Dynamic Feature-based Newsvendor

Zexing Xu [1] **Ziyi Chen** [2] **Xin Chen** [3]

Abstract

In this paper, we proposed a dynamic contextual newsvendor model that combines the significance of feature information with a multi-period inventory control framework. To solve this model, we propose the Contextual Value Iteration (CVI) algorithm and obtain its convergence rate to the optimal solution as well as sample complexity result. Our experimental results demonstrate that our CVI is more efficient and practical than value iteration for the vanilla Markovian Decision Process (MDP).

1. Introduction

Inventory control plays a pivotal role in supply chain management. The newsvendor problem, a classic single-period inventory control problem, has been a cornerstone in the literature of operations research and management science (Arrow et al. 1951, Scarf 1958, Karlin & Scarf 1959, Eppen 1979, Gallego 1993).

A significant extension of the single-period newsvendor problem is the multi-period newsvendor problem, which accounts for inventory control decisions over multiple periods. This extension has been studied in various contexts such as stochastic demands (Veinott 1965), perishable items (Nahmias 1975), and backlogged demands (Levi et al. 2006).

In today's era of big data, decision-makers, or sellers, are furnished with an extensive array of relevant information, such as customer demographics, weather forecasts, seasonality factors, economic indicators, and past demands (McAfee & Brynjolfsson 2013, Ban & Rudin 2019). This plethora of information calls for an investigation into the incorporation of observed side information, or features, into the decision-making process (Harrison & Zeevi 2011).

[1]Department of Industrial & Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL, US [2]Department of Electrical & Computer Engineering, University of Utah, Salt Lake City, UT, US [3]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA. Correspondence to: Zexing Xu <zexingx2@illinois.edu>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

While previous studies have delved into multi-period inventory control and feature-based uncertainty reduction, they have mostly considered these desirable properties separately. Specifically, some studies focus on multi-period inventory control, neglecting the available features (Veinott 1965, Nahmias 1975, Levi et al. 2006), while others center their attention on feature-based inventory models, concentrating solely on single-period inventory control (Agarwal et al. 2011, Li et al. 2015, Cohen et al. 2016).

To better adapt to practical applications, this paper proposes a dynamic contextual newsvendor model for a multi-period inventory control problem with backlogged demands, which incorporates observed side information, or features, into the decision-making process. The model enables the seller to consider the current inventory level and observed features to determine the optimal order quantity for each time period. This dynamic feature-based newsvendor model extends the classic newsvendor problem, incorporating the observed features into the decision-making process to make more informed and efficient inventory control decisions (Gaur et al. 2005). In this work, we focus on the policy optimization decision-making problem, aiming to find a policy (decision rule) that minimizes the accumulated discounted cost by providing an ordering decision for every feature value.

For this generalized problem formulation, we propose a value iteration algorithm and prove that it converges to the optimal solution at the rate $\mathcal{O}\left(\gamma^K + \frac{1}{n}\right)$ ($\gamma \in (0,1)$ is the discount factor), which is exponentially fast with the number of iterations K and scales sublinearly with sample size n. Compared with the value iteration algorithm for reinforcement learning (Agarwal et al. 2022) which only uses state and action samples to estimate state transition distribution, we leverage the structure of the state transition as well as the available samples of feature (contextual information) X_t and demand D_t to estimate the distribution of X_t, D_t, from which we can obtain not only the transition distribution of the inventory level (state) Y_t but also additional information on the underlying dynamic process. As a result, the sample approximation error in our convergence results from estimating the joint distribution $\mathbb{P}_{joint}$ of X_t, D_t rather than estimating the transition kernel of Y_t.

2. Related Works

Alongside these developments, there is growing interest in contextual Markov decision processes (CMDPs) and contextual reinforcement learning (CRL) in recent years. CMDPs extend the traditional MDP framework by incorporating side information, or context, into the state space (Lazaric 2010, Abel et al. 2016). CRL can be approached using either model-based or model-free methods, depending on the availability of a model for the environment (Langford & Strehl 2007, Tang & Singh 2010, Zhang et al. 2013, Silver et al. 2014).

Returning to the main thread of our work, we bring new advancements to the dynamic contextual newsvendor models and contextual reinforcement learning by introducing a multi-period inventory control problem with backlogged demands that incorporates observed features into the decision-making process (Levi et al. 2018). Building on the existing body of work on contextual newsvendor problems and CMDPs, our model provides a practical framework for understanding and solving such issues (Besbes & Zeevi 2015).

