}, 1 \leq s < T \quad (14g)$$

$$\sum_{t=s}^{\tau} \Delta_m^t \geq \sum_{t=s+1}^{\tau+1} \delta_m^t, \quad \forall m \in \mathcal{M}^{s^*}, 1 \leq s \leq \tau < \min\{s + \theta, T\} \quad (14h)$$

$$\mathcal{V}_k^t + \sum_{s=\max\{1, t-\theta\}}^{t-1} \left(\sum_{m \in \mathcal{M}_{k^*}^{s^*}: \tau \neq t} \delta_m^t - \sum_{m \in \mathcal{M}_{k^*}^{s^*}} (d_m - \delta_m^t) \right) \leq \sum_p K_p (\mathcal{N}_{kp}^t + r_{kp+}^t - r_{kp-}^t) + z_{k+}^t - z_{k-}^t, \quad \forall k \in D, 1 < t \leq T \quad (14i)$$

$$I_{im+}^t - I_{im-}^t = \sum_{\iota=t+1}^{\min\{s+\theta, T\}-1} \Delta_m^\iota, \quad \forall m \in \mathcal{M}_i^{s^*}, i \in G, 1 \leq s \leq \tau \leq t < \min\{s + \theta, T\} \quad (14j)$$

$$I_{im-}^t - I_{im+}^t = \sum_{\iota=s}^t \Delta_m^\iota, \quad \forall m \in \mathcal{M}_i^{s^*}, i \in G, 1 \leq s \leq t < \tau < \min\{s + \theta, T\} \quad (14k)$$

$$I_{0m}^t = \sum_{\tau=s}^{t-1} \Delta_m^\tau - \sum_{\tau=s+1}^t \delta_m^\tau, \quad \forall m \in \mathcal{M}^{s^*}, 1 \leq s < t \leq \min\{s + \theta, T\} \quad (14l)$$

$$r_{kp-}^t \leq \mathcal{N}_{kp}^t, \quad \forall k \in D, p \in \{F, L\}, 1 < t \leq T \quad (14m)$$

$$z_{k-}^t \leq \sum_{m \in \mathcal{M}_{k^*}^t} d_m - \mathcal{V}_k^t, \quad \forall k \in D, 1 < t \leq T \quad (14n)$$

$$\pi \in \{0, 1\}, \rho \in \{0, 1\}, \nu \in \{0, 1\}, r \in \mathbb{Z}_+, \Delta \geq 0, \delta \geq 0, z \geq 0, I \geq 0 \quad (14o)$$

The objective (14a) is to minimize total net rerouting and reconsolidation costs. Note that we calculate net inventory cost changes at the growers, but only consider inventory costs after reassignment at the center. This is because the growers' inventory changes depend on the IRP solution before reassignment, whereas the center's inventory of each MMD has the same pattern regardless of when it was shipped in the subproblem solutions. (14b)-(14d) ensure that the binary rerouting variables are correctly related for a new route, i.e., pickup can occur only if a grower is visited; insertion can occur only if the grower was not visited in the original IRP subproblem solution; removal can occur only if the grower was visited. (14e)-(14f) are short-haul transportation capacity and duration constraints, i.e., the net increase of reassigned pickup volume does not exceed the residual regular vehicle capacity plus alternative capacity in each period; similarly, the net increase of duration after reassignment does not exceed the residual time plus the length of one period. (14g)-(14h) are demand satisfaction constraints redefined for each MMD, i.e., the total short-haul pickup volume equals the total long-haul delivery volume after reassignment; at any point before the product spoils, the total pickup volume to date is no less than the total delivery volume by the next period. (14i) are aggregated direct shipping capacity constraints after canceling out the courier volume, i.e., the net increase of reassigned delivery volume to a seller does not exceed the residual long-haul transportation capacity plus the extra capacity in each period. (14j)-(14l) are inventory balance constraints, i.e., the grower's inventory of an MMD in period t increases by the total later reassigned pickup volume if it was shipped by period t in the original IRP solution; the grower's inventory of an MMD in period t decreases by the total pickup volume reassigned earlier than or to period t if it was shipped after that in the original IRP solution; the central inventory of an MMD in period t equals the total pickup volume that has arrived minus the total volume that has been delivered. (14m)-(14n) and (14o) are boundary conditions and domain, respectively.

At first glance, Model (14) may appear complicated with many variables and constraints. However, it exhibits several features that enable efficient search: first, decisions for matched demand are fixed so the problem size is smaller than the full MIP; second, combinatorial rerouting costs are linearly approximated; third, complex subtour elimination constraints are circumvented with the introduction of simple binary variables. In our experiments, CPLEX almost solves it instantly compared to the subproblems.

3.4.3 An Iterative Framework

We have set up an optimization problem in hope of eliminating subproblem MMD at the lowest cost. There are caveats, though:

- We fix the subproblem decisions for matched demand before solving the re-assignment problem, which may not be the case. For instance, if a grower is removed from a tour but only a fraction of the shipments were mismatched, then the matched part should leave the tour as well and the residual vehicle capacity would be larger, but Model (14) cannot capture this.
- The changes that Model (14) does reflect may not be accurate. For instance, the effect of multiple grower rerouting is the corresponding TSP tour cost change, which generally is not the summation of that incurred by each single grower. In fact, the savings of removing a single grower may also deviate from the actual amount if the new tour is formed by simply joining the predecessor and the successor.
- The input that Model (14) inherits from the subproblem solutions may be problematic. For instance, the subproblems start with a predefined central demand/supply assignment, but the pickup/delivery times are not revealed until the solutions are out; hence the inventory calculation for fixed commodities may differ with the true values.

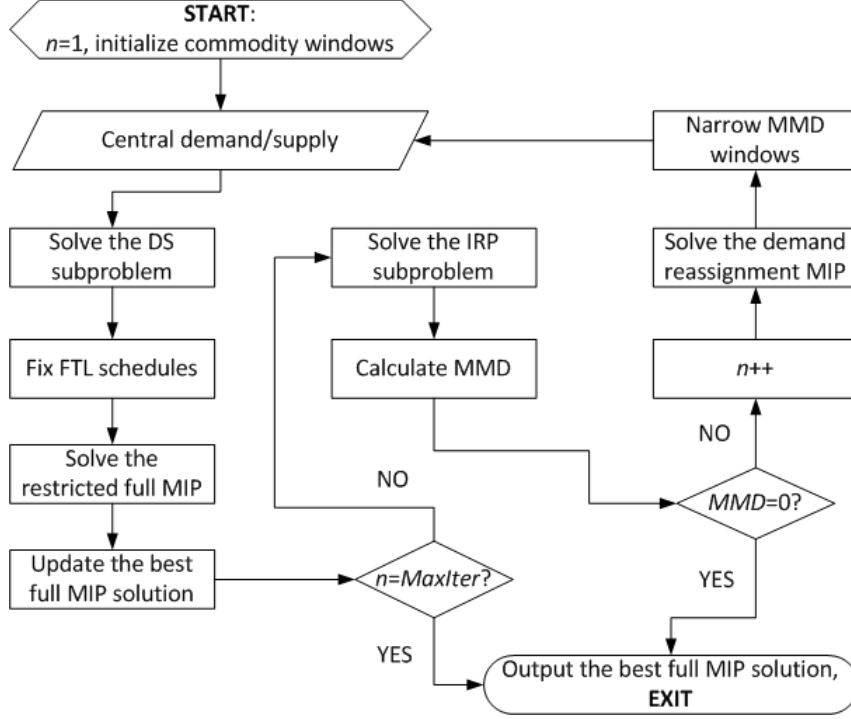


Figure 10: Flowchart of the iterative framework

For these reasons, we do not count on finding the optimal solution by solving Model (14) once. Instead, we propose to solve the full MIP by embedding the decomposition and the local search in an iterative framework (Figure 10). Throughout the process, we maintain a recorder that keeps the best full MIP solution that has been found, and a calculator that determines when to exit. The initial iteration starts with earliest possible central supply, under which we solve the DS subproblem first and the restricted full MIP subsequently while fixing the FTL schedules. If a predefined maximum allowable number of iterations is not reached, we then solve the IRP subproblem with latest possible central demand and calculate the induced MMD. When the value of MMD is positive, it invokes the reassignment MIP to trigger the next iteration. We assign the new central demand/supply by narrowing the pickup/delivery windows associated with MMD so that mismatched demands take the reassigned values whereas matched demands the initial values. At the end of each iteration, we also update the incumbent solution pool to keep track of the current best full MIP

solution. The process is repeated until the maximum number of iterations is met or MMD no longer exists, where we output the best full MIP solution and exit.

3.5 Computational Study

3.5.1 Benchmark Approaches

We have implemented our approach and two sets of benchmarks on CONDOR, the Georgia Tech ISYE computing system. All the MIPs are solved with CPLEX 12.5, but each approach involves different models and CPU time settings. The MIP emphasis parameter is set to HIDDENFEAS so that the solver searches on paths with potentially high quality feasible solutions that are otherwise very difficult to obtain.

In our approach, we solve the three subproblems derived from the full MIP as well as Model (14) in a total of 5 hours. In single iteration tests, the DS subproblem is solved to generate FTL schedules, followed by the corresponding restricted full MIP. In multiple iteration tests, the above models are solved in the same fashion, but Model (14) is called after the restricted full MIP at each iteration except the final round, and the resulting central supply triggers a new iteration. We report the best full MIP solutions found as the procedure ends. CPU time allocation for each individual MIP depends on instance sizes and number of iterations, with details described in Sections 3.5.3 and 3.5.4.

In the *CPLEX benchmarks*, we solve the full MIP in 5 hours for small instances, and report results in both 5 hours and 10 hours for medium instances.

In the *DS-guided benchmarks*, we first solve the DS subproblem assuming earliest possible central supply to obtain the long-haul shipping decisions, then calculate the central demand implied by the FTL, LTL and express volumes, and finally go to the IRP subproblem. We report the feasible full MIP solutions formed by the subproblem solutions. The CPU time limit is 5 hours in total, with the DS subproblem in 1 hour, and the IRP subproblem in 4 hours.

3.5.2 Test Instances

We construct test instances based on real data from the California cut flower industry for the year 2010. Our initial trials indicated that problem difficulty increases most significantly as the number of growers increases, and CPLEX is unable to find non-trivial upper bounds for the full MIP in 72 hours when there are 20 growers. Hence we distinguish instance classes by the size of the grower set.

- *Small instances:* $|G| = 10$, $|D| = 5$, $T = 15$, $\theta = 3$, $h_i = 1 \forall i \in G$.
- *Medium instances:* $|G| = 15$, $|D| = 5$, $T = 15$, $\theta = 3$, $h_i = 1 \forall i \in G$.

Within each class, we create eight inventory cost combinations where $h_0/h_i \in [0, 4]$ to examine the impact of central inventory cost on routing and consolidation decisions. Since the real-world demand is seasonal and cyclic, we further develop empirical demand distributions as in [95], and test the approaches in a simulation environment. Each grower-retailer pair follows a unique distribution pattern (Figure 11), with peak periods and nonpeak periods characterized separately. We generate 10 peak samples and 10 nonpeak samples for each pair.

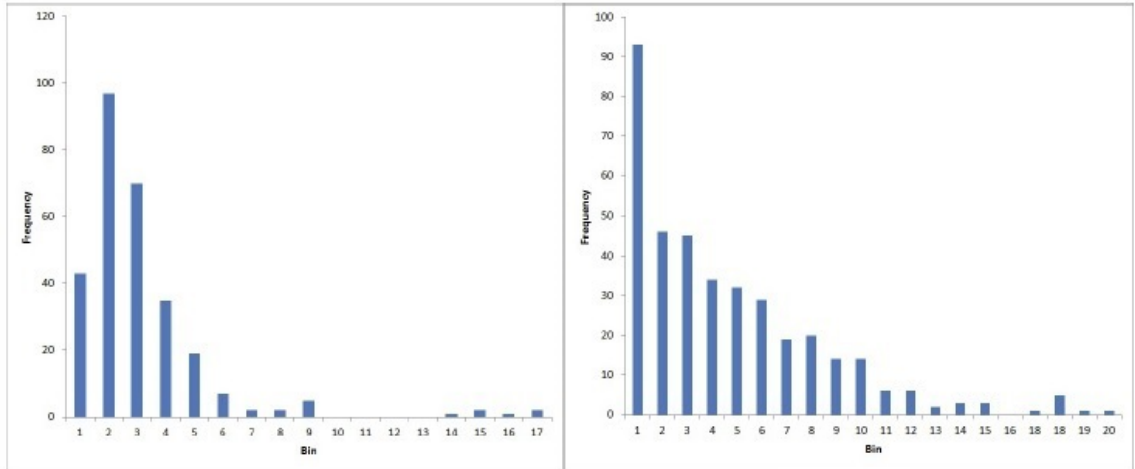


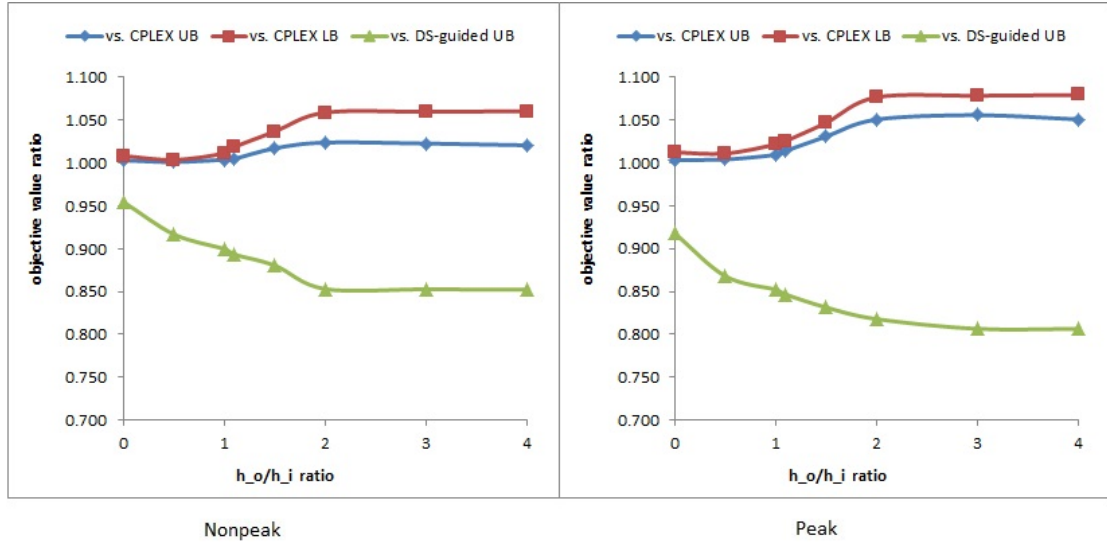
Figure 11: Demand distribution for two grower-retailer pairs

3.5.3 Single Iteration Results

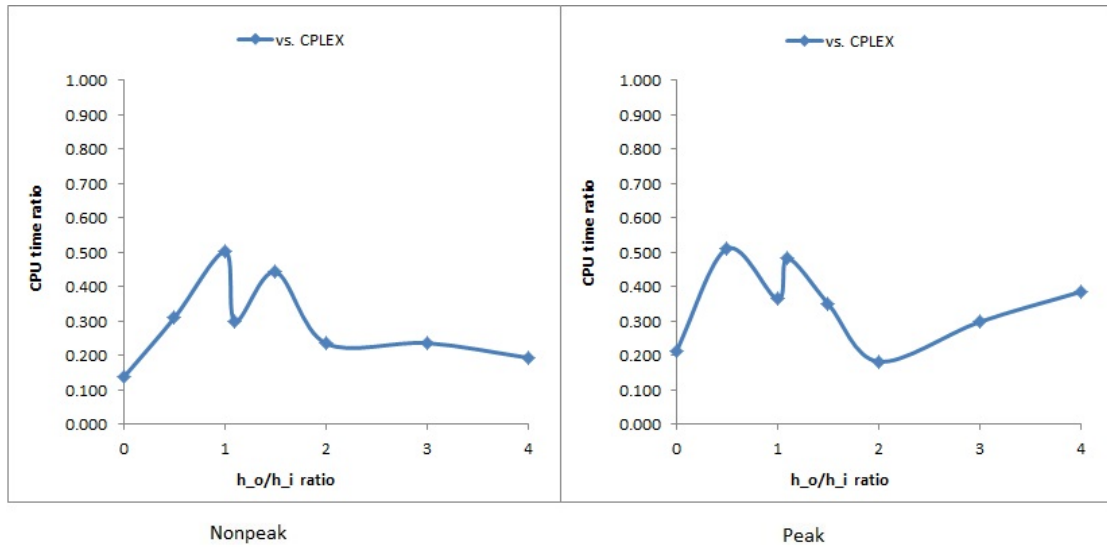
We run a single iteration of our approach with the DS subproblem in 1 hour and the restricted full MIP in 4 hours. Since the reassignment MIP is not employed, this allows us to examine potential behavior of the decomposition. Figure 12 and Figure 13 show the average performance in the small instances and the medium instances, respectively.

In 132 out of the 160 samples we have tested for the small instances, the CPLEX approach fully utilized the CPU, whereas our decomposition solved both subproblems to optimality in 35% of the total time limit on average, and only required 10% of the CPLEX CPU in 77% samples. For both nonpeak and peak demand, our decomposition found solutions within 8% gap of the lower bounds and 6% of the upper bounds CPLEX obtained. The DS-guided approach was even faster, but could only find solutions at least 15% worse than the CPLEX upper bounds.

The solution quality of our decomposition and the DS-guided approach were more affected as the center-grower inventory cost ratio increased, though the final solutions became stable when the ratio was sufficiently large. This can be explained by the critical role of central inventory in DS-IRP trade-offs. When the center inventory cost is very low, both DS and IRP benefit from storage at the center which allows for consolidation opportunities and grower inventory savings. Given the dominance of long-haul shipping costs, the “naive” DS-guided approach is well justified since it strongly controls the major costs without over sacrificing the others. As h_0/h_i increases, however, the balance between long-haul and local shipping costs interacts with that between inventory and transportation in a more complicated way, and thus the approach should be judicious in neighborhood selection to avoid extremes. Both our decomposition and the DS-guided approach prioritize long-haul costs, but by fixing only FTL schedules we allow for flexibility in other trade-offs, so it is not surprising that the DS-guided approach significantly deviates from CPLEX while our



(a) Solution quality

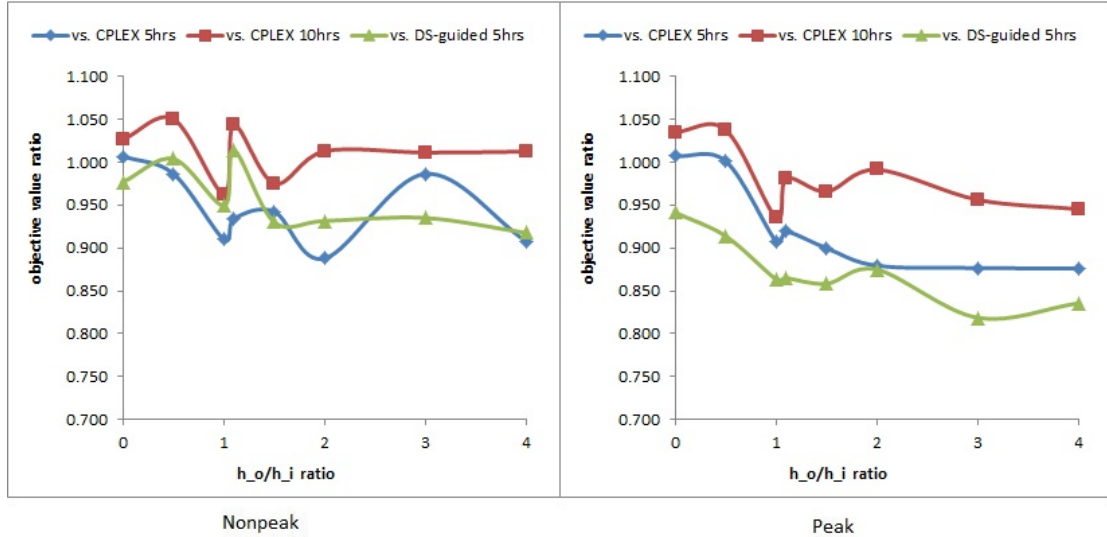


(b) CPU time

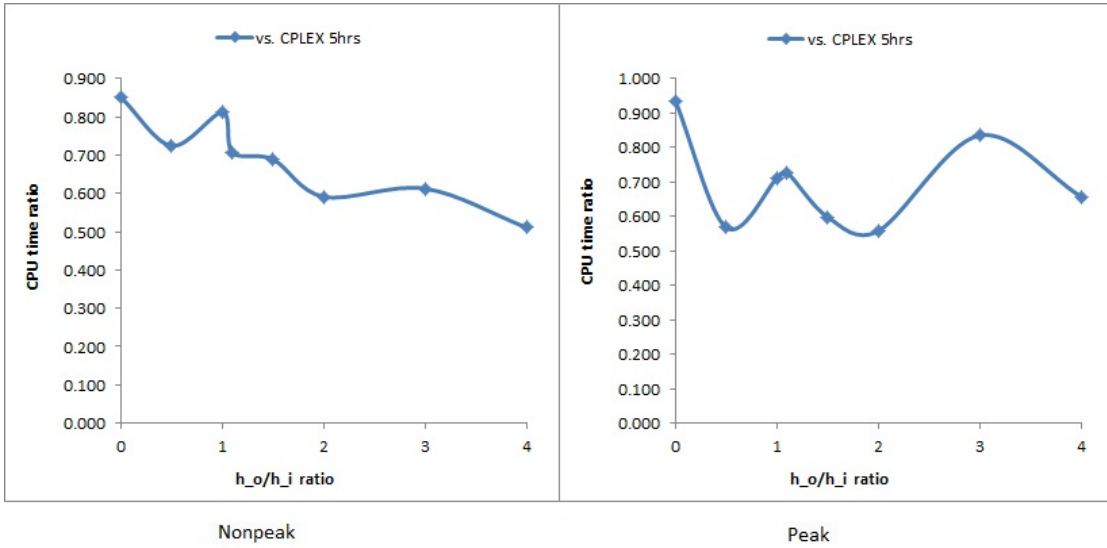
Figure 12: Average performance of a single iteration for small instances

approach tends to be robust. Finally, when the center inventory cost is very high, the center acts like a crossdock and the global optima heavily depend on the coordination between DS and IRP. Both our decomposition and the DS-guided approach provided solutions without central storage, but failed to capture the best such coordination.

The results for medium instances are more interesting. CPLEX could not find good lower bounds in any sample, so we report CPLEX upper bounds in 10 hours as



(a) Solution quality



(b) CPU time

Figure 13: Average performance of a single iteration for medium instances

an alternative. For both nonpeak and peak demand, our approach and the DS-guided approach were closer to CPLEX than before. In fact, our solutions beat CPLEX 5-hour upper bounds in 115 out of the 160 samples with respect to both quality and efficiency. On average, our approach found solutions over 5% better than CPLEX in 5 hours except three instances where it did less than 1% worse. Furthermore, our solutions in around 3.5 hours generally matched up to CPLEX in 10 hours. As for

sensitivity to demand distribution and cost parameters, our decomposition turned out more favorable for peak demand and when h_0/h_i was high. DS-guided still performed worst for peak demand, but could slightly beat CPLEX in 5 hours and even our decomposition for nonpeak demand. On one hand, this demonstrates that model-based decompositions are promising in handling larger instances of the problem. On the other hand, it implies that we may easily get stuck in bad local optima for more complicated systems, which motivates our next experiments.

3.5.4 Multiple Iteration Results

We select samples where a single iteration of our approach was over 5% worse than CPLEX or the average in Figure 12 and Figure 13 to test the effectiveness of our approach with multiple iterations. For each sample, we run 5 iterations with the DS subproblem in 1 hour, the IRP subproblem in 2 hours, and the restricted full MIP in 2 hours per iteration. The DS subproblem time limit is the same as in single iteration tests to ensure the first iteration replicates the FTL schedules before. We report the best solution as the procedure ends, and the total time actually spent in finding the best upper bound per iteration.

Table 3: Multiple iteration statistics

Instance class Demand type	Small Nonpeak	Small Peak	Medium Nonpeak	Medium Peak
# Samples	15	19	20	14
Avg. 1-iter. UB ratio	1.028	1.071	1.070	1.074
Avg. 1-iter. CPU ratio	0.075	0.158	0.230	0.263
Avg. m-iter. UB ratio	1.005	1.008	0.983	1.000
Avg. m-iter. CPU ratio	0.340	0.350	0.620	0.580
Avg. # iter.	2.1	2.3	2.8	2.7

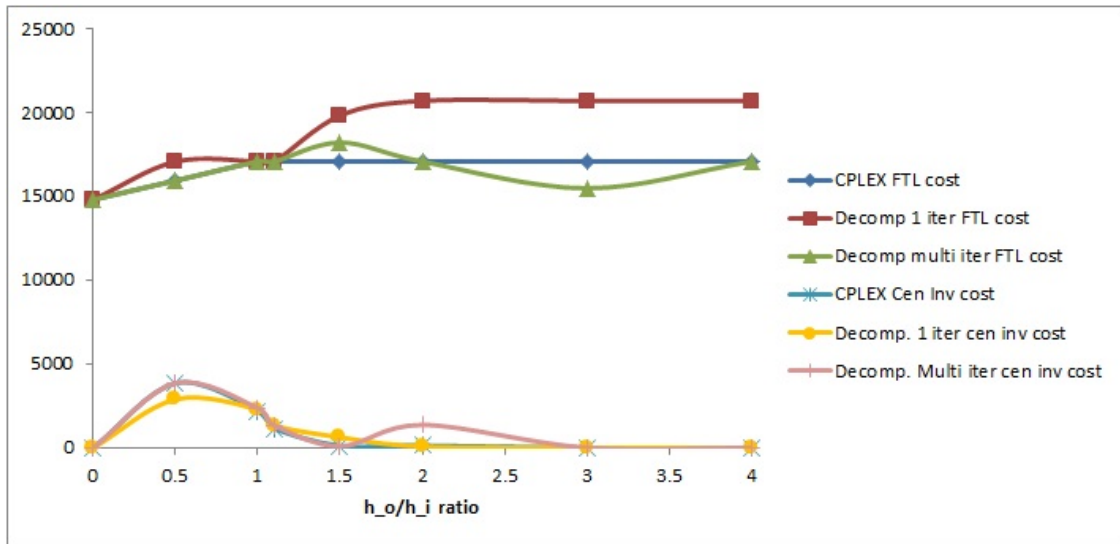
Benchmarks: CPLEX 5 hours for small instances, CPLEX 10 hours for medium instances

Table 3 lists a comparison of the average results among the chosen samples. With multiple iterations, our approach quickly found solutions that matched CPLEX upper

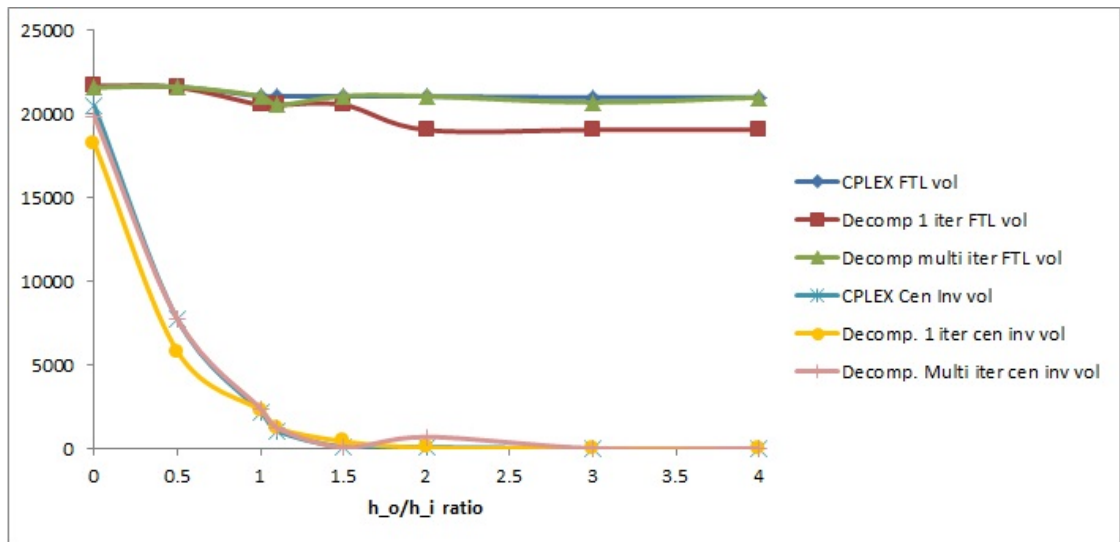
bounds for small instances, and reduced the objective value from 7% above CPLEX 10 hours to zero gap or below for medium instances.

3.5.5 Discussion

It is interesting to see that the best solutions were generally obtained in fewer iterations for the small samples. A scrutiny of detailed solutions reveals various possible reasons that motivate future work. For illustrative purpose we discuss two cases here.



(a) FTL and center inventory costs



(b) FTL and center inventory volumes

Figure 14: Cost and volume breakdown for a small-sized sample

Figure 14 shows the cost and volume breakdown of a small sample where our approach reduced the upper bound difference from over 10% to below 1% in two iterations for $h_0/h_i = 2, 3, 4$. The center inventory costs and volumes were close for all the three approaches under all cost parameters, whereas the single iteration solutions shipped less volume with more FTL trucks in these settings. This indicates that utilization of the FTL capacity can be a key index for solution quality, and Model (14) essentially reallocated the direct shipping volume to realize higher utilization. It is not generally true, however. Figure 15 shows the per iteration breakdown for a medium sample where our approach reduced the upper bound ratio from 1.05 to 0.80 in four iterations, whereas the FTL volume-cost ratio went from 1.58 to 1.55 and fluctuated between 1.40 and 1.58. An explanation is that the solutions in the first and the third iterations used expensive direct shipping alternatives for local pick-up. Consequently Model (14) revised routing decisions before triggering the second and the fourth iterations. Since Model (14) is developed with estimated routing parameters, the resulting neighborhoods may deviate more from the true global optimum than in the previous case, and thus more iterations would be desirable. In any case, we expect a better understanding of systemwise tradeoffs to help strengthen our methodology.

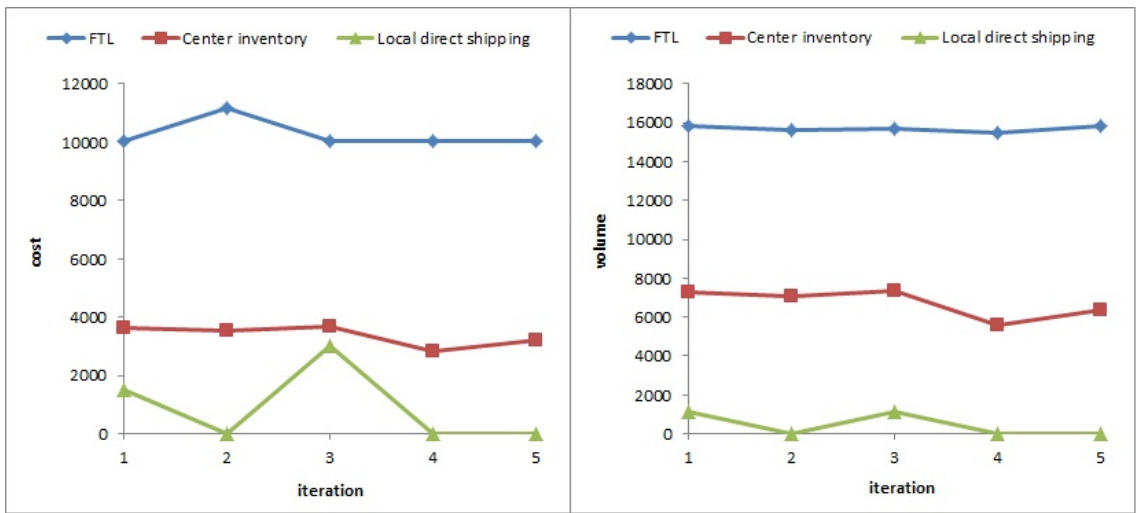


Figure 15: Per iteration cost and volume breakdown for a medium-sized sample

3.6 Conclusions

This chapter summarizes our work on a novel integrated inventory routing and freight consolidation problem. We have built a large-scale MIP model for the problem, and developed an iterative solution framework consisting of a decomposition procedure and an optimization-based local search scheme. Extensive experiments with small to medium sized problem instances demonstrate the effectiveness of both solution approaches, and direct us in possible extensions.

Currently, we are conducting experiments to show the scalability of the proposed framework. As previously mentioned, we do not need to solve each subproblem to optimality in multiple iterations since the decisions interact with each other and vary from iteration to iteration. In light of this, we can obtain fast solutions by further decomposing the IRP subproblem and the restricted full MIP subproblem: 1) determine shipping quantities, inventory levels and tentative routes with approximated routing costs; 2) calculate the actual routing costs by solving a TSP with the predetermined growers to visit in each period. The justification is that with an appropriate usage of CPU time, the final integrated solution is hopefully better than the decomposition could gain with a brute force implementation of CPLEX algorithms. Therefore, we are applying naive routing cost approximations to larger problem instances, and expect to see that despite the simplification, on average the iterative framework can still generate superior solutions to the standard DS-guided approach and the best CPLEX upper bounds given the same CPU time.

