**中 山 大 学**

攻读硕士学位研究生

**硕士学位论文开题报告**

题 目：多产品联合补货决策: 机器学习的方法

|  |  |
| --- | --- |
| 姓 名： | 辛睿哲 |
| 学 号： | 20210135 |
| 院 系： | 岭南学院 |
| 专 业： | 管理科学与工程 |
| 研究方向： |  |
| 指导教师： | 冯灏霖 |

中山大学研究生院

2021年 10 月 29 日

## 研究背景和意义

库存管理是现代供应链管理的重要概念之一。随机库存控制指的是对于一个库存系统而言，它的需求是随机的。在实际的生产工作当中，往往会碰到库存中不仅仅有一种产品的，在很多情况下，同一仓库储存的不同的产品之间的需求具有一定的关系。例如，如果两者是可替代的商品，当一个产品的需求上升，另一个产品的需求会下降；如果两者是互补的商品，当一个产品需求上升的时候，另一种产品的需求也会上升。同时，在进货的过程中，如果两种产品的进货量超过一定的数量，往往会获得一定的优惠。

在生产当中，如何管理好需求具有相关关系的库存问题具有非常重要的现实意义。例如，对于一个手机销售商来说，当它的手机需求变大的时候，往往它的配件例如充电器、手机壳等的需求会上升，它们的需求是正相关的关系，被称为是互补商品；又例如，对于一个电脑销售商来说，当它的手提电脑需求上升的时候，往往它的台式机需求会下降，它们的需求是负相关的关系，被称为是可替代商品。在当今的电商时代，人们会接触到更好的导购系统和个性化推荐系统。当我们购买了一件需要的商品之后，网站往往会弹出购买互补商品的链接，这会让互补商品的需求的正相关性变大；或者在我们挑选商品的时候，网站会对同类商品进行详细的比较，这会让可替代商品的需求的负相关性变大。同时，在企业的经营过程中，存在多种产品来源于同一家供应商，企业每次补货过程中都会产生固定订购费，其中，固定订购费与订购产品的数量无关，而是与订购的次数相关，例如，通讯费、人员差旅费等费用。为了有效降低固定费用，联合补货是一种可行而且有效的解决方法，对于多种产品进行联合补货可以使固定成本分摊到每种产品上面，降低固定订购费，实现规模经济，Jayaraman and Tabucanon （1984)在研究中表明通过多产品联合补货策略可以比单独补货策略的补货费用降低15%-30%。所以说，在当今电商发展极为迅速的中国，有很多需求具有一定相关关系的产品，而且利润的竞争愈发激烈，联合补货的问题成为了一个对于很多企业来说很重要的话题。

近十几年来，数字化技术推动了商业领域重大发展，越来越多的企业在进行供应链管理的时候开始应用大数据和数字化技术来优化他们的管理策略，库存管理相关研究人员也开始从数字化、数据驱动的角度研究库存管理问题。在现存的大部分文献当中，机器学习相关算法应用在库存问题上的文献非常多，但是应用在多产品联合补货问题当中的论文非常稀缺，而多产品联合补货问题在大零售、网络平台横行的背景下，很多公司都不是单一产品的，所以多产品的联合进货对于公司来说极其重要。

## 文献综述

对于一个单产品的库存问题，一般有一个(s,S)策略，当库存少于或者等于s的时候，通过补货把库存增加到S。单产品每期回顾的库存问题的最优(s,S)策略被Iglehart (1963)证明了，Veinott and Wagner (1965)证明了单产品在库存具有上下界限的最优策略，(Zheng and Federgruen (1991)提出了一种基于有限时间长度和有限制的库存上下界的最优策略。首先他们确定s或者S，然后迭代另一个参数来降低价值函数使他们的结果重合，他们认为每个周期都进行决策的模型有一个平均的长期的支出，他们同时也把模型推广到无限周期问题以及产品在库存中会折价的问题中。