Notably, our work's uniqueness lies in the problem formulation that generalizes the inventory control problem in several aspects. First, our model and analysis do not require a specific form of the cost function $q(y)$ where y denotes inventory level, making our framework more versatile than most newsvendor models (Gallego & Moon 1993). Secondly, the state transition rule in our model can be more general, such as $Y_{t+1} = g(Y_t, A_t, D_t)$ with a known transition mapping g. This transition rule is more informative than those of MDPs, as it uses not only the state Y_t and the action A_t, but also the dynamic feature X_t (Puterman 1994).

Moreover, for this generalized problem formulation, we proposed a value iteration algorithm and proved its convergence to the optimal solution at a rate that is exponentially fast with the number of iterations and scales sublinearly with the number of samples. We leveraged the structure of the state transition and the available samples of feature (contextual information) and demand to estimate the distribution of these variables, providing additional information on the underlying dynamic process. Consequently, the sample approximation error in our convergence results arises from estimating the joint distribution of the feature and demand, rather than the transition kernel of the state. This reflects a significant improvement over the existing reinforcement learning value iteration algorithm, which only uses state and action samples to estimate the state transition distribution (Agarwal et al. 2022).

Our problem and model differ from the traditional CMDP framework in several key aspects. In our problem, the inventory level Y_t and observed side information X_t are used to determine the optimal order quantity A_t, whereas CMDPs typically involve a context-dependent state space. Moreover, our contextual information X_t varies with time point t, which makes our framework more general and challenging to learn, compared to CMDPs where all the time points within a trajectory share the same contextual information (Hallak et al. 2015). This distinction underscores the novelty of our approach, as we provide a fresh framework for understanding and solving dynamic contextual newsvendor problems that incorporate backlogged demands.

The remainder of the paper is structured as follows. In Section 3, we present a detailed problem formulation ofthe feature-based newsvendor problem. Section 4 introduces our proposed algorithm (CVI) to solve this feature-based newsvendor problem. In Section 5, we offer an convergence analysis of CVI algorithm. In Section 6, we validate our theoretical results through a series of numerical experiments, highlighting the effectiveness of our proposed algorithms in various aspects. Finally, we conclude the paper in Section 7 with a summary of our contributions and potential directions for future research.

3. Problem Setting

This work focuses on a dynamic, multi-period inventory control problem, often referred to as the multi-period newsvendor problem, with backlogged demands. This model not only applies to inventory control but can also serve a variety of applications such as capacity planning, supply chain management, and demand response in energy grids, underlining its versatile utility.

In each time period, denoted by t, a seller is tasked with deciding on a non-negative quantity $A_t \in \mathcal{A}$ of inventory to order. This decision is guided by the current inventory level $Y_t \in \mathcal{Y}$ and observed side information (also referred to as feature) $X_t \in \mathcal{X} \subset \mathbb{R}^d$. After the order quantity A_t is set, a random demand $D_t \in \mathcal{D}$ is realized. For simplicity, we consider finite sets $\mathcal{A}$, $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{D}$. The inventory level then transitions according to the rule $Y_{t+1} = g(Y_t, A_t, D_t)$, where g is a transition function. In traditional inventory control, the transition function is typically given by $g(Y_t, A_t, D_t) = Y_t + A_t - D_t$ (Puterman 1994), where the new inventory level is the current level plus ordered inventory minus the demand.

However, our model allows for a more general transition function g. This expands the applicability of our problem setting beyond conventional inventory control, accommodating contexts such as capacity planning where the transition could consider factors like depreciation, supply chain management where transit delays might alter the function g, or demand response in energy grids where the interaction of supply, demand, and storage capacities can be modeled in a complex fashion.

Negative inventory levels in this model correspond to backlogged demand, which is fulfilled when additional inventory becomes available. We assume that the pairings $(X_t, D_t)_t$ are independently and identically distributed according to distribution $\mathbb{P}$.

We define $\gamma \in (0,1)$ as the discounting factor and denote $c, h, b \geq 0$ as the unit costs of ordering, holding, and backlogging items, respectively. The time-dependent cost, which includes holding and shortage costs, is represented as $q(Y_t)$, where q is the cost function. Many existing models require a specific form of this cost function to develop effective algorithms. Often, a piece-wise linear function is employed, given by $q(y) = h\max(0,-y)+b\max(0,y)$, where $h, b \geq 0$ represent constants (Scarf 1958). However, our model does not mandate specific forms of the cost function, thereby offering increased flexibility and more comprehensive real-world applications. This framework thus broadens the utility and adaptability of our model, permitting it to cater to a multitude of problem scenarios and settings.

The total expected cost to be minimized is

$$V_f(y) = \mathbb{E}\left\{ \sum_{t=0}^{\infty} \gamma^t \Big(cA_t + q[g(Y_t, A_t, D_t)] \Big) \Big| Y_0 = y \right\}, \quad (1)$$

where $A_t = f_t(Y_t, X_t)$.