We will further investigate the performance of the standard approach and a seemingly counterintuitive approach for very large problem instances. Enlightened by the work in Chapter 4, we can prove that as the number of growers goes to infinity, it is asymptotically optimal to first solve the short-haul IRP, and then use the pickup quantities as the supply of the long-haul consolidation decision. However, this advantage is heavily offset by the computational burden with large-scale IRPs, and we

observe that the standard approach can yield more benefits with the routing cost approximation for the IRP subproblem.

We note that variants of Model (14) can be used for diversification mechanisms in the iterative framework. The presented formulation aims to minimize net reassignment cost of mismatched commodities. As discussed in Section 3.5.5, other metrics e.g. FTL utilization and percentage of alternative local direct shipping, may also be key performance indices of the local search scheme. Hence it is worth searching the solution space under their guidance, e.g. by incorporating them in the objective function (14a), considering them for matched commodities, etc. Moreover, slight revision of constraints (14g)-(14h), e.g. summing the shipping quantities over a different time interval, directly results in alternative reassignment strategies. Therefore, we intend to explore larger neighborhoods and provide more valuable information for future decomposition. This is the flexibility we may enjoy with the optimization-based local search approach.

We make a final comment on the stopping criteria in the iterative framework. While the current experiments use the maximum number of iterations for ease of implementation, other appealing candidates include the ratio of MMD to total demand, the gap between the incumbent best solution and a theoretical lower bound (similar to that in Proposition 4.5.6), etc.

CHAPTER IV

THE “REVERSED” ONE-WAREHOUSE MULTI-RETAILER PROBLEM WITH MIXED TRUCKLOAD COSTS

4.1 *Introduction*

In this chapter, we study another two-echelon distribution problem which also consists of a short-haul echelon from multiple growers to a consolidation center, and a long-haul echelon from the consolidation center to each retailer. Unlike the previous chapter, we do not allow local routing among the growers, but assume mixed direct shipping options for both echelons. The situation naturally arises when the growers operate independently, e.g. each has its own fleet or uses dedicated resources via a strategic contract with a transportation provider. To simplify analysis, we consider one retailer, a non-perishable product, as well as static cost parameters in the basic version, and extend the results to more complicated settings when they are applicable.

The basic problem is stated as follows. A set of growers (denoted by G) ship a single product to a common retailer via a consolidation center (denoted by 0) in the network depicted in Figure 16. Each grower’s demand ($d_i^t, \forall i \in G, t = 1 \dots T$) is deterministic and known over a discretized finite horizon of length T . Both local pickup and final delivery are conducted by volume-dependent services including fixed *full-truckload* (FTL) rates, fixed *less-than-truckload* (LTL) rates, and linear courier rates. The transportation capacity is the same per FTL or LTL at all facilities, whereas the associated costs vary from location to location. We assume zero transportation lead times without loss of generality. The growers and the center may keep inventory to leverage economies of scale in the transportation cost, and each unit of

inventory incurs a facility-specific holding cost. All demands must be fulfilled without backorder.

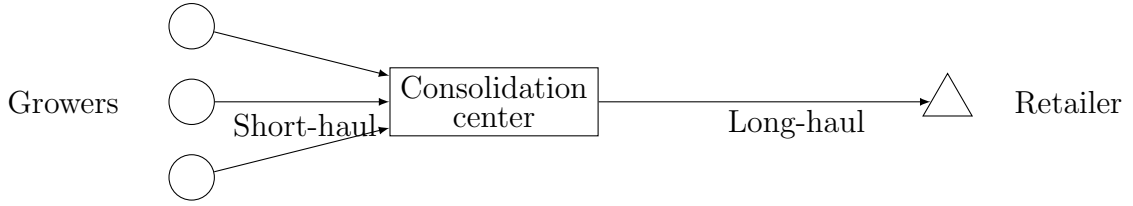


Figure 16: The distribution network: “reversed” OWMR

The problem can be viewed as a *one-warehouse multi-retailer* (OWMR) system reversed in time flows. In a conventional OWMR problem, a warehouse orders a commodity from an external supplier and then serves the demand of distinct retailers, assuming direct and instant transportation at both echelons. When the warehouse operates as a cross-docking station, the OWMR problem becomes the *joint replenishment problem* (JRP) where a major setup cost is incurred each time a subset of retailers order in addition to separable minor setup costs for each of them. Thus, a grower’s demand ready at period t in our problem corresponds to the same volume for a retailer due by period $(T - t)$ in a conventional OWMR problem, or a conventional JRP problem when the central holding cost is prohibitively high. To clarify, we will substitute *retailer* with *grower*, *warehouse* with *consolidation center*, and *supplier* with *retailer* when interpreting conventional OWMR results. Also, we will interchangeably refer to *ordering cost*, *transportation cost* and *setup cost* as the same cost element for a problem in various contexts.

The contribution of our study is that we consider more realistic transportation costs rather than the fixed setup costs in the majority of the literature. Specifically, we analyze various combinations of the aforementioned shipping options (Figure 17). Let K_F, K_L be the FTL and LTL capacities, and c_F, c_L, c_U the costs per truck, per LTL unit and per courier unit (facility subscript omitted), respectively. In the case of FTL costs, only trucks are dispatched, any truck costs c_F and carries a volume no

higher than K_F . In the case of FTL+stepwise LTL costs, both trucks and LTL units can be dispatched, each truck costs c_F and each LTL unit costs c_L when the volume is within the corresponding capacity limits. In the case of FTL+linearized LTL costs, again both trucks and LTL units are allowed, each truck costs c_F if its capacity is not violated, whereas the cost of an LTL shipment is linear in the volume and increases at a rate of $\frac{c_L}{K_L}$. Finally in the case of FTL+LTL+courier costs, all three shipping options are allowed, each truck costs c_F and each LTL unit costs c_L , whereas the cost of a courier shipment linearly increases at the rate of c_U . The cost breakpoints determine the volume where the cheapest option shifts when two or three services are involved. In Figure 17(b)-(c), for instance, a shipment uses LTL until its volume reaches an FTL-LTL breakpoint $b_F = \frac{c_F}{c_L}$. Similarly, a shipment that uses courier for very small volume switches to LTL at an LTL-courier breakpoint $b_L = \frac{c_L}{c_U}$, and further changes to FTL at some FTL-LTL breakpoint b_F in Figure 17(d). Note, however, that the calculation of b_F varies. Henceforth, we will call our problem *the “reversed” OWMR problem with mixed truckload costs*, while noting that our methodologies can be applied to conventional OWMR problems with the same cost structures, and vice versa.

The problem is clearly \mathcal{NP} -hard since the JRP is strongly \mathcal{NP} -hard under discrete-time dynamic demand, a special case of an OWMR instance in [19], which further reduces to our problem with FTL costs when the truck capacity exceeds the total demand. Our main contribution is to provide theoretical results from two perspectives of mixed truckload costs: first, special piecewise linear (PWL) non-convex functions on \mathbb{R}_+ ; second, *soft-capacitated* transportation with multiple modes, i.e., each mode may impose a capacity constraint but there is no limit on the total shipping volume allowed in a period. The former perspective links this class of costs with the literature, whereas the latter allows us to directly analyze the special case of FTL costs and asymptotic behavior of the system under mild assumptions.

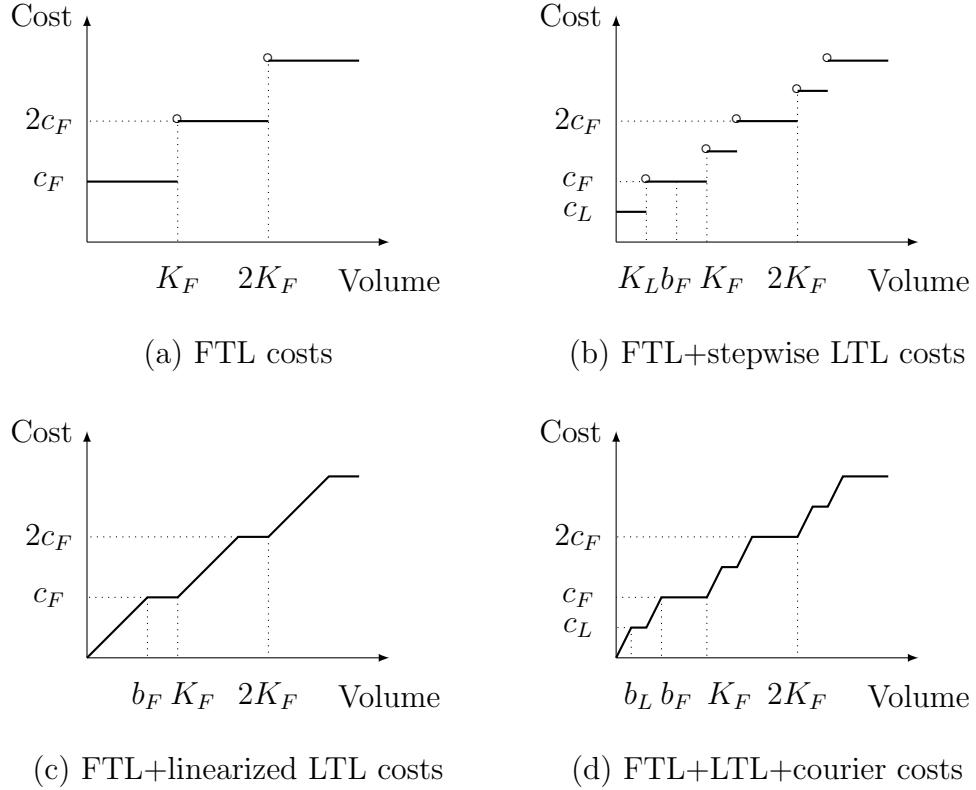


Figure 17: Mixed truckload costs

The remainder of the chapter is organized as follows. We briefly review the literature in Section 4.2. We formulate the problem and show that classical OWMR solution approaches may fail in Section 4.3. We present a technique that converts our problem with several cost structures to existing models and derive the resulting approximation guarantees in Section 4.4. We then decompose the problem into single-echelon subproblems in Section 4.5, and propose a combinatorial algorithm that generates an improved approximation for FTL costs in Section 4.6. The subproblem structural properties motivate us to further analyze asymptotic optimality of a decentralized approach for mixed truckload costs and more general batch ordering costs in Section 4.7. We extend some results in Section 4.8, and conclude the chapter with potential future research in Section 4.9.

4.2 Literature Review

The study of inventory models with deterministic dynamic demand dates back to [127], who introduced an efficient *dynamic programming* (DP) algorithm to solve an uncapacitated *single-item lot sizing problem* (LSP) with fixed setup costs and linear holding costs. Since then, the importance of this problem has been widely recognized in both academia and practice. See [73, 39, 37] for comprehensive surveys on numerous variants and extensions.

The OWMR problem and its special case the JRP problem are two-echelon generalizations of the LSP and have received increasing attention in recent years. Since both problems are well-known as strongly \mathcal{NP} -hard, much research interest centers on heuristics including those with theoretical solution guarantees. We explain some terminology before going to the details. An algorithm is said to be ϵ -optimal or a *polynomial-time approximation scheme* (PTAS) of a minimization problem if for any instance and a given parameter $\epsilon > 0$, it produces a solution within a factor of $(1 + \epsilon)$ from optimality in polynomial time of the input size and $1/\epsilon$. A polynomial-time approximation scheme is efficient (EPTAS) if the exponent of the input size parameters is independent of $1/\epsilon$ in the polynomial. An algorithm is said to be β -optimal or a β -approximation ($\beta \geq 1$) if for any instance of a minimization problem, it generates a solution with a cost no higher than β times the optimal cost in polynomial time of the input size. It is a constant approximation if the ratio β is a constant.

In a standard setting of OWMR/JRP problems, transportation, production or procurement is uncapacitated, the shipping, setup or ordering cost is fixed, whereas the inventory cost is non-speculative, linear in quantities held and additive over periods. Standard OWMR results largely rely on the dominance of *zero-inventory-ordering* (ZIO) policies, i.e., there always exists an optimal solution such that each demand is fully satisfied by a certain order. [58] proposed a DP-based heuristic for the OWMR and demonstrated ϵ -optimality as well as asymptotic optimality assuming bounded

cost parameters and demands. [44] proved that the OWMR problem with time-varying growers' ordering costs is as hard as the set cover problem, which is unlikely to be polynomial-time approximable within a sublogarithmic factor of optimality. Therefore, OWMR approximation algorithms usually assume static ordering costs at the growers. [89] relaxed the standard holding cost assumptions by redefining it per (demand, order times) tuple and imposing the *Monge property* which basically favors *first-in-first-out* (FIFO) inventory depletion. They formulated a *mixed integer linear program* (MIP) for the resulting OWMR, and gave the first constant approximation which achieves a worst-case bound of 1.8-optimality based on *linear programming* (LP) relaxation and rounding techniques. [119] considered the same definition, and slightly relaxed [89]'s holding cost structure to the so-called *metric holding costs* which replace the Monge property with the *triangle inequality*. They also built a similar MIP model of the OWMR problem, but developed a simpler 2-approximation by recombining decomposed single-echelon solutions, and extended the results to some non-linear holding costs. [68] applied a revision of the technique to the case where demand shortages are allowed, and obtained a 3-approximation for the OWMR problem with backlog, a 2-approximation for the JRP with backlog and a 2-approximation for the OWMR problem with lost-sales, respectively. The techniques in [119] and [68] are somewhat similar to the approach in Section 4.6, but our subproblems and recombination procedure are necessarily more complicated because of the transportation cost structures.

The classical JRP literature focus on stationary demand ([81]), and the few approximation algorithms under dynamic demand all use randomized rounding techniques with a natural LP relaxation of the problem. The best-known upper bounds in the finite horizon setting are a PTAS in [112] and an EPTAS in [97] for stationary demand, a 1.791-approximation in [31] for dynamic demand, and a 1.574-approximation in [30] for a dynamic demand variant *JRP-D* which considers no holding cost but

associates a deadline with each order. Another stream of standard OWMR/JRP research is the infinite horizon counterparts of both problems, where demands occur at constant rates on a continuous time basis. [107]’s *power-of-two policies* gain 98% effectiveness (i.e. a 1.02-approximation) for these variants, and [118] attained a 1.275-approximation by recombining single-echelon *economic order quantity* (EOQ) formulae. An open question is the computational difficulty with stationary demand, which [48] partially answered by showing that the *periodic JRP* (PJRP), where the ordering cycles of the commodities are required to be integral multiples of the joint order cycle time, is strongly \mathcal{NP} -hard for both finite and infinite horizons.

Fundamental properties like ZIO can be easily violated for OWMR/JRP problems with general ordering costs, and therefore the computational difficulty slows research in this area. Nevertheless, an interesting application has been cost discounts, which carriers or suppliers usually offer as an incentive for large shipments or ordering quantities. According to [93], three types of cost discount schemes are common in practice: the *all-unit discounts* where different rates apply to all units of different ordering quantities; the *incremental discounts* where different rates apply to incremental ranges of the total ordering quantity; and the *truckload discounts* where an LTL rate is charged linearly with respect to the ordering quantities until a threshold is reached when the customer is willing to pay for the cost of an FTL rate, and the pattern repeats itself once a truck is full so the next truck is loaded. All three types result in PWL costs and have been investigated for OWMR problems to some extent under finite-horizon dynamic demand. [89] proposed a primal-dual method, which together with their heuristic for the standard setting implies a 3.6-approximation for the OWMR problem with static FTL costs. Our transformation technique in Section 4.4 matches this worst-case ratio, which is further improved in Section 4.6. When the holding cost rate is lower at the center than the growers, [79] developed an exponential-sized DP as well as an iterative Lagrangian heuristic, and proved asymptotic optimality of a

decentralized solution as the number of growers goes to infinity. The DP does not apply here since we relax their holding cost assumptions, but the asymptotic results are extended to more general cost structures in Section 4.7. [44] studied an OWMR problem where each grower is associated with a *modified all-unit discount* cost function which alternates sections with positive slopes and flat sections. Assuming incremental discounts at the center, they proved that the effectiveness of ZIO policies can be as high as $\frac{4}{3}$ or $\frac{5.6}{4.6}$ under time-varying and static costs, respectively. They also designed non-polynomial DP and LP-based heuristics that output optimal or near-optimal ZIO policies. [115] studied a JRP problem with a class of PWL costs which generalize short-haul modified all-unit discounts but maintain long-haul incremental discounts. They converted the problem by a factor of $(1 + \epsilon)$ to a large-scale PWL concave cost multi-commodity network flow problem, which is further approximable to logarithmic optimality in polynomial time.

Our mixed truckload costs differ from all-unit discounts and incremental discounts in that they can be neither diminishing in unit cost nor concave on \mathbb{R}_+ . We give examples of relevant cost structures in Figure 18(a)-(c), and introduce in Figure 18(d) a new class of ordering costs, the *near-concave batch costs*, which refers to a cost structure approximable by concave batch costs to a sublinear ratio in $|G|$ with the technique in Section 4.4. We will see that this is a more general concept than mixed truckload costs and concave batch ordering costs. Table 4 compares our results to the best known in the OWMR literature.

Since we decompose the “reversed” OWMR problem to single-echelon subproblems in Section 4.5, some lot sizing problems are also closely related. [90], [59] and [57] discussed the solvability of concave batch-ordering costs, concave costs under *hard* capacities (which impose a limit on the total volume allowed in each period), and quantity discounts, respectively. For FTL costs, in particular, [90] gave a DP that runs in $\mathcal{O}(T^3)$, [100] studied formulations and valid inequalities, whereas [79] improved

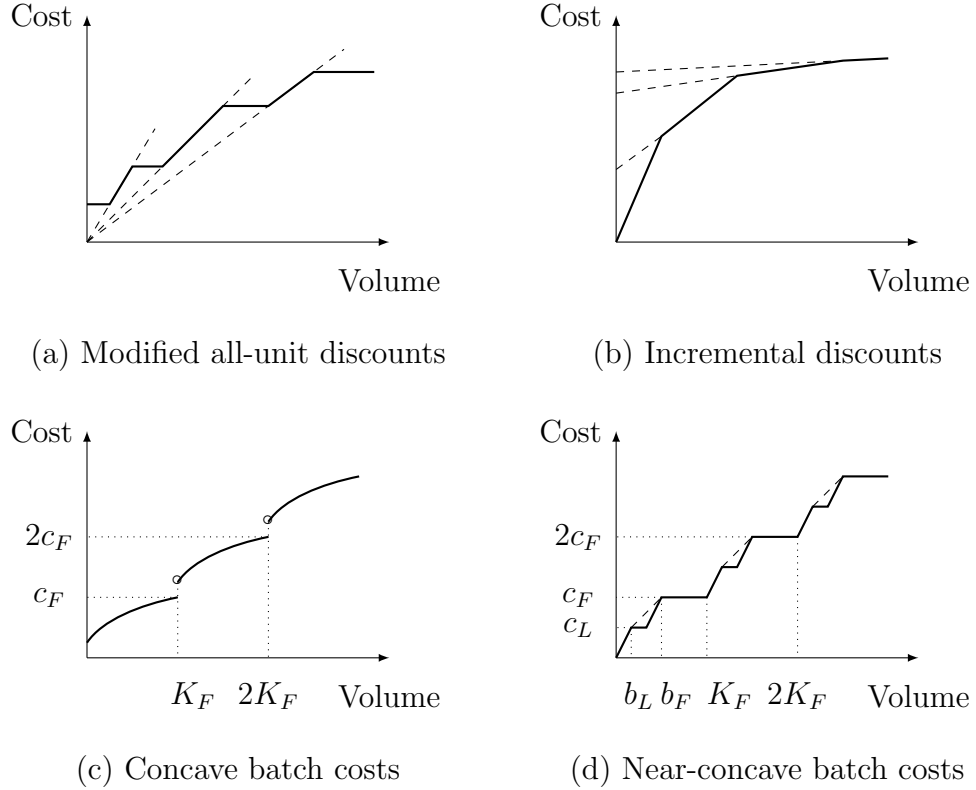


Figure 18: Relevant cost structures

the complexity to $\mathcal{O}(T^2)$ with a DP that utilizes structural properties of the problem. For modified all-unit discounts, [43] proved that the LSP is \mathcal{NP} -hard when either the cost function is time dependent or the number of breakpoints is not bounded by a constant, and obtained the same efficacy ratios of ZIO policies as [44]. [18] identified polynomial cases of modified all-unit discounts and incremental discounts. LSP with general PWL and piecewise concave production costs can be found in [114] and [82], respectively.

Multi-item lot sizing models involve the joint decision of multiple items using the same resource. In these problems, a joint setup cost is incurred each time a subset of items order, whereas the holding cost depends on which items are ordered and how much the quantities are. [88] proved strong \mathcal{NP} -hardness and developed a 2-approximation under hard ordering capacities and time-varying linear holding costs. [14] proposed a DP to solve a *multi-item multi-vehicle* (MIMV) model, which

Table 4: Comparison of our study to the OWMR literature

Papers	Ordering cost structure	Results
[89]	FTL	3.6-approximation
[79]	FTL	Asymptotic optimal decentralization
[44]	Modified all-unit discounts	$\frac{4}{3}$ -optimality of ZIO policies
[115]	A class of PWL costs, JRP	$((1 + \epsilon) \log n)$ -approximation
Our study	FTL	2-approximation
	A modified all-unit discounts	4-approximation
	Short-haul FTL+stepwise LTL	$\frac{8}{3}$ -optimality of ZIO policies
	Mixed truckload costs, JRP	$((1 + \epsilon) \log n)$ -approximation
	(Near-)Concave batch costs	Asymptotic optimal decentralization

captures a multi-item lot sizing problem with static FTL costs and indexed items such that each item has higher holding costs than all the items of smaller indices in all periods. The complexity of the DP is polynomial if the number of items is fixed, but exponential otherwise. [15] proposed techniques to enhance the DP and design fast heuristics for large problem sizes, and demonstrated empirical efficiency as well as effectiveness of both algorithms. Later on, [16] studied the convex hull of this MIMV model under time-varying costs, and gave an equivalent description with a system of linear inequalities.

4.3 Preliminaries

In this section, we present two formulations of the “reversed” OWMR problem with mixed truckload costs, and demonstrate possible inefficacy of previous OWMR solution methods. The first formulation is attained by treating it as a general “reversed” OWMR problem, which also applies to other cost structures and can be modeled as follows.

$$(R\text{-OWMR}) \quad \min \sum_{t=1}^T \sum_{i \in G \cup \{0\}} c_i(q_i^t) + \sum_{t=1}^T \sum_{i \in G \cup \{0\}} h_i(I_i^t) \quad (15a)$$

$$\text{s.t. } I_i^t + q_i^t - I_i^{t-1} = d_i^t, \quad \forall i \in G, t = 1 \dots T \quad (15b)$$

$$I_0^t + q_0^t - I_0^{t-1} - \sum_{i \in G} q_i^t = 0, \quad \forall t = 1 \dots T \quad (15c)$$

$$I_i^0 = 0, \quad \forall i \in G \cup \{0\} \quad (15d)$$

$$I_i^T = 0, \quad \forall i \in G \cup \{0\} \quad (15e)$$

$$q \geq 0, I \geq 0, \quad (15f)$$

where I and q are inventory and shipping quantities, respectively. Functions $c_i(\cdot), h_i(\cdot) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are the transportation and inventory costs of interest. The objective (15a) is to minimize total distribution cost of all the demand over the planning horizon. Constraints (15b)-(15c), (15d)-(15e), (15f) specify inventory balance, inventory boundary condition and domain, respectively.

To explicitly capture the volume-dependent mixed truckload costs, we introduce integer variables representing FTL numbers or LTL units dispatched in each period, and redefine shipping quantities accordingly, which gives an alternative MIP formulation:

$$\min \sum_{t=1}^T \sum_{i \in G \cup \{0\}} \sum_{p \in \{F, L\}} c_{i,p} r_{i,p}^t + \sum_{t=1}^T \sum_{i \in G \cup \{0\}} c_{i,U} q_{i,U}^t + \sum_{t=1}^T \sum_{i \in G \cup \{0\}} h_i I_i^t \quad (16a)$$

$$\text{s.t. } I_i^t - I_i^{t-1} + \sum_{p \in \{F, L, U\}} q_{i,p}^t = d_i^t, \quad \forall i \in G, t = 1 \dots T \quad (16b)$$

$$I_0^t + \sum_{p \in \{F, L, U\}} q_{0,p}^t - I_0^{t-1} - \sum_{i \in G} \sum_{p \in \{F, L, U\}} q_{i,p}^t = 0, \quad \forall t = 1 \dots T \quad (16c)$$

$$q_{i,p}^t - K_p r_{i,p}^t \leq 0, \quad \forall i \in G \cup \{0\}, p \in \{F, L\}, t = 1 \dots T \quad (16d)$$

$$I_i^0 = 0, \quad \forall i \in G \cup \{0\} \quad (16e)$$

$$I_i^T = 0, \quad \forall i \in G \cup \{0\} \quad (16f)$$

$$q \geq 0, I \geq 0, r \in \mathbb{Z}_+, \quad (16g)$$

where $c_{i,F}, c_{i,L}, c_{i,U}$ are the transportation cost per FTL, per LTL, and per courier unit, respectively, at each facility. K are the FTL/LTL capacity parameters. The newly added r variables are the dispatched FTL/LTL numbers, and (16d) are transportation capacity constraints.

As mentioned in Section 4.2, many existing inventory models count on the dominance of ZIO policies when the cost functions are concave over \mathbb{R}_+ . Another fact is that the standard OWMR problem reduces to JRP as long as $0 \leq h_i \leq h_0 \forall i \in G$. Unfortunately, both properties may not hold in our case.

Example 7 (Suboptimality of ZIO, an extension, and cross-docking). Table 5 lists a few numerical examples for FTL costs. Columns short-haul and long-haul refer to the shipping quantities. Given $|G| = 1$ and $h_0 > c_{0,F}$, Instance 1 reduces to an LSP with batch ordering cost $c_F = c_{0,F} + c_{1,F}$. The best ZIO policy is then found with DP. For Instance 2 and Instance 3, the best solutions under the restrictive ordering rules are found by enumeration.

Table 5: “Reversed” OWMR-FTL instances

Common settings: $K_F = 10, c_{i,F} = 10 \forall i \in G, c_{0,F} = 20$							
Instance 1: $ G = 1, T = 4$							
h_i	d_i^t	Best ZIO solution			A better non-ZIO solution		
$h_0 = 40$		short-haul	long-haul	obj.	short-haul	long-haul	obj.
$h_1 = 2$	(4,4,6,6)	(0,8,6,6)	(0,8,6,6)	98	(0,0,10,10)	(0,0,10,10)	92
Instance 2: $ G = 2, T = 2$							
h_i	d_i^t	Best RPP solution			A better non-RPP solution		
$h_0 = 40$		short-haul	long-haul	obj.	short-haul	long-haul	obj.
$h_1 = 1$	(4,6)	(0,10)	(0,10)		(3,7)	(3,7)	
$h_2 = 3$	(7,3)	(0,10)	(0,10)	85	(7,3)	(7,3)	81
Instance 3: $ G = 3, T = 2$							
h_i	d_i^t	Best JRP solution			A better non-JRP solution		
$h_0 = 4$		short-haul	long-haul	obj.	short-haul	long-haul	obj.
$h_1 = 1$	(4,0)	(3,1)	(3,1)		(4,0)	(4,0)	
$h_2 = 3$	(7,0)	(7,0)	(7,0)		(7,0)	(6,1)	
$h_3 = 1$	(0,9)	(0,9)	(0,9)	81	(0,9)	(0,9)	74

Instance 1 shows that ZIO is not necessarily optimal for our problem.

As an extension of ZIO to soft transportation capacities, the *regeneration point property* (RPP), i.e. $(q^t \bmod K)I^t = 0$ (subscripts omitted), has been introduced to both single-echelon LSP e.g. [90] and two-echelon OWMR e.g. [79]. However, Instance 2 is a counterexample against global optimality of RPP when $h_0 > h_i$.

The possible violation occurs because the “reversed” echelon holding cost $H'_i(I_i) = (h_i - h_0)I_i$ is a decreasing function, which no longer favors earlier short-haul FTL shipments; also see Corollary 4.7.4.

Instance 3 shows that for our problem, the consolidation center could hold inventory to better utilize truck capacity even if $h_0 > h_i \forall i \in G$.

Since FTL can be viewed as a special case of more complicated costs with LTL and/or courier for small enough cost breakpoints between the transportation modes, Example 7 is valid in these cases as well. Therefore, we may not find good solutions to our problem with a straightforward application of existing OWMR methods. As an illustration, ZIO enables [89] and [119] to model the system-wise holding cost of each demand with one single parameter, which is not true here since we may want to split the demands. The relationship between OWMR and JRP further motivates these papers to partition the growers into two sets where a grower with $h_i < h_0$ always assumes cross-docking and central inventory is held only if $h_i \geq h_0$. The simplification is without loss of optimality in standard settings, but does not hold for us. On the other hand, we may assume FIFO inventory depletion given linear holding costs.

4.4 Approximations for Mixed Truckload Costs

One way to analyze the problem is converting it to previous OWMR models by modifying the cost functions $c_i(\cdot)$. The goal is to control solution quality within a moderate loss of optimality while not changing the feasible region. This section summarizes approximations we can thus achieve for various mixed truckload costs. Proposition 4.4.1 and Corollary 4.4.2 establish the relationship between our problem and those in the literature.

Proposition 4.4.1. *Consider our problem P and an OWMR variant Q with the same short-haul transportation and holding cost structures. Denote the long-haul transportation cost function by $g(\cdot)$ with subscripts indicating the associated problems.*

If $g_Q(\cdot) \geq g_P(\cdot) \geq g_Q(\cdot)/\alpha$ for some $\alpha > 1$, then any β -approximation to problem Q implies an $\alpha\beta$ -approximation to problem P .

Proof. Denote a solution and its corresponding objective value by x and $f(x)$, respectively. Let $f(\cdot) = g(\cdot) + h(\cdot)$, where $h(\cdot)$ is the sum of short-haul costs and central holding costs. Define x_P^* and x_Q^* as the optimal solutions of problem P and problem Q , respectively. Let x_Q be a solution which yields a β -approximation of x_Q^* . Since P and Q have the same solution set and $h_P(\cdot) = h_Q(\cdot)$, it follows that

$$\begin{aligned}
f_P(x_Q) &= g_P(x_Q) + h_Q(x_Q) \\
&\leq f_Q(x_Q) && \text{by upper bound} \\
&\leq \beta f_Q(x_Q^*) && \text{by approximation} \\
&\leq \beta f_Q(x_P^*) && \text{by optimality} \\
&= \beta g_Q(x_P^*) + \beta h_P(x_P^*) \\
&\leq \alpha \beta g_P(x_P^*) + \beta h_P(x_P^*) && \text{by lower bound} \\
&\leq \alpha \beta f_P(x_P^*). \quad \square
\end{aligned}$$

Corollary 4.4.2. *Proposition 4.4.1 is true for any facility-wise separation of $f(\cdot)$ to $g(\cdot)$ and $h(\cdot)$.*

These properties suggest a natural approach to attain worst-case guarantees of our problem:

Step 1. Find a target cost structure with a known OWMR approximation ratio i.e. β in Proposition 4.4.1.

Step 2. Derive a valid conversion from our cost structure(s) to the target cost structure, i.e. bound our cost structure(s) with some α in Proposition 4.4.1.

Step 3. Apply an approximation algorithm for the target problem to the converted problem.

Since Step 3 is a one-time application, the total runtime of the procedure equals the sum in constructing the conversion and applying the algorithm. Next we discuss converted approximations for various mixed truckload costs. We will refer to α as the *transformation ratio*, and temporarily ignore the facility symbol in subscripts to simplify notation.

4.4.1 FTL

We relate the “reversed” OWMR problem with FTL costs to the standard OWMR problem with fixed setup costs.

Proposition 4.4.3. *The cost function of our problem with FTL costs can be transformed to that of an OWMR problem with fixed setup costs and linear variable ordering costs in constant time. The transformation ratio is $\alpha = 2$ and tight. Furthermore, the converted problem is solvable with a standard OWMR problem in the sense that they only differ by a constant in the objective values.*

Proof. Figure 19 shows a valid pair of upper bound and lower bound for each facility, which converts our problem with FTL costs to OWMR problems with fixed setup costs and variable ordering costs that are proportional to volume. Given FTL cost c_F and capacity K_F , the converted transportation cost functions $\tilde{c}(\cdot)$ (upper bound) and $\underline{c}(\cdot)$ (lower bound) are

$$\tilde{c}(x) = 2\underline{c}(x) = \begin{cases} 0, & x = 0 \\ c_F + \frac{c_F}{K_F}x, & x > 0 \end{cases}. \quad (17)$$

Since all demands must be satisfied and the cost parameters are static, variable costs $\frac{c_F}{K_F}x$ are ignorable and the transformed problems with $\tilde{c}(\cdot)$ and $\underline{c}(\cdot)$ reduce to standard OWMR problems with fixed setup costs c_F and $0.5c_F$, respectively.