但是，对于一个多产品的库存问题，就没有一个比较简单具有概括性的模型可以直接被应用了。Ignall (1969)给出了一个简单的两个产品的例子来展示对于多产品库存问题来说，比较难有一个简单的模型作为最优策略。一类简单的控制联合产品的库存问题策略，是由Balintfy and Joseph (1964)提出来的(s,c,S)策略，对于每一个产品都有对应的s、c和S，其中s<c<S，这个策略提出当一个产品的库存少于s的时候进行补货，当进行补货的时候，如果另一个产品的库存少于c，那么这个产品加入到补货的清单中去，所有产品的补货都补到上限S。

本文还和用数据驱动的方法解决库存问题相关联。对于多产品的库存问题来说，有各种各样的方法来解决不同的问题，比如Huh and Rusmevichientong (2009)和(Smith and Agrawal (2000)研究了多项式的逻辑模型，Sauré and Zeevi (2013)设计了一个随机的效用模型，Shi, Chen, and Duenyas (2016)在考虑仓库上限的问题中，将凸函数的优化方法和需求预测两种方法进行库存管理放在一起进行比较。但是这些文章中都没有考虑到缺货的问题。在单产品的库存问题当中， Besbes and Muharremoglu (2013)考虑了不同的具体的需求函数，评价了对不同的需求函数变化对于不同的机器学习算法应用过程中的最优参数和成本函数的影响。Huh and Rusmevichientong (2009)将随机梯度下降的算法应用在这个问题上，抽样平均的方法在Anton et al. (2002)以及Huh et al. (2011)的论文又被提到过，卡普兰-梅尔估计在Huh et al. (2011)的论文中论证过。Ban and Rudin (2019)在大规模数据的报童模型上应用了机器学习的算法。Ban, Gallien, and Mersereau (2019)利用残差树的方法用相似产品的需求量来预测一个新产品的需求量变化，从而通过预测的需求来进行最优的库存管理。当库存问题和商品定价结合在一起研究的时候，Burnetas et al. (2000)使用了近似动态规划当中多臂Bandit算法和逼近理论。Chen, Chao, and Ahn (2019)利用线性回归的方法来逼近真实的需求函数。其他一些在同样在库存管理问题的文献中应用过的方法有，譬如，Bookbinder and Lordahl (1989)使用了重抽样的方法，Lariviere and Porteus (1999)使用了贝叶斯学习的方法。

另外，近似动态规划这个方法在数据驱动领域同样扮演了重要的角色，这个方法对每个参数静态固定一个静态策略之后，通过梯度下降的方法寻找每个参数的最优解，对解决多维度的库存问题也有较大的帮助。Godfrey and Powell (2001)以及Powell, Ruszczyński, and Topaloglu (2004)使用了近似动态规划的方法应用在大规模的交通网络分配的问题上，这两篇研究是通过非线性的方法来逼近的，他们通过近似动态规划的方法，只找到了一些可以逼近的规律和方法，但是没有得出可以解决问题的具体策略。能直接得出具体策略似乎是不可行的，因为问题的维度较高，在动态规划的每个阶段当中每个动作都是连续的，每个动作的结果都是精细化的，不像是二元调节器（LQR）那样比较好解决。近似动态规划的方法还依赖于价值函数的分析形式，需要明确状态转移概率和每个动作的成本和奖励。一些对于近似动态规划的方法的变化，例如Shervais (2000)和Si et al. (2004)适应性评价控制系统，以及Papadaki and Powell (2003)的自适应性动态规划系统已经用在了很多不同种类的库存问题上面。