Here the expectation is taken with respect to the randomness for D_t and possibly random action A_t for $t = 0, \ldots, T-1$. For simplicity, we consider the time-stationary policy, i.e., $f_t \equiv f$ for any t. By the convexity of the expectation with respect to its input, we can assume that f is a deterministic transition function. Therefore, the problem above reduces to the MDP problem

$$\min_{f:\mathcal{Y}\times\mathcal{X}\to\mathcal{A}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(cf(Y_t,X_t) + q[g(Y_t, f(Y_t,X_t), D_t)] \Big) \right]. \quad (2)$$

4. Algorithm

In this section, we propose Contextual Value Iteration (CVI) algorithm (see Algorithm 1) to find the optimal value function $V^*(y) := \min_f V_f(y)$ for our proposed general dynamic contextual newsvendor model in Section 3. Compared with value iteration for vanilla Markovian Decision Process (MDP) which estimates the transition kernel of the state y_t without using contextual information x_t, this CVI algorithm averages over x_t and estimates the conditional distribution $P(\cdot|x_t)$ of the demand D which underlies the transition kernel of y_t.

The Contextual Value Iteration (CVI) Algorithm starts by initializing the value function estimates $V_k(y)$ for each state $y \in \mathcal{Y}$. It then obtains the data $\{x_t, d_t\}_{t=1}^T \sim (X, D)$ and estimates the conditional distribution $\mathbb{P}(D|X)$ using maximum likelihood estimation (MLE). This is performed over a distribution family $\mathcal{P}$.

Algorithm 1 Contextual Value Iteration (CVI) Algorithm

Initialize $V^0(y) = 0$ for all $y \in \mathcal{Y}$.
Obtain data $\{x_t, d_t\}_{t=1}^T \sim (X, D)$.
Estimate the conditional distribution $\mathbb{P}(D|X)$ using maximum likelihood estimation (MLE) from a distribution family $\mathcal{P}$, i.e., $\hat{\mathbb{P}} := \max_{p\in\mathcal{P}} \sum_{t=1}^T \ln p(d_t|x_t)$.
for $k = 0, 1, \ldots, K-1$ **do**
  for each $y \in \mathcal{Y}$ **do**

$$\begin{aligned} V^{k+1}(y) =& \hat{\mathcal{T}}(V^k)(y) \\ :=& \frac{1}{T}\sum_{t=0}^{T-1} \min_a \mathbb{E}_{D\sim\hat{\mathbb{P}}(\cdot|x_t)}\big\{ ca + q(y,a,D) \\ &+ \gamma V^k[g(y,a,D)] \big| x_t \big\}. \end{aligned} \quad (3)$$

  end for
end for

The main iterative procedure is performed over K iterations. Within each iteration k, the algorithm sweeps over each state $y \in \mathcal{Y}$ and computes an updated value function $V^{k+1}(y)$ via the Bellman optimality operator $\mathcal{T}$ (referenced as $\hat{\mathcal{T}}$ in the algorithm to reflect the estimated transition probabilities). This computation involves minimizing the action space and taking an expectation over the estimated joint distribution of the random variables D and $\mathbf{X}$.

The algorithm continues until the difference between consecutive value function estimates is less than a predefined threshold δ, ensuring that the algorithm ceases once the estimates have adequately converged. After convergence, the algorithm yields the optimal policy by mapping each state to the action that minimizes the expected cost, incorporating the value of subsequent states as computed by the final value function.

5. Analysis

In this section, we analyze the convergence of Algorithm 1 and obtain its sample complexity. First, we will introduce the following important notations for the convergence analysis.

Given the data $\{x_t, d_t\}_{t=1}^n$, denote $\hat{\mathbb{P}}_1$ as the empirical estimation of the true distribution $\mathbb{P}_1$ of the feature X_t, and $\hat{\mathbb{P}}$ as the MLE of the true conditional distribution $\mathbb{P}(\cdot|x)$. Denote $\mathcal{M}$ as the real environment where $X_t \sim \mathbb{P}_1$ and $D_t \sim \mathbb{P}(\cdot|X_t)$ and $\hat{\mathcal{M}}$ as the estimated environment where $X_t \sim \hat{\mathbb{P}}_1$ and $D_t \sim \hat{\mathbb{P}}(\cdot|X_t)$.