Tightness of α : Clearly $\alpha \leq 2$ with the above construction. Meanwhile, the best guarantee of this kind is restricted by the ratio of the upper bound versus the lower bound when we ship one full truck, i.e. $\alpha \geq \frac{\tilde{c}(K_F)}{\underline{c}(K_F)} = \frac{2c_F}{c_F} = 2$. \square

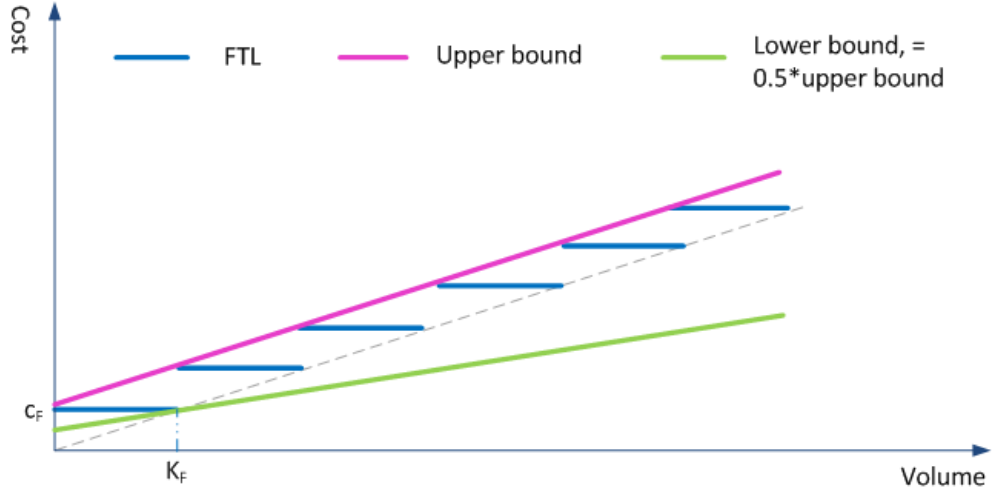


Figure 19: Transformation of FTL costs

As a direct consequence of Proposition 4.4.3, we can obtain polynomial-time approximations for a variety of OWMR/JRP problems with FTL costs. By Corollary 4.4.2, the desired approximation ratio for each variant is 2 times the best-known result under fixed setup costs. Table 6 lists the details.

Table 6: Approximations for OWMR/JRP variants with FTL costs

Original papers	Demand assumptions	Implied FTL results
<i>Discrete-time finite-horizon OWMR, linear holding costs:</i>		
[89]	Dynamic	3.6-approximation
<i>Discrete-time finite-horizon JRP, linear holding costs:</i>		
[31]	Dynamic	3.582-approximation
[97]	Stationary over time	$(2 + \epsilon)$ -scheme
<i>Discrete-time finite-horizon JRP, no holding costs:</i>		
[30]	Dynamic with deadlines	3.148-approximation
<i>Discrete-time finite-horizon OWMR/JRP, linear holding costs and shortage costs:</i>		
[68], OWMR	Dynamic, backlog	6-approximation
[68], OWMR	Dynamic, lost-sales	4-approximation
[119], JRP	Dynamic, backlog	4-approximation
<i>Continuous-time infinite-horizon OWMR/JRP, linear holding costs:</i>		
[107]	Constant rates	2.04-approximation

Note that 1) we may count on better *a posteriori* guarantees than the listed results since the variable costs $\frac{c_F}{K_F}x > 0$ in formula (17) ensure that any β -approximation algorithm in the standard setting yields a solution with a cost less than β times the

optimal cost for the converted OWMR problem; 2) we abuse PTAS terminology using $(2+\epsilon)$ -*scheme* to mean an algorithm which outputs a solution with a cost within $(2+\epsilon)$ times the optimum in polynomial time of the input size and $1/\epsilon$ for any instance and a given parameter $\epsilon > 0$; 3) we do not impose capacities on backorders or lost-sales since they are modeled with linear penalty costs in the original papers; 4) the JRP with deadlines can be modeled with threshold holding cost functions.

4.4.2 Short-haul FTL+stepwise LTL and long-haul FTL

We relate the “reversed” OWMR problem with short-haul FTL+stepwise LTL and long-haul FTL costs to [44]’s OWMR problem. As mentioned in Section 4.2, [44] considered modified all-unit discounts for the growers and incremental discounts for the center.

The modified all-unit discounts can be represented by the following PWL cost function:

$$c(x) = \begin{cases} 0, & x = 0 \\ \eta_1 M_1, & 0 < x < M_1 \\ \min\{\eta_i x, \eta_{i+1} M_{i+1}\}, & M_i \leq x < M_{i+1}, \quad i = 1, \dots, n-1 \\ \eta_n x, & M_n \leq x \end{cases}, \quad (18)$$

where $\eta_1 > \eta_2 > \dots \geq 0$, M_i are the breakpoints between the i -th flat section and the i -th non-flat section, and $\eta_1 M_1 \geq 0$ is a minimum charge for shipping a small volume.

Since FTL costs can be converted to fixed setups which are a special case of incremental discounts, Proposition 4.4.4 is sufficient for the desired transformation.

Proposition 4.4.4. *FTL+stepwise LTL costs can be transformed to [44]’s modified all-unit discounts with $\alpha = 2$. Furthermore, the transformation ratio is tight if the FTL-LTL breakpoints satisfy $b_F \geq 2K_L$.*

Proof. We illustrate the transformation for FTL+stepwise LTL costs in Figure 20.

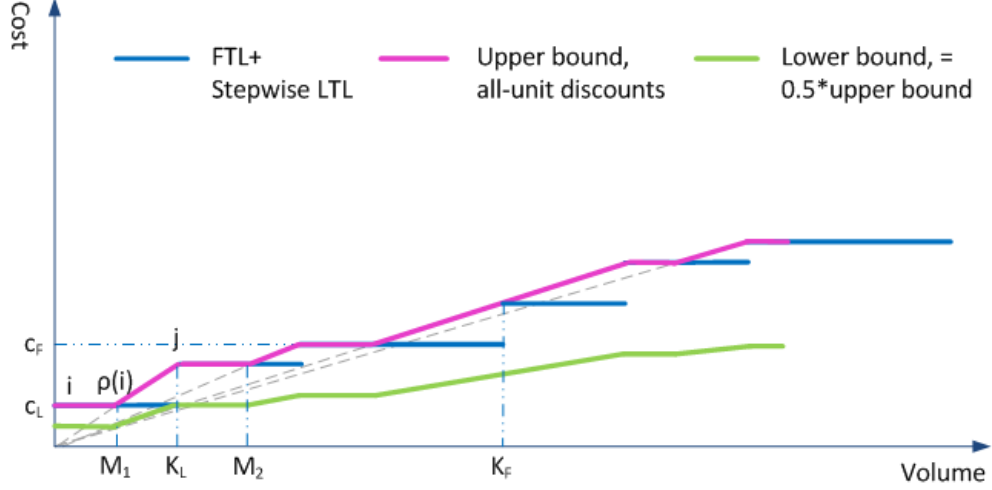


Figure 20: Transformation of FTL+stepwise LTL costs

Upper bound construction: Order the jump points from smallest to largest possible where the true cost function becomes discontinuous. In Figure 20, for instance, the first jump point occurs at $(0, c(0+)) = (0, c_L)$, the second is $(K_L, 2c_L)$, etc. The last jump point is such that the corresponding volume bounds the total demand from above, e.g., this can be point $(3K_F, 3c_F + c_L)$ if $\sum_{i \in G, t \leq T} d_i^t = 3K_F - K_L$. Our target cost function alternates flat sections which start with a subset of the jump points and non-flat sections with decreasing slopes as the volume increases. Let $\pi(i)$ be the predecessor of jump point i in the partially converted cost function, e.g. $\pi(2) = 1$ in Figure 20. Also, let $\rho(i)$ be the breakpoint between the flat section incident to point i and the successive non-flat section. Initialize $i \leftarrow 1, j \leftarrow 2$.

Step 1. For jump points pair (i, j) , compare the associated unit costs u_i and u_j , which are defined as the slopes at these points, e.g. $u_1 = \infty, u_2 = \frac{c_L}{K_L}$ in Figure 20. Go to Step 2 if $u_i \geq u_j$, Step 5 o.w.

Step 2. Exit if j is the largest jump point; go to Step 3 o.w.

Step 3. Find point $\rho(i)$ in the flat section incident to point i and such that $u_{\rho(i)} = u_j$ in the true cost function. Plot line segments $i - \rho(i) - j$.

Step 4. $\pi(j) \leftarrow i, i \leftarrow j, j \leftarrow j + 1$. Go to Step 1.

Step 5. $i \leftarrow \pi(i)$. Go to Step 1.

The above construction yields a PWL upper bound which alternates flat and non-flat sections. Each flat section starts with an original jump point i in the true cost function, and ends with a breakpoint $\rho(i)$ found in Step 3. Delete the breakpoints that overlap with the corresponding jump points, i.e. ignore $\rho(i)$ if $\rho(i) = i$ or the length of flat section $i - \rho(i)$ is zero in the constructed function. Order the final breakpoints $\rho(i)$ by volume from the smallest to the largest. Denote the resulting set by \mathcal{K} , and let M_k be the accumulated capacity as the k -th flat section ends, a mathematical description of the converted function is then

$$\tilde{c}(x) = \begin{cases} 0, & x = 0 \\ c_L, & 0 < x < M_1 \\ \min\{u_k x, u_{k+1} M_{k+1}\}, & M_k \leq x < M_{k+1}, k = 1, \dots, |\mathcal{K}| \end{cases}, \quad (19)$$

where each $k \in \mathcal{K}$ is associated with some $\rho(i)$ in the construction procedure. Since Step 3 is executed only if $u_i \geq u_j$, this results in a flat section followed by a non-flat section with decreasing slopes. Step 4 defines $\pi(i)$ for each pair of jump points (i, j) considered in Step 3, which guarantees that a valid pair of (i, j) such that $u_i \geq u_j$ will be reached when Step 5 iterates. Therefore, $u_1 > u_2 > \dots > u_{|\mathcal{K}|} \geq 0$ in the final upper bound cost function, and (19) is a special case of the modified all-unit discounts given by (18), where $M_1 \leq K_L$ and an appropriate value of $M_{|\mathcal{K}|+1}$ can be the total demand plus a small positive volume. Scaling the upper bound down by factor 2 gives a valid lower bound since $\tilde{c}(x)$ is a non-decreasing function and the cost ratio between consecutive jump points does not exceed $\frac{\min\{2c_L, c_F\}}{c_L} \leq 2$.

Tightness of α : The construction implies a transformation ratio no higher than 2. Meanwhile, the best guarantee of this kind is restricted by the cost ratio of the

upper bound versus the lower bound when we ship one LTL unit, i.e. $\alpha \geq \frac{c_L}{0.5c_L} = 2$ if $b_F \geq 2K_L$. \square

We give the approximation results in Corollaries 4.4.5-4.4.6, which are direct consequences of Proposition 4.4.3, Proposition 4.4.4, and [44, 115].

Corollary 4.4.5. *There exists a ZIO policy with objective value no more than $\frac{8}{3} \approx 2.67$ and $\frac{5.6}{2.3} \approx 2.43$ times the optimal cost for our problem with short-haul FTL+stepwise LTL and long-haul FTL in the cases of time-varying and stationary cost parameters, respectively. However, finding the best ZIO policy is \mathcal{NP} -hard.*

We note that [44]’s bound is tight for the transformed OWMR problem. They also proved that for any instance of the problem, it is not possible to develop a polynomial-time algorithm which generates a solution better than a factor of $\mathcal{O}(\log |G|)$ from optimality, unless $\mathcal{P} = \mathcal{NP}$. Hence, these are the best guarantees we can achieve by converting FTL+stepwise LTL costs to modified all-unit discounts.

Corollary 4.4.6. *If the center operates as a cross-docking station and the number of jump points is bounded above by a polynomial function of the problem size parameters in the transformation, then our OWMR problem with long-haul FTL and short-haul FTL+stepwise LTL costs can be approximated within a factor of $(2 + \epsilon)$ by solving a related PWL concave cost multi-commodity network flow problem with [115]’s approach.*

[44]’s cost functions are special cases of the structures considered in [115]. As the latter pointed out, the PWL concave cost multi-commodity network flow problem is \mathcal{NP} -hard, but can be approximated within a logarithmic factor by solving a polymatroid reformulation. Given Corollary 4.4.6, we can thus obtain a $((1 + \epsilon) \log n)$ -approximation for our problem with long-haul FTL and short-haul FTL+stepwise LTL costs, where $\epsilon > 0$ and n is a polynomial function of $|G|, T$. The assumption of

polynomially-sized jump points ensures that the transformation can finish in polynomial time.

4.4.3 FTL+linearized LTL or FTL+LTL+courier

We relate the “reversed” OWMR problem with FTL+linearized LTL or FTL+LTL+courier costs to incremental discounts, and give a $((1 + \epsilon) \log n)$ -approximation.

Proposition 4.4.7. *Assuming cross-docking at the consolidation center, our OWMR problem with FTL+linearized LTL or FTL+LTL+courier costs can be transformed to [115]’s OWMR problem with $\alpha \leq 2$.*

Proof. [115] considered a class of general PWL costs, among which the incremental discounts are our target cost structure and can be described with a PWL concave function:

$$c(x) = \begin{cases} \eta_1 x, & 0 \leq x < M_1 \\ \eta_1 M_1 + \eta_2(M_2 - M_1) + \dots + \eta_k(x - M_{k-1}), & M_{k-1} \leq x < M_k, \ 2 \leq k < n, \\ \eta_1 M_1 + \sum_{k=2}^n (M_k - M_{k-1}) + \eta_{n+1}(x - M_n), & M_n \leq x \end{cases} \quad (20)$$

where $\eta_1 > \eta_2 > \dots > \eta_{n+1} \geq 0$, and M_k are the breakpoints for charging new incremental discounts, $1 \leq k \leq n$.

For FTL+linearized LTL costs, two linear pieces of the concave function are enough to validate the bounds (Figure 21). The converted cost functions are

$$\tilde{c}(x) = 2\underline{c}(x) = \begin{cases} \frac{c_L}{K_L} x, & 0 \leq x < b_F \\ c_F(1 - \frac{b_F}{K_F}) + \frac{c_F}{K_F} x, & b_F \leq x \end{cases}, \quad (21)$$

where $\frac{c_L}{K_L} = \frac{c_F}{b_F} > \frac{c_F}{K_F} \geq 0$.

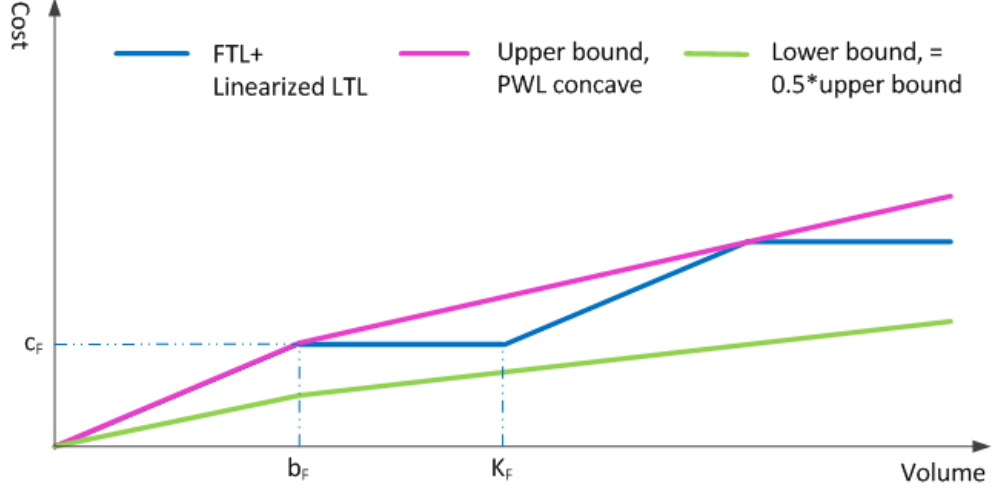


Figure 21: Transformation of FTL+linearized LTL costs

For FTL+LTL+courier costs, the desired PWL concave function contains more linear pieces (Figure 22), and is constructed in a way similar to the proof of Proposition 4.4.4, where the jump points are replaced by the actual cost breakpoints e.g. b_L , b_F , etc. Let the breakpoints be an ordered set \mathcal{K} for the final upper concave envelope. A mathematical description of the converted costs is

$$\tilde{c}(x) = 2\underline{c}(x) = \begin{cases} u_1 x, & 0 \leq x < M_1 \\ u_1 M_1 + u_2(M_2 - M_1) + \dots + u_k(x - M_{k-1}), & M_{k-1} \leq x < M_k, \quad 2 \leq k < |\mathcal{K}| \\ u_1 M_1 + \sum_{k=2}^{|\mathcal{K}|} u_k(M_k - M_{k-1}) + u_{|\mathcal{K}|+1}(x - M_{|\mathcal{K}|}), & M_{|\mathcal{K}|} \leq x \end{cases} \quad (22)$$

where the unit costs satisfy $u_1 = c_U = \frac{c_L}{b_L} > u_2 > \dots > u_{|\mathcal{K}|+1} \geq 0$ by definition of breakpoints in the true cost function, $M_1 = b_L$, and $M_{|\mathcal{K}|}$ may be bounded by the total demand.

(21)-(22) are special cases of (20), hence the transformations are valid. For both FTL+linearized LTL and FTL+LTL+courier costs, the tightest possible transformation ratio α depends on the cost parameters, but is bounded above by 2. \square

Corollary 4.4.8. *Corollary 4.4.6 applies to our problem with FTL+linearized LTL*

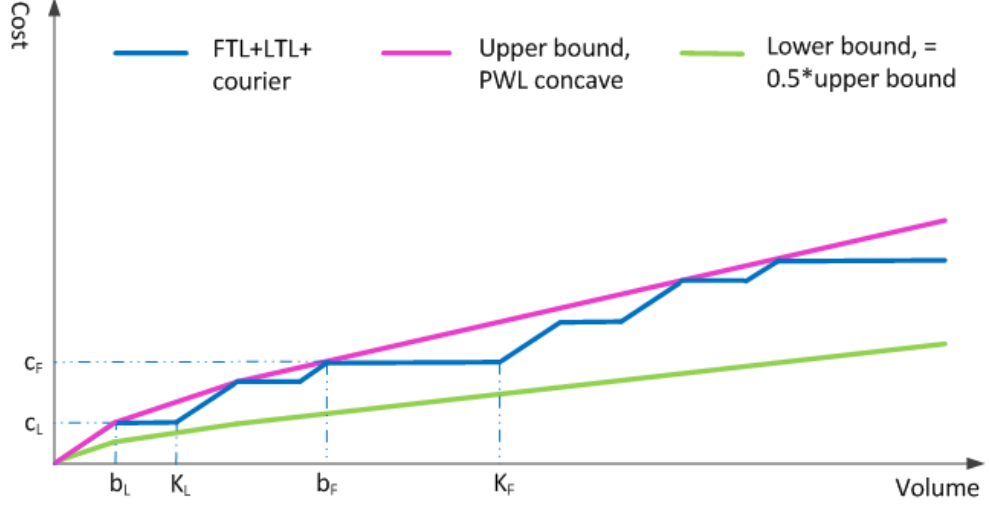


Figure 22: Transformation of FTL+LTL+courier costs

or *FTL+LTL+courier costs*, either universal or mixed with short-haul *FTL+stepwise LTL costs*.

We conclude this section by noting that existing approximations on all the transformed OWMR problems, including [89, 44, 115], are built upon ZIO policies or LP relaxations. Although the optimal ZIO policy is effective amongst the approximation ratios we derive for mixed truckload costs, it suffers from theoretical intractability and thus may be hard to implement. In addition, the LP relaxation of Model (16) can be very weak since it assumes linear shipping costs at unit rates of $\frac{c_{i,F}}{K_F} \forall i \in G \cup \{0\}$ ([50]). Therefore, we are motivated to develop alternative approaches.

4.5 *Single-Echelon Subproblems*

In this section, we take a different perspective and explore structural properties of pertinent subproblems to facilitate subsequent analysis. The idea is to 1) decompose R-OWMR into single-echelon lot sizing problems which imply a lower bound of the two-echelon problem, and 2) approximate mixed truckload costs with concave batch costs under which the lot sizing problems are tractable with off-the-shelf methods.

As the name suggests, concave batch costs are a class of cost structures where

items are produced/ordered/shipped in batches of identical capacities and the cost per batch is concave in volume. In the remainder of the chapter, we will assume the batch capacity to be K_F . Model (15) formulates the “reversed” OWMR problem with concave batch costs if

$$c_i(q_i^t) := c_{i,B}(q_i^t \bmod K_F) + K_F^{-1} \cdot (q_i^t - q_i^t \bmod K_F) \cdot c_{i,B}(K_F), \quad \forall t = 1 \dots T, \quad (23)$$

where $c_{i,B}(\cdot)$ is a concave cost function on $[0, K_F]$, $\forall i \in G \cup \{0\}$. We will refer to this customized formulation as the *C-OWMR model*.

Observations 4.5.1-4.5.2 provide the basis of a uniform analytical framework for all the mixed truckload costs. Occasionally we will call a cost structure *near-concave* if it can be transformed to concave costs with some $\alpha \sim o(|G|)$ in Proposition 4.4.1.

Observation 4.5.1. *The following cost structures are special cases of PWL concave batch costs: FTL costs, FTL+linearized LTL costs.*

Observation 4.5.2. *The following cost structures can be approximated by PWL concave batch costs with transformation ratio $\alpha \leq 2$: FTL+stepwise LTL costs, FTL+LTL+courier costs.*

Proof. We ignore facility subscripts for notational convenience. Figure 23 illustrates a valid transformation for FTL+stepwise LTL costs. The transformation ratio is $\alpha = 2$ and tight when the volume is K_L if $b_F \geq 2K_L$. The converted cost functions for each batch are PWL concave on $[0, K_F]$:

$$\tilde{c}_B(x) = 2\underline{c}_B(x) = \begin{cases} 0, & x = 0 \\ c_L + \frac{c_L}{K_L}x, & 0 < x \leq b_F \\ c_F, & b_F < x \leq K_F \end{cases} \quad (24)$$

Figure 24 illustrates a transformation with $\alpha < 2$ for FTL+LTL+courier costs. The target bounds are obtained by repeating the construction on $[0, K_F]$ for the same

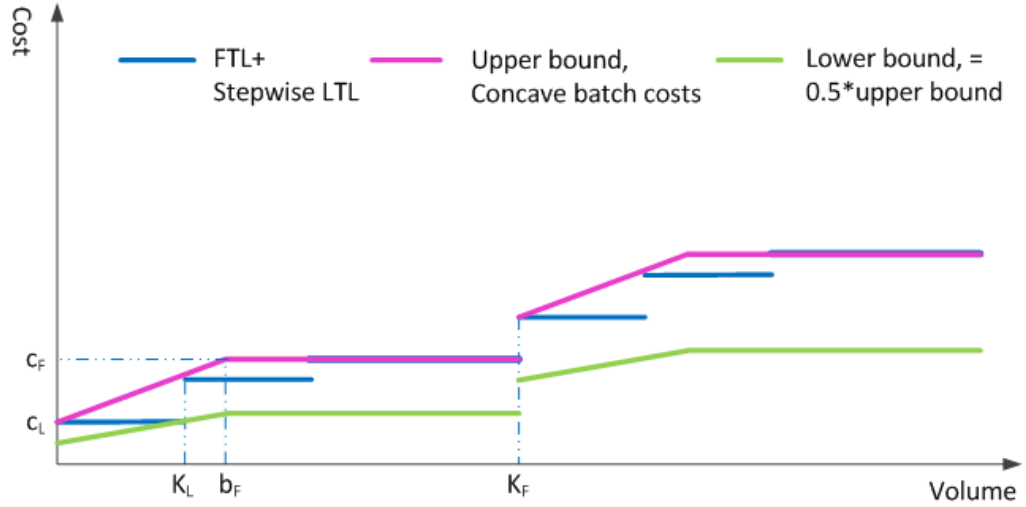


Figure 23: Concave batch cost transformation of FTL+stepwise LTL costs

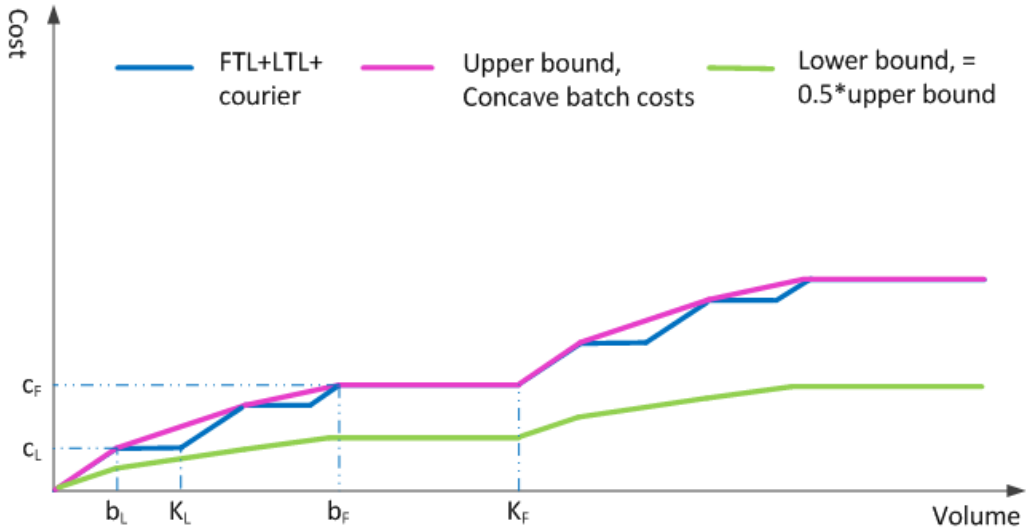


Figure 24: Concave batch cost transformation of FTL+LTL+courier costs

cost structure in the proof of Proposition 4.4.7, which results in PWL concave batch costs. □

With Observations 4.5.1-4.5.2, we may assume the transportation to be conducted by a single mode, i.e. the converted *batch truck* for each facility. In the corresponding Model (16), we can further disaggregate capacity constraints by truck if the number of batches is properly bounded. The resulting objective function is concave instead

of linear compared to (16a):

$$\sum_{t=1}^T \sum_{i \in G \cup \{0\}} \sum_{p=1}^{\mathcal{N}_{i,t}} c_{i,B}(q_{i,p}^t) + \sum_{t=1}^T \sum_{i \in G \cup \{0\}} h_i I_i^t, \quad (25)$$

where $\mathcal{N}_{i,t}$ is the maximum possible number of batch trucks at facility $i \in G \cup \{0\}$ in period $t = 1 \dots T$. Meanwhile, the feasible region is a network flow polytope since we can remove the integer variables and rewrite (16d) as

$$q_{i,p}^t \leq K_F, \quad \forall i \in G \cup \{0\}, p = 1 \dots \mathcal{N}_{i,t}, t = 1 \dots T. \quad (26)$$

Following the spirit of decomposition, we consider two subproblems of C-OWMR: SHP as the short-haul problem where demands are not due until the planning horizon ends, and LHP as the long-haul problem where demands are available upon the ready period. Depending on our purposes, sometimes we also revise the inventory holding cost coefficients, which are denoted by $h_{i,i}$ for SHP and $h_{0,i}$ for LHP, $\forall i \in G$.

We formulate the SHP by dropping the central inventory and delivery variables from C-OWMR, and replacing h_i with $h_{i,i}$, $\forall i \in G$.

$$\text{(SHP)} \quad \min \sum_{t=1}^T \sum_{i \in G} c_i(q_i^t) + \sum_{t=1}^T \sum_{i \in G} h_{i,i} I_i^t \quad (27a)$$

$$\text{s.t. } I_i^t + q_i^t - I_i^{t-1} = d_i^t, \quad \forall i \in G, t = 1 \dots T \quad (27b)$$

$$I_i^0 = 0, \quad \forall i \in G \quad (27c)$$

$$I_i^T = 0, \quad \forall i \in G \quad (27d)$$

$$q \geq 0, I \geq 0, \quad (27e)$$

where $c_i(\cdot)$ is specified by (23), $\forall i \in G$.

Observation 4.5.3 is immediate by the independence of SHP decisions for each grower.

Observation 4.5.3. *Model (27) can be decomposed, without loss of optimality, into $|G|$ single-item lot sizing problems with concave batch ordering costs where each grower is handled as an individual item.*

The $|G|$ lot sizing problems can be solved in parallel. Hence SHP is as difficult as the LSP under static concave batch ordering costs.

We formulate the LHP by dropping the growers' inventory and pickup variables from C-OWMR, and splitting the central variables by grower to reflect the revised central holding cost rates $h_{0,i}$.

$$(LHP) \quad \min \sum_{t=1}^T c_0 \left(\sum_{i \in G} q_{0,i}^t \right) + \sum_{t=1}^T \sum_{i \in G} h_{0,i} I_{0,i}^t \quad (28a)$$

$$\text{s.t. } I_{0,i}^t + q_{0,i}^t - I_{0,i}^{t-1} = d_i^t, \quad \forall i \in G, t = 1 \dots T \quad (28b)$$

$$I_{0,i}^0 = 0, \quad \forall i \in G \quad (28c)$$

$$I_{0,i}^T = 0, \quad \forall i \in G \quad (28d)$$

$$q \geq 0, I \geq 0, \quad (28e)$$

where $c_0(\cdot)$ is specified by (23). The center (i.e. 0) can be removed from variable subscripts $(0, i)$ since any variable is associated with some grower i . However, we keep it in the model to avoid confusion with (SHP).

Because the growers jointly decide on long-haul shipments, LHP is a multi-item lot sizing problem under concave batch ordering costs. If two items have the same revised central holding cost rate, i.e. $h_{0,i} = h_{0,j}$ for some $i, j \in G$, they are identical and can be combined into one item in LHP. Hence we have the following observation, which establishes the computational equivalence of the subproblems in a special case.

Observation 4.5.4. *LHP reduces to a single-item lot sizing problem with concave batch ordering costs if $h_{0,i} = h_{0,j}, \forall i, j \in G$.*

Capacitated lot sizing problems are hard in general. Under concave batch ordering costs and linear holding costs, however, the literature solves all the variants involved later.

As Table 7 shows, SHP with concave batch costs (including the special case of FTL costs) and LHP with FTL costs can be solved by DP, a class of algorithms which

Table 7: Exact subproblem solutions

Papers	Subproblems & settings	Methods
[79]	SHP with FTL costs	DP, $\mathcal{O}(T^2)$
[90]	SHP with concave batch costs	DP, $\mathcal{O}(T^5)$
[16]	LHP with FTL costs	LP, $\mathcal{O}(G T^2)$
[90]	LHP with concave batch costs, $h_{0,i} = h_{0,j} \forall i, j \in G$	DP, $\mathcal{O}(T^5)$

implicitly enumerate feasible solutions via an underlying shortest path network. For lot sizing problems, this network typically consists of $\mathcal{O}(T)$ nodes, each representing a period, and a forward arc between each pair of nodes. The weight of arc $(s, t), \forall s < t$ equals the minimum cost to fulfill all the demands in the traversed periods, conditional on inventory boundaries in period s and period t . Since the number of arcs is polynomially bounded by the problem size, so is the runtime of the DP if there is an efficient way to calculate the arc costs. For the classical LSP, this process can be done in quadratic time by recursively applying the ZIO property and Bellman’s optimal equations to smaller intervals. Similarly, the regeneration property enables polynomial solvability of the LSP under concave batch ordering costs.

Proposition 4.5.5 (Regeneration property for single-item lot sizing problems with concave batch ordering costs). *There exists an optimal SHP solution where each grower i ’s decision is such that*

- (a) *in each period there is at most one partially filled batch truck;*
- (b) *between consecutive periods $1 \leq s < t \leq T$ with partially filled batch truck, there is a regeneration point, i.e. $I_i^\tau = 0$ for some $s < \tau \leq t$;*
- (c) *between consecutive regeneration points $1 \leq s < t \leq T$, there is at most one period $\tau \in (s, t]$ with a partially filled batch truck;*
- (d) $I_i^t < K_F, \forall t = 1 \dots T$;

(e) when the batch truck is identical to an FTL, if there is a partially filled truck between consecutive regeneration points $s < t$, it must occur in the second regeneration point t .

Proof. We may consider each grower i 's problem separately by Observation 4.5.3. (a) is trivial by (23) and definition of concave batch costs. (b)-(d) are results of [90]'s concave cost model for a single-item lot sizing problem under time-varying concave batch ordering costs and concave holding costs, which fits our SHP. For the special case of FTL costs, if the partially filled shipment period $\tau \neq t$ in a regeneration interval, we can obtain a solution with no-higher holding costs by adding volume to the partial truck until it is fully utilized in period τ and repeat the process until we reach the first new regeneration point afterwards ($\leq t$). The transportation cost does not increase by fixed dispatch cost of a truck. Thus (e) follows. \square

Proposition 4.5.5 suggests that the total transportation cost between consecutive regeneration points is readily determined since the number of FTLs and the volume carried by the possible partial shipment are fixed. The corresponding inventory cost can be calculated by enumerating the periods when the partial truck can be dispatched, which takes polynomial number of iterations for concave batch ordering costs and constant time for FTL costs in each regeneration interval. Hence efficient DP algorithms are available by merely considering feasible partitions of the planning horizon to successive regeneration intervals in the underlying shortest path network. These are proposed in [90] and [79].

For the multi-item counterpart, [14, 15] utilized similar properties to develop DP as well as heuristics for the MIMV problem. However, the calculation of inventory costs is much more involved to coordinate multiple items and the DP runs in exponential time unless $|G|$ is fixed. Instead, [16] reformulated the MIMV model into an LP, which is polynomially solvable with the Ellipsoid method. These approaches directly apply to our LHP with FTL costs.