在使用近似动态规划解决的问题当中，文献中要么简化了价值函数，要么加入了神经网络的使用，但是这两者都影响到了方法的可延展性和通用性，这两个问题在其他论文中往往会用强化学习的方法来解决。强化学习算法已成功应用于文献中广泛的问题，包括Mnih et al. (2015)的虚拟游戏，以及不同形式的物理系统，如Shalev-Shwartz, Shammah, and Shashua (2016)的自动驾驶和Mnih et al. (2015)的飞行动力学，以及运筹学领域， Khadilkar (2018)应用在了火车线路的规划上，Verma et al. (2019)应用在了轮船的装卸和箱子的选择上，Zhang and Dietterich (1995)应用在了程序规划上面。这些应用程序的一个基本特征是能够规划在约束条件下最大化长期折扣未来回报的策略。这种特性使强化学习的算法成为供应链决策的中顺其自然的方法，其中关键的复杂性是在每个时间段需要计算的大量（可能是可变的）并发决策。以前有一些学者，例如Giannoccaro and Pontrandolfo (2002)， Jiang and Sheng (2009)， Kara and Dogan (2018)，但是他们都仅仅应用在了单一产品的库存管理问题当中。同时，尽管强化学习应用于库存问题可以追朔到几十年前，学者们也对问题进行了一系列的扩展研究，例如Giannoccaro and Pontrandolfo (2002)考虑到将利润最大化，Kara and Dogan (2018)将运营成本最小化，Jiang and Sheng (2009)将库存的上下限控制在一定的范围之内。

从上述中可以发现，机器学习应用在多产品的库存补货问题的文献极为稀少，大多数文献都是用启发式算法解决在多产品补货问题中动态规划的维度灾问题，所以，在本研究当中，主要应用近似动态规划和强化学习两种算法，近似动态规划中的不同状态下不同行动的价值的近似值用KNN算法进行拟合，而强化学习主要应用DQN和PPO两种算法。文献中会着重对这几种算法进行一定的修改和测试，这几种机器学习的算法能否很好的应用在多产品的补货问题当中。另外，文献中基本没有研究在随机需求的情况下，如果补货指令下达之后补货产品不能马上到货，而是几个周期之后才能到货的情况，本文在拓展情况中会对这个问题进行一定的研究。

## 问题描述

考虑现在一个库存管理系统里面，一共管理着 N种产品的库存，假设每个阶段库存当中的状态为X（n）={X1 (n), X2 (n)……X N(n)}，其中Xi(n)代表产品i在第n周期时候库存的数量，每个阶段都可以给每个产品进行补货。对于每个产品i来说，有每单位的购买成本为，每次购买的固定成本为，在仓库每单位的持有成本为，而缺货的每单位惩罚成本为。但是在多产品的库存管理问题当中，联合进货的优点在于当企业在一家供应商或者有合作的供应商种订购一系列的产品的时候，订购的固定费用能被几个产品一起平摊，同时，当订购的数量大于一定的数量的时候，往往能获得一系列的优惠或者回购。在论文中，假设当联合补货的时候，若n个产品同时补货，同时补货产生的固定成本K会满足max{K1 , K2 ……K3}<K< K1 ,+K2 ……+K3 .

让代表第n周期不同产品库存数量的变量，让f代表第n周期的价值函数，其中f包括了缺货成本和持有成本，则有：

其中代表第i个产品在该周期的持有成本和缺货成本之和，用式子可以表示为：

在第n周期当中，库存的数量受需求的影响，用 来代表在第n周期不同产品的需求，不同的需求之间是独立不相关的关系。

让代表补货的决策，其中代表第n周期，代表第j周期不同产品的补货数量，当r趋向于正无穷的时候，就是无限周期的补货问题中的全部补货决策，而在第n阶段的补货成本（剔除固定成本）可以表示为：

如果第j周期的库存为，需求为补货的决策是，那么根据动态规划，在第j+1期的库存数量为：

根据上述设定，当给定库存的初始存量x(n)和补货策略Y的时候，可以得出折旧费用的总成本，而总成本是累加的而且是按照几何折旧的，其中折旧率为，.