Under the estimated environment $\hat{\mathcal{M}}$, define the value func-

tion as follows

$$\hat{V}_f(y) := \mathbb{E}_{X_t\sim\hat{\mathbb{P}}_1, D_t\sim\hat{\mathbb{P}}(\cdot|X_t)} \left\{\sum_{t=0}^{\infty}\gamma^t\Big(cA_t+q[g(Y_t,f(Y_t,X_t),D_t)]\Big)\Big|Y_0=y\right\}. \quad (4)$$

which is similar to the value function defined by eq. (1) under the real environment $\mathcal{M}$ with the only difference in the distribution of X_t, D_t. To facilitate the convergence analysis, the Bellman operator $\hat{\mathcal{T}}$ defined by eq. (3) under $\hat{\mathcal{M}}$ can be equivalenly written as follows.

$$\hat{\mathcal{T}}(V)(y) = \min_f \mathbb{E}_{(X,D)\sim\hat{\mathbb{P}}_{joint}}\big\{cf(y,X) + q\big[g\big(y,f(y,X),D\big)\big]+\gamma V\big[g\big(y,f(y,X),D\big)\big]\big\} \quad (5)$$

where $\hat{\mathbb{P}}_{joint}(x,d) := \hat{\mathbb{P}}_1(x)\hat{\mathbb{P}}(d|x)$ is the MLE of the joint distribution of (X,D). Similarly, under the real environment $\mathcal{M}$, we denote the optimal value function $V^* = \inf_f V_f(y)$ and define the Bellman operator $\mathcal{T}$ as follows.

$$\mathcal{T}(V)(y) := \min_{f\in\mathcal{F}} \mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}}\big\{cf(y,X) + q\big[g\big(y,f(y,X),D\big)\big]+\gamma V\big[g\big(y,f(y,X),D\big)\big]\big\} \quad (6)$$

where $\mathbb{P}_{joint}(x,d) = \mathbb{P}_1(x)\mathbb{P}(d|x)$ is the true joint distribution of (X,D).

We make the following assumption, which is widely used in reinforcement learning.

Assumption 5.1. There exists constants $a_{\max}, q_{\max} > 0$ such that $0 \le A_t \le a_{\max}$, $0 \le q(y) \le q_{\max}$ for all $y \in \mathcal{Y}$.

Then we obtain the following convergence result for Algorithm 1.

Theorem 5.2. *Under Assumption 5.1, the value function V^K obtained by Algorithm 1 converges to the optimal value function V^* at the following rate with probability at least $1-\delta$ ($\delta \in (0,1)$),*

$$\|V^K - V^*\|_\infty \le \frac{2(ca_{\max}+q_{\max})}{1-\gamma}\left(\gamma^K + \frac{\log(|\mathcal{P}_{joint}|/\delta)}{n(1-\gamma)}\right) \quad (7)$$

where $|\mathcal{P}_{joint}|$ denotes the cardinality of $\mathcal{P}_{joint}$, the candidate set of $\hat{\mathbb{P}}_{joint}$. Consequently, for any $\epsilon > 0$, we can achieve $\|V^K - V^*\|_\infty \le \epsilon$ using hyper-parameters $K \ge \frac{1}{\ln(\gamma^{-1})}\ln\left(\frac{4(ca_{\max}+q_{\max})}{\epsilon(1-\gamma)}\right)$ and $n \ge \frac{4(ca_{\max}+q_{\max})}{\epsilon(1-\gamma)^2}\log(|\mathcal{P}_{joint}|/\delta)$, which requires sample complexity $\widetilde{O}\big((1-\gamma)^{-3}\epsilon^{-1}\big)$.

Remark: The proof of Theorem 5.2 is in Appendix A.

The above convergence rate (7) consists of two terms. The first term $\frac{2\gamma^K(ca_{\max}+q_{\max})}{1-\gamma}$ shows the exponential convergence of value iteration to the optimal value function $\hat{V}^*$ under the estimated environment $\hat{M}$. The second term $\frac{2(ca_{\max}+q_{\max})}{n(1-\gamma)^2}\log(|\mathcal{P}_{joint}|/\delta)$ results from the error in estimating the joint distribution $\mathbb{P}_{joint}$ of (X,D). Hence, the distribution family $\mathbb{P}_{joint}$ should be selected such that on one hand, it is representative enough so that the true solution is close to $\mathbb{P}_{joint}$. On the other hand, $|\mathcal{P}_{joint}|$ should not be to large to control the convergence rate (7).