While it is always optimal to solve SHP with separate lot sizing problems, the same does not hold in the decomposition of C-OWMR to SHP and LHP. In fact, the subproblem solutions could be infeasible for C-OWMR due to the lack of coordination in the SHP pickup time and the LHP delivery time of a demand. On the other hand, we claim that a feasible C-OWMR solution cannot cost less than the total subproblem optimal objective values if $h_{i,i}$ and $h_{0,i}$ are properly chosen.

Proposition 4.5.6. *Take $h_{i,i} = 0.5h_i$ and $h_{0,i} = 0.5 \min\{h_i, h_0\}$, $\forall i \in G$. Let $Z^*(P)$ be the optimal objective function value of problem P ; then $Z^*(SHP) + Z^*(LHP) \leq Z^*(C-OWMR)$. That is, the subproblems bound from below the minimum total cost to distribute the entire demand in the C-OWMR model.*

Proof. Introduce variables $I_{0,i}^t \geq 0, q_{0,i}^t \geq 0, S_i^t \geq 0$ to Model (15), where $I_{0,i}^t, q_{0,i}^t$ correspond to the central inventory and shipping quantities in (LHP), and S_i^t represent the actual central inventory quantities for grower i in the C-OWMR model. Add valid equalities $I_{0,i}^t = S_i^t + I_i^t$, $\sum_{i \in G} S_i^t = I_0^t$, $\sum_{i \in G} q_{0,i}^t = q_0^t$; the optimal objective value of Model (15) remains the same. Given an arbitrary feasible C-OWMR solution, we abuse notation and let $I_m^\nu = d_m \forall \nu = s \dots \tau - 1$, $S_m^\nu = d_m \forall \nu = \tau \dots t$, where m is a tuple that consists of a grower i and ready-pickup-delivery times s, τ, t such that each split demand d_m ($d_m \leq d_i^s$) is served by the shipments in a unique pair of periods (τ, t) . The actual holding cost for d_m is

$$\begin{aligned}
H(d_m) &= [h_i(\tau - s) + h_0(t - \tau)]d_m \\
&= \sum_{\nu=s}^{\tau-1} h_i I_m^\nu + \sum_{\nu=\tau}^t h_0 S_m^\nu \\
&\geq 0.5 \sum_{\nu=s}^{\tau-1} h_i I_m^\nu + 0.5 \sum_{\nu=s}^{\tau-1} \min\{h_i, h_0\} I_m^\nu + 0.5 \sum_{\nu=\tau}^t \min\{h_i, h_0\} S_m^\nu \\
&= \sum_{\nu=s}^{\tau-1} h_{i,i} I_m^\nu + \sum_{\nu=s}^t h_{0,i} I_{0,m}^\nu, \tag{29}
\end{aligned}$$

where $I_m, I_{0,m}$ are the induced subproblem inventory quantities for d_m . (27a)+(28a) \leq (15a) results from summing (29) over all split demands in the C-OWMR solution.

Furthermore, (27b)-(27e), (28b)-(28e) are consequences of (15b)-(15f) and the valid equalities. When put together, the subproblems thus give a relaxation of the C-OWMR model. The lower bound then follows. \square

Model (27) and Model (28) can be reformulated as concave cost network flow problems in light of (25)-(26). Some of the properties can be extended to more complicated settings with the alternative network flow formulation. We will revisit them in Section 4.8.

4.6 An Improved Approximation for FTL Costs

In this section, we aim to answer the following question: Given optimal SHP and LHP solutions, can we construct a reasonably good solution to C-OWMR? We address the special case of FTL costs with a simple 2-approximation algorithm that recombines the subproblem solutions into a feasible C-OWMR solution whose total cost is within 2 times the optimal objective value. We also give a tightness example which shows that the ratio of 2 is in fact the best the algorithm can achieve. Similar ideas can be found in [119] and [68] for OWMR problems with fixed setup costs. The main difference is that they imposed ZIO throughout the construction so the solution merely depends on when the orders are placed, whereas we further split demands associated with subproblem shipments so the decision of shipping time as well as quantities is more involved, and the resulting C-OWMR solution may not preserve regeneration properties, either.

4.6.1 Algorithm

Our algorithm consists of two phases:

Phase 1. Solve SHP with $h_{i,i} = 0.5h_i$, LHP with $h_{0,i} = 0.5 \min\{h_0, h_i\}$, $\forall i \in G$.

Phase 2. Convert the subproblem solutions to a globally feasible C-OWMR solution.

Now that both SHP and LHP can be solved for non-identical growers' holding cost rates, the remaining issue is Phase 2. A solution is globally feasible if all demands are satisfied, and the short-haul pickup of any (split) demand occurs no later than the long-haul delivery. The former condition automatically holds in any feasible SHP and LHP solutions. Since the latter condition may be violated when we solve SHP and LHP separately, a natural fix is to judiciously split, merge or move the shipments originally planned in Phase 1. We propose such a procedure that adjusts the subproblem solutions in T iterations, where each iteration t yields a partial C-OWMR solution feasible for periods $1 \dots t$ and so the final output is globally feasible. Meanwhile, the adjusted subproblem solutions maintain feasibility from iteration to iteration. We control the location and timing of the fix operations to bound the objective value of the resulting C-OWMR solution. Some notation follows.

Since inventory quantities are uniquely determined once shipping quantities are known, both C-OWMR and subproblem solutions can be fully described by the q values in Models (27)-(28). Let

- $q_i^\tau, \forall i \in G, \tau = 1 \dots T$: short-haul shipping quantities in the temporary SHP solution at the beginning of an iteration;
- $q_{0,i}^\tau, \forall i \in G, \tau = 1 \dots T$: long-haul shipping quantities in the temporary LHP solution at the beginning of an iteration.

The q values are updated from iteration to iteration. To track the changes throughout the procedure, let

- $\hat{q}_i^\tau, \forall i \in G, \tau = 1 \dots T$: short-haul shipping quantities in the temporary SHP solution at the end of an iteration;
- $\hat{q}_{0,i}^\tau, \forall i \in G, \tau = 1 \dots T$: long-haul shipping quantities in the temporary LHP solution at the end of an iteration.

For illustrative convenience, let

- $G^\tau := \{i \in G : q_{0,i}^\tau > 0\}$, $\forall \tau = 1 \dots T$: set of growers with positive long-haul shipping quantities for each period in the temporary LHP solution.

The input SHP and LHP solutions from Phase 1 give the initial values of q_i^τ and $q_{0,i}^\tau$, respectively, $\forall i \in G, \tau = 1 \dots T$. The temporary solutions between consecutive iterations satisfy $q_i^{\tau+1} = \hat{q}_i^\tau$, $q_{0,i}^{\tau+1} = \hat{q}_{0,i}^\tau$, $\forall i \in G, 1 \leq \tau < T$. The temporary LHP solution implies the sets G^τ at any point of the procedure. Our goal is globally feasible \hat{q} values when the final iteration T ends. As previously mentioned, a pair of subproblem solutions are globally feasible for the C-OWMR problem if demands are satisfied in both echelons and the short-haul shipments arrive in time at the center for long-haul shipping:

$$\sum_{t=s}^T q_i^t = d_i^s, \quad \forall i \in G, s = 1 \dots T, \quad (30a)$$

$$\sum_{t=s}^T q_{0,i}^t = d_i^s, \quad \forall i \in G, s = 1 \dots T, \quad (30b)$$

$$\sum_{\tau=1}^t q_i^\tau \geq \sum_{\tau=1}^t q_{0,i}^\tau, \quad \forall i \in G, t = 1 \dots T, \quad (30c)$$

where (30a) is the short-haul demand satisfaction condition, (30b) long-haul demand satisfaction, and (30c) the link condition. Note that (30c) guarantees in-time central arrivals for every split demand because of FIFO. Any such pair of subproblem solutions induce a feasible C-OWMR solution to Model (15) where the short-haul quantities equal q_i^t and the long-haul quantities are $q_0^t = \sum_{i \in G} q_{0,i}^t = \sum_{i \in G^t} q_{0,i}^t, \forall t = 1 \dots T$.

At the beginning of iteration t , we have the current subproblem solutions. Suppose the temporary q values are such that

$$(30a) - (30b),$$

$$\sum_{\iota=1}^{\tau} q_i^\iota \geq \sum_{\iota=1}^{\tau} q_{0,i}^\iota, \quad \forall i \in G, \tau = 1 \dots t-1. \quad (31)$$

Again (30a)-(30b) ensure SHP and LHP subproblem feasibility. (31) is a subset of (30c) and states that there are sufficient short-haul arrivals at the center for long-haul shipping in periods $1 \dots (t - 1)$. Hence, the induced C-OWMR solution is partially feasible in these periods. Initially the conditions hold for $t = 1$. If they continue to hold for every iteration, then we will obtain a desired C-OWMR solution which is feasible in all the periods $1 \dots T$. While there are many possible ways to construct such a solution, we base our fix operations on an additional condition:

$$\sum_{\iota=1}^{\tau} q_i^{\iota} = \sum_{\iota=1}^{\tau} q_{0,i}^{\iota}, \quad \forall i \in G^{\tau}, \tau = 1 \dots t. \quad (32)$$

In other words, we impose ZIO at the consolidation center. Consequently, all the short-haul arrivals from a grower leave the center if it is chosen for period t and some earlier period τ as well in the LHP solution when iteration t ends, i.e. $\sum_{\iota=\tau+1}^t q_i^{\iota} = \sum_{\iota=\tau+1}^t q_{0,i}^{\iota}$ if $i \in G^t \cap G^{\tau}$; and the final C-OWMR solution satisfies (32) in all periods. It is interesting to see that while the condition may not hold in the initial LHP solution (which instead satisfies a regeneration property similar to Proposition 4.5.5, [14]), imposing central ZIO in the C-OWMR solution still yields a reasonable worst-case guarantee. On the other hand, the final SHP solution may violate both ZIO and the regeneration property, which significantly differs from [119] and [68] who assumed ZIO at every facility.

Now we compare the q values as of period t in the current SHP and LHP solutions. If $G^t = \emptyset$, no LHP shipping is planned in period t so we set $\hat{q}_i^{\tau} = q_i^{\tau}$, $\hat{q}_{0,i}^{\tau} = q_{0,i}^{\tau} \forall i \in G, 1 \leq \tau \leq T$, and jump to the next iteration. This maintains feasibility conditions (30a)-(30b) and (31) for periods $\tau = 1 \dots t$ without incurring extra costs. For the same reason, we do not change the current solutions related to growers $i \notin G^t \neq \emptyset$. For each grower $i \in G^t \neq \emptyset$, either of the following cases can happen.

Case 1. $\sum_{\tau=1}^t q_i^{\tau} \geq \sum_{\tau=1}^t q_{0,i}^{\tau}$.

Consider the last time grower i 's demand is delivered in the long-haul echelon; by (32) we have $\sum_{\tau=1}^{\iota} q_i^{\tau} = \sum_{\tau=1}^{\iota} q_{0,i}^{\tau}$ where $\iota = \max\{\tau < t : i \in G^{\tau}\}$. Hence $(\sum_{\tau=1}^t q_i^{\tau} - \sum_{\tau=1}^{\iota} q_{0,i}^{\tau})$ units are shipped from grower i to the consolidation center since then. Besides, the global feasibility condition (31) continues to hold if we ship out all these units from the center in period t . Hence we want to exhaust them. The question is when to pick them up from the grower so that the holding cost is relatively small. If $h_i \geq h_0$, it makes sense to obey the current SHP schedule and keep possible inventory at the center until period t ; otherwise, we would rather postpone them to period t in the SHP solution and then deliver them immediately in the LHP solution (Figure 25).

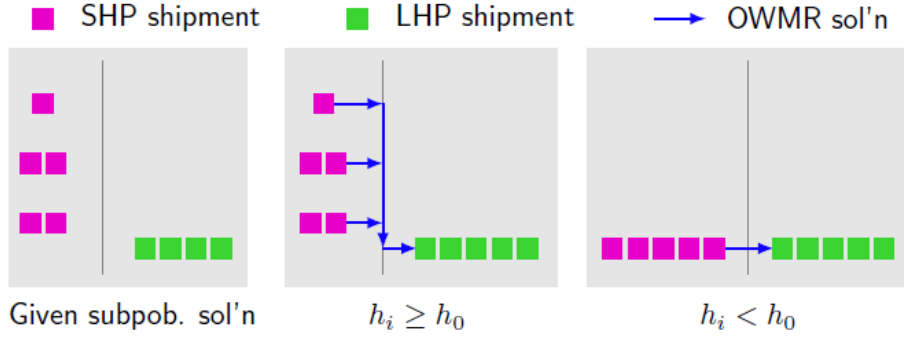


Figure 25: Recombination strategies for Case 1

Therefore, we update the solutions for periods $\max\{\iota < t : i \in G^{\iota}\} < \tau \leq t$ as

$$\hat{q}_{0,i}^t = \sum_{\tau=1}^t q_i^{\tau} - \sum_{\tau=1}^{\iota} q_{0,i}^{\tau}, \quad (33a)$$

$$\hat{q}_i^{\tau} = \begin{cases} q_i^{\tau}, & h_i \geq h_0 \\ 0, & o.w. \end{cases}, \quad \max\{\iota < t : i \in G^{\iota}\} < \tau < t, \quad (33b)$$

$$\hat{q}_i^t = \begin{cases} q_i^t, & h_i \geq h_0 \\ \sum_{\tau=\max\{\iota < t : i \in G^{\iota}\}+1}^t q_i^{\tau}, & o.w. \end{cases}. \quad (33c)$$

(33a) changes the original LHP shipping quantity for grower i to the total short-haul shipping volume since the last time i is chosen for long-haul delivery, which

ensures (32) in period t . (33b)-(33c) obey or shift the original SHP schedules depending on grower i 's holding cost rate. These maintain SHP demand satisfaction (30a) and in-time central demand arrivals, i.e., (31) remains valid for period $\tau = t$.

Meanwhile, depleting all the SHP quantities like (33a) requires us to possibly move some original LHP quantities up to period t . Hence we further adjust the LHP solution in adherence to demand satisfaction condition (30b): deduct by FIFO a total volume of $\tilde{q}_{0,i}^t = (\hat{q}_{0,i}^t - q_{0,i}^t)$ from later periods, i.e., iteratively calculate

$$\hat{q}_{0,i}^\tau = \max\{0, q_{0,i}^\tau - \tilde{q}_{0,i}^{\tau-1}\}, \quad (34a)$$

$$\tilde{q}_{0,i}^\tau = \tilde{q}_{0,i}^{\tau-1} - (q_{0,i}^\tau - \hat{q}_{0,i}^\tau), \quad (34b)$$

where $\tau = t + 1 \dots T$, until $\tilde{q}_{0,i}^\tau = 0$.

All the other quantities remain unchanged from period $t + 1$ on.

In Case 1, we do not split a single SHP shipment, only adjust LHP shipments which originally carry positive volumes, and reasonably account for inventory decisions. Hence we expect the resulting costs to be acceptable.

Case 2. $\sum_{\tau=1}^t q_i^\tau < \sum_{\tau=1}^t q_{0,i}^\tau$.

Given (31)-(32), this implies $\sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau < q_{0,i}^t$, i.e., the total SHP shipping volume since the last time grower i is chosen for long-haul delivery is insufficient for the current LHP solution in period t . The *late* volume equals $q_{0,i}^t - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau$, and by SHP feasibility condition (30a) is expected to arrive at the center in some period(s) after t . Hence we may well move up the corresponding late short-haul shipments. Specifically, we need to decide which shipment(s) to adjust, how much of the quantities to move up, as well as when to ship these quantities. We first observe the following.

Observation 4.6.1. *An optimal SHP solution never sends a shipment that only contains the demand ready in earlier periods.*

Proof. Assume that the condition is violated somewhere in an optimal SHP solution, e.g., grower i sends a shipment full of earlier demands in period t . We construct a feasible solution by moving up the entire shipment to the last period when grower i 's demand is ready but not shipped until period t . Repeating the construction for all growers $i \in G$ and periods $t = T \dots 1$ gives a solution which satisfies the condition. The solution incurs lower holding cost and no higher transportation cost for the SHP subproblem, which is a contradiction. \square

Observation 4.6.1 suggests that whenever Case 2 occurs for some grower i in iteration t , there could be only one late shipment for the LHP schedule in period t , which by FIFO is the first SHP shipment after period t . This is true since otherwise the first late shipment would only carry demands in periods $1 \dots t$, and the initial SHP solution should be suboptimal. On the other hand, this late SHP shipment may correspond to an arbitrary number of LHP shipments, so the same situation could arise again if we move up a volume of $(q_{0,i}^t - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau)$ to period t . Naively repeating such operations in future iterations would split the late SHP shipment many times and result in undesirable transportation cost. Hence it is somewhat subtle. We propose to fix this case in six steps.

Step 1. Deplete all the SHP shipments since the last time grower i is chosen for LHP shipping till period $(t - 1)$. That is, treat the central arrivals prior to period t as Case 1, and update the SHP \hat{q}_i^τ values with (33b) for periods $\max\{\iota < t : i \in G^\iota\} < \tau < t$.

Step 2. Search down the current list of SHP shipments for the first shipment after period t , i.e., find $\ell(t) := \min\{\tau > t : q_i^\tau > 0\}$. This locates the late SHP shipment.

Step 3. Search down the current list of LHP shipments for the last period when the late SHP shipment remains late or becomes just-in-time, i.e., find

$$e(t) := \max\{\iota : q_{0,i}^\iota > 0, \sum_{\tau=1}^\iota q_{0,i}^\tau \leq \sum_{\tau=1}^{\ell(t)} q_i^\tau, t \leq \iota \leq \ell(t)\}.$$

By Observation 4.6.1, the late SHP shipment contains some demand which is ready in period $\ell(t)$ and could not be scheduled for LHP shipping until then. Hence there exists a unique $e(t)$ as defined. Suppose $e(t) < \ell(t)$ and we move up the total late volume for periods $t \dots e(t)$, i.e. $\sum_{\tau=t}^{e(t)} q_{0,i}^\tau - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau$, from period $\ell(t)$ to period $e(t)$ in the SHP solution. Also, we postpone all these quantities to period $e(t)$ in the LHP solution. Immediately (31) holds for periods $\tau = t \dots e(t)$ and the partial C-OWMR infeasibility is resolved therein. Moreover, the definition of $e(t)$ implies $\min\{\iota > e(t) : i \in G^\iota\} \geq \ell(t)$, i.e., the next time grower i is chosen for LHP shipping after period $e(t)$ is no earlier than period $\ell(t)$. Therefore, the residual SHP shipment may temporarily stay in period $\ell(t)$ (subject to possible moves in future iterations) without causing C-OWMR infeasibility later on. On the other hand, we do not need to change the SHP shipment if $e(t) = \ell(t)$, so we exhaust it in the LHP solution. See Figure 26 for an illustration.

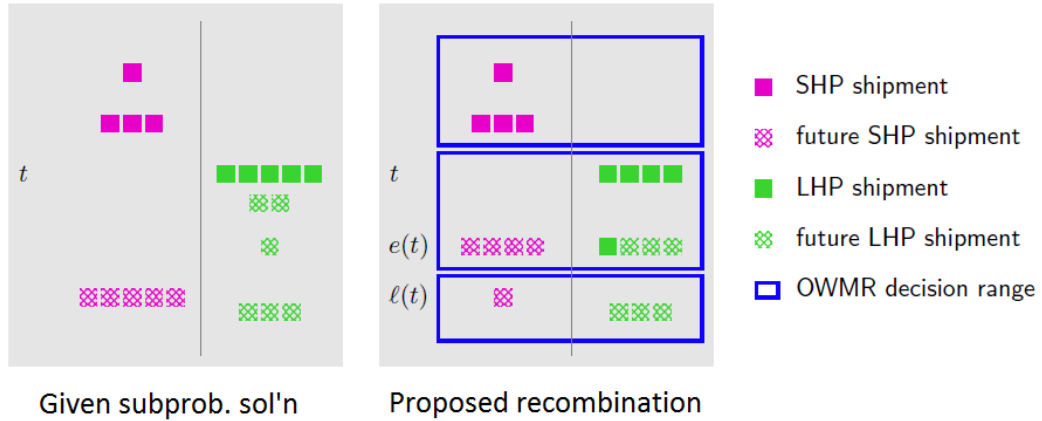


Figure 26: Recombination strategies for Case 2

Next we detail the formulae for these fix operations.

Step 4. Update SHP shipping quantities for periods $t \dots \ell(t)$:

$$\hat{q}_i^t = \begin{cases} q_i^t, & h_i \geq h_0, t < e(t) \\ \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau, & h_i < h_0, t < e(t) \\ q_{0,i}^t - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^{t-1} q_i^\tau, & h_i \geq h_0, t = e(t) \\ q_{0,i}^t, & h_i < h_0, t = e(t) \end{cases}, \quad (35a)$$

$$\hat{q}_i^{e(t)} = \begin{cases} \sum_{\tau=t}^{e(t)} q_{0,i}^\tau - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau, & t < e(t) < \ell(t) \\ q_i^{\ell(t)}, & e(t) = \ell(t) \end{cases}, \quad (35b)$$

$$\hat{q}_i^{\ell(t)} = q_i^{\ell(t)} - \sum_{\tau=t}^{e(t)} q_{0,i}^\tau + \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau, \quad \ell(t) > e(t), \quad (35c)$$

$$\hat{q}_i^\tau = 0, \quad t < \tau < e(t) \text{ or } e(t) < \tau < \ell(t). \quad (35d)$$

(35a) adjusts the SHP shipping quantity in period t depending on whether it overlaps with $e(t)$ and how grower i 's holding cost rate compares to the center. If $t < e(t)$, then the original units are shipped in addition to possible quantities shifted from periods $\max\{\iota < t : i \in G^\iota\} < \tau < t$ in Step 1, which equals q_i^t and $\sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau$ when $h_i \geq h_0$ and $h_i < h_0$, respectively. If $t = e(t)$, then \hat{q}_i^t also includes the original late volume in period t , which is $(q_{0,i}^t - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau)$ and yields the final quantities when added to the corresponding values for $t < e(t)$.

(35b) adjusts the SHP shipping quantity in period $e(t)$ when it is located later than t . Since $\ell(t)$ is the first period with positive shipping volume after period t in the original SHP solution, we have $q_i^{e(t)} = 0$ at the beginning of iteration t if $e(t) < \ell(t)$, and $\hat{q}_i^{e(t)}$ is exactly the original total late volume for periods $t \dots e(t)$, i.e. $\sum_{\tau=t}^{e(t)} q_{0,i}^\tau - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau$. Otherwise, all $q_i^{\ell(t)}$ units are depleted as mentioned above.

(35c) adjusts the SHP shipping quantity in period $\ell(t)$ when it is located later than $e(t)$. Clearly, $\hat{q}_i^{\ell(t)}$ is the residual quantity of the late SHP shipment, and equals

the original quantity minus the portion we move up to period $e(t)$.

(35d) gives the SHP shipping quantities in periods other than $t, e(t), \ell(t)$. Since there are no shipments before or after the iteration, we assign zero values.

Step 5. Update LHP shipping quantities for periods $t \dots e(t)$:

$$\hat{q}_{0,i}^t = \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau, \quad t < e(t), \quad (36a)$$

$$\hat{q}_{0,i}^\tau = 0, \quad t < \tau < e(t), \quad (36b)$$

$$\hat{q}_{0,i}^{e(t)} = \begin{cases} q_{0,i}^t, & e(t) = t \\ \sum_{\tau=t}^{e(t)} q_{0,i}^\tau - \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau, & t < e(t) < \ell(t) \\ q_i^{\ell(t)}, & e(t) = \ell(t) \end{cases} \quad (36c)$$

(36a) adjusts the LHP shipping quantity in period t when it does not overlap with $e(t)$. Since we delay the late volume for period t to period $e(t)$, the residual volume sent equals the total SHP volume in periods $\max\{\iota < t : i \in G^\iota\} < \tau \leq t$, which is the same before or after the iteration.

(36b) adjusts the LHP shipping quantities in periods $t < \tau < e(t)$. Since we aggregate the shipments to period $e(t)$, these all equal zero.

(36c) adjusts the LHP shipping quantity in period $e(t)$. If $e(t) = t$, it equals the original LHP quantity. If $t < e(t) < \ell(t)$, this is exactly the original total late volume for periods $t \dots e(t)$. If $e(t) = \ell(t)$, we set it to be the total volume of the original late SHP shipment based on central ZIO condition (32). This could result in an extra LHP volume of $(q_i^{\ell(t)} - \sum_{\tau=t}^{e(t)} q_{0,i}^\tau + \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau)$ in period $\ell(t)$, which we finally account for by revising future LHP quantities with LHP demand satisfaction condition (30b).

Step 6. If $e(t) = \ell(t)$, deduct by FIFO a total volume of $\tilde{q}_{0,i}^{\ell(t)} = (q_i^{\ell(t)} - \sum_{\tau=t}^{e(t)} q_{0,i}^\tau + \sum_{\tau=\max\{\iota < t: i \in G^\iota\}+1}^t q_i^\tau)$ from the LHP quantities in later periods, i.e. iteratively

update

$$\hat{q}_{0,i}^\tau = \max\{0, q_{0,i}^\tau - \tilde{q}_{0,i}^{\tau-1}\}, \quad (37a)$$

$$\tilde{q}_{0,i}^\tau = \tilde{q}_{0,i}^{\tau-1} - (q_{0,i}^\tau - \hat{q}_{0,i}^\tau), \quad (37b)$$

where $\tau = (\ell(t) + 1) \dots T$, until $\tilde{q}_{0,i}^\tau = 0$.

All the other quantities remain unchanged from period $e(t) + 1$ on.

In Case 2, we split the late SHP shipment at most twice to ensure sufficient central arrivals for LHP shipping, so the extra short-haul transportation cost should be moderate. The adjusted LHP shipments take place in periods with positive original LHP shipping volume, so hopefully the long-haul transportation cost is under control, too. As for the holding cost, we will count on bounding it by somehow relating the revised holding cost rates and the actual amount.

Algorithm 1 outlines the construction procedure for all possible situations. Below we formally prove its feasibility and computational efficiency.

Proposition 4.6.2. *Algorithm 1 generates a feasible solution to C-OWMR.*

Proof. The global feasibility conditions (30) are guaranteed if (30a)-(30b) and (31) are valid from iteration to iteration. This is evident in the way we construct the C-OWMR solution. Specifically, the initial subproblem solutions automatically satisfy (30a)-(30b); and in each iteration $t = 1 \dots T$ where the shipping quantities are updated, we have

- For Case 1: (33a) extends the link condition (31) to period $\tau = t$. (33b)-(33c) imply $\sum_{\tau=1}^T \hat{q}_i^\tau = \sum_{\tau=1}^T q_i^\tau$ and $\hat{q}_i \geq 0$, which by FIFO guarantee short-haul demand satisfaction condition (30a). (34), (33a) and the definition of $\tilde{q}_{0,i}^t$ give $\sum_{\tau=1}^T \hat{q}_{0,i}^\tau = \sum_{\tau=1}^T q_{0,i}^\tau$ and $\hat{q}_{0,i} \geq 0$, so the long-haul demand satisfaction condition (30b) is also true.

Algorithm 1 Subproblem-based OWMR heuristic

```
1: Input: SHP and LHP solutions, i.e. optimal  $q$  values in Models (27)-(28)
2:  $G^\tau \leftarrow \{i \in G : q_{0,i}^\tau > 0\}$ ,  $\tau = 1 \dots T$ 
3: for  $t = 1 \dots T$  do
4:   for  $i \in G^t$  do
5:     if  $\sum_{\tau=1}^t q_i^\tau \geq \sum_{\tau=1}^t q_{0,i}^\tau$  then ▷ Case 1
6:       for  $\tau = (\max\{\iota < t : i \in G^\iota\} + 1) \dots t$  do
7:         Update shipping quantities with (33)
8:       end for
9:       for  $\tau = (t + 1) \dots T$  do
10:        Update LHP shipping quantities with (34)
11:         $G^\tau \leftarrow G^\tau \setminus \{i\}$  if  $\hat{q}_{0,i}^\tau = 0$ 
12:      end for
13:     else ▷ Case 2
14:       for  $\tau = (\max\{\iota < t : i \in G^\iota\} + 1) \dots (t - 1)$  do
15:        Update SHP shipping quantities with (33b)
16:      end for
17:       $\ell(t) \leftarrow \min\{\tau > t : q_i^\tau > 0\}$ 
18:       $e(t) \leftarrow \max\{\iota : q_{0,i}^\iota > 0, \sum_{\tau=1}^\iota q_{0,i}^\tau \leq \sum_{\tau=1}^{\ell(t)} q_i^\tau, t \leq \iota \leq \ell(t)\}$ 
19:      for  $\tau = t \dots \ell(t)$  do
20:        Update SHP shipping quantities with (35)
21:      end for
22:      for  $\tau = t \dots e(t)$  do
23:        Update LHP shipping quantities with (36)
24:         $G^\tau \leftarrow G^\tau \setminus \{i\}$  if  $\hat{q}_{0,i}^\tau = 0$ 
25:      end for
26:      if  $e(t) = \ell(t)$  then
27:        for  $\tau = (e(t) + 1) \dots T$  do
28:          Update LHP shipping quantities with (37)
29:           $G^\tau \leftarrow G^\tau \setminus \{i\}$  if  $\hat{q}_{0,i}^\tau = 0$ 
30:        end for
31:      end if
32:    end if
33:  end for
34: end for
35: Output: Feasible  $q$  values for Model (15), where  $q_0^\tau = \sum_{i \in G} q_{0,i}^\tau$ ,  $\tau = 1 \dots T$ 
```

- For Case 2: (33b) and (35) imply $\sum_{\tau=1}^T \hat{q}_i^\tau = \sum_{\tau=1}^T q_i^\tau$ and $\hat{q}_i \geq 0$. (36)-(37) imply $\sum_{\tau=1}^T \hat{q}_{0,i}^\tau = \sum_{\tau=1}^T q_{0,i}^\tau$ and $\hat{q}_{0,i} \geq 0$. By FIFO subproblem demand satisfaction conditions (30a)-(30b) hold. (35)-(36) guarantee (31) for periods $\tau = t \dots e(t)$.

By induction, (30a)-(30b) and (31) hold throughout the procedure. □

Proposition 4.6.3. *Algorithm 1 runs in $\mathcal{O}(|G|T^2)$ time.*

Proof. The initial construction of G^τ , i.e. Line 2 takes $\mathcal{O}(|G|T)$ time. The outer *for* loop involves T iterations. Whenever Case 1 arises for a grower, Lines 6-12 take $\mathcal{O}(T)$ time to finish, same for Lines 14-16 and Lines 19-30 in Case 2. Given FIFO, periods $\ell(t)$ and $e(t)$ can be found in $\mathcal{O}(\log T)$ time via binary search for each grower in each iteration. Since T dominates $\log T$, the per iteration runtime for each grower is $\mathcal{O}(T)$. The total runtime is then $\mathcal{O}(|G|T^2)$. □

The complexity can be reduced to $\mathcal{O}(|G|T \log T)$ by using priority queues. At the beginning of iteration t , the partially fixed solutions have been finalized for periods $1 \dots \max\{t < t : i \in G^t\}$. Therefore, if we keep a list of these points for each grower, then it suffices to truncate the list and move down in future iterations. The length of the list is $\mathcal{O}(T)$, so the total runtime of Lines 6-12, 14-16 and 19-30 is $\mathcal{O}(T)$ for each grower as the procedure ends. Meanwhile, the search of $\ell(t)$ and $e(t)$ in Lines 17-18 still takes $\mathcal{O}(\log T)$ for each grower in each iteration because of the updated solutions. Hence the total runtime becomes $\mathcal{O}(|G|T \log T)$, which is $\mathcal{O}(\log T)$ times that of [119]’s procedure for the standard OWMR problem with fixed setup costs. Again, the additional complexity is due to our efforts in identifying $e(t)$ and $\ell(t)$, which are unnecessary if ZIO holds for all facilities.

4.6.2 Approximation Ratio

Next we demonstrate the efficacy of Algorithm 1’s output. We bound the total cost of the final C-OWMR solution by comparing each cost element with the objective

values of the initial SHP and LHP solutions.