其中代表第n阶段的补货的固定成本之和。本研究就是需要根据需求寻找合适的补货决策Y，来使成本函数最小化。

## 研究方法

这次研究主要使用了数据驱动的方法去解决多产品的联合补货问题。

首先,为了能将机器学习算法应用在多产品联合补货问题当中,如果直接使用动态规划,很容易因为维度比较高而产生维度灾的问题,所以,为了能让本研究的模型适用于更大的数据集,会对动态规划中每个动作对最后的价值函数带来的变化进行近似的拟合,而在价值函数的近似拟合当中,论文参考Powell (2011)书中的方法，使用深度学习的KNN进行动作价值的近似拟合，这就是研究中的第一个模型，把它命名为ADP-1.接下来，不同于Powell中的方法,论文中会把损失函数改写成和库存问题相关,这个方法参考了Oroojlooyjadid, Snyder, and Taká (2016)应用在报童问题上的方法,在这篇论文当中，研究者主要使用了KNN模型，将观察到的需求和不同的进货量和最后的收益输入到模型当中，能直接预测下一个周期当中进货的最优解，而这篇文章的一个亮眼之处就是把KNN算法的损失函数改写成和现实问题有关的函数，而不是使用简单的最小二乘法，在这次研究中也模仿他们改写损失函数，让它涵盖了持有成本和缺货成本的影响,提高模型的可解释性.这个修改也能让直接得出补货时所需要的货品量,而不是先预测出需求,再进行补货量的计算,这可以避免一系列的动态规划的计算，减少运算的时间，提高运算的速度.在这个模型中，只需要把观测到的需求分布放进去DNN网络中,就可以得到每个阶段需要补货的数量.当然,在文中,为了能让模型获得更好的效果,也在后续的实验中把其他一系列的特征例如上一阶段预测补货量,每个阶段剩余库存等特征放进DNN网络的输入当中，这个方法在论文中被称为ADP-2.

其次,为了能解决近似动态规划中碰到的延伸性和扩展性的问题,所以引入了强化学习的算法,强化学习的算法能在进行一段时间的自动模拟之后,自己进行一系列的特征的收集和多重神经网络的拟合.首先,文章中测试了两个最常见的强化学习模型DQN和PPO,其中DQN是一种离线算法,而PPO是一种在线的策略,两种策略基本代表了强化学习当中两大类的方法，这样可以得出模型DQN-1和PPO-1.同样的,像在ADP-2方法的处理一样，论文中通过修改算法当中的损失函数,加入了持有成本和缺货成本的影响,来观察修改后的模型是否更具有可信性和预测的效果会不会更好，由此可以得到模型DQN-2和PPO-2.

最后,在方法测试的阶段，首先，本研究将会在一个维度比较低的情况下进行一些测试。本研究将会对M’Hallah, Benkherouf, and Al-Kandari(2020)中二维产品补货问题的最优策略作为标准策略，将上述提到的不同策略与该策略进行比较。本研究将会针对不同的产品,有不同的需求函数,测试当不同的产品之间具有相同的需求函数,或是具有不同的需求函数等情况,来比较上述的几种方法在计算速度,计算准确度,预测结果和模拟的最优结果之间的差距波动等结果上的不同.接下来，会扩大产品的维度，测试不同方法随着产品的增多，维度的增加，预测结果的差异。最后，在论文中我使用了一个Kaggle（2018）上的一个公开的数据集进行不同模型的测试，这个数据集囊括了50000种产品和6000个具体客户的具体购买记录，但是由于数据集当中没有具体的购买日期，只是统计了每个客户每隔多少时间进行了同一种产品的购买，所以在论文中假设所有产品都是在一年内的第一天开始被购买的，这样就会得出一系列具有时间戳的数据，就可以作为我们现实中的数据。由此形成的数据集跨度为349天。我们假设库存补充每天发生4次，导致1396个时间段。其中，我们使用前900个时间段进行培训，使用最后496个时间段进行测试。使用这个数据集会让论文的模型在更贴合现实生产中进行测试。