An alternative way to solve the dynamic contextual newsvendor model in Section 3 is to formulate it as a vanilla MDP with joint state (Y_t, X_t), named as contextual-state MDP (cMDP). In this MDP, the value iteration update becomes $V^{k+1}(y,x) = \min_a \mathbb{E}_{Y'\sim\hat{\mathbb{P}}_{Y|XYA}(\cdot|x,y,a),X'\sim\hat{\mathbb{P}}_1}[ca + q(Y') + \gamma V^k(Y',X')]$ where $\hat{\mathbb{P}}_{Y|XYA} \in \mathcal{P}_{Y|XYA}$ and $\hat{\mathbb{P}}_1 \in \mathcal{P}_1$ are obtained using MLE. This update rule is similar to Algorithm 1 with the major difference that the value function $V^k(y)$ is replaced with $V^k(y,x)$. Hence, the vanilla MDP requires to compute V^k for each joint state (y,x), which requires more computation than Algorithm 1 given $\hat{\mathbb{P}}_{joint}$. On the other hand, we can obtain the same sample complexity result by following the same proof logic. In addition, value iteration for this vanilla MDP requires the joint state (y,x) to have finitely many values, while Algorithm 1 only requires y to have finitely many values. In a similar way, it can be proved that the sample complexity of this value iteration algorithm for MDP is almost the same as that in Theorem 5.2, with the only difference that $|\mathcal{P}_{joint}|$ is replaced by $|\mathcal{P}_{Y|XYA}||\mathcal{P}_1|$.

For vanilla MDP with state Y_t, named as general MDP (gMDP) (Puterman 1994), we can use the value iteration update $V^{k+1}(y) = \min_a \mathbb{E}_{Y'\sim\hat{P}_{Y|YA}(\cdot|y,a)}[ca+q(Y')+\gamma V^k(Y')]$ where $\hat{P}_{Y|YA}$ is the estimated transition kernel. The sample complexity is almost the same as that in Theorem 5.2, with the only difference that $|\mathcal{P}_{joint}|$ is replaced by $|\mathcal{P}_{Y|YA}|$ where $\mathcal{P}_{Y|YA}$ is the function class to estimate the transition kernel $\hat{P}_{Y|YA}$.

6. Experiment

6.1. Experiment Setup

We adopt the experimental setup delineated in Zhu et al. (2012), Zhang et al. (2021) for synthetic datasets generation. Specifically, the feature-demand pair (X,D) arises from the high-dimensional quantile regression relation:

$$D = 1.7 * \big[\sin\big(2\langle X,\beta_0\rangle\big) + 2\exp\big(-16\langle X,\beta_0\rangle^2\big) + 1\big)\big] + \epsilon,$$

where $X \sim \mathcal{N}(0,\Sigma)$, with $\Sigma_{i,j} = (199/200)^{|i-j|}, i,j \in [d]$ with dimensionality $d = 5000$, $\beta_0 = [200,-200,199,-199,\ldots,1,-1,0,\ldots,0] \in \mathbb{R}^d$, and $\epsilon \sim \mathcal{N}(0,I)$. The associated demand D for each observed feature is then rounded to the closest integer in the set {0, 1, 2, ..., 9}. The objective is to investigate the performance of the system under different parameter settings. We vary the holding cost parameter h between 0.1 and 1, the

backlogging cost parameter b between 0.1 and 1, the ordering cost parameter c between 0 and 1, and the discount factor parameter γ between 0.1 and 0.9. By exploring these parameter ranges, we aim to gain insights into how different cost factors and discounting impact the optimal inventory control policy and overall system performance.

6.2. Baseline Models

For comparison, we utilized two baseline models: a generalized Markov Decision Process (gMDP)and a vanilla contextual-state MDP formulation (cMDP).These models served as benchmarks against our dynamic contextual MDP approach, which was implemented with the value iteration method and applied to a multi-period inventory control problem that accommodates backlogged demands.

6.3. Experimental Results

6.3.1. Computation Time Comparison

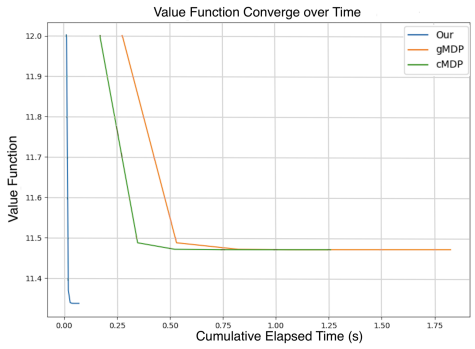

Figure 1. Comparison of the value function convergence over time for different models

We assessed the computational efficiency of the models by measuring the convergence time to obtain an optimal policy. Figure 1 illustrates the convergence comparison, demonstrating that our CMDP model utilizing value iteration surpasses both the non-contextual and vanilla MDP models in terms of convergence speed, thereby highlighting its computational efficiency. Moreover, when compared with the gMDP approach, our model, which incorporates the inventory model, exhibits accelerated convergence. Furthermore, in comparison with the cMDP approach, our model demonstrates reduced computational complexity, as discussed in Section 5, resulting in faster convergence.