Proposition 4.6.4 (Bounded short-haul transportation cost). *Let V^a, V^b be the total short-haul transportation cost incurred by Algorithm 1's final solution and the initial SHP solution, respectively, then $V^a \leq 2V^b$.*

Proof. We do not change the quantities of SHP shipments which arrive in time so their transportation costs remain the same unless they are merged after possible shifts. Any late shipment is split into at most two suborders in Case 2. Whenever this happens, the delivered part leaves the system, whereas the residual part becomes early in future iterations so we do not further split it. Each suborder costs no more than the original order by monotonicity of $c_i(\cdot)$. Possible merging of shipments caused by shifting and splitting does not increase the transportation cost by definition of concave batch costs. The inequality then follows. \square

Proposition 4.6.5 (Bounded long-haul transportation cost). *Let W^a, W^b be the total long-haul transportation cost incurred by Algorithm 1's final solution and the initial LHP solution, respectively, then $W^a \leq 2W^b$ for the R-OWMR problem with FTL costs.*

Proof. Let \mathcal{N}_F^b and \mathcal{N}_F^a be the total numbers of full long-haul trucks before and after applying Algorithm 1, respectively, and $\mathcal{N}_L^b, \mathcal{N}_L^a$ the numbers of partially filled trucks. We have $W^a = c_F(\mathcal{N}_F^a + \mathcal{N}_L^a)$ and $W^b = c_F(\mathcal{N}_F^b + \mathcal{N}_L^b)$. The inequality $W^a \leq 2W^b$ is true because

- $\mathcal{N}_F^a \leq \mathcal{N}_F^{\max} \leq \mathcal{N}_F^b + \mathcal{N}_L^b$, where $\mathcal{N}_F^{\max} = \lceil \sum_{i \in G} \sum_{\tau=1}^T d_i^\tau / K_F \rceil$ is the maximum possible number of full long-haul trucks in any feasible LHP solution, which is further bounded above by the total number of trucks dispatched in the initial LHP solution.
- $\mathcal{N}_L^a \leq \mathcal{N}_F^b + \mathcal{N}_L^b$: We only modify long-haul shipments in periods with positive initial LHP shipping quantities. The FTL cost structure ensures that at most

one partially filled truck is dispatched in each period, which costs the same as any truck initially dispatched. \square

Proposition 4.6.6 (Bounded holding cost). *Let H^a , H^b be the total inventory cost incurred by Algorithm 1's final solution and that evaluated with $h_{i,i}, h_{0,i}$ in the initial subproblem solutions, respectively, then $H^a \leq 2H^b$.*

Proof. Consider an arbitrary (split) demand d_m which is a portion of d_i^s , shipped in a unique pair of SHP-LHP periods (τ, t) in the initial subproblem solutions, and also shipped in a unique pair of SHP-LHP periods in the final C-OWMR solution. Figure 27 depicts all possible distribution flows based on the iterative decisions.

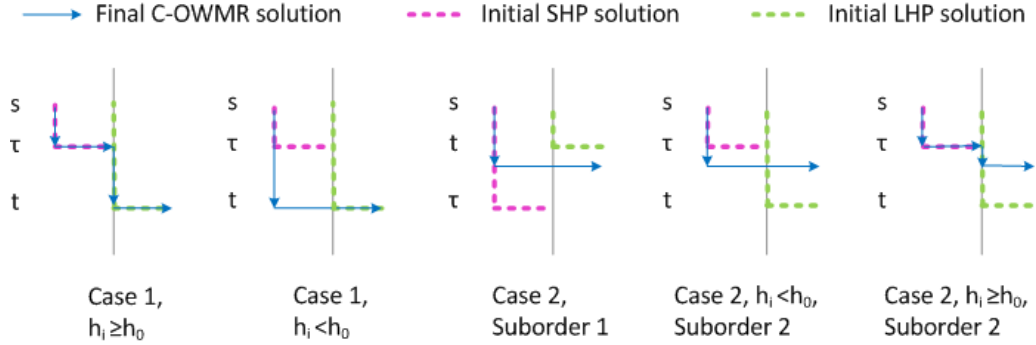


Figure 27: Patterns of holding cost changes

- Case 1, $h_i \geq h_0$: $H^a(d_m) = h_i(\tau - s)d_m + h_0(t - \tau)d_m = 2h_{i,i}(\tau - s)d_m + 2h_{0,i}(t - \tau)d_m \leq 2h_{i,i}(\tau - s)d_m + 2h_{0,i}(t - s)d_m = 2H^b(d_m)$.
- Case 1, $h_i < h_0$: $H^a(d_m) = h_i(t - s)d_m = 2h_{i,i}(\tau - s)d_m + 2h_{0,i}(t - \tau)d_m \leq 2h_{i,i}(\tau - s)d_m + 2h_{0,i}(t - s)d_m = 2H^b(d_m)$.
- Case 2, first suborder, i.e. the portion we move up from the original late shipment to period $e(t)$: $H^a(d_m) = h_i(e(t) - s)d_m \leq h_i(\tau - s)d_m \leq 2H^b(d_m)$.
- Case 2, $h_i < h_0$, second suborder, i.e. the residual portion of the original late shipment after we split it: $H^a(d_m) \leq h_i(\tau - s)d_m + h_i(t - \tau)d_m \leq 2h_{i,i}(\tau - s)d_m + 2h_{0,i}(t - s)d_m \leq 2H^b(d_m)$.

- Case 2, $h_i \geq h_0$, second suborder, i.e. the residual portion of the original late shipment after we split it: $H^a(d_m) \leq h_i(\tau - s)d_m + h_0(t - \tau)d_m \leq 2h_{i,i}(\tau - s)d_m + 2h_{0,i}(t - s)d_m \leq 2H^b(d_m)$.

Summing $H^a(d_m)$ over all the demands, we obtain $H^a \leq 2H^b$. □

Finally, we show the overall solution quality, which improves [89]’s worst-case ratio from 3.6 to 2 for R-OWMR-FTL, the “reversed” OWMR problem with FTL costs.

Proposition 4.6.7. *Algorithm 1, combined with [79]’s DP for SHP and [16]’s LP reformulation for LHP, gives an efficient 2-approximation for the “reversed” OWMR problem with FTL costs.*

Proof. Let $Z^b(P)$ and $Z^a(P)$ be the objective values of problem P before and after Algorithm 1 applies, respectively. Then

$$\begin{aligned}
Z^a(\text{R-OWMR-FTL}) &= V^a + W^a + H^a \\
&\leq 2(V^b + W^b + H^b) && \text{by Propositions 4.6.4-4.6.6} \\
&= 2(Z^b(\text{SHP}) + Z^b(\text{LHP})) && \text{by (29)} \\
&= 2(Z^*(\text{SHP}) + Z^*(\text{LHP})) \\
&\leq 2Z^*(\text{R-OWMR-FTL}) && \text{by Proposition 4.5.6. } \square
\end{aligned}$$

4.6.3 Tightness

The above approximation ratio equals the upper bound given by [119]’s algorithm for fixed setup costs. Since the center only sends shipments when there are positive deliveries in the initial LHP solution, the final long-haul transportation cost would not increase should the transportation be uncapacitated. [119] thus argued that the a posteriori guarantee of their procedure is superior to 2. Here we show that Algorithm 1, though in a similar spirit to [119], cannot do better than 2-optimality in the worst case. This gap is the price we shall pay to account for batch capacities.

Example 8 (Tight approximation ratio). Consider an instance with $|G|$ growers and $T = |G| + 1$ periods. Table 8 lists the demand and cost assumptions, where $\epsilon > 0$ is a small positive value.

Table 8: Worst-case example

Parameters: $T = G + 1, h_1 < h_2 < \dots < h_0, h_i \ll c_{i,F} \forall i \in G \cup \{0\}$						
Demand	Period 1	Period 2	Period 3	Period 4	...	Period T
Grower 1	ϵ	$K_F - \epsilon$				
Grower 2		2ϵ	$K_F - 2\epsilon$			
Grower 3			3ϵ	$K_F - 3\epsilon$		
\vdots				\ddots	\ddots	
Grower $ G $						$K_F - G \epsilon$

For each grower i , the total demand is $\sum_{s=1}^T d_i^s = d_i^i + d_i^{i+1} = K_F$, and the costs satisfy $h_i \ll c_{i,F}$. Hence the optimal SHP solution is to delay period i 's demand and then ship a full truck in period $i + 1$, i.e. $q_i^{i+1} = K_F, \forall i \in G$. The consequent SHP objective value is

$$Z^*(\text{SHP}) = \sum_{i=1}^{|G|} (c_{i,F} + ih_{i,i}\epsilon) = \sum_{i=1}^{|G|} (c_{i,F} + 0.5ih_i\epsilon). \quad (38)$$

On the other hand, the LHP demand arrivals in each period are such that

$$\left(\sum_{i=1}^{|G|} d_i^s \right) \bmod K_F = (d_{s-1}^s + d_s^s) \bmod K_F = (K_F + \epsilon) \bmod K_F = \epsilon, \forall s = 2 \dots |G|.$$

Since $h_{0,i} = 0.5h_i < 0.5h_{i+1} = h_{0,i+1} \ll c_{0,F} \forall i \in G$, the optimal LHP solution is to ship a full truck in each period $t = 2 \dots T$, keep ϵ volume of item 1 in periods $1 \dots |G|$ while holding extra ϵ volume of items $1 \dots \tau - 1$ in periods $\tau = 2 \dots |G|$. That is, $q_{0,i}^{i+1} = K_F - (i+1)\epsilon, q_{0,i+1}^{i+1} = (i+1)\epsilon$ for all $i \in G$ and $q_{0,1}^T = |G|\epsilon$. The consequent LHP objective value is

$$Z^*(\text{LHP}) = |G|(c_{0,F} + h_{0,1}\epsilon) + \sum_{\tau=2}^{|G|} \sum_{i=1}^{\tau-1} h_{0,i}\epsilon = |G|(c_{0,F} + 0.5h_1\epsilon) + \sum_{i=1}^{|G|-1} 0.5(|G| - i)h_i\epsilon. \quad (39)$$

Now we apply Algorithm 1. Grower 1's SHP shipment arrives in time for period 2, hence we deplete it, i.e. $\hat{q}_1^2 = \hat{q}_{0,1}^2 = K_F$. Meanwhile, the LHP solution plans a shipment for grower 2 of volume 2ϵ , which does not arrive until period 3 in the SHP solution. Hence we split this late SHP shipment: 2ϵ to be shipped in period 2, whereas the residual $(K_F - 2\epsilon)$ stays for period 3. The adjusted long-haul shipping volume in period 2 is then $K_F + 2\epsilon > K_F$. This results in an extra truck dispatched from grower 2 as well as the center. As the procedure iterates, we see the following pattern in the final C-OWMR solution.

- Grower i 's short-haul shipments are conducted by two partially filled trucks:
 $\hat{q}_i^i = i\epsilon$, $\hat{q}_i^{i+1} = K_F - i\epsilon$, $\forall i = 2 \dots |G|$.
- Central long-haul shipments exceed the capacity of an FTL except the last period: $\hat{q}_0^t = K_F - (t-1)\epsilon + t\epsilon = K_F + \epsilon$, $\forall t = 3 \dots |G|$. The end period ships $\hat{q}_0^T = K_F - |G|\epsilon$.
- No inventory is held in periods $2 \dots T$.

The consequent C-OWMR objective value is

$$Z^a(\text{R-OWMR-FTL}) = c_{1,F} + 2 \sum_{i=2}^{|G|} c_{i,F} + (2|G| - 1)c_{0,F} + h_1\epsilon. \quad (40)$$

In fact, the optimal C-OWMR solution is to not split the SHP shipments given the holding cost assumptions. The consequent objective value is

$$Z^*(\text{R-OWMR-FTL}) = \sum_{i=1}^{|G|} (c_{i,F} + ih_i\epsilon) + |G|c_{0,F}. \quad (41)$$

Combining (38)-(41), we have

$$Z^a(\text{R-OWMR-FTL}) \rightarrow 2(Z^*(\text{SHP}) + Z^*(\text{LHP})) \rightarrow 2Z^*(\text{R-OWMR-FTL})$$

as $|G| \rightarrow \infty$ and $\epsilon \rightarrow 0$, where a proper value of ϵ may be $1/|G|^2$.

For the concerned instance, the above comparison suggests that we may obtain a posteriori better solutions with other recombination rules in Algorithm 1. However, this relies on the holding cost assumptions and does not imply a guarantee for an arbitrary case.

4.7 Asymptotic Analysis

An interesting question about the two-echelon system is how it will behave as the number of growers increases, which may give insights on whether or when it is worth integrating both echelons. In this section, we generalize a result in [79] to OWMR problems with concave and near-concave batch costs as defined in Section 4.5.

Definition 4.7.1 (Decentralized system, [79]). The OWMR system could be managed in a decentralized fashion, in which each of the individual members makes its own decisions based on its local demands and costs. In this setting, grower i observes its demands d_i^t , $t = 1 \dots T$, and minimizes its total transportation and inventory costs with $c_i(\cdot)$ and h_i . The grower then ships its cost-minimizing plans, q_i^t , $t = 1 \dots T$, to the consolidation center. As a result, the center faces demands $d_0^t = \sum_{i \in G} q_i^t$, $t = 1 \dots T$, and finds its corresponding cost-minimizing shipments q_0^t with $c_0(\cdot)$ and h_0 .

Theorem 4.7.2 (Theorem 7.1, [79]). *Let demands d_i^t be i.i.d. with a mean of $E[d_i^t] = \bar{d} > 0$, and let Z^* and Z^{DC} denote the cost of the optimal and decentralized solutions, respectively. With probability one,*

$$\lim_{|G| \rightarrow \infty} \frac{Z^{DC} - Z^*}{Z^*} = 0.$$

In other words, the decentralized approach is asymptotically optimal for our OWMR problem with FTL costs and $h_0 \leq h_i \forall i \in G$.

Theorem 4.7.2 results from regeneration properties under FTL costs and the assumption of $h_0 \leq h_i \forall i \in G$. We first prove the regeneration property for the

decentralized system under concave batch ordering costs and arbitrary static linear holding costs.

Proposition 4.7.3 (Regeneration property for decentralized C-OWMR). *Proposition 4.5.5 holds for all the facilities in a decentralized C-OWMR system.*

Proof. By Definition 4.7.1, the decentralized short-haul subproblem and long-haul subproblem can be modeled by (SHP) with $h_{i,i} = h_i$ and (LHP) with $h_{0,i} = h_0$, respectively, $\forall i \in G$. The claim thus follows by Observation 4.5.4. \square

As a byproduct, we obtain a “partial” regeneration property for the centralized system.

Corollary 4.7.4. *In a centralized C-OWMR system, Proposition 4.5.5 holds at the consolidation center and the growers such that $h_i \geq h_0, i \in G$.*

Proof. The proof of Proposition 4.5.5 applies to the center given any demand arrivals. For any i with $h_i \geq h_0$, it is never better to keep more inventory than K_F at the grower since otherwise a full batch truck can be move up without incurring additional costs. The remainder of the proposition follows by [100]’s extreme flow arguments since the forward arcs associated with i incur non-negative costs in their regeneration subtree. \square

Corollary 4.7.4 is unnecessary to develop asymptotic results, but explains why our problem with FTL costs cannot be solved to optimality with [79]’s DP regardless of computational tractability: Proposition 4.5.5 could be violated at growers i with $h_i < h_0$ in the centralized system; also see Example 7.

We now present the main result in this section, the asymptotic optimality of decentralized C-OWMR systems.

Proposition 4.7.5. *The decentralized solution is asymptotically optimal for OWMR problems with concave batch ordering costs under finite-horizon dynamic demand, static parameters, and linear holding costs.*

Proof. See [79] Theorem 7.1. Proposition 4.7.3 guarantees all the conditions therein. Note that the proof only requires the regeneration property at the center, which bounds from above extra central inventory volume and long-haul batch numbers at an order lower than $|G|$ in the decentralized solution. \square

We further generalize the result to near-concave batch costs with the transformation technique in Section 4.4.

Proposition 4.7.6. *The decentralized solution is asymptotically optimal for two-echelon distribution problems that can be approximated by the OWMR problems in Proposition 4.7.5 with a transformation ratio $\alpha \sim o(|G|)$ in Proposition 4.4.1.*

Proof. Let Z^* be the long-haul cost of the global optimum, Z^{DC} be $f_P(x_Q^*)$ in Proposition 4.4.1, where P and Q are the decentralized LHP with the true transportation cost function and that approximated by concave batch costs, respectively. By Proposition 4.7.3, the upper bound on Z^{DC} and lower bound on Z^* are valid in [79] Theorem 7.1. Since $f_P(x_Q^*) \leq \alpha f_P(x_P^*) \leq \alpha f_P(x_Q^*)$, the limit in Theorem 4.7.2 goes to zero if $\alpha \sim o(|G|)$. \square

Conceptually, Proposition 4.7.6 indicates that the integrated problem is as difficult as the subproblems when $|G|$ is very large, since the decentralized solution is arguably good and there is little incentive to coordinate both echelons.

Finally, the asymptotic optimality of decentralized C-OWMR systems under mixed truckload costs is straightforward based on previous discussions.

Proposition 4.7.7. *Asymptotically optimal solutions are attainable in polynomial time for our reversed OWMR problem with any mixed truckload cost structure described in Section 4.1 for long-haul transportation, and short-haul FTL or FTL+linearized LTL costs.*

Proof. The SHP of this kind can be solved efficiently to optimality by Observation 4.5.1 and the DP algorithms listed in Table 7. By Observation 4.5.2, the decentralized

LHP under mixed truckload costs can be approximated to 2-optimality with single-item lot sizing problems under PWL concave batch costs, which are solvable with the same algorithms. Proposition 4.7.6’s proof applies. \square

4.8 Extensions

In this section, we extend some results from the basic R-OWMR problem to more complicated settings. Section 4.8.1 relates our study to the literature in an opposite direction. Section 4.8.2 discusses asymptotic optimality of the decentralized OWMR system for perishable products with a fixed lifetime.

4.8.1 Regular Modified All-Unit Discounts

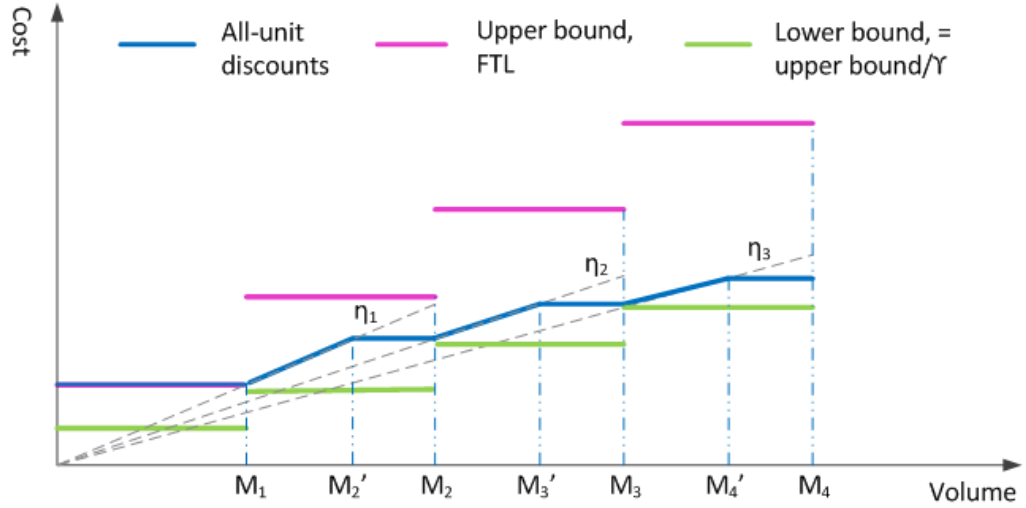
[18] developed a polynomial-time DP for the LSP with two types of *regular modified all-unit discounts*, which are special all-unit discounts where the sections of the cost function with positive slopes have identical length and so do the flat sections. Mathematically, this cost structure can be described with (18) by setting $M_1 = M_2 - M_1 = \dots = M_n - M_{n-1}$ and $M'_1 = M'_2 - M_1 = \dots = M'_n - M_{n-1}$, where $M'_{k+1} := \{x : \eta_k x = \eta_{k+1} M_{k+1}\}$ are the cost breakpoints between consecutive non-flat and flat sections. We demonstrate constant-ratio approximations and asymptotically optimal solutions for the corresponding OWMR problem, which complements the work in Section 4.4.

Proposition 4.8.1. *The OWMR with regular modified all-unit-discounted transportation costs can be transformed to C-OWMR with $\alpha = \gamma$ if there exists a constant γ such that*

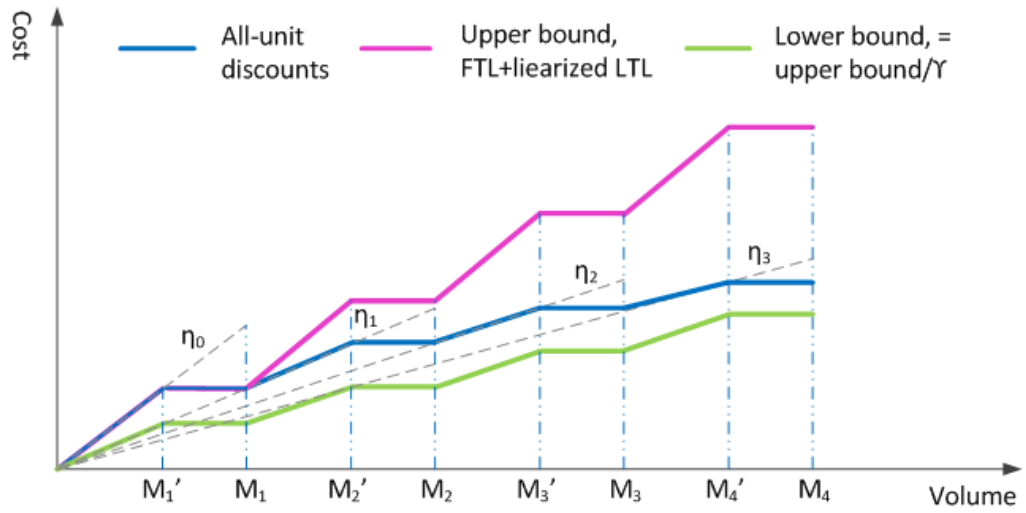
- *when the discount cost function starts with a flat section whose length equals the sum of a normal flat section and a non-flat section, the positive slopes can be bounded as $\eta_k \geq \frac{k+1}{\gamma^k} \eta_1, \forall k \geq 2$;*

- when the discount cost function starts with a non-flat section, the positive slopes can be bounded as $\eta_k \geq \frac{1}{\gamma}\eta_1, \forall k \geq 2$.

Proof. See Figure 28. The first type can be transformed to our problem with FTL costs, where $K_F = M_1, c_F = \eta_1 M_1$ and $\alpha = \gamma$. The second type can be transformed to our problem with FTL+linearized LTL, where $K_F = M_1, c_F = \eta_1 M_1, c_L = \eta_0 M_1'$, and $\alpha = \gamma$.



(a) First type to FTL costs



(b) Second type to FTL+linearized LTL costs

Figure 28: Transformations of regular modified all-unit discounts to concave batch costs

In either case, the target cost structure is PWL concave batch costs, and the

bounds are valid by the restriction of γ . \square

A direct result of Proposition 4.8.1 and Proposition 4.6.7 is a 2γ -approximation for the first type of regular modified all-unit discounts. Intuitively, if the largest possible volume discount is less than 50% and a normal flat section is not too long compared to a non-flat section in the first type of costs, then $\gamma = 2$, which combined with Algorithm 1 gives a 4-approximation.

[18]’s DP was designed for the second type of regular all-unit discounts based on properties similar to Proposition 4.5.5. The authors intended to disprove them for the first type with a counterexample. However, we note that the example violates the definition of the cost function and the DP works in this case as well, which results in the following.

Proposition 4.8.2. *Asymptotically optimal solutions are available in polynomial time for the OWMR with regular modified all-unit discounted transportation costs.*

Proof. Decentralize the system, solve the SHP with [18]’s DP, and then approximate the induced LHP with LSP under concave batch ordering costs as in Proposition 4.8.1. Proposition 4.7.6’s proof applies. \square

We comment that the short-haul echelon of this OWMR is a special case of [44], whereas the long-haul echelon is different since [44] considered incremental discounts on \mathbb{R}_+ . Nevertheless, Proposition 4.8.1 and Proposition 4.8.2 can be extended to variants of [44]’s long-haul echelon e.g. soft-capacitated incremental discounts or near-concave batch costs.

4.8.2 Perishable Products

The following statement extends Section 4.7’s asymptotic results to perishable goods.

Proposition 4.8.3. *The decentralized system is asymptotically optimal for OWMR problems under near-concave batch ordering costs and perishable products with an identical fixed lifetime.*

Proof. Essentially, we need to 1) argue subproblem regeneration properties and solvability for C-OWMR with perishable products, and 2) bound the possible extra costs of a decentralized solution when perishability exists.

Redefine a regeneration point t such that no commodity due in periods $t+1, \dots, T$ are shipped in periods $1, \dots, t$. These regeneration points separate the decision over the entire planning horizon to subintervals. Enlightened by equations (25)-(26), we give a network representation of each subproblem in Figure 29, where (k, τ) refers to the commodity ready in period τ at facility $k \in G \cup \{0\}$. The network is finite since the number of batches is bounded by $\lceil \sum_{i \in G, \tau \leq t} d_i^\tau / K_F \rceil, \forall t = 1 \dots T$.

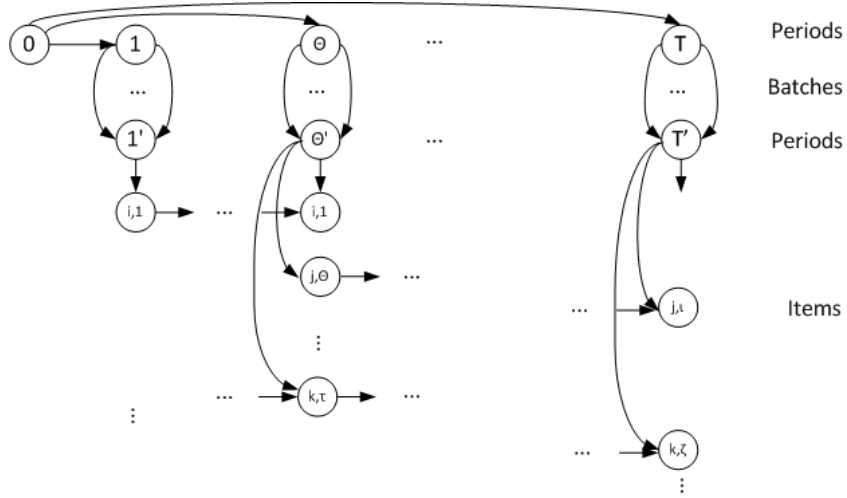


Figure 29: A network flow representation of subproblems for decentralized perishable C-OWMR

- **Bounds on transportation costs:** In each regeneration interval, there is at most one period with a partially filled batch. This is true because otherwise we can construct an extreme solution that satisfies this condition by circulating flows in the above network. These feasible flows exist since the above graph is connected

for the “items” part within a redefined regeneration interval. Therefore, the maximum possible total extra transportation cost is less than or equal to TK_F for each subproblem.

- Bounds on inventory costs: Since transportation cost parameters are static and the holding costs for all (fresh) items are the same in each decentralized subproblem, [90]’s perturbation still applies, which indicates $I^t < K_F, \forall t$.

Recall that in a decentralized system, $h_{0,i} = h_{0,j} = h_0 \forall i, j \in G$. Given a fixed lifetime $\theta \in \mathbb{Z}_{++}$, the LHP problem is equivalent to a non-perishable multi-item lot-sizing problem with threshold central holding costs

$$h_{0,i}^t = \begin{cases} 0, & \text{if } t < s_i \\ h_0, & \text{if } s_i \leq t < s_i + \theta, \\ \infty & \text{o.w.} \end{cases}$$

where each item i is specified by its ready period s_i . The new holding costs are time-varying, but satisfy the nonspeculative and item-wise dominant assumptions in [16]. Hence we can solve the subproblems under FTL costs in polynomial time with their LP reformulation. For concave batch costs, we could only solve the subproblems in pseudo-polynomial time with directly available methods, but the asymptotic optimality still holds. □

4.9 Conclusions

In this chapter, we study a two-echelon distribution problem which consists of multiple growers, a consolidation center and an external retailer. Assuming direct shipping at both echelons, this results in an OWMR problem reversed in time flows. Unlike the classical literature which focus on fixed setup costs, we consider various combinations of shipping options, including FTL, LTL and courier services. The transportation cost is a PWL non-convex function of the shipping volume, and repeats itself when

the volume exceeds the capacity of a truck. We first bound these mixed truckload costs with fixed setup costs and/or quantity discounts, which give us a bunch of approximation results by applying existing OWMR methods. We then decompose the system to single-echelon lot sizing problems with batch capacities and derive structural properties of the subproblems. For the special case of FTL costs, we devise a simple 2-approximation algorithm based on the optimal subproblem solutions, and demonstrate tightness of the approximation ratio with a worst-case example. We further analyze asymptotic optimality of a decentralized approach under a wider range of transportation costs approximable by concave batch ordering costs. Finally, we extend some results to quantity discounts as well as perishable products. All the techniques fit conventional OWMR problems under the same cost structures.

We conclude the chapter with ongoing and potential future research. A natural question is whether we can modify Algorithm 1 to accommodate more complicated volume-dependent transportation costs, e.g. truckload costs involving both LTL and courier services, repetitive PWL costs with an arbitrary number of breakpoints, and/or concave batch ordering costs. While the output solution of Algorithm 1 remains feasible for the OWMR problem under any transportation cost structure, the issue now is we may not be able to bound the resulting long-haul transportation cost since it depends on not only the number of batch trucks sent in each period, but also how much volume each truck carries. Therefore, we are motivated to generalize the approach by more judiciously recombining subproblem shipments. In the case of FTL+linearized LTL costs, for instance, we may further split the suborders related to the late shipment in a period, say the pair in periods $e(t)$ and $\ell(t)$ as defined earlier, at the first period with full trucks upwards and downwards, respectively. Enlightened by [14] and [90], we can prove a similar regeneration property for LHP under concave batch costs, which guarantees the existence of those further split periods, bounds the final transportation cost by 4 times that of the initial subproblem solutions, and thus

gives a 4-approximation for the OWMR problem. More generally, we can utilize a similar logic to get the following results:

- For PWL concave batch costs with a fixed number of breakpoints, constant approximations are attainable.
- For generic concave batch costs, approximations are attainable within $(\log T)$ -optimality.

We are yet to give tight worst-case examples for these cost structures. On the other hand, our attempts pose additional questions on the LHP subproblems under the corresponding transportation costs. When the growers' holding cost rates differ and are lower than the central holding cost rate, those problems can be viewed as variants of [16]'s MIMV problem, but the LP reformulation is no longer equivalent to the original MIP model because of the revision in the objective function. To the best of our knowledge, neither solvability nor approximability has been studied for such problems, though a naive combination of existing methods e.g. [15], [90] may perform well experimentally. It also interests us to customize Algorithm 1 in the setting of infinite horizon and constant demand rates, as [118] did for fixed setup costs. Other directions may include extensions to demand shortages and/or non-linear holding costs, computational studies of the solution approaches with empirical data, etc.

CHAPTER V

CONCLUSION

In this thesis, we study three multi-period supply chain optimization problems, and provide some flavors of methodologies to handle these and similar problems.

Chapter 2 aims at understanding the long-term implications of training, hiring and promotion decisions in large public health care systems. We take a linear programming approach to model the problem with a single worker type in an infinite horizon frame. Under mild technical and realistic assumptions, we derive common-sense structural properties of the model, and further use them to prove the optimality of a simple lookahead policy. This provides a theoretical foundation for relevant workforce management policies implemented in practice. By simulating lookahead policies with real data, we demonstrate effectiveness and robustness of the approach in various scenarios including uncertain demand growth. Our next steps are 1) to directly model and analyze the system's uncertainty; 2) to apply the work to crucial yet less studied health systems e.g. the U.S. pediatric workforce, health workforce providers in rural areas, etc.

Chapter 3 focuses on the economic benefits and technical challenges of integrating two classical problems in transportation and logistics: the inventory routing problem and the freight consolidation problem. Since each problem can be hard itself with real-world complications, they are typically solved separately by individual supply chain functions. We propose a large-scale mixed-integer programming model for the integrated problem, and develop an iterative solution framework with a decomposition procedure and an optimization-based local search scheme. Extensive experiments with empirical demand distributions based on real data demonstrate that 1) the

integration can gain remarkable system efficiency compared to a standard sequential approach in practice; 2) both the decomposition and the local search are effective in solving realistically sized problem instances. We are conducting more experiments with approximation techniques to show the scalability of the decomposition, and expect simple diversification mechanisms to help strengthen the local search as well.

Chapter 4 centers around a variant of the one-warehouse multi-retailer problem, where the distribution network is reversed in time flows, and the transportation costs are special piecewise linear non-convex functions which repeat themselves when the shipping volume exceeds a batch capacity. We derive approximation guarantees by converting the problem to existing models with a technique that bounds the cost functions. We also devise a polynomial-time 2-approximation algorithm based on single-echelon lot sizing subproblems for the special case of full truckload costs. We further prove the asymptotic optimality of a decentralized solution in a wider range of settings. We are bound to give constant or logarithmic approximations for more complicated cost structures from two-mode transportation costs to generic concave batch ordering costs. Meanwhile, we note possible computational difficulties with the resulting subproblems to ensure polynomial runtime of the approximation algorithms.

We refer the reader to the conclusions of each chapter for more detailed ongoing work, and hope the study can motivate more interesting research in the promising world of supply chain engineering.

APPENDIX A

COMPLEX PROOFS

A.1 Proof of Lemma 2.3.2

All the arguments below apply to solutions that satisfy Lemma 2.3.1.