## 文章结构

具体来说本文共分为五章，分别是绪论、文献综述、模型构建、模型测试、结论与展望。

第一章主要阐述问题提出的背景，点出解决这个问题的实际意义。

第二章对现有文献当中处理多产品库存问题的方法进行总结，从大部分文献钟使用的启发式算法开始进行总结，再到数据驱动领域的不同方法，再到与本文两个模型相关的近似动态规划的方法，以及对于动态规划方法的一定改进，即强化学习的内容和强化学习仔库存问题领域当中的文献进行总结。

第三章将进行四个模型的构造，分别是用KNN方法进行每个阶段不同动作的value的拟合的ADP-1；在ADP-1基础上修改损失函数，在损失函数钟加入库存管理当中持有的成本以及缺失的成本的方法ADP-2；将强化学习当中的PPQ和DNQ方法应用在多产品库存问题当中的PPQ-1和DNQ-1；在PPQ-1和DNQ-1当中根据现实情况，像ADP-2一样，改写损失函数来尝试提高模型预测的精确度，而改进出模型PPQ-2和DNQ-2.

第四章通过输入不同的需求分布，包括泊松分布和正太分布，并改变需求分布的参数，对上述六个模型进行实验，从产品维度由少变多，分别比较不同模型在需求分布的参数变化时出现的变化规律、不同分布对于模型的影响、不同的产品对应不同的需求函数的时候。当问题是双产品问题的时候，用M’Hallah, Benkherouf, and Al-Kandari(2020)的方法作为基准进行比较。最后也使用了一个kaggle上的公开数据集，测试在需求更加不确定，更加接近真实的生产活动的时候，不同模型的预测效果的不同。

第六章将基于前面所有的研究内容得出本研究的最终结论，提出本研究中存在的一些不足，并为之后的研究指出方向。

## 文章的创新型

这篇论文的目的在于建立一个库存管理系统，对需求具有相关关系的产品进行库存的管理。因为需要证明该论文当中的系统是一种比较好的方法，所以首先，在维度比较少的状态下，文章建立了一个基于马尔科夫链的库存系统，其中系统中包括了不同种类且需求具有相关关系的产品，联合补货的费用。论文中通过应用机器学习的方法，包括了近似动态规划的方法和强化学习算法，与该库存问题中最优的结果进行对比，本文的创新点主要如下:

1. 在文献当中基本上使用了启发式算法的来解决多产品的补货问题，而本文当中制定了近似动态规划和强化学习的模型来解决多产品联合捕获问题，在针对不同的实际生产中的运营目标，例如在运营过程中产品的损失最小或者是因为缺货造成的损失最小，训练不同的模型，以及不需要重新训练就可以把模型移植到具有相同的需求函数的情况当中。
2. 在文献当中的，应用在联合补货库存问题上的模型，都没有对损失函数进行按照实际问题对其进行修改。而在这次研究当中，在损失函数的设计当中，为了能让近似动态规划和强化学习两个方法适应多产品联合补货问题，所以重新进行了损失函数的设计，让损失函数覆盖了持有成本和缺失成本，这能让模型不是先进行需求的预测，再进行补货数量的计算，而是直接拟合出不同产品需要补货的数量。在文章的后面，除了一开始使用的数据特征包括观测到的之前的需求量、每期的库存量，也加入了其他特征，这比较符合在现在的生产中，生产者可以更加容易的观测到更多维的数据。
3. 文章会对不同的需求函数在不同的模型中进行对比，观察不同模型应对变化时候的反应快慢，也能表明在面对现实中比较多变的内外部环境的时候，怎样的方法是比较适合管理者进行参考和使用的。

## 可能碰到的问题和解决方法

1. 由于文章中使用了一系列机器学习的内容，还对传统模型进行了一系列的修改，所以需要一定的运算时间和电脑容量进行最优化的工作，使用自己的电脑进行实验可能会运到运算空间不足和运算速度不足的问题，所以会使用学院提供的远程电脑或者阿里云的远程运算机器进行实验。
2. 模型在模拟数据中能得到比较好的结果，但是在文章应用的kaggle数据集中难以取得比较好的结果。由于数据集中还提供了产品保质时间和产品的体积和重量的数据，所以可以通过这三个数据对产品进行一定的分类，然后再用不同的模型进行测试，可能会让结果更加容易让人理解。
3. 代码可能涉及一系列平时接触的比较少的强化学习的领域。进行应用DQN，PPO方法的经典论文的复现，争取能把这两个方法在多产品库存问题上应用做出最优解，这样有利于论文后续对这几个方法的比较。