6.3.2. Convergence Rate Analysis

In addition to computation time, we also compare the convergence rate in terms of iterations. The experimental results in Figure 2 indicate that our proposed model has a faster convergence rate compared to gMDP, which reveals the advantage of incorporating contextual information in our model. In addition, our model has comparable convergence

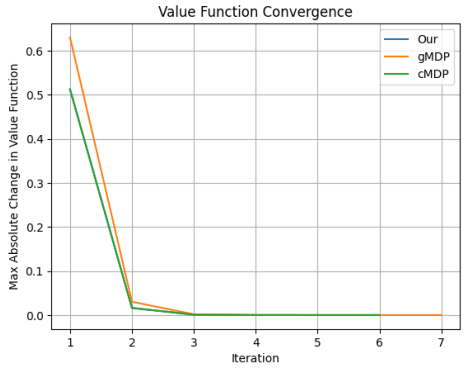

Figure 2. Comparison of maximum change of the value function across iterations for different models

rate to cMDP. Note that Section 5 compares our model with vanilla MDP, not cMDP.

6.3.3. Policy Customization

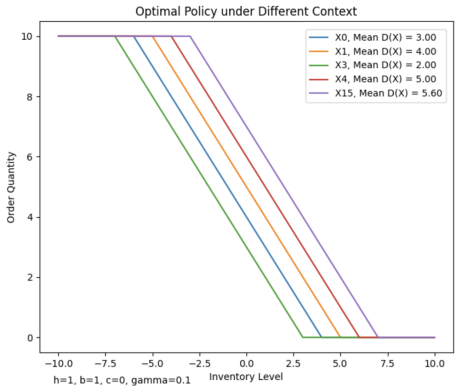

Figure 3. Optimal policies under varied contexts as generated by our model

Our algorithm stands out from conventional models due to its ability to leverage contextual information. Traditional inventory management models typically ignore varying contexts and instead depend on a fixed policy across all scenarios. This lack of adaptability could potentially lead to inefficient inventory decisions, such as stockouts or overstocking, and by extension, an increase in costs. Our algorithm addresses these limitations by using context-specific information, enabling the system to adapt to changing environmental conditions effectively and make more precise policy recommendations.

As depicted in Figure 3, our algorithm generates policies that are tailored to different contexts with different average demand. The X-axis represents the initial inventory level, while the Y-axis demonstrates the optimal order quantity under different contexts represented by the colored lines.

Let's consider an illustrative example. Suppose we are dealing with a product such as an umbrella, and the context is the weather forecast. As we all know, the demand for umbrellas would significantly vary depending upon whether it's a sunny or a rainy day. A rainy day, defined as context X_{15}, would cause a shift in the demand distribution for umbrellas, with an increased mean demand of 5.6 units. In such a scenario, our algorithm would advise maintaining a higher initial inventory level, thereby preparing for the surge in demand. In contrast, on sunny days or in other weather contexts with a lower demand for umbrellas, the algorithm intelligently suggests maintaining a lower inventory level. Such context-sensitive adjustments allow for efficient inventory management and provide a hedge against unpredictable demand fluctuations.

Overall, the ability to customize policies to fit specific contexts provides significant advantages in inventory management. This contextual awareness of our algorithm offers a dynamic solution that helps meet customer demand more effectively while minimizing costs associated with unnecessary stockpile in scenarios of lower demand. This, in essence, contributes towards building a more sustainable and cost-efficient supply chain.

7. Conclusion

In our paper, we delved into the complex landscape of the dynamic feature-based newsvendor problem within a multi-period inventory control setting with backlogged demands. Recognizing the importance of incorporating feature information into a multi-stage decision-making process, we introduced a versatile dynamic contextual newsvendor model. In the face of the complex and dynamic nature of this model, we developed the Contextual Value Iteration (CVI) algorithm. Theoretical examination of this approach has not only yielded insights into its convergence rate towards an optimal solution but also allowed us to ascertain its sample complexity. Furthermore, our experimental evaluations have underscored the superior efficiency of our proposed CVI when compared to the traditional value iteration employed in the vanilla Markovian Decision Process (MDP). This performance edge, combined with the robustness and flexibility of the dynamic contextual newsvendor model, establishes a compelling case for the use of our approach in practical multi-period inventory control scenarios.

In the future, we will extend our dynamic contextual newsvendor model to non-Markovian features, and to the setting where there are two actions respectively with immediate effect and delayed effect (Agarwal et al. 2020).

References

Abel, D., MacGlashan, J., Agarwal, A., Tellex, S., and Konidaris, G. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 10–19, 2016.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. Contextual multi-armed bandits. *Journal of Machine Learning Research*, 3:1–34, 2011.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2022.