A.1.1 Proof when $n = 2$

A.1.1.1 Proof of (4b)

Assume a feasible solution is given for which (4b) is violated in some period. We start from the earliest such period, relabeling it as period 1 without loss of generality, and make the following changes:

$$\Delta x_{1,2}^t = \begin{cases} -\epsilon, & t = 1 \\ \frac{p_2 + p_1 q_{1,2}}{1 + q_{1,2}} \epsilon, & t = 2 \\ \frac{(p_1 + p_2 q_{1,2})^{t-3}}{(1 + q_{1,2})^{t-1}} (p_1 - p_2)^2 q_{1,2} \epsilon, & t = 3, \dots \end{cases},$$

$$\Delta s_1^t = \begin{cases} \epsilon, & t = 1 \\ \frac{(p_1 + p_2 q_{1,2})^{t-2}}{(1 + q_{1,2})^{t-1}} (p_1 - p_2) \epsilon, & t = 2, \dots \end{cases},$$

$$\Delta s_2^t = \begin{cases} -\epsilon, & t = 1 \\ \frac{(p_1 + p_2 q_{1,2})^{t-2}}{(1 + q_{1,2})^{t-1}} (p_1 - p_2) q_{1,2} \epsilon, & t = 2, \dots \end{cases}.$$

The resulting solution is feasible for small positive ϵ . Furthermore, we achieve an objective improvement

$$\begin{aligned} \Delta C &= c_{1,2} \left(-1 + \gamma \frac{p_2 + p_1 q_{1,2}}{1 + q_{1,2}} + \sum_{t=3}^{\infty} \frac{(p_1 + p_2 q_{1,2})^{t-3}}{(1 + q_{1,2})^{t-1}} q_{1,2} (p_1 - p_2)^2 \gamma^{t-1} \right) \epsilon \\ &\quad + h_1 \left(1 + \sum_{t=2}^{\infty} \frac{(p_1 + p_2 q_{1,2})^{t-2}}{(1 + q_{1,2})^{t-1}} (p_1 - p_2) \gamma^{t-1} \right) \epsilon \end{aligned}$$

$$\begin{aligned}
& +h_2 \left(-1 + q_{1,2} \sum_{t=2}^{\infty} \frac{(p_1 + p_2 q_{1,2})^{t-2}}{(1 + q_{1,2})^{t-1}} (p_1 - p_2) \gamma^{t-1} \right) \epsilon \\
= & c_{1,2} \frac{(1 + q_{1,2})(1 - p_1 \gamma)(p_2 \gamma - 1)}{1 + q_{1,2} - (p_1 + p_2 q_{1,2}) \gamma} \epsilon \\
& + \frac{(1 + q_{1,2})((1 - \gamma p_2)h_1 - (1 - \gamma p_1)h_2)}{1 + q_{1,2} - (p_1 + p_2 q_{1,2}) \gamma} \epsilon \\
< & 0,
\end{aligned}$$

where the last inequality follows by Assumption 2.2.5.

The rationale behind the construction is to choose training and promotion perturbations so that staff at the two levels increase proportionally in later periods, which implies feasibility; on the other hand, the cost decrease exceeds the increase when discounts and monotonic payrolls are applied, which leads to the lower objective.

A.1.1.2 Proof of (4a)

Since (4b) can be achieved without resorting to (4a), we consider the solutions where (4b) is satisfied while (4a) is not. Again we rename the earliest such period to be period 1. We have $x_0^0 > 0$.

Case 1: $x_{1,2}^2 < p_1 s_1^1$.

Construct a new feasible solution with the formulas below:

$$\Delta x_0^t = \begin{cases} -\frac{\epsilon}{p_0}, & t = 0 \\ \frac{p_1}{p_0} \epsilon, & t = 1 \\ 0, & t = 2, \dots \end{cases},$$

$$\Delta s_1^t = \begin{cases} -\epsilon, & t = 1 \\ 0, & t = 2, \dots \end{cases}.$$

The resulting objective improvement is

$$\begin{aligned}
\Delta C &= c_0 \left(-\frac{1}{p_0 \gamma} + \frac{p_1}{p_0} \right) \epsilon - h_1 \epsilon \\
&< 0.
\end{aligned}$$

Case 2: $x_{1,2}^2 = p_1 s_1^1$.

We first note that at most one of $x_{1,2}^{t+1} = p_1 s_1^t$ and $s_1^{t+1} = d_{t+1}$ can be true provided (4b) for any $t \geq 1$. Assuming both equalities hold for some t , we then have $s_2^{t+1} \geq p_2 s_2^t + x_{1,2}^{t+1} = p_2 s_2^t + p_1 s_1^t$. Since $x_{1,2}^{t+1} > 0$ implies $s_2^{t+1} = q_{1,2} s_1^{t+1}$ by (4b), we further have $d_{t+1} = s_1^{t+1} \geq \frac{p_2 s_2^t + p_1 s_1^t}{q_{1,2}} \geq \frac{p_2 q_{1,2} d_t + p_1 d_t}{q_{1,2}} = \left(p_2 + \frac{p_1}{q_{1,2}}\right) d_t$, but this contradicts Assumption 2.2.6.

Let i be the smallest possible period with $x_{1,2}^i < p_1 s_1^{i-1}$. From the above observation we have $s_1^t > d_t, t \leq i - 1$. Thus we can perturb as follows to obtain a new feasible solution:

$$\Delta x_0^t = \begin{cases} -\frac{\epsilon}{p_0}, & t = 0 \\ -\frac{p_1(p_1 + p_2 q_{1,2})^{t-1}}{p_0 q_{1,2}^t} \epsilon, & t = 1, \dots, i-2 \\ \frac{p_1(p_1 + p_2 q_{1,2})^{i-2}}{p_0 q_{1,2}^{i-2}} \epsilon, & t = i-1 \\ 0, & t = i, \dots \end{cases},$$

$$\Delta x_{1,2}^t = \begin{cases} -p_1 \epsilon, & t = 2 \\ -\frac{p_1^2(p_1 + p_2 q_{1,2})^{t-3}}{q_{1,2}^{t-2}} \epsilon, & t = 3, \dots, i-1 \\ \frac{p_1 p_2(p_1 + p_2 q_{1,2})^{i-3}}{q_{1,2}^{i-3}} \epsilon, & t = i \\ 0, & t = i+1, \dots \end{cases},$$

$$\Delta s_1^t = \begin{cases} -\epsilon, & t = 1 \\ -\frac{p_1(p_1 + p_2 q_{1,2})^{t-2}}{q_{1,2}^{t-1}} \epsilon, & t = 2, \dots, i-1 \\ 0, & t = i, \dots \end{cases},$$

$$\Delta s_2^t = \begin{cases} -\frac{p_1(p_1 + p_2 q_{1,2})^{t-2}}{q_{1,2}^{t-2}} \epsilon, & t = 2, \dots, i-1 \\ 0, & t = i, \dots \end{cases}.$$

The corresponding objective improvement is

$$\begin{aligned}
\Delta C &= c_0 \left(-\frac{1}{p_0 \gamma} - \sum_{t=1}^{i-2} \frac{p_1 (p_1 + p_2 q_{1,2})^{t-1} \gamma^{t-1}}{p_0 q_{1,2}^t} + \frac{p_1 (p_1 + p_2 q_{1,2})^{i-2} \gamma^{i-2}}{p_0 q_{1,2}^{i-2}} \right) \epsilon \\
&\quad + h_1 \left(-1 - \sum_{t=2}^{i-1} \frac{p_1 (p_1 + p_2 q_{1,2})^{t-2} \gamma^{t-1}}{q_{1,2}^{t-1}} \right) \epsilon \\
&\quad + h_2 \left(-\sum_{t=2}^{i-1} \frac{p_1 (p_1 + p_2 q_{1,2})^{t-2} \gamma^{t-1}}{q_{1,2}^{t-2}} \right) \epsilon \\
&\quad + c_{1,2} \left(-p_1 \gamma - \sum_{t=3}^{i-1} \frac{p_1^2 (p_1 + p_2 q_{1,2})^{t-3} \gamma^{t-1}}{q_{1,2}^{t-2}} + \frac{p_1 p_2 (p_1 + p_2 q_{1,2})^{i-3} \gamma^{i-1}}{q_{1,2}^{i-3}} \right) \epsilon \\
&< c_0 \left(-\frac{p_1 p_2^{i-2} \gamma^{i-2}}{p_0} - \sum_{t=1}^{i-2} \frac{p_1^2 (p_1 + p_2 q_{1,2})^{t-1} p_2^{i-2-t} \gamma^{i-2}}{p_0 q_{1,2}^t} + \frac{p_1 (p_1 + p_2 q_{1,2})^{i-2} \gamma^{i-2}}{p_0 q_{1,2}^{i-2}} \right) \epsilon \\
&\quad + c_{1,2} \left(-p_1 p_2^{i-2} \gamma^{i-1} - \sum_{t=3}^{i-1} \frac{p_1^2 (p_1 + p_2 q_{1,2})^{t-3} p_2^{i-t} \gamma^{i-1}}{q_{1,2}^{t-2}} + \frac{p_1 p_2 (p_1 + p_2 q_{1,2})^{i-3} \gamma^{i-1}}{q_{1,2}^{i-3}} \right) \epsilon \\
&= 0.
\end{aligned}$$

A.1.2 Problem structure when $n \geq 3$

Given a solution that violates Lemma 2.3.2, our goal is to construct a new solution that is both feasible and incurs a lower total cost. While the big picture appears similar to the proof when $n = 2$, things are much more complicated here: the effect of Δs_k^t may not end at level $k + 1$; instead it can force $x_{k+1, k+2}^{t+2}$ and thus s_{k+2}^{t+2} to change, which will propagate to higher levels; even worse, lower levels may also be influenced since there may be multiple violated levels and the perturbation may not start from level 1. Therefore, it is unlikely that we can rely on one-time substitutions as before.

Instead, our strategy is to construct a perturbation period by period. To develop such a dynamic approach we first introduce three sets of new variables:

- $r_k^t = \frac{s_k^t}{p^t}, \quad \forall k = 0, \dots, n, \quad \forall t = 1, \dots,$
- $z_k^t = \frac{x_k^t}{p^t}, \quad \forall k = 0, \dots, n, \quad \forall t = 1, \dots,$

- $z_{k,k+1}^t = \frac{x_{k,k+1}^t}{p^t}, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots$

The original problem can be reformulated as follows.

$$\inf W(r, z) = \sum_{t=1}^{\infty} \gamma^{t-1} p^t \left(\sum_{k=0}^n c_k z_k^t + \sum_{k=1}^{n-1} c_{k,k+1} z_{k,k+1}^t + \sum_{k=1}^n h_k r_k^t \right) \quad (42a)$$

$$\text{s.t. } r_1^t \geq d_t/p^t \quad \forall t = 1, \dots \quad (42b)$$

$$r_{k+1}^t \geq q_{k,k+1} r_k^t, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (42c)$$

$$r_1^{t-1} - r_1^t + z_0^{t-1} - z_{1,2}^t + z_1^t = 0, \quad \forall t = 1, \dots \quad (42d)$$

$$r_k^{t-1} - r_k^t + z_{k-1,k}^t - z_{k,k+1}^t + z_k^t = 0, \quad \forall k = 2, \dots, n-1, \quad \forall t = 1, \dots \quad (42e)$$

$$r_n^{t-1} - r_n^t + z_{n-1,n}^t + z_n^t = 0, \quad \forall t = 1, \dots \quad (42f)$$

$$z_{k,k+1}^t \leq r_k^{t-1}, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots \quad (42g)$$

$$r^t, z^t \geq 0, \quad \forall t = 1, \dots, \quad (42h)$$

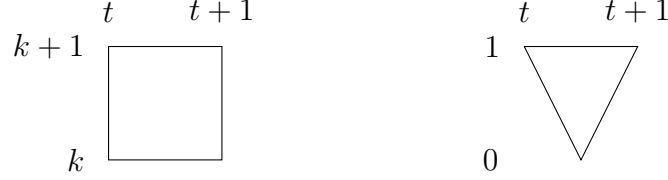
where $r_0^t = z_0^t$ for $t = 0, \dots$. The constraints above can be divided into three sets: demand/ratio constraints (42b–42c), promotion bounds (42g), and network flow constraints (including flow conservation (42d–42f) and nonnegativity (42h)). Graphically, if we consider the r variables as flows between successive periods and the z variables as flows between successive levels, a feasible solution can be represented by an infinite time-space network. The equivalence of the reformulated problem and the original problem stems from a one-to-one correspondence between their solutions. Therefore, any result obtained from one version applies to the other as well.

Next we identify four structural characteristics of our problem(s). Claim A.1.1 describes the cost of certain structures and will help justify the superiority of a perturbed solution; Claim A.1.2 is a dominance property and will enable us to consider a relatively small set of solutions for perturbation; Claims A.1.3 and A.1.4 analyze necessary conditions for feasibility and will shed light on how to perturb.

Claim A.1.1. *For the reformulated problem, any flow circulating counterclockwise*

(either in a cycle, on a doubly-infinite path, or on a one-way infinite path) incurs negative cost.

Proof. $z_0^t, \forall t = 1, \dots$ can be reduced to a common super source node representing level 0 in the network. Define a *basic unit* in the grid-like network as either case below:



The corresponding total costs per unit counterclockwise flow are

$$\begin{aligned} \gamma^{t-1}p^t(h_k - h_{k+1}) + p^t c_{k,k+1}(p\gamma^t - \gamma^{t-1}) &< 0, \\ -\gamma^{t-1}h_1p^t + (-\gamma^{t-1}c_0p^t + \gamma^t c_0p^{t+1}) &< 0, \end{aligned}$$

respectively. We will refer to the two types of basic units as *basic square* and *basic triangle*, respectively. Any cycle can be decomposed into a finite number of basic squares and/or triangles; any doubly-infinite path can be decomposed into a countable number of basic squares; and any one-way infinite path can be decomposed into a countable number of basic squares and/or triangles. Since counterclockwise flows around both basic units incur negative costs, the same is true for arbitrary cycles/infinite paths. \square

Claim A.1.2. For $t \geq 1$, let ℓ_t and ℓ_{t+1} be levels such that $s_{\ell_{t+1}}^t > q_{\ell_t, \ell_{t+1}+1} s_{\ell_t}^t$, $x_{\ell_t, \ell_{t+1}}^t > 0$, and $s_{\ell_{t+1}+1}^{t+1} > q_{\ell_{t+1}, \ell_{t+1}+1} s_{\ell_{t+1}}^{t+1}$. Assuming ℓ_t and ℓ_{t+1} exist, for any k with $\min\{\ell_t, \ell_{t+1}\} \leq k \leq \max\{\ell_t, \ell_{t+1}\}$ there exists some $t' \leq t$ such that $x_{k, k+1}^{t'} > 0$. A solution cannot be optimal if both $s_{\ell_{t+1}}^{t+1} = q_{\ell_t, \ell_{t+1}+1} s_{\ell_t}^{t+1}$ and $x_{k, k+1}^{t+1} = 0$ hold for some k with $\min\{\ell_t, \ell_{t+1}\} \leq k \leq \max\{\ell_t, \ell_{t+1}\}$.

Proof. Clearly $\ell_t \neq \ell_{t+1}$. Consider level i , the largest such k if $\ell_t > \ell_{t+1}$, or the smallest such k if $\ell_t < \ell_{t+1}$.

Case 1: $i = \ell_t$.

Let $\Delta z_{\ell_t, \ell_t+1}^{t+1} = -\Delta z_{\ell_t, \ell_t+1}^t = \epsilon$. Since $x_{\ell_t, \ell_t+1}^{t+1} = 0$, we know $x_{\ell_t+1, \ell_t+2}^{t+1} < ps_{\ell_t+1}^t$ and thus feasibility is not violated. By Claim A.1.1 this corresponds to a counterclockwise flow around a basic square and incurs less total cost.

Case 2: $i \neq \ell_t$.

We first have

$$s_{i+1}^{t+1} = ps_{i+1}^t - x_{i+1, i+2}^{t+1} \leq ps_{i+1}^t,$$

$$s_i^{t+1} \geq ps_i^t + x_{i-1, i}^{t+1} \geq ps_i^t.$$

If $\ell_t > \ell_{t+1}$, then $x_{i+1, i+2}^{t+1} > 0$ by definition of i , and hence $s_{i+1}^{t+1} < ps_{i+1}^t$, which together with $s_{i+1}^{t+1} \geq q_{i, i+1} s_i^{t+1}$ indicates that $s_{i+1}^t > q_{i, i+1} s_i^t$. Similarly, if $\ell_t < \ell_{t+1}$, then $x_{i-1, i}^{t+1} > 0$ and again $s_i^{t+1} > ps_i^t$ indicates that $s_{i+1}^t > q_{i, i+1} s_i^t$. If $x_{i, i+1}^t = 0$, then $s_{i+1}^t > q_{i, i+1} s_i^t$ further indicates that $s_{i+1}^{t-1} > q_{i, i+1} s_i^{t-1}$. Recursively utilizing this fact for $t-1, t-2, \dots$, finally we can find a period $t_0 \geq t'$ ($t_0 = t$ if $x_{i, i+1}^t > 0$) where $s_{i+1}^{t_0} > q_{i, i+1} s_i^{t_0}$ and $x_{i, i+1}^{t_0} > 0$. Now construct a new solution by letting $\Delta z_{i, i+1}^{t+1} = -\Delta z_{i, i+1}^{t_0} = \epsilon$; it is feasible due to slack and zero promotions at level i in periods $t_0 + 1, \dots, t$, and its lower cost is guaranteed by Claim A.1.1. \square

Claim A.1.3. For $t \geq 1$, For $t \geq 1$, let g be a level such that $x_{g, g+1}^{t+1} = 0$, $g \leq n$. If $s_g^{t+1} = q_{g-1, g} s_{g-1}^{t+1} = \dots = q_{1, g} s_1^{t+1}$, where $q_{k, \ell} = q_{k, k+1} q_{k+1, k+2} \dots q_{\ell-1, \ell}$, then at most one of $x_{k, k+1}^{t+1} = ps_k^t$ and $s_1^{t+1} = d_{t+1}$ can be true for each $k \leq g-1$.

Proof. Assume both equalities hold. Adding together equations (1e) for levels $k+1, \dots, g$ in period $t+1$, plugging in $x_{k, k+1}^{t+1} = ps_k^t$ and $s_g^{t+1} = q_{g-1, g} s_{g-1}^{t+1} = \dots = q_{1, g} s_1^{t+1}$, we have

$$s_1^{t+1} \sum_{i=k+1}^g q_{1, i} \geq ps_1^t \sum_{i=k}^g q_{1, i}. \quad (43)$$

Note that

$$\begin{aligned}
\frac{\sum_{i=k}^g q_{1,i}}{\sum_{i=k+1}^g q_{1,i}} &= \frac{q_{1,k}}{\sum_{i=k+1}^g q_{1,i}} + 1 \\
&= \frac{1}{\sum_{i=k+1}^g q_{k,i}} + 1 \\
&\geq \frac{1}{\sum_{i=1}^{g-k} q_{\max}^i} + 1 \\
&= \frac{1 + \sum_{i=1}^{g-k} q_{\max}^i}{q_{\max}(1 + \sum_{i=1}^{g-k-1} q_{\max}^i)}.
\end{aligned}$$

Combined with (43), this results in

$$s_1^{t+1} \geq \frac{1 + \sum_{i=1}^{g-k} q_{\max}^i}{q_{\max}(1 + \sum_{i=1}^{g-k-1} q_{\max}^i)} p s_1^t > \frac{p}{q_{\max}} s_1^t \geq \frac{p}{q_{\max}} d_t \geq d_{t+1}$$

by Assumption 2.2.6. We have arrived at a contradiction. \square

Claim A.1.4. For $t \geq 1$, let k be an arbitrary level, and $g = \min\{i : x_{i,i+1}^{t+1} = 0, k + 1 \leq i \leq n\}$. If $x_{k,k+1}^{t+1} = p s_k^t$ and $s_g^{t+1} = q_{g-1,g} s_{g-1}^{t+1} = \dots = q_{\ell+1,g} s_{\ell+1}^{t+1}$, $0 \leq \ell \leq k - 1$, then $x_{\ell,\ell+1}^{t+1} > 0$.

Proof. The result follows directly by Lemma 2.3.1 if $x_{\ell+1}^{t+1} > 0$. Consider when $x_{\ell+1}^{t+1} = 0$. $\ell = k - 1$ is trivial since $x_{k-1,k}^{t+1} = s_k^{t+1} > 0$. For $\ell \leq k - 2$, we show that $x_{\ell,\ell+1}^{t+1} > p q_{\ell+1,k} s_{\ell+1}^t$ by induction.

Base case: Here $\ell = k - 2$. $s_{k-1}^{t+1} = p s_{k-1}^t + x_{k-2,k-1}^{t+1} - s_k^{t+1} = p s_{k-1}^t + x_{k-2,k-1}^{t+1} - q_{k-1,k} s_{k-1}^{t+1}$ implies $s_{k-1}^{t+1} = \frac{p s_{k-1}^t + x_{k-2,k-1}^{t+1}}{1 + q_{k-1,k}}$. $q_{k-1,g} s_{k-1}^{t+1} = s_g^{t+1} = p s_g^t + x_{g-1,g}^{t+1} > p s_g^t \geq p q_{k-1,g} s_{k-1}^t$ implies $s_{k-1}^{t+1} > p s_{k-1}^t$. Therefore, $\frac{p s_{k-1}^t + x_{k-2,k-1}^{t+1}}{1 + q_{k-1,k}} > p s_{k-1}^t$, and thus $x_{k-2,k-1}^{t+1} > p q_{k-1,k} s_{k-1}^t > 0$.

Induction: Assume that the claim holds for ℓ , $1 \leq \ell \leq k - 2$, i.e. $x_{\ell,\ell+1}^{t+1} > p q_{\ell+1,k} s_{\ell+1}^t$, then

$$\begin{aligned}
s_{\ell}^{t+1} &= p s_{\ell}^t + x_{\ell-1,\ell}^{t+1} - x_{\ell,\ell+1}^{t+1} \\
&< p s_{\ell}^t + x_{\ell-1,\ell}^{t+1} - p q_{\ell+1,k} s_{\ell+1}^t \\
&< p s_{\ell}^t + x_{\ell-1,\ell}^{t+1} - p q_{\ell,k} s_{\ell}^t.
\end{aligned}$$

On the other hand, $q_{\ell,g}s_{\ell}^{t+1} = s_g^{t+1} > ps_g^t \geq pq_{\ell,g}s_{\ell}^t$ implies $s_{\ell}^{t+1} > ps_{\ell}^t$. Therefore, $x_{\ell-1,\ell}^{t+1} > pq_{\ell,k}s_{\ell}^t > 0$, i.e. the claim holds for $\ell - 1$ as well. \square

A.1.3 A perturbation procedure for $n \geq 3$

(4a) is a special case of (4b) if we define $s_0^t = d_t$, $x_{0,1}^t = px_0^{t-1}$, $q_{0,1} = 1$; the only difference is that $x_{0,1}^t$ have no upper bound. Pick the earliest period where (4b) is violated, as before we rewrite it as period 1 and redefine successive periods as 2, 3, \dots

Let m be any violated level in period 1. Below is the key notation we will use:

- j_t : a level that has ever seen staff reduction and has full promotions in period t ; mathematically this means $x_{j_t,j_{t+1}}^t = ps_{j_t}^{t-1}$ and $\Delta s_{j_t}^{t'} < 0$ for some $t' < t$.
- ℓ_t : a level with ratio slack and for which there is a j_t where all levels in between have positive promotions in period t , i.e. $s_{\ell_t+1}^t > q_{\ell_t,\ell_t+1}s_{\ell_t}^t$ and $x_{k,k+1}^t > 0$, $\forall k = \ell_t, \dots, j_t$.
- $\bar{\ell}_t$: an ℓ_t where the ratio relationship would be violated if not perturbed in period $t+1$, i.e. $s_{\bar{\ell}_t+1}^{t+1} + p\Delta s_{\bar{\ell}_t+1}^t < q_{\bar{\ell}_t,\bar{\ell}_t+1}(s_{\bar{\ell}_t}^{t+1} + p\Delta s_{\bar{\ell}_t}^t)$ and $\bar{\ell}_t \in L_t$. An implicit constraint is $s_{\bar{\ell}_t+1}^{t+1} = q_{\bar{\ell}_t,\bar{\ell}_t+1}s_{\bar{\ell}_t}^{t+1}$.
- J_t, L_t, \bar{L}_t : the set of all j_t , the set of all ℓ_t , and the set of all $\bar{\ell}_t$, respectively. $\bar{L}_t \subseteq L_t$.
- $\max j_t$: the largest element in J_t , i.e. $\max\{j : j \in J_t\}$, with other maxima and minima defined analogously.

Suppose we have perturbed periods $1, \dots, t-1$ and the current perturbed solution satisfies constraints in these periods. Clearly all j_t should be perturbed to guarantee feasibility. By Claims A.1.3 and A.1.4 there must exist an ℓ_t for each j_t , and thus it is a candidate for the perturbation in period t to stop. The perturbation in period $t-1$ also causes infeasibility at $\bar{\ell}_{t-1}$ in period t , which constitutes an additional source for

further perturbation. Obviously we can set $J_1 = L_1 = \{m\}$, $\bar{L}_0 = \emptyset$. Algorithm 2 illustrates how to identify J_t , L_t and \bar{L}_{t-1} when $t \geq 2$.

Algorithm 2 J_t , L_t , and \bar{L}_{t-1} when $t \geq 2$

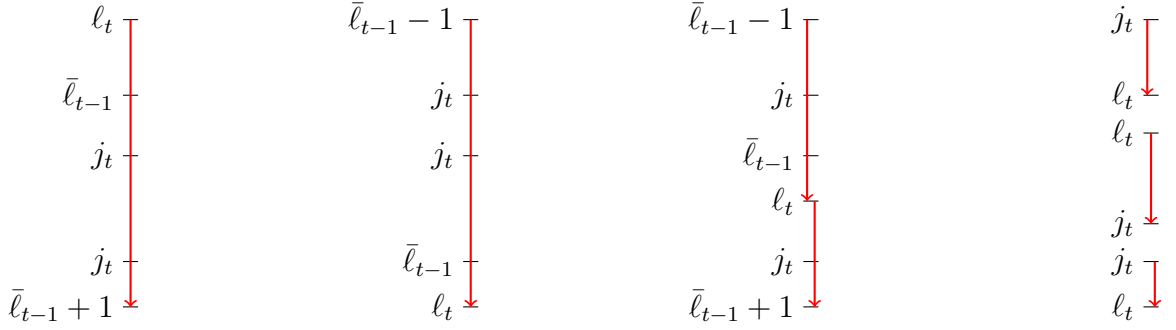
```

1:  $J_t = L_t = \bar{L}_{t-1} = \emptyset, L'_{t-1} = \{\ell_{t-1} : \Delta x_{\ell_{t-1}, \ell_{t-1}+1}^{t-1} < 0\}$ 
2: for  $\ell_{t-1} \in L_{t-1}$  do
3:   if  $s_{\ell_{t-1}+1}^t + p\Delta s_{\ell_{t-1}+1}^{t-1} < q_{\ell_{t-1}, \ell_{t-1}+1}(s_{\ell_{t-1}}^t + p\Delta s_{\ell_{t-1}}^{t-1})$  then
4:      $\bar{L}_{t-1} = \bar{L}_{t-1} \cup \{\ell_{t-1}\}$ 
5:   end if
6: end for
7: for  $k = \min\{i : i \in \bigcup_{t' \leq t-1} J_{t'} \cup L'_{t'}\} \dots \max\{i : i \in \bigcup_{t' \leq t-1} J_{t'} \cup L'_{t'}\} + 1$  do
8:   if  $x_{k, k+1}^t = ps_k^{t-1}$  and  $\Delta s_k^{t'} < 0$  for some  $t' < t$  then
9:      $J_t = J_t \cup \{k\}$ 
10:     $g = \min\{i : x_{i, i+1}^{t+1} = 0, k+1 \leq i \leq n\}$ 
11:    for  $i \leq g-1$  do
12:      if  $s_{i+1}^t > q_{i, i+1}s_i^t$  and  $x_{k, k+1}^t > 0, \forall k = i, \dots, g-1$  then
13:         $L_t = L_t \cup \{i\}$ 
14:      end if
15:    end for
16:  end if
17: end for

```

We are now ready to construct perturbing operations. Since Lemma 2.3.1 and Claim A.1.2 have identified several non-optimal cases, we only consider solutions that satisfy the conditions therein.

Perturbation for period 1 is trivial: $\Delta r_{m+1}^1 = -\Delta r_m^1 = \Delta z_{m, m+1}^1 = -\epsilon$. For an arbitrary $t \geq 2$ before the possible end period, $\Delta z_{k, k+1}^t$ consists of two parts: a change due to full promotions at j_t , and a change due to tight ratio relationships at $\bar{\ell}_{t-1}$. The calculation of Δz consequently depends on the locations of j_t , ℓ_t and $\bar{\ell}_{t-1}$. If there is an $\bar{\ell}_{t-1} \notin L_t$, then the levels between any pair of j_t and $\bar{\ell}_{t-1}$ must have been reached at some point before t , thus by Claim A.1.2 we know $x_{k, k+1}^t > 0, \forall k = \bar{\ell}_{t-1}, \dots, \ell_t, \forall \ell_t \in L_t$. Otherwise $\bar{L}_{t-1} = \emptyset$. Hence although there may be multiple j_t , ℓ_t and $\bar{\ell}_{t-1}$, and a large number of possible locations, it suffices to check the following four cases:



In the above graph, promotions are positive at levels connected by the vertical lines, a gap between levels indicates zero promotions, and the direction of the arrows is consistent with the perturbation flows. Claims A.1.5 through A.1.8 validate the operations we use in each case.

Claim A.1.5. *Assuming a perturbed solution is feasible for periods $1, \dots, t-1$, if there exists an $\ell_t \in L_t$ with $\ell_t \geq \max\{\max j_t, \max \bar{\ell}_{t-1} + 1\}$ and $x_{k,k+1}^t > 0$ for k with $\min\{\min j_t, \min \bar{\ell}_{t-1} + 1\} \leq k \leq \ell_t$, then a solution also feasible for period t can be obtained by sequentially applying equations*

$$\Delta z_{k,k+1}^t = \min\{\Delta z_{k-1,k}^t + \Delta r_k^{t-1} - \Delta r_{k-1}^t q_{k-1,k}, \Delta r_k^{t-1}\}, \quad (44a)$$

$$\Delta r_k^t = \Delta z_{k-1,k}^t + \Delta r_k^{t-1} - \Delta z_{k,k+1}^t, \quad (44b)$$

to $k = k_b, \dots, k_e$, and finally letting

$$\Delta r_{k_e+1}^t = \Delta z_{k_e,k_e+1}^t + \Delta r_{k_e+1}^{t-1}, \quad (44c)$$

where $k_b = \min\{\min j_t, \min \bar{\ell}_{t-1} + 1\}$, $k_e = \min\{\ell_t : \ell_t \geq \max\{\max j_t, \max \bar{\ell}_{t-1} + 1\}\}$.

Proof. We can choose the smallest such ℓ_t as a common level for the perturbation driven by all j_t and $\bar{\ell}_{t-1}$ to stop in period t . To fix potential infeasibility caused by full promotions or tight ratio relationships, it is reasonable to decrease the promotions at these levels and update employment accordingly. The level to start such operations

should of course be $\min\{\min j_t, \min \bar{\ell}_{t-1} + 1\}$. The decreasing effect will finally be conveyed to the ℓ_t we choose.

We consider level k to determine $\Delta z_{k,k+1}^t$. Since Δr_k^{t-1} , Δr_{k-1}^t , and $\Delta z_{k-1,k}^t$ (in particular, $\Delta z_{k_b-1,k_b}^t = 0$) are all known as we reach node (k, t) , we can solve

$$\begin{cases} \Delta z_{k-1,k}^t + \Delta r_k^{t-1} = \Delta z_{k,k+1}^t + \Delta r_k^t \\ \Delta r_k^t \geq \Delta r_{k-1}^t q_{k-1,k} \\ \Delta z_{k,k+1}^t \leq \Delta r_k^{t-1} \end{cases},$$

and choose the largest possible $\Delta z_{k,k+1}^t$ (so that $|\Delta z_{k,k+1}^t|$ is as small as possible), which yields (44a). The three constraints above represent flow conservation, ratio relationships, and promotion bounds respectively. Δr_k^t can then be determined via flow conservation, i.e. (44b). (44c) is a result of stopping perturbation at k_e . \square

Claim A.1.6. *Assuming a perturbed solution is feasible for periods $1, \dots, t-1$, if there exists an $\ell_t \in L_t$ with $\ell_t \leq \min\{\min j_t, \min \bar{\ell}_{t-1} - 1\}$ and $x_{k,k+1}^t > 0$ for k with $\ell_t \leq k \leq \max\{\max j_t, \max \bar{\ell}_{t-1} - 1\}$, then a solution also feasible for period t can be obtained by sequentially applying equations*

$$\Delta z_{k,k+1}^t = \min\{\Delta z_{k+1,k+2}^t + \Delta r_{k+2}^t / q_{k+1,k+2} - \Delta r_{k+1}^{t-1}, \Delta r_k^{t-1}\}, \quad (45a)$$

$$\Delta r_{k+1}^t = \Delta z_{k,k+1}^t + \Delta r_{k+1}^{t-1} - \Delta z_{k+1,k+2}^t, \quad (45b)$$

to $k = k_b, \dots, k_e$, and finally letting

$$\Delta r_{k_e}^t = \Delta r_{k_e}^{t-1} - \Delta z_{k_e,k_e+1}^t, \quad (45c)$$

where $k_b = \max\{\max j_t, \max \bar{\ell}_{t-1} - 1\}$, $k_e = \max\{\ell_t : \ell_t \leq \min\{\min j_t, \min \bar{\ell}_{t-1} - 1\}\}$.

Proof. We can choose the largest such ℓ_t as a common level for the perturbation driven by all j_t and $\bar{\ell}_{t-1}$ to stop in period t . Again we fix potential infeasibility by

decreasing promotions. But unlike the previous claim, we operate in a top-down fashion since the perturbation is expected to end at the $\ell_t \in L_t$ we choose.