## 进度安排

2021.9.1-2021.10.29 完成开题报告

2021.10.29-2021.11.9 汇报开题报告

2021.11.1-2022.2.1 完成论文初稿

2022.2.1-2022.3.1 完成论文终稿

参考文献

Anton, J., Kleywegt, Alexander, Shapiro, Tito, and Homem-de-Mello. 2002. ‘The Sample Average Approximation Method for Stochastic Discrete Optimization’. *SIAM Journal on Optimization* 12(2):479–502.

Balintfy, and L. Joseph. 1964. ‘On a Basic Class of Multi-Item Inventory Problems’. *Management Science* 10(2):287–97.

Ban, Gah℡i, J. Gallien, and A. Mersereau. n.d. ‘Dynamic Procurement of New Products with Covariate Information: The Residual Tree Method’. *Social Science Electronic Publishing*.

Ban, Gah-Yi, and Cynthia Rudin. 2019. ‘The Big Data Newsvendor: Practical Insights from Machine Learning’. *Operations Research* 67(1):90–108.

Besbes, O., and A. Muharremoglu. 2013. ‘On Implications of Demand Censoring in the Newsvendor Problem’. *Management Science* 59(6):págs. 1407-1424.

Bookbinder, James H., and Anne E. Lordahl. 1989. ‘Estimation of Inventory Re-Order Levels Using the Bootstrap Statistical Procedure’. *IIE Transactions* 21(4):302–12.

Burnetas, N. Apostolos, Smith, and E. Craig. 2000. ‘ADAPTIVE ORDERING AND PRICING FOR PERISHABLE PRODUCTS.’ *Operations Research*.

Chen, B., X. Chao, and H. S. Ahn. n.d. ‘Coordinating Pricing and Inventory Replenishment with Nonparametric Demand Learning’. *Social Science Electronic Publishing*.

Giannoccaro, Ilaria, and Pierpaolo Pontrandolfo. 2002. ‘Inventory Management in Supply Chains: A Reinforcement Learning Approach’. *International Journal of Production Economics* 78(2):153–61.

Godfrey, Gregory A., and Warren B. Powell. 2001. ‘An Adaptive, Distribution-Free Algorithm for the Newsvendor Problem with Censored Demands, with Applications to Inventory and Distribution’. *Management Science* 47(8):1101–12.

Huh, Woonghee Tim, Retsef Levi, Paat Rusmevichientong, and James B. Orlin. 2011. ‘Adaptive Data-Driven Inventory Control with Censored Demand Based on Kaplan-Meier Estimator’. *Operations Research*.

Huh, Woonghee Tim, and Paat Rusmevichientong. 2009. ‘A Nonparametric Asymptotic Analysis of Inventory Planning with Censored Demand’. *Mathematics of Operations Research* 34(1):103–23. doi: 10.1287/moor.1080.0355.

Iglehart, Donald L. 1963. ‘Optimality of (s, S) Policies in the Infinite Horizon Dynamic Inventory Problem’. *Management Science* 9(2):259–67. doi: 10.1287/mnsc.9.2.259.

Ignall, E. 1969. ‘Theory Series || Optimal Continuous Review Policies for Two Product Inventory Systems with Joint Setup Costs’. *Management Science* 15(5):278–83.

Jayaraman, R., and M. T. Tabucanon. 1984. ‘Co‐ordinated Versus Independent Replenishment Inventory Control’. *International Journal of Operations & Production Management* 4(1):61–68.

Jiang, Chengzhi, and Zhaohan Sheng. 2009. ‘Case-Based Reinforcement Learning for Dynamic Inventory Control in a Multi-Agent Supply-Chain System’. *Expert Systems with Applications* 36(3):6520–26.