Arrow, K. J., Harris, T., and Marschak, J. An optimal inventory policy. *Econometrica*, 19(3):250–272, 1951.

Ban, G. Y. and Rudin, C. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67 (1):90–108, 2019.

Besbes, O. and Zeevi, A. Contextual pricing: The death of list price and the new era of negotiation. *Marketing Science*, 34 (1):66–86, 2015.

Cohen, M. C., Lobel, I., and Perakis, G. Contextual learning in a newsvendor setting: Algorithms, bounds, and a new model. *Operations Research*, 64(2):366–384, 2016.

Eppen, G. D. A note on multi-item deterministic inventory models. *Management Science*, 25(3):284–286, 1979.

Gallego, G. Multi-product inventory systems under stochastic demand. *Operations Research*, 41(4):760–771, 1993.

Gallego, G. and Moon, I. A newsvendor's procurement problem when suppliers are unreliable. *Management Science*, 39(5): pp. 573–581, 1993.

Gaur, V., Kesavan, S., and Raman, A. Dynamic newsvendor model with pricing: Properties, algorithms, and simulation. *Management Science*, 51(3):451–466, 2005.

Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *ArXiv:1502.02259*, 2015.

Harrison, J. M. and Zeevi, A. A newsvendor model for a case when the demand distribution is not known. *Operations Research*, 59(2):363–376, 2011.

Karlin, S. and Scarf, H. Dynamic inventory policy with varying stochastic demands. *Management Science*, 6(1):1–22, 1959.

Langford, J. and Strehl, A. L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2256, 2007.

Lazaric, A. Transfer in reinforcement learning: A framework and a survey. In *Reinforcement Learning*, pp. 143–173. Springer, 2010.

Levi, R., Roundy, R., and Shmoys, D. The two-period newsvendor problem with partially observed markovian demand. *Operations Research*, 54(6):1066–1079, 2006.

Levi, R., Perakis, G., and Zhao, J. Learning the newsvendor: A counterfactual analysis. *Operations Research*, 66(1):50–67, 2018.

Li, L., Wei, Y., and Sanner, S. Contextual newsvendor problems: A contextual bandit approach. *Journal of Machine Learning Research*, 17(95):1–39, 2015.

McAfee, A. and Brynjolfsson, E. Big data's biggest challenge? convincing people not to trust their judgment. *Harvard Business Review*, 12, 2013.

Nahmias, S. Perishable inventory theory: A review. *Operations Research*, 30(4):680–708, 1975.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming. 1994.

Scarf, H. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, pp. 201–216, 1958.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *ICML*, volume 14, pp. 387–395, 2014.

Tang, L. and Singh, S. P. Optimization transfer for contextual markov decision processes. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1023–1030, 2010.

Veinott, A. F. Optimal policy for a multi-product, dynamic, non-stationary inventory problem. *Management Science*, 12(3): 206–222, 1965.

Zhang, H., P'erez, D., Carvalho, A. K., and Sanner, S. Trading off exploration and exploitation in a topological world. In *Advances in Neural Information Processing Systems*, pp. 1909–1917, 2013.

Zhang, L., Yang, J., and Gao, R. Optimal robust policy for feature-based newsvendor. *Optimization Online*, 2021.

Zhu, L., Huang, M., and Li, R. Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*, 22(4): 1379, 2012.

A. Proof of Theorem 5.2

We will prove Theorem 5.2 in the following steps.

Step 1: Proving that $V^* = \mathcal{T}(V^*)$ and $\hat{V}^* = \hat{\mathcal{T}}(\hat{V}^*)$

For any policy f, the Bellman equation of the value function V_f can be derived as follows.

$$
\begin{aligned}
V_f(y) \overset{(i)}{=} & \mathbb{E}_{\mathbb{P}_{joint}}\left\{\sum_{t=0}^{\infty}\gamma^t\Big(cf(Y_t,X_t)+q[g(Y_t,f(Y_t,X_t),D_t)]\Big)\Big|Y_0=y\right\}\\
=&\mathbb{E}_{\mathbb{P}_{joint}}\Big\{cf(Y_0,X_0)+q[g(Y_0,f(Y_0,X_0),D_0)]\\
&+\gamma\sum_{t=1}^{\infty}\gamma^{t-1}\Big(cf(Y_t,X_t)+q[g(Y_t,f(y,X_t),D_t)]\Big)\Big|Y_0=y\Big\}\\
=&\mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}}\big[cf(y,X)+q[g(y,f(y,X),D)]\big]+\gamma\sum_{y'\in\mathcal{Y}}\Pr(Y_1=y'|Y_0=y)\\
&\mathbb{E}_{\mathbb{P}_{joint}}\left\{\sum_{t=0}^{\infty}\gamma^t\Big(cf(Y_{t+1},X_{t+1})+q[g(Y_{t+1},f(Y_{t+1},X_{t+1}),D_{t+1})]\Big)\Big|Y_1=y'\right\}\\
\overset{(ii)}{=}&\mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}}\big[cf(y,X)+q[g(y,f(y,X),D)]\big]+\gamma\sum_{y'\in\mathcal{Y}}\sum_{x,d:g(y,f(y,x),d)=y'}\mathbb{P}_{joint}(x,d)V_f(y')\\
=&\mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}}\big[cf(y,X)+q[g(y,f(y,X),D)]\big]+\gamma\sum_{x\in\mathcal{X},d\in\mathcal{D}}\mathbb{P}_{joint}(x,d)V_f[g(y,f(y,x),d)]\\
=&\mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}}\big[cf(y,X)+q[g(y,f(y,X),D)]+\gamma V_f[g(y,f(y,X),D)]\big] \qquad (8)
\end{aligned}
$$