We consider level $k + 1$ to determine $\Delta z_{k,k+1}^t$. Since Δr_{k+1}^{t-1} , Δr_k^{t-1} , Δr_{k+2}^t , and $\Delta z_{k+1,k+2}^t$ (in particular, $\Delta z_{k_b+1,k_b+2}^t = 0$) are all known as we reach node $(k + 1, t)$, we can solve

$$\begin{cases} \Delta z_{k,k+1}^t + \Delta r_{k+1}^{t-1} = \Delta z_{k+1,k+2}^t + \Delta r_{k+1}^t \\ \Delta r_{k+1}^t \leq \Delta r_{k+2}^t / q_{k+1,k+2} \\ \Delta z_{k,k+1}^t \leq \Delta r_k^{t-1} \end{cases},$$

and choose the largest possible $\Delta z_{k,k+1}^t$, which yields (45a). Δr can then be determined via flow conservation, i.e. (45b) and (45c). \square

Claim A.1.7. *Assuming a perturbed solution is feasible for periods $1, \dots, t - 1$, if there exists an $\ell_t \in L_t$ with $\min\{\min j_t, \min \bar{\ell}_{t-1} + 1\} \leq \ell_t \leq \max\{\max j_t, \max \bar{\ell}_{t-1} - 1\}$ and $x_{k,k+1}^t > 0$ for k with $\min\{\min j_t, \min \bar{\ell}_{t-1} + 1\} \leq k \leq \max\{\max j_t, \max \bar{\ell}_{t-1} - 1\}$, then a solution also feasible for period t can be obtained by sequentially applying equations (44) to levels $\min\{\min j_t, \min \bar{\ell}_{t-1} + 1\} \leq k \leq \ell_t - 1$, equations (45) to levels $\ell_t + 1 \leq k \leq \max\{\max j_t, \max \bar{\ell}_{t-1} - 1\}$, and finally letting*

$$\Delta z_{\ell_t, \ell_t+1}^t =$$

$$\min\{\Delta r_{\ell_t}^{t-1} + \Delta z_{\ell_t-1, \ell_t}^t - \Delta r_{\ell_t-1}^t q_{\ell_t-1, \ell_t}, \Delta z_{\ell_t+1, \ell_t+2}^t + \Delta r_{\ell_t+2}^t / q_{\ell_t+1, \ell_t+2} - \Delta r_{\ell_t+1}^{t-1}, \Delta r_{\ell_t}^{t-1}\} \quad (46a)$$

$$\Delta r_{\ell_t}^t = \Delta z_{\ell_t-1, \ell_t}^t + \Delta r_{\ell_t}^{t-1} - \Delta z_{\ell_t, \ell_t+1}^t \quad (46b)$$

$$\Delta r_{\ell_t+1}^t = \Delta z_{\ell_t, \ell_t+1}^t + \Delta r_{\ell_t+1}^{t-1} - \Delta z_{\ell_t+1, \ell_t+2}^t. \quad (46c)$$

Proof. This is a hybrid of the previous two claims. By a similar analysis the perturbation should only cover levels from $\min\{\min j_t, \min \bar{\ell}_{t-1} + 1\}$ to $\max\{\max j_t, \max \bar{\ell}_{t-1} - 1\}$. We can treat any such ℓ_t as a breakpoint above which Claim A.1.6 applies and

below which Claim A.1.5 applies. It only remains to perturb level ℓ_t itself. To guarantee feasibility we can choose the minimum of the values provided by Claims A.1.5 and A.1.6 to determine $\Delta z_{\ell_t, \ell_t+1}^t$, and $\Delta r_{\ell_t+1}^t$ and $\Delta r_{\ell_t}^t$ can then be calculated by flow conservation. This yields (46). \square

Claim A.1.8. *Assuming a perturbed solution is feasible for periods $1, \dots, t-1$, if there are zero promotions between successive levels ℓ_t, \dots, j_t or j_t, \dots, ℓ_t , then a solution also feasible for period t can be obtained by applying (44) to each succession ℓ_t, \dots, j_t or (45) to each succession j_t, \dots, ℓ_t , as long as the j_t values partition J_t .*

Proof. By Claim A.1.2 there is no $\bar{\ell}_{t-1}$ and so we only consider the impact of full promotions at j_t . The applicability of the claims follows immediately from the fact that each succession here is an instance of Claim A.1.5 or A.1.6.

We need to ensure, though, that each j_t is included in exactly one succession. Starting from $\max j_t$, if there exists some ℓ_t with $\ell_t \geq \max j_t$ and $x_{k, k+1}^t > 0$ for k with $\max j_t \leq k \leq \ell_t$, there may be a lower j_t that satisfies this condition as well; hence we can decrease promotions from the lowest such j_t to any such ℓ_t , and use the formulas from Claim A.1.5 to determine $\Delta z_{k, k+1}^t$ and Δr_k^t . Otherwise, by Claims A.1.3 and A.1.4 there must exist some ℓ_t with $\ell_t < \max j_t$ and $x_{k, k+1}^t > 0$ for k with $\ell_t \leq k \leq \max j_t$, and so we can use the formulas from Claim A.1.6 to determine $\Delta z_{k, k+1}^t$ and Δr_{k+1}^t until reaching the highest such ℓ_t . After either case is done, we can move downwards to the next j_t that has not been visited, and apply the same argument again. This process goes on until reaching $\min j_t$. \square

We now state our perturbation procedure as Algorithm 3. Note that if in some period we find $x_{k+1, k+2}^t < ps_{k+1}^{t-1}$ for each perturbed level k (including when $J_t = \bar{L}_{t-1} = \emptyset$), the procedure can end in this period, and from then on the perturbed solution will remain the same as the initial solution; otherwise, the procedure will iterate forward infinitely but converge to a new feasible solution. Claim A.1.9 justifies

the lower cost of the final perturbed solution.

Claim A.1.9. *Algorithm 3 modifies the given solution by adding to it a series of negative cost cycles or infinite paths in the time-space network.*

Proof. Except for the possible end period, all the perturbations are initiated by decreasing the z values. Pick any node (k, t) with negative Δr flows in the perturbation network. Each time we conduct an operation as in Claims A.1.5 to A.1.8, by flow conservation Δr_k^t is passed to either Δr_k^{t-1} or $\Delta z_{k-1,k}^t$, resulting in a left arc and a downward arc, respectively. As this propagates, two cases may occur.

Case 1: We reach some $\ell_{\underline{t}}$ ($\underline{t} \leq t$) where the flow turns right, then follows a right-down-right pattern, and finally turns upwards at the perturbation's end period.

This constitutes a counterclockwise cycle as described in Claim A.1.1.

Case 2: The flow may continue shifting in a right-down-right pattern perpetually, which constitutes a counterclockwise infinite path as described in Claim A.1.1.

□

A.1.4 A technical note on the perturbation amount ϵ

We have constructed perturbation operations that are feasible for small enough ϵ . To obtain a valid perturbed solution, however, we need to guarantee that $\epsilon > 0$. Because the perturbation may range over infinitely many periods, it could be that the required ϵ eventually converges to zero. We next argue why this is not the case.

If the perturbation ends in some period, it essentially works in finite dimensions and thus $\epsilon > 0$. On the other hand, if the procedure iterates infinitely, ϵ depends on the x and s values. In particular, the values of the training variables, promotion variables, and slack between consecutive levels matter since we are decreasing them. To eliminate the possibility of ϵ converging to zero, it suffices to bound those values from below wherever they are perturbed.

Algorithm 3 Perturbation when $n \geq 3$ and $p_k = p_{k+1}, \forall k = 1, \dots, n-1$

```

1:  $\Delta r_{m+1}^1 = -\Delta r_m^1 = \Delta z_{m,m+1}^1 = -\epsilon, J_1 = L_1 = \{m\}, \bar{L}_0 = \emptyset, t = 2, ep = 0$ 
2: while  $ep = 0$  do
3:   if  $x_{k+1,k+2}^t < ps_{k+1}^{t-1}$  for every perturbed level  $k$  then
4:      $ep = 1$ 
5:   else
6:     Run Algorithm 2 to identify  $J_t, L_t,$  and  $\bar{L}_{t-1}$ 
7:     if Claim A.1.5 is applicable then
8:       Perturb according to Claim A.1.5
9:     else if Claim A.1.6 is applicable then
10:      Perturb according to Claim A.1.6
11:    else if Claim A.1.7 is applicable then
12:      Perturb according to Claim A.1.7
13:    else
14:      Perturb according to Claim A.1.8
15:    end if
16:    Update  $\Delta x, \Delta s, x, s$ 
17:     $t = t + 1$ 
18:  end if
19: end while
20: for perturbed levels  $k$  in increasing order do
21:    $\Delta z_{k,k+1}^t = \Delta r_k^{t-1} + \Delta z_{k-1,k}^t$ 
22:   Update  $\Delta x, \Delta s, x, s$ 
23: end for

```

When $n = 2$, the proof of (4b) reduces the promotion in period 1 and increases it in later periods. The increments depend on the reduction in period 1 and so depend on how much we can decrease there, which clearly is positive. In the proof of (4a), ϵ depends on $s_1^{t+1} - d_{t+1}$ if $x_{1,2}^{t+1} = p_1 s_1^t$, and $p_1 s_1^t - x_{1,2}^{t+1}$ if $s_1^{t+1} = d_{t+1}$. In the former case, $s_1^{t+1} - d_{t+1} = \frac{p_2 s_2^t + p_1 s_1^t}{q_{1,2}} - d_{t+1} \geq (p_{min} + \frac{p_{min}}{q_{max}})d_t - d_{t+1} \geq p_{min}d_t$, which is bounded away from zero since $d_t \geq d_1 > 0$. In the latter case, the perturbation ends in period $t + 1$ so the perturbation is finite.

Now consider when $n \geq 3$. Recall that in all the cases considered for perturbation, we perturb levels between j_t and ℓ_t , or $\bar{\ell}_{t-1}$ and ℓ_t , or both. For any level k with $\ell_t \leq k \leq j_t$, the proof of Claim A.1.4 actually provides a lower bound independent of t , i.e. $x_{k,k+1}^{t+1} > pq_{k+1,j_t} s_{k+1}^t \geq pq_{1,j_t} d_1$. Furthermore, if we redefine $g = \min \{i : x_{i,i+1}^{t+1} \leq C, j_t + 1 \leq i \leq n\}$, where $C > 0$ can be any constant less than $\min_{t'=1,\dots,t} \{ps_i^{t'}\}$ for each i (such as $C = pq_{1,n}d_1$), the bound still holds, and hence the x variables at levels k with $j_t \leq k \leq \ell_t$ are also bounded below by a constant independent of t .

The same trick can be applied to any level k between ℓ_t and $\bar{\ell}_{t-1}$, i.e. the lower bounds on the x variables in Claim A.1.2 can be strengthened from zero to the same constant C above. Finally, Claim A.1.3 is still correct for $x_{g,g+1}^{t+1} \leq C$ as long as $C < pq_{max}^{n-1}d_1$. This enables us to use a g that still satisfies the properties in both claims as the level starting from which a search of ℓ_t is conducted in Algorithm 2.

Since the proofs of the claims only utilize the linear relationships between s and x , we can obtain similar bounds for the slack between consecutive s variables. All the lower bounds depend only on d_1 and fixed parameters like n , p and q . It follows that the ϵ in the infinite case is indeed positive.

A.2 Proof of Theorem 2.3.3

We will construct a one-period lookahead policy based on Lemmas 2.3.1 and 2.3.2, and then demonstrate that the resulting solution is unique.

A.2.1 A one-period lookahead policy

The notation we use is summarized below:

- $B_{i,i+1}^t$: upper promotion bound for level i in period t .
- I : a list of levels where promotion bounds would be violated if hiring were not allowed.

Claim A.2.1. *Consider a subsystem consisting of levels from j to $k+1$, $0 \leq j \leq k \leq n-1$ in period t . Assume s_i^{t-1} , $i = j, \dots, k+1$ and s_j^t are known. If $p_{j+1}s_{j+1}^{t-1} < q_{j,j+1}s_j^t$ and $p_{i+1}s_{i+1}^{t-1} \leq q_{i,i+1}p_i s_i^{t-1}$, $i = j+1, \dots, k$, then the unique solution to the following equations provides a solution that satisfies promotion bounds at levels in I :*

$$p_{j+1}s_{j+1}^{t-1} + x_{j,j+1}^t - x_{j+1,j+2}^t + x_{j+1}^t = q_{j,j+1}(s_j^t - x_{j,j+1}^t) \quad (47a)$$

$$p_{i+1}s_{i+1}^{t-1} + x_{i,i+1}^t - x_{i+1,i+2}^t + x_{i+1}^t = q_{i,i+1}(p_i s_i^{t-1} + x_{i-1,i}^t - x_{i,i+1}^t + x_i^t), \quad (47b)$$

$$i = j+1, \dots, k-1$$

$$p_{k+1}s_{k+1}^{t-1} + x_{k,k+1}^t + x_{k+1}^t = q_{k,k+1}(p_k s_k^{t-1} + x_{k-1,k}^t - x_{k,k+1}^t + x_k^t), \quad (47c)$$

where $x_i^t = 0$ if $i \notin I$ and $x_{i,i+1}^t = B_{i,i+1}^t$ if $i \in I$.

Proof. Clearly any solution to the above linear system satisfies promotion bounds at any $i \in I$. Nonnegativity of the x variables is guaranteed by the deficiency of staff at level j and the tight ratio relationships (with respect to retention from period $t-1$)

at levels j, \dots, k . For a specific solution x , a policy can be obtained by letting

$$s_i^t = \begin{cases} s_i^t - x_{i,i+1}^t, & i = j \\ p_i s_i^{t-1} + x_{i-1,i}^t - x_{i,i+1}^t + x_i^t, & i = j + 1, \dots, k \\ p_i s_i^{t-1} + x_{i-1,i}^t + x_i^t, & i = k + 1 \end{cases} \quad (48)$$

Let the coefficient matrix be such that column ℓ ($\ell = 1, \dots, k - j + 1$) records the coefficients of $x_{j+\ell-1,j+\ell}^t$ if $j + \ell - 1 \notin I$ and $x_{j+\ell}^t$ otherwise. The elements are

$$a_{m\ell} = \begin{cases} -q_{j+\ell,j+\ell+1}, & m = \ell + 1 \\ 1 \text{ if } j + \ell - 1 \in I \text{ and } 1 + q_{j+\ell-1,j+\ell} \text{ otherwise,} & m = \ell \\ 0 \text{ if } j + \ell - 1 \in I \text{ and } -1 \text{ otherwise,} & m = \ell - 1 \\ 0, & \text{otherwise} \end{cases}.$$

Define D_m ($m = 1, \dots, k - j + 1$) as the determinant of the submatrix composed of the first m rows and the first m columns. For $m \geq 3$, we have the recursion

$$D_m = \begin{cases} D_{m-1}, & j + m - 1 \in I \\ (1 + q_{m+j-1,m+j})D_{m-1} - q_{m+j-1,m+j}D_{m-2}, & \text{otherwise} \end{cases}.$$

By induction we know $D_{k-j+1} \geq D_{k-j} \geq \dots \geq D_1 > 0$, and hence the solution is unique. \square

We now construct a feasible one-period lookahead policy by solving subproblems composed of levels $1, \dots, k + 1$ sequentially until $k = n - 1$. During each loop, we first check if the resulting solution is feasible without promotion at level k , if yes then we are done. Otherwise, we try to get a solution which uses only promotions, i.e. solve (47) with $I = \emptyset$. If this happens to be feasible, then we update the s^t and x^t values and exit; otherwise we calculate a feasible solution by allowing hiring, i.e. solve (47) with $I \neq \emptyset$. For each k , we keep iterating these steps for subproblems composed of

levels $j, \dots, k + 1$ so that we can stop at the highest j and the lower levels are not affected. When determining promotion and hiring, we force the ratio constraints to be tight so that we use the smallest possible x^t . In other words, we promote and hire only if necessary. A formal statement is described in Algorithm 4.

Algorithm 4 A one-period lookahead policy

```

1:  $s_i^t = d_t$  if  $i = 0$  and  $p_i s_i^{t-1}$  if  $i \geq 1$ ,  $B_{0,1}^t = 0$  if  $t = 1$  and  $\infty$  if  $t \geq 2$ ,  $B_{i,i+1}^t =$ 
    $p_i s_i^{t-1}$ ,  $i = 1, \dots, n - 1$ 
2:  $k = 0$ 
3: while  $k \leq n - 1$  do
4:    $j = k$ 
5:   while  $j \geq 0$  and  $s_{j+1}^t < q_{j,j+1} s_j^t$  do
6:      $I = \emptyset$ 
7:      $B_{j,j+1}^t = B_{j,j+1}^t - x_{j,j+1}^t$ 
8:     Solve (47)
9:      $I = \{i : x_{i,i+1}^t > B_{i,i+1}^t, j \leq i \leq k\}$ 
10:    if  $I \neq \emptyset$  then
11:      while  $I$  is changed do
12:        Solve (47)
13:        Update  $I$ 
14:      end while
15:    end if
16:    Update  $s$  with (48)
17:     $j = j - 1$ 
18:  end while
19:   $k = k + 1$ 
20: end while

```

We end this section with two comments. First, the procedure is applicable to both $n = 2$ and $n \geq 3$. Second, once i enters I at some iteration, it will be there forever: The first time i enters I , $x_{i,i+1}^t$ must be decreased (from infeasibility) to full promotion and so s_i^t must be increased in the next iteration (which is true since the only possibility for s_i^t not to be increased is then to decrease x_i^t or $x_{i-1,i}^t$, but this would induce infeasibility between s_i^t and lower levels). To further satisfy the ratio relationships at levels i to k , $s_{\ell+1}^t$ ($i \leq \ell \leq k$) cannot be decreased either, which in turn forces the promotions at these levels to be full if they were. This implies the procedure terminates.

A.2.2 Optimality of the one-period lookahead policy

Claim A.2.2. *Recursively applying Algorithm 4 yields the unique solution that satisfies Lemmas 2.3.1 and 2.3.2.*

Proof. Clearly the solution satisfies the lemmas. Suppose there are multiple feasible solutions for which the lemmas hold. We compare an arbitrary one of them, say (u, y) , with (s, x) obtained from Algorithm 4. By Lemma 2.3.1 $y_{k+1}^t > 0$ only if $y_{k,k+1}^t = p_k u_k^{t-1}$. Start from the earliest period, say t , where there is a difference between x^t and y^t . Pick the lowest different level, say i . We have $s_k^t = u_k^t, \forall k \leq i-1$, and $s^{t'} = u^{t'}, \forall t' \leq t-1$.

We first note that $y_{i,i+1}^t < x_{i,i+1}^t$ or $y_i^t < x_i^t$ cannot be true; otherwise we should be able to obtain a smaller x^t as Algorithm 4 finishes since a feasible solution must satisfy the ratio and bound constraints at every level.

We next show that $y_{i,i+1}^t > x_{i,i+1}^t$ or $y_i^t > x_i^t$ cannot be true, either. Since (u, y) is feasible, by (1e) $y_{i,i+1}^t > x_{i,i+1}^t$ or $y_i^t > x_i^t$ implies $u_k^t > s_k^t$ for some $k \geq i$, so there is over promotion/hiring and the lemmas must be violated somewhere in (u, y) . \square

APPENDIX B

EXTENDED WORKFORCE MODEL

We modified (1) for our computational examples as follows. Let:

- $T \geq 1$: Length of planning horizon.
- $L \geq 1$: Length of the training program.
- a : Age of the student or worker, $a_l \leq a \leq a_u$.
- $p_{0,i} \in (0, 1)$: Per-period rate of continuing education for students in school year $i = 1, \dots, L - 1$, or per-period rate of graduating and going to the workforce for students in school year $i = L$.
- $p_{k,a} \in (0, 1)$: Per-period retention rate of workers of age $a = a_l, \dots, a_u$ that stay in the system at level $k = 0, \dots, n$ from one period to the next.
- $m_{k,a}$: The age distribution of students ($k = 0$) or workers ($k = 1, \dots, n$) of age $a = a_l, \dots, a_u$.
- $s_{0,i,a}^t$: Students of age $a = a_l, \dots, a_u$ in school year $i = 1, \dots, L$ at end of period $t = 1, \dots, T$.
- $s_{k,a}^t$: Workers of age $a = a_l, \dots, a_u$ in level $k = 1, \dots, n$ at end of period $t = 1, \dots, T$.

Our modified problem has the following formulation.

$$\min C(s, x) = \sum_{t=1}^T \gamma^{t-1} \left(\sum_{k=0}^n c_k x_k^t + \sum_{k=1}^{n-1} c_{k,k+1} x_{k,k+1}^t + \sum_{k=1}^n h_k \sum_{a=a_l}^{a_u} s_{k,a}^t \right)$$

$$\begin{aligned}
& \text{s.t. } \sum_{a=a_l}^{a_u} s_{1,a}^t \geq d_t, \quad \forall t = 1, \dots, T \\
& \sum_{a=a_l}^{a_u} s_{k+1,a}^t - q_{k,k+1} \sum_{a=a_l}^{a_u} s_{k,a}^t \geq 0, \quad \forall k = 1, \dots, n-1, \quad \forall t = 1, \dots, T \\
& s_{0,1,a}^t - m_{0,a} x_0^t = 0, \quad \forall a = a_l, \dots, a_u, \quad \forall t = 1, \dots, T \\
& s_{0,i,a}^t - p_{0,i-1} s_{0,i-1,a-1}^{t-1} = 0, \quad \forall i = 2, \dots, L \quad \forall a = a_l + 1, \dots, a_u - 1, \quad \forall t = 1, \dots, T \\
& s_{0,i,a_u}^t - p_{0,i-1} (s_{0,i-1,a_u}^{t-1} + s_{0,i-1,a_u-1}^{t-1}) = 0, \quad \forall i = 2, \dots, L, \quad \forall t = 1, \dots, T \\
& s_{1,a}^t - p_{1,a-1} s_{1,a-1}^{t-1} - m_{1,a} x_1^t - p_{0,L} s_{0,L,a-1}^{t-1} + m_{2,a} x_{1,2}^t = 0, \\
& \quad \forall a = a_l + 1, \dots, a_u - 1, \quad \forall t = 1, \dots, T \\
& s_{1,a_u}^t - p_{1,a_u} s_{1,a_u}^{t-1} - p_{1,a_u-1} s_{1,a_u-1}^{t-1} - m_{1,a_u} x_1^t - p_{0,L} (s_{0,L,a_u}^{t-1} + s_{0,L,a_u-1}^{t-1}) \\
& \quad + m_{2,a} x_{1,2}^t = 0, \quad \forall t = 1, \dots, T \\
& p_{k,a-1} s_{k,a-1}^{t-1} - s_{k,a}^t + m_{k,a} x_{k-1,k}^t - m_{k+1,a} x_{k,k+1}^t + m_{k,a} x_k^t = 0, \quad \forall k = 2 \dots, n-1, \\
& \quad \forall a = a_l + 1, \dots, a_u - 1, \quad \forall t = 1, \dots, T \\
& p_{k,a_u-1} s_{k,a_u-1}^{t-1} + p_{k,a_u} s_{k,a_u}^{t-1} - s_{k,a_u}^t + m_{k,a_u} x_{k-1,k}^t - m_{k+1,a_u} x_{k,k+1}^t + m_{k,a_u} x_k^t = 0, \\
& \quad \forall k = 2 \dots, n-1, \quad \forall t = 1, \dots, T \\
& p_{n,a-1} s_{n,a-1}^{t-1} + m_{n,a} (x_{n-1,n}^t + x_n^t) - s_{n,a}^t = 0, \quad \forall a = a_l + 1, \dots, a_u - 1, \\
& \quad \forall t = 1, \dots, T \\
& p_{n,a_u-1} s_{n,a_u-1}^{t-1} + p_{n,a_u} s_{n,a_u}^{t-1} + m_{n,a_u} (x_{n-1,n}^t + x_n^t) - s_{n,a_u}^t = 0, \quad \forall t = 1, \dots, T \\
& p_{k,a-1} s_{k,a-1}^{t-1} - m_{k+1,a} x_{k,k+1}^t \geq 0, \quad \forall k = 1, \dots, n-1, \\
& \quad \forall a = a_l + 1, \dots, a_u, \quad \forall t = 1, \dots, T \\
& s_{0,i,a}^t = 0, \quad \forall i = 2, \dots, L \quad \forall a = a_l, \dots, a_l + i - 2, \quad \forall t = 1, \dots, T \\
& s_{k,a}^t = 0, \quad \forall k = 2, \dots, n-1, \quad \forall a = a_l, \dots, a_l + k + L - 2, \quad \forall t = 1, \dots, T \\
& x^t, s^t \geq 0, \quad \forall t = 1, \dots, T
\end{aligned}$$

APPENDIX C

DETAILED DECOMPOSITION RESULTS

C.1 Abbreviations

- hs: Grower inventory cost per unit per period.
- hc: Center inventory cost per unit per period.
- UB: Upper bound.
- LB: Lower bound.
- Decomp.: Our decomposition, either single or multiple iterations.

C.2 Iteration Results

- Tables 9-12: Each entry represents the average result over 10 samples generated from the corresponding empirical demand distribution.
- Tables 13-16: Each entry represents the result of a specific sample defined by the corresponding empirical demand distribution and cost parameters.

C.3 Non-Trivial Lower Bounds for Medium Instances

- Table 17: Each entry represents the result of a specific sample defined by the corresponding empirical demand distribution and cost parameters. The lower bounds were returned by CPLEX in 72 hours. Column “Decomp. UB” refers to the best upper bounds obtained by our approach.

Table 9: Single iteration: small instances, nonpeak demand

Parameters		Decomp.		vs. CPLEX			vs. DS-guided	
hs	hc	UB	CPU	UB ratio	LB ratio	CPU raito	UB ratio	CPU ratio
1	0	54444	1653	1.003	1.008	0.137	0.955	4.514
1	0.5	58250	3017	1.001	1.004	0.309	0.917	6.675
1	1	60600	8474	1.004	1.011	0.501	0.900	28.950
1	1.1	60838	5192	1.005	1.019	0.299	0.893	17.475
1	1.5	62122	7967	1.017	1.037	0.443	0.880	39.091
1	2	63027	4229	1.024	1.059	0.235	0.853	21.589
1	3	63027	4249	1.023	1.060	0.236	0.853	24.773
1	4	63027	3480	1.021	1.060	0.193	0.852	23.937

Table 10: Single iteration: small instances, peak demand

Parameters		Decomp.		vs. CPLEX			vs. DS-guided	
hs	hc	UB	CPU	UB ratio	LB ratio	CPU raito	UB ratio	CPU ratio
1	0	43844	3411	1.003	1.012	0.213	0.918	4.444
1	0.5	49383	5689	1.004	1.011	0.511	0.868	24.004
1	1	52472	5205	1.010	1.022	0.364	0.852	17.077
1	1.1	52849	6457	1.014	1.026	0.482	0.847	33.247
1	1.5	54042	5627	1.031	1.047	0.351	0.832	29.081
1	2	55464	3055	1.051	1.077	0.182	0.818	27.673
1	3	55599	4736	1.056	1.079	0.298	0.807	43.970
1	4	55599	6188	1.051	1.079	0.385	0.807	64.933

Table 11: Single iteration: medium instances, nonpeak demand

Parameters		Decomp.		vs. CPLEX 5hrs		vs. CPLEX 10 hrs		vs. DS-guided	
hs	hc	UB	CPU	UB	CPU	UB	CPU	UB	CPU
1	0	57454	15324	1.007	0.851	1.027	0.426	0.977	1.017
1	0.5	63973	13046	0.987	0.725	1.050	0.362	1.004	1.553
1	1	64085	14606	0.910	0.811	0.963	0.406	0.950	1.465
1	1.1	69081	12728	0.934	0.707	1.043	0.354	1.014	1.224
1	1.5	65886	12401	0.942	0.689	0.975	0.344	0.930	3.118
1	2	68578	10646	0.888	0.591	1.013	0.296	0.932	9.321
1	3	68854	11031	0.987	0.613	1.011	0.306	0.935	19.177
1	4	67639	9196	0.908	0.511	1.013	0.255	0.918	9.143

Table 12: Single iteration: medium instances, peak demand

Parameters		Decomp.		vs. CPLEX 5hrs		vs. CPLEX 10 hrs		vs. DS-guided	
hs	hc	UB	CPU	UB	CPU	UB	CPU	UB	CPU
1	0	47538	16813	1.007	0.934	1.035	0.467	0.941	1.093
1	0.5	53270	10218	1.001	0.568	1.038	0.284	0.914	1.284
1	1	54477	12809	0.907	0.712	0.935	0.356	0.863	1.949
1	1.1	54952	13084	0.919	0.727	0.981	0.363	0.865	3.225
1	1.5	56367	10765	0.900	0.598	0.966	0.299	0.858	2.670
1	2	58871	10052	0.879	0.558	0.992	0.279	0.875	4.826
1	3	56026	15035	0.877	0.835	0.956	0.418	0.819	11.476
1	4	57179	11836	0.876	0.658	0.945	0.329	0.836	4.803

Table 13: Multiple iterations: small instances, nonpeak demand

Parameters		CPLEX		Decomp. 1-iter.		Decomp. m-iter.		
hs	hc	UB	CPU	UB ratio	CPU ratio	#iter.	UB ratio	CPU ratio
Nonpeak instance 3								
1	2	62081	18000	1.041	0.038	2	1.001	0.340
1	3	62092	18000	1.040	0.046	2	1.002	0.120
1	4	62140	18000	1.040	0.040	2	1.007	0.106
Nonpeak instance 5								
1	2	57943	18000	1.034	0.065	2	1.020	0.155
1	3	58032	18000	1.033	0.073	2	1.024	0.246
Nonpeak instance 9								
1	2	62866	18000	1.038	0.024	2	1.009	0.094
1	3	62889	18000	1.038	0.022	2	0.999	0.176
1	4	63195	18000	1.033	0.023	2	1.011	0.422
Nonpeak instance 10								
1	0	53320	18000	1.007	0.148	2	1.000	0.341
1	1	58765	18000	1.002	0.025	4	1.000	0.783
1	1.1	59018	18000	1.001	0.168	2	0.999	0.328
1	1.5	59690	18000	1.019	0.325	2	1.000	0.651
1	2	60103	18000	1.033	0.046	2	0.998	0.456
1	3	60689	18000	1.023	0.044	2	0.989	0.436
1	4	60170	18000	1.032	0.034	2	1.012	0.450

Table 14: Multiple iterations: small instances, peak demand

Parameters		CPLEX		Decomp. 1-iter.		Decomp. m-iter.		
hs	hc	UB	CPU	UB ratio	CPU ratio	#iter.	UB ratio	CPU ratio
Peak instance 1								
1	1.5	57004	18000	1.038	0.040	2	1.019	0.283
1	3	57382	18000	1.043	0.041	2	1.007	0.744
Peak instance 2								
1	3	49632	18000	1.056	0.093	2	1.003	0.272
1	4	49902	18000	1.050	0.029	2	1.010	0.097
Peak instance 3								
1	2	52341	18000	1.062	0.118	2	0.994	0.408
1	3	52046	18000	1.068	0.054	2	0.998	0.456
1	4	52191	18000	1.065	0.091	2	1.019	0.544
Peak instance 4								
1	1.5	50111	14800	1.070	0.074	2	1.024	0.180
1	2	50115	7017	1.076	0.169	4	1.000	0.406
1	3	50119	8292	1.076	0.129	2	1.000	0.210
1	4	50119	10911	1.076	0.148	2	1.000	0.179
Peak instance 6								
1	2	53245	18000	1.115	0.088	2	1.005	0.112
1	3	52963	12780	1.121	0.181	2	1.010	0.157
1	4	53478	18000	1.110	0.162	2	0.990	0.154
Peak instance 8								
1	3	55037	18000	1.061	0.325	2	1.013	0.521
1	4	55025	18000	1.061	1.000	4	1.005	1.000
Peak instance 9								
1	2	48176	16464	1.064	0.027	2	1.016	0.121
1	3	48182	12101	1.064	0.033	2	1.021	0.089
1	4	48183	5723	1.064	0.070	4	1.019	0.560

Table 15: Multiple iterations: medium instances, nonpeak demand

Parameters		CPLEX 10hrs		Decomp. 1-iter.		Decomp. m-iter.		
hs	hc	UB	CPU	UB ratio	CPU ratio	#iter.	UB ratio	CPU ratio
Nonpeak instance 1								
1	1.5	64430	36000	1.001	0.214	4	0.995	0.800
1	4	64560	36000	1.205	0.273	2	1.030	0.500
Nonpeak instance 3								
1	0	56764	36000	1.060	0.364	4	0.994	0.800
1	1.1	68052	36000	1.168	0.203	2	1.007	0.500
1	2	73607	36000	1.151	0.157	2	0.903	0.500
1	3	67639	36000	1.156	0.230	2	0.964	0.500
Nonpeak instance 4								
1	1.1	60873	36000	1.049	0.300	5	0.997	1.000
Nonpeak instance 5								
1	0.5	63813	36000	1.051	0.206	4	0.970	0.800
Nonpeak instance 6								
1	0.5	60108	36000	0.991	0.201	3	0.977	0.583
1	1	63734	36000	1.229	0.209	2	1.021	0.500
1	1.1	67724	36000	0.942	0.203	2	0.934	0.500
Nonpeak instance 7								
1	0.5	59965	36000	1.051	0.194	2	1.037	0.500
Nonpeak instance 8								
1	2	63652	36000	1.117	0.107	2	1.059	0.500
1	2	63652	36000	1.117	0.107	4	1.032	0.800
1	4	63392	36000	1.059	0.300	4	0.993	0.800
Nonpeak instance 9								
1	4	64921	36000	1.011	0.242	4	0.958	0.800
Nonpeak instance 10								
1	0	56785	36000	1.098	0.214	3	0.995	0.583
1	0.5	58738	36000	1.048	0.300	2	1.010	0.500
1	1.1	64724	36000	1.008	0.300	2	0.972	0.500
1	3	77411	36000	0.952	0.300	2	0.871	0.500