Kaggle: Instacart market basket analysis data. https://www.kaggle.com/c/instacart-market-basket-analysis/data. Accessed Aug 2018

Kara, Ahmet, and Ibrahim Dogan. 2018. ‘Reinforcement Learning Approaches for Specifying Ordering Policies of Perishable Inventory Systems’. *Expert Systems with Applications* 91:150–58.

Khadilkar, Harshad. 2018. ‘A Scalable Reinforcement Learning Algorithm for Scheduling Railway Lines’. *IEEE Transactions on Intelligent Transportation Systems* 20(2):727–36.

Lariviere, Martin A., and Evan L. Porteus. 1999. ‘Stalking Information: Bayesian Inventory Management with Unobserved Lost Sales’. *Management Science* 45(3):346–63.

M’Hallah, R., L. Benkherouf, and A. Al-Kandari. n.d. ‘Optimal Inventory Policies for a Two-Dimensional Stochastic Inventory Model: A Numerical Investigation - ScienceDirect’. *Computers & Operations Research* 119.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, and others. 2015. ‘Human-Level Control through Deep Reinforcement Learning’. *Nature* 518(7540):529–33.

Oroojlooyjadid, A., L. Snyder, and Martin Taká. 2016. ‘Applying Deep Learning to the Newsvendor Problem’.

Papadaki, Katerina P., and Warren B. Powell. 2003. ‘An Adaptive Dynamic Programming Algorithm for a Stochastic Multiproduct Batch Dispatch Problem’. *Naval Research Logistics (NRL)* 50(7):742–69.

Powell, Warren, Andrzej Ruszczyński, and Huseyin Topaloglu. 2004. ‘Learning Algorithms for Separable Approximations of Discrete Stochastic Optimization Problems’. *Mathematics of Operations Research* 29(4):814–36.

Sauré, D., and A. Zeevi. 2013. ‘Optimal Dynamic Assortment Planning with Demand Learning’. *Manufacturing and Service Operations Management*.

Shalev-Shwartz, Shai, Shaked Shammah, and Amnon Shashua. 2016. ‘Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving’. *ArXiv Preprint ArXiv:1610.03295*.

Shervais, Stephen. 2000. *Adaptive Critic Design of Control Policies for a Multi-Echelon Inventory System*. Portland State University.

Shi, Cong, Weidong Chen, and Izak Duenyas. 2016. ‘Nonparametric Data-Driven Algorithms for Multiproduct Inventory Systems with Censored Demand’. *Operations Research* 64(2):362–70.

Si, Jennie, Andrew G. Barto, Warren B. Powell, and Don Wunsch. 2004. *Handbook of Learning and Approximate Dynamic Programming*. Vol. 2. John Wiley & Sons.

Smith, S. A., and N. Agrawal. 2000. ‘Management of Multi-Item Retail Inventory Systems with Demand Substitution’. *Operations Research* 48(1):50–64.

Veinott, A. F., and H. M. Wagner. 1965. ‘Computing Optimal ( *s, S* ) Inventory Policies’. *Management Science* 11(5):525–52.

Verma, Richa, Sarmimala Saikia, Harshad Khadilkar, Puneet Agarwal, Gautam Shroff, and Ashwin Srinivasan. 2019. ‘A Reinforcement Learning Framework for Container Selection and Ship Load Sequencing in Ports’. Pp. 2250–52 in *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*.

Zhang, Wei, and Thomas G. Dietterich. 1995. ‘A Reinforcement Learning Approach to Job-Shop Scheduling’. Pp. 1114–20 in *IJCAI*. Vol. 95. Citeseer.

Zheng, Yu Sheng, and A. Federgruen. 1991. ‘FINDING OPTIMAL (s, S) POLICIES IS ABOUT AS SIMPLE AS EVALUATING A SINGLE POLICY’. *Operations Research* 40(1):192–192.