Table 16: Multiple iterations: medium instances, peak demand

Parameters		CPLEX 10hrs		Decomp. 1-iter.		Decomp. m-iter.		
hs	hc	UB	CPU	UB ratio	CPU ratio	#iter.	UB ratio	CPU ratio
Peak instance 1								
1	0.5	54346	36000	1.323	0.021	2	1.065	0.500
1	0.5	54346	36000	1.323	0.021	3	0.999	0.500
Peak instance 3								
1	0	45569	36000	1.060	0.500	3	1.007	0.583
1	0.5	52917	36000	1.064	0.300	3	0.993	0.583
Peak instance 4								
1	0	45175	36000	1.100	0.500	4	1.000	0.800
1	2	57028	36000	1.053	0.245	2	0.985	0.500
Peak instance 5								
1	1	53116	36000	1.070	0.156	2	0.991	0.500
Peak instance 6								
1	0.5	48294	36000	1.077	0.201	2	1.028	0.500
1	3	52272	36000	1.053	0.186	2	0.991	0.500
1	4	53147	36000	1.004	0.300	2	0.987	0.500
Peak instance 7								
1	0	44923	36000	1.014	0.200	3	1.006	0.583
Peak instance 8								
1	0	44162	36000	1.025	0.450	5	0.992	1.000
Peak instance 10								
1	2	53771	36000	1.032	0.200	2	0.993	0.500
1	4	53560	36000	1.079	0.200	2	1.027	0.500

Table 17: Lower bounds and optimality gaps, medium instances

hs	hc	CPLEX LB	Decomp. UB	Gap	CPLEX LB	Decomp. UB	Gap
Nonpeak instance 2				Nonpeak instance 3			
1	0	52781	55188	4.56%	48297	56397	16.77%
1	0.5	57154	58071	1.60%	60131	62927	4.65%
1	1	57579	62772	9.02%	60838	65097	7.00%
1	1.5	57405	66153	15.24%	61104	67541	10.53%
1	2	57515	63443	10.31%	60827	66488	9.31%
1	3	57293	63962	11.64%	61135	65210	6.67%
1	4	57561	62856	9.20%	60772	69950	15.10%
Nonpeak instance 4				Nonpeak instance 5			
1	0	47175	54218	14.93%	55297	58956	6.62%
1	0.5	56688	62122	9.59%	59902	61896	3.33%
1	1	58859	60696	3.12%	61072	69447	13.71%
1	1.5	58522	63911	9.21%	57002	66661	16.95%
1	2	58046	67711	16.65%	57325	68579	19.63%
1	3	58400	64405	10.28%	61030	68654	12.49%
1	4	57904	66359	14.60%	60785	67339	10.78%
Peak instance 6				Peak instance 7			
1	0	42581	44876	5.39%	42659	45175	5.90%
1	0.5	47056	49637	5.48%	48399	49250	1.76%
1	1	49931	51885	3.91%	50442	53292	5.65%
1	1.5	49154	54439	10.75%	51976	54706	5.25%
1	2	50777	56301	10.88%	46561	57654	23.82%
1	3	49149	51778	5.35%	52059	54575	4.83%
1	4	50240	52337	4.17%	51386	59151	15.11%
Peak instance 9				Peak instance 10			
1	0	43759	45879	4.84%	41372	42794	3.44%
1	0.5	48919	50919	4.09%	47042	49623	5.49%
1	1	51307	56881	10.86%	45794	51110	11.61%
1	1.5	50676	54855	8.25%	49420	52546	6.33%
1	2	50722	57679	13.72%	50155	53368	6.41%
1	3	51696	54854	6.11%	50180	55436	10.47%
1	4	50936	55902	9.75%	49490	54990	11.11%

REFERENCES

- [1] ABERNATHY, W. J., BALOFF, N., HERSHEY, J. C., and WANDEL, S., “A three-stage manpower planning and scheduling model—A service-sector example,” *Operations Research*, vol. 21, pp. 693–711, 1973.
- [2] ABSI, N., ARCHETTI, C., DAUZERE-PERES, S., and FEILLET, D., “A two-phase iterative heuristic approach for the production routing problem,” *Transportation Science*, vol. 49, no. 4, pp. 784 – 795, 2015.
- [3] ADELMAN, D., “Price-directed replenishment of subsets: Methodology and its application to inventory routing,” *Manufacturing and Service Operations Management*, vol. 5, pp. 348–371, 2003.
- [4] ADELMAN, D., “A price-directed approach to stochastic inventory/routing,” *Operations Research*, vol. 52, pp. 499–514, 2004.
- [5] ADELMAN, D. and KLABJAN, D., “Duality and existence of optimal policies in generalized joint replenishment,” *Mathematics of Operations Research*, vol. 30, pp. 28–50, 2005.
- [6] ADELMAN, D. and KLABJAN, D., “Computing near-optimal policies in generalized joint replenishment,” *INFORMS Journal on Computing*, vol. 24, pp. 148–164, 2011.
- [7] ADULYASAK, Y., CORDEAU, J. F., and JANS, R., “Formulations and branch-and-cut algorithms for multivehicle production and inventory routing problems,” *INFORMS Journal on Computing*, vol. 26, no. 1, pp. 103–120, 2014.
- [8] ADULYASAK, Y., CORDEAU, J. F., and JANS, R., “Optimization-based adaptive large neighborhood search for the production routing problem,” *Transportation Science*, vol. 48, no. 1, pp. 20–45, 2014.
- [9] ADULYASAK, Y., CORDEAU, J. F., and JANS, R., “The production routing problem: A review of formulations and solution algorithms,” *Computers & Operations Research*, vol. 55, no. 0, pp. 141–152, 2015.
- [10] AHN, H. S., RIGHTER, R., and SHANTHIKUMAR, J. G., “Staffing decisions for heterogeneous workers with turnover,” *Mathematical Methods of Operations Research*, vol. 62, pp. 499–514, 2005.
- [11] AKBALIK, A. and RAPINE, C., “Polynomial time algorithms for the constant capacitated single-item lot sizing problem with stepwise production cost,” *Operations Research Letters*, vol. 40, pp. 390–397, 2012.

- [12] ANDERSON, E. J. and NASH, P., *Linear Programming in Infinite-Dimensional Spaces*. Chichester: John Wiley & Sons, Inc., 1987.
- [13] ANDERSSON, H., CHRISTIANSEN, A. H. M., HASLE, G., and LØKKETANGEN, A., “Industrial aspects and literature survey: Combined inventory management and routing,” *Computers & Operations Research*, vol. 37, no. 9, pp. 1515–1536, 2010.
- [14] ANILY, S. and TZUR, M., “Shipping multiple items by capacitated vehicles: An optimal dynamic programming approach,” *Transportation Science*, vol. 39, no. 2, pp. 233–248, 2005.
- [15] ANILY, S. and TZUR, M., “Algorithms for the multi-item multi-vehicles dynamic lot sizing problem,” *Naval Research Logistics*, vol. 53, no. 2, pp. 157–169, 2006.
- [16] ANILY, S., TZUR, M., and WOLSEY, L. A., “Multi-item lot-sizing with joint set-up costs,” *Mathematical Programming Series A*, vol. 119, pp. 79–94, 2009.
- [17] ARCHETTI, C., BERTAZZI, L., PALETTA, G., and SPERANZA, M. G., “Analysis of the maximum level policy in a production-distribution system,” *Computers & Operations Research*, vol. 38, pp. 1731–1746, 2011.
- [18] ARCHETTI, C., BERTAZZI, L., and SPERANZA, M. G., “Polynomial cases of the economic lot sizing problem with cost discounts,” *European Journal of Operations Research*, vol. 237, pp. 519–527, 2014.
- [19] ARKIN, E., JONEJA, D., and ROUNDY, R., “Computational complexity of uncapacitated multi-echelon production planning problems,” *Operations Research Letters*, vol. 8, no. 2, pp. 61–66, 1989.
- [20] ARLOTTO, A., CHICK, S. E., and GANS, N., “Optimal hiring and retention policies for heterogeneous workers who learn,” *Management Science*, vol. 60, no. 1, pp. 110–129, 2014.
- [21] ARMENTANO, V. A., SHIGUEMOTO, A. L., and LØKKETANGEN, A., “Tabu search with path relinking for an integrated production-distribution problem,” *Computers & Operations Research*, vol. 38, no. 8, pp. 1199–1209, 2011.
- [22] BALINSKY, W. and REISMAN, A., “Some manpower planning models based on levels of educational attainment,” *Management Science*, vol. 18, pp. 691–705, 1972.
- [23] BARD, J. F. and NANANUKUL, N., “Heuristics for a multiperiod inventory routing problem with production decisions,” *Computers & Industrial Engineering*, vol. 57, no. 3, pp. 713–723, 2009.

- [24] BARD, J. F. and NANANUKUL, N., “The integrated production-inventory-distribution-routing problem,” *Journal of Scheduling*, vol. 12, no. 3, pp. 257–280, 2009.
- [25] BARD, J. F. and NANANUKUL, N., “A branch-and-price algorithm for an integrated production and inventory routing problem,” *Computers & Operations Research*, vol. 37, no. 12, pp. 2202–2217, 2010.
- [26] BARD, J. F. and WAN, L., “Workforce design with movement restrictions between workstation groups,” *Manufacturing and Service Operations Management*, vol. 10, pp. 24–42, 2008.
- [27] BARTHOLOMEW, D. J., FORBES, A. F., and MCCLEAN, S. I., *Statistical techniques for manpower planning*. John Wiley & Sons, Inc., second ed., 1991.
- [28] BATUN, S. and BEGEN, M. A., “Optimization in healthcare delivery modeling: Methods and applications,” in *Handbook of Healthcare Operations Management* (DENTON, B. T., ed.), vol. 184 of *International Series in Operations Research & Management Science*, pp. 75–119, Springer, 2013.
- [29] BC STATISTICS, “Population estimates, British Columbia.” Available on-line at <http://www.bcstats.gov.bc.ca>, April 2013.
- [30] BIENKOWSKI, M., BYRKA, J., CHROBAK, M., DOBBS, N., NOWICKI, T., SVIRIDENKO, M., SWIRSZCZ, G., and YOUNG, N. E., “Approximation algorithms for the joint replenishment problem with deadlines,” *Journal of Scheduling*, vol. 18, no. 6, pp. 545–560, 2015.
- [31] BIENKOWSKI, M., BYRKA, J., CHROBAK, M., JEZ, L., and SGALL, J., “Better approximation bounds for the joint replenishment problem,” *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 42–54, 2013.
- [32] BIENSTOCK, D. and ZENTENO, A. C., “Models for managing the impact of an epidemic.” Working paper, Department of Industrial Engineering and Operations Research, Columbia University. Version Sat.Mar.10.153610.2012. Available at <http://www.columbia.edu/~dano/papers/bz.pdf>, 2012.
- [33] BLUMENFELD, D. E., BURNS, L. D., DILTZ, D. J., and DAGANZO, C. F., “Analyzing trade-offs between transportation, inventory and production costs on freight networks,” *Transportation Research Part B: Methodological*, vol. 19, no. 5, pp. 361–380, 1985.
- [34] BOUDIA, M., LOULY, M. A. O., and PRINS, C., “A reactive GRASP and path relinking for a combined production-distribution problem,” *Computers & Operations Research*, vol. 34, no. 11, pp. 3402–3419, 2007.

- [35] BOUDIA, M. and PRINS, C., “A memetic algorithm with dynamic population management for an integrated production-distribution problem,” *European Journal of Operational Research*, vol. 195, no. 3, pp. 703–715, 2009.
- [36] BOUDREAU, J., HOPP, W., MCCLAIN, J. O., and THOMAS, L. J., “On the interface between operations and human resources management,” *Manufacturing and Service Operations Management*, vol. 5, pp. 179–202, 2003.
- [37] BRAHIMI, N., DAUZERE-PERES, S., NAJID, N. M., and NORDLI, A., “Single item lot sizing problems,” *European Journal of Operational Research*, vol. 168, no. 1, pp. 1–16, 2006.
- [38] BRANDEAU, M. L., HOPKINS, D. S. P., and MELMON, K. L., “An integrated budget model for medical school financial planning,” *Operations Research*, vol. 35, pp. 684–703, 1987.
- [39] BUSCHKÜHL, L., SAHLING, F., HELBER, S., and TEMPELMEIER, H., “Dynamic capacitated lot-sizing problems: A classification and review of solution approaches,” *OR Spectrum*, vol. 32, no. 2, pp. 231–261, 2010.
- [40] CAMPBELL, J., DUSSAULT, G., BUCHAN, J., POZO-MARTIN, F., GUERRA ARIAS, M., LEONE, C., SIYAM, A., and COMETTO, G., “A universal truth: No health without a workforce,” tech. rep., Global Health Workforce Alliance and World Health Organization, Geneva. Third Global Forum on Human Resources for Health, Recife, Brazil, 2013.
- [41] CANADIAN INSTITUTE FOR HEALTH INFORMATION, “Workforce trends of registered nurses in Canada, 2006.” Available on-line at <http://www.cihi.ca>, April 2013.
- [42] ÇETINKAYA, S., “Coordination of inventory and shipment consolidation decisions: A review of premises, models, and justification,” in *Applications of Supply Chain Management and E-Commerce Research* (JOSEPH, G., AKÇALI, E., PARDALOS, P. M., ROMEIJN, H. E., and SHEN, Z. J., eds.), vol. 92 of *Applied Optimization*, pp. 3–51, US: Springer, 2005.
- [43] CHAN, L. M. A., MURIEL, A., SHEN, Z. J., and SIMCHI-LEVI, D., “On the effectiveness of zero-inventory-ordering policies for the economic lot sizing model with a class of piecewise linear cost structures,” *Operations Research*, vol. 50, no. 6, pp. 1058–1067, 2002.
- [44] CHAN, L. M. A., MURIEL, A., SHEN, Z.-J., SIMCHI-LEVI, D., and TEO, C.-P., “Effective zero-inventory-ordering policies for the single-warehouse multi-retailer problem with piecewise linear cost structures,” *Management Science*, vol. 48, no. 11, pp. 1446–1460, 2002.

- [45] CHEEVAPRAWATDOMRONG, T., SCHOCHETMAN, I. E., SMITH, R. L., and GARCIA, A., “Solution and forecast horizons for infinite-horizon nonhomogeneous Markov decision processes,” *Mathematics of Operations Research*, vol. 32, pp. 51–72, 2007.
- [46] COELHO, L. C., CORDEAU, J. F., and LAPORTE, G., “The inventory-routing problem with transshipment,” *Computers & Operations Research*, vol. 39, no. 11, pp. 2537–2548, 2012.
- [47] COELHO, L. C., CORDEAU, J. F., and LAPORTE, G., “Thirty years of inventory routing,” *Transportation Science*, vol. 48, no. 1, pp. 1–19, 2014.
- [48] COHEN, T. and YEDIDSON, L., “The periodic joint replenishment problem is strongly NP-hard,” *arXiv preprint arXiv:1511.02454*.
- [49] CROSS, W. P., ROMELJN, H. E., and SMITH, R. L., “Approximating extreme points of infinite dimensional convex sets,” *Mathematics of Operations Research*, vol. 23, pp. 433–442, 1998.
- [50] CROXTON, K. L., GENDRON, B., and MAGNANTI, T. L., “A comparison of mixed-integer programming models for nonconvex piecewise linear cost minimization problems,” *Management Science*, vol. 49, no. 9, pp. 1268–1273, 2003.
- [51] DAL POZ, M. R., GUPTA, N., QUAIN, E., and SOUCAT, A. L. B., *Handbook on Monitoring and Evaluation of Human Resources for Health with Special Applications for Low and Middle-Income Countries*. World Health Organization, 2009.
- [52] DALL, T. M., GALLO, P. D., CHAKRABARTI, R., WEST, T., SEMILLA, A. P., and STORM, M. V., “An aging population and growing disease burden will require a large and specialized health care workforce by 2025,” *Health Affairs*, vol. 32, pp. 2013–2020, 2013.
- [53] DAVIS, A., MEHROTRA, S., HOLL, J., and DASKIN, M. S., “Nurse staffing under demand uncertainty to reduce costs and enhance patient safety,” *Asia-Pacific Journal of Operational Research*, vol. 31, no. 1, pp. 1–19, 2014.
- [54] DE VÉRICOURT, F. and JENNINGS, O. B., “Nurse staffing in medical units: A queueing perspective,” *Operations Research*, vol. 59, no. 6, pp. 1320–1331, 2011.
- [55] DUCHSCHER, J. E. B., “Transition shock: The initial stage of role adaptation for newly graduated registered nurses,” *Journal of Advanced Nursing*, vol. 65, pp. 1103–1113, 2009.
- [56] ERNST, A. T., JIANG, H., KRISHNAMOORTHY, M., OWENS, B., and SIER, D., “An annotated bibliography of personnel scheduling and rostering,” *Annals of Operations Research*, vol. 127, pp. 21–144, 2004.

- [57] FEDERGRUEN, A. and LEE, C.-Y., “The dynamic lot size model with quantity discount,” *Naval Research Logistics*, vol. 37, pp. 707–713, 1990.
- [58] FEDERGRUEN, A. and TZUR, M., “Time-partitioning heuristics: Application to one warehouse, multi-item, multi-retailer lot-sizing problems,” *Naval Research Logistics*, vol. 46, pp. 463–486, 1999.
- [59] FLORIAN, M. and KLEIN, M., “Deterministic production planning with concave costs and capacity constraints,” *Management Science*, vol. 18, pp. 12–20, 1971.
- [60] FLORIAN, M., LENSTRA, J. K., and KAN, A. H. G. R., “Deterministic production planning: Algorithms and complexity,” *Management Science*, vol. 26, no. 7, pp. 669–679, 1980.
- [61] FREED, G. L., ABRAHAM, L. M., and BRZOZNOWSKI, K. F., “Inactive physicians: The state of our understanding,” *Journal of Pediatrics*, vol. 151, no. 4, pp. 431–434, 2007.
- [62] FREED, G. L., DUNHAM, K. M., and SWITALSKI, K. E., “Clinical inactivity among pediatricians: Prevalence and perspectives,” *Pediatrics*, vol. 123, no. 2, pp. 605–610, 2009.
- [63] FRY, M. J., MAGAZINE, M. J., and RAO, U. S., “Firefighter staffing including temporary absences and wastage,” *Operations Research*, vol. 54, no. 2, pp. 353–365, 2006.
- [64] FUMERO, F. and VERCELLIS, C., “Synchronized development of production, inventory, and distribution schedules,” *Transportation Science*, vol. 33, pp. 330–340, 1999.
- [65] GANS, N., KOOLE, G., and MANDELBAUM, A., “Telephone call centers: Tutorial, review, and research prospects,” *Manufacturing and Service Operations Management*, vol. 5, pp. 79–141, 2003.
- [66] GANS, N. and ZHOU, Y. P., “Managing learning and turnover in employee staffing,” *Operations Research*, pp. 991–1006, 2002.
- [67] GASS, S. I., COLLINS, R. W., MEINHARDT, C. W., LEMON, D. M., and GILLETTE, M. D., “The army manpower long-range planning system,” *Operations Research*, vol. 36, pp. 5–17, 1988.
- [68] GAYON, J.-P., MASSONNET, G., RAPINE, C., and STAUFFER, G., “Constant approximation algorithms for the one warehouse multiple retailers problem with backlog or lost-sales,” *European Journal of Operational Research*, vol. 250, pp. 155–163, 2016.

- [69] GHATE, A., “Infinite horizon problems,” in *Wiley Encyclopedia of Operations Research and Management Science* (COCHRAN, J. J., COX, L. A., KESKINOÇAK, P., KHAROUFEH, J. P., and SMITH, J. C., eds.), John Wiley & Sons, Inc., 2010.
- [70] GHATE, A., SHARMA, D., and SMITH, R. L., “A shadow simplex method for infinite linear programs,” *Operations Research*, vol. 58, pp. 865–877, 2010.
- [71] GHATE, A. and SMITH, R. L., “Characterizing extreme points as basic feasible solutions in infinite linear programs,” *Operations Research Letters*, vol. 37, pp. 7–10, 2009.
- [72] GHATE, A. and SMITH, R. L., “A linear programming approach to nonstationary infinite-horizon markov decision processes,” *Operations Research*, vol. 61, pp. 413–425, 2013.
- [73] GLOCK, C. H., GROSSE, E. H., , and RIES, J. M., “The lot sizing problem: A tertiary study,” *International Journal of Production Economics*, vol. 155, pp. 39–51, 2014.
- [74] GOETSCHALCKX, M., *Supply Chain Engineering*. London: Springer, 2011.
- [75] GREEN, L. V., SAVIN, S., and SAVVA, N., “Nurse vendor problem: Personnel staffing in the presence of endogenous absenteeism,” *Management Science*, vol. 56, no. 10, pp. 2237–2256, 2013.
- [76] GRINOLD, R. C., “Manpower planning with uncertain requirements,” *Operations Research*, vol. 24, pp. 387–399, 1976.
- [77] HALL, R. W., “Consolidation strategy : Inventory, vehicles and terminals,” *Journal of Business Logistics*, vol. 8, no. 2, pp. 57–73, 1987.
- [78] JEWETT, E. A., BROTHERTON, S. E., RUCH-ROSS, H., SPERO, J. C., MCQUIDE, P. A., MATTE, R., HEYWOOD, P., HARAHAP, N. P., ARYANI, S., CASTILLO-LABORDE, C., and OTHERS, “A national survey of ‘inactive’ physicians in the United States of America: Enticements to reentry,” *Human Resources for Health*, vol. 2011, no. 9, p. 7, 2011.
- [79] JIN, Y. and MURIEL, A., “Single-warehouse multi-retailer inventory systems with full truckload shipments,” *Naval Research Logistics*, vol. 56, no. 5, pp. 450–464, 2009.
- [80] JONES, P. C., ZYDIAK, J. L., and HOPP, W. J., “Stationary dual prices and depreciation,” *Mathematical Programming*, vol. 41, pp. 357–366, 1988.
- [81] KHOUJA, M. and GOYAL, S., “A review of the joint replenishment problem literature: 1989-2005,” *European Journal of Operational Research*, vol. 186, pp. 1–16, 2008.

- [82] KOCA, E. and YAMAN, H., “Lot sizing with piecewise concave production costs,” *INFORMS Journal on Computing*, vol. 26, no. 4, pp. 767–779, 2014.
- [83] KOCHER, R. and SAHNI, N. R., “Rethinking health care labor,” *New England Journal of Medicine*, vol. 365, pp. 1370–1372, 2011.
- [84] LAVIERI, M. S., *Nursing workforce planning and radiation therapy treatment decision making: Two healthcare operations research applications*. PhD thesis, University of British Columbia, 2009.
- [85] LAVIERI, M. S. and PUTERMAN, M. L., “Optimizing nursing human resource planning in British Columbia,” *Health Care Management Science*, vol. 12, pp. 119–128, 2009.
- [86] LAVIERI, M. S., REGAN, S., PUTERMAN, M. L., and RATNER, P. A., “Using operations research to plan the British Columbia registered nurses’ workforce,” *Health Care Policy*, vol. 4, pp. 113–131, 2008.
- [87] LEE, H. L., PIERSKALLA, W. P., KISSICK, W. L., LEVY, J. H., GLICK, H. A., and BLOOM, B. S., “Policy decision modeling of the costs and outputs of education in medical schools,” *Operations Research*, vol. 35, pp. 667–683, 1987.
- [88] LEVI, R., LODI, A., and SVIRIDENKO, M., “Approximation algorithms for the capacitated multi-item lot-sizing problem via flow-cover inequalities,” *Mathematics of Operations Research*, vol. 33, no. 2, pp. 461–474, 2008.
- [89] LEVI, R., ROUNDY, R., SHMOYS, D., and SVIRIDENKO, M., “A constant approximation algorithm for the one-warehouse multi-retailer problem,” *Management Science*, vol. 54, no. 4, pp. 763–776, 2008.
- [90] LIPPMAN, S. A., “Optimal inventory policy with multiple set-up costs,” *Management Science*, vol. 16, no. 1, pp. 118–138, 1969.
- [91] MARTEL, A. and PRICE, W., “Stochastic programming applied to human resource planning,” *Journal of the Operational Research Society*, vol. 32, pp. 187–196, 1981.
- [92] MASURSKY, D., DEXTER, F., O’LEARY, C. E., APPELEGET, C., and NUSSMEIER, N. A., “Long-term forecasting of anesthesia workload in operating rooms from changes in a hospital’s local population can be inaccurate,” *Anesthesia and Analgesia*, vol. 106, pp. 1223–1231, 2008.
- [93] NAHMIAS, S., *Production and Operations Analysis*. Boston, MA: McGraw-Hill, 2001.
- [94] NEEDLEMAN, J., BUERHAUS, P., MATTKE, S., STEWART, M., and ZELEVINSKY, K., “Nurse-staffing levels and the quality of care in hospitals,” *New England Journal of Medicine*, vol. 346, no. 22, pp. 1715–1722, 2002.

- [95] NGUYEN, C., DESSOUKY, M., and TORIELLO, A., “Consolidation strategies for the delivery of perishable products,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 69, pp. 108–121, 2014.
- [96] NGUYEN, C., TORIELLO, A., DESSOUKY, M., and MOORE, J. E., “Evaluation of transportation practices in the california cut flower industry,” *Interfaces*, vol. 43, no. 2, pp. 182–193, 2013.
- [97] NONNER, A. and SVIRIDENKO, M., “An efficient polynomial-time approximation scheme for the joint replenishment problem,” in *Integer Programming and Combinatorial Optimization* (HUTCHISON, D., KANADE, T., KITTLER, J., KLEINBERG, J. M., MATTERN, F., MITCHELL, J. C., NAOR, M., PANDU RANGAN, C., STEFFEN, B., TERZOPOULOS, D., TYGAR, D., and WEIKUM, G., eds.), vol. 7801 of *Lecture Notes in Computer Science*, pp. 314–323, Springer, 2013.
- [98] PAPAGEORGIOU, D. J., NEMHAUSER, G. L., SOKOL, J., CHEON, M.-S., and KEHA, A. B., “MIRPLib – A library of maritime inventory routing problem instances: Survey, core model, and benchmark results,” *European Journal of Operational Research*, vol. 235, no. 2, pp. 350 – 366, 2014.
- [99] PINKER, E. J. and SHUMSKY, R. A., “The efficiency-quality trade-off of cross-trained workers,” *Manufacturing and Service Operations Management*, vol. 2, pp. 32–48, 2000.
- [100] POCHE, Y. and WOLSEY, L. A., “Lot-sizing with constant batches: Formulation and valid inequalities,” *Mathematics of Operations Research*, vol. 18, pp. 767–785, 1993.
- [101] POWELL, W. B., *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, Inc., second ed., 2011.
- [102] PUTERMAN, M. L., *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 2005.
- [103] RAO, P. P., “A dynamic programming approach to determine optimal manpower recruitment policies,” *Journal of the Operational Research Society*, vol. 41, pp. 983–988, 1990.
- [104] ROMEIJN, H. E., SHARMA, D., and SMITH, R. L., “Extreme point characterizations for infinite network flow problems,” *Networks*, vol. 48, pp. 209–222, 2006.
- [105] ROMEIJN, H. E. and SMITH, R. L., “Shadow prices in infinite-dimensional linear programming,” *Mathematics of Operations Research*, vol. 23, pp. 239–256, 1998.

- [106] ROMELJN, H. E., SMITH, R. L., and BEAN, J. C., “Duality in infinite dimensional linear programming,” *Mathematical Programming*, vol. 53, pp. 79–97, 1992.
- [107] ROUNDY, R., “98%-effective integer-ratio lot-sizing for one-warehouse multi-retailer systems,” *Management Science*, vol. 31, no. 11, pp. 1416–1430, 1985.
- [108] RUOKOKOSKI, M., SOLYALI, O., JANS, R., and SÜRAL, H., “Efficient formulations and a branch-and-cut algorithm for a production-routing problem,” GERAD Technical Report G-2010-66, HEC Montréal, Canada, 2010.
- [109] SALSBERG, E. and GROVER, A., “Physician workforce shortages: Implications and issues for academic health centers and policymakers,” *Academic Medicine*, vol. 81, no. 9, pp. 782–787, 2006.
- [110] SCHELL, G. J., LI, X., LAVIERI, M. S., TORIELLO, A., MARTYN, K. K., and FREED, G. L., “Strategic modeling of the pediatric nurse practitioner workforce,” *Pediatrics*, vol. 135, no. 2, pp. 298–306, 2015.
- [111] SCHNEIDER, D. P. and KILPATRICK, K. E., “An optimum manpower utilization model for health maintenance organizations,” *Operations Research*, vol. 23, pp. 869–889, 1975.
- [112] SEGEV, D., “An approximate dynamic-programming approach to the joint replenishment problem,” *Mathematics of Operations Research*, vol. 39, no. 2, pp. 432–444, 2014.
- [113] SHARKEY, T. C. and ROMELJN, H. E., “A simplex algorithm for minimum-cost network-flow problems in infinite networks,” *Networks*, vol. 52, pp. 14–31, 2008.
- [114] SHAW, D. X. and WAGELMANS, A. P. M., “An algorithm for single-item capacitated economic lot sizing with piecewise linear production costs and general holding costs,” *Management Science*, vol. 44, no. 6, pp. 831–838, 1998.
- [115] SHEN, Z.-J., SHU, J., SIMCHI-LEVI, D., TEO, C.-P., and ZHANG, J., “Approximation algorithms for general one-warehouse multi-retailer systems,” *Naval Research Logistics*, vol. 56, no. 7, pp. 642–658, 2009.
- [116] SOLYALI, O. and SÜRAL, H., “A relaxation based solution approach for the inventory control and vehicle routing problem in vendor managed systems,” in *Modeling, Computation and Optimization* (NEOGY, S. K., DAS, A. K., and BAPAT, R. B., eds.), vol. 6 of *Statistical Science and Interdisciplinary Research*, Singapore: World Scientific, 171-189, 2009.
- [117] SONG, H. and HUANG, H. C., “A successive convex approximation method for multistage workforce capacity planning problem with turnover,” *European Journal of Operational Research*, vol. 188, no. 1, pp. 29–48, 2008.

- [118] STAUFFER, G., “Using the economical order quantity formula for inventory control in one-warehouse multiretailer systems,” *Naval Research Logistics*, vol. 59, no. 3-4, pp. 285–297, 2012.
- [119] STAUFFER, G., MASSONNET, G., RAPINE, C., and GAYON, J.-P., “A simple and fast 2-approximation algorithm for the one-warehouse multi-retailers problem,” *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 67–79, 2011.
- [120] TURNER, J., MEHROTRA, S., and DASKIN, M. S., “Perspectives on health-care resource management problems,” in *A Long View of Research and Practice in Operations Research and Management Science* (SODHI, M. and TANG, C., eds.), vol. 148 of *International Series in Operations Research & Management Science*, pp. 231–247, New York, NY: Springer, 2010.
- [121] U.S. CENSUS BUREAU, “U.S. population projections.” Available on-line at <http://www.census.gov/population/www/projections/>, January 2012.
- [122] U.S. DEPARTMENT OF AGRICULTURE and U.S. DEPARTMENT OF TRANSPORTATION, “Study of rural transportation issues,” tech. rep., Washington DC, April 2010.
- [123] U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES, “National center for health workforce analysis.” <http://bhpr.hrsa.gov/healthworkforce/>, January 2012.
- [124] VAIRAKTARAKIS, G. and WINCH, J. K., “Worker cross-training in paced assembly lines,” *Manufacturing and Service Operations Management*, vol. 1, no. 2, pp. 112–131, 1999.
- [125] VANDERBY, S. A., CARTER, M. W., LATHAM, T., and FEINDEL, C., “Modeling the future of the Canadian cardiac surgery workforce using system dynamics,” *Journal of the Operational Research Society*, vol. 65, no. 9, pp. 1325–1335, 2014.
- [126] VANDERBY, S. A., CARTER, M. W., LATHAM, T., OUZOUNIAN, M., HASSAN, A., TANG, G. H., TENG, C. J., KINGSBURY, K., and FEINDEL, C. M., “Modeling the cardiac surgery workforce in Canada,” *The Annals of Thoracic Surgery*, vol. 90, pp. 467–473, 2010.
- [127] WAGNER, H. M. and WHITIN, T. M., “Dynamic version of the economic lot sizing model,” *Management Science*, vol. 5, pp. 89–96, 1958.
- [128] WERKER, G. R. and PUTERMAN, M. L., “Strategic workforce planning in healthcare under uncertainty.” Working paper, 2013.
- [129] YANKOVIC, N. and GREEN, L. V., “Identifying good nursing levels: A queuing approach,” *Operations Research*, vol. 59, no. 4, pp. 942–955, 2011.

[130] ZIPKIN, P. H., *Foundations of Inventory Management*. McGraw-Hill, 2000.