where (i) uses eq. (1), (ii) uses $\Pr(Y_1=y'|Y_0=y)=\sum_{x,d:g(y,f(y,x),d)=y'}\mathbb{P}_{joint}(x,d)$. By taking infimum of the above equality over $f\in\mathcal{F}$, we obtain that

$$
\begin{aligned}
V^*(y) &= \inf_{f\in\mathcal{F}} V_f(y)\\
&= \inf_{f\in\mathcal{F}} \mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}}\big[cf(y,X)+q[g(y,f(y,X),D)]+\gamma V_f[g(y,f(y,X),D)]\big]\\
&= \mathcal{T}(V^*)(y),
\end{aligned}
$$

where the last step uses eq. (6).

Similarly, $\hat{V}^* = \hat{\mathcal{T}}(\hat{V}^*)$ can be proved in the same way, with the only difference that the joint distribution of (X,D) changes from $\mathbb{P}_{joint}$ to $\hat{\mathbb{P}}_{joint}$.

Step 2: Bounding $\|V^K - \hat{V}^*\|_\infty$

For any $y\in\mathcal{Y}$ and V functions $V_1, V_2 : \mathcal{Y}\times\mathcal{X}\to\mathcal{A}$, we have

$$
\begin{aligned}
&\Big|\hat{\mathcal{T}}(V_1)(y)-\hat{\mathcal{T}}(V_2)(y)\Big|\\
&\overset{(i)}{=}\Big|\inf_{f\in\mathcal{F}}\mathbb{E}_{(X,D)\sim\hat{\mathbb{P}}_{joint}}\{cf(y,X)+q[g(y,f(y,X),D)]+\gamma V_1[g(y,f(y,X),D)]\}\\
&\quad-\inf_{f\in\mathcal{F}}\mathbb{E}_{(X,D)\sim\hat{\mathbb{P}}_{joint}}\{cf(y,X)+q[g(y,f(y,X),D)]+\gamma V_2[g(y,f(y,X),D)]\}\Big|\\
&\le\gamma\|V_1-V_2\|_\infty,
\end{aligned}
$$

where (i) uses the last step of eq. (5). Therefore, $\|\hat{\mathcal{T}}(V_1)-\hat{\mathcal{T}}(V_2)\|_\infty\le\gamma\|V_1-V_2\|_\infty$, i.e., $\hat{\mathcal{T}}$ is a γ-contraction mapping. Therefore, based on the Banach fixed-point theorem, the value iteration process (3) (i.e., $V^{k+1}=\hat{\mathcal{T}}(V^k)$) converges exponentially fast to the fixed point $\hat{V}^*$ of $\mathcal{T}$, i.e.,

$$
\|V^K-\hat{V}^*\|_\infty\le\gamma^K\|V^0-\hat{V}^*\|_\infty\le\frac{2\gamma^K(ca_{\max}+q_{\max})}{1-\gamma}, \qquad (9)
$$

where the second $\leq$ uses $V_0 = 0$, $\hat{V}^* = \inf_{f\in\mathcal{F}} \hat{V}_f$, eq. (4) and Assumption 5.1.

Step 3: Bounding $\|\hat{V}^* - V^*\|_\infty$

For any $y \in \mathcal{Y}$, we have

$$
\begin{aligned}
&|\hat{V}^*(y) - V^*(y)| \\
=&|\hat{\mathcal{T}}(\hat{V}^*)(y) - \mathcal{T}(V^*)(y)| \\
=&\Big| \inf_{f\in\mathcal{F}} \mathbb{E}_{(X,D)\sim\hat{\mathbb{P}}_{joint}} \big\{cf(y,X) + q[g(y,f(y,X),D)] + \gamma\hat{V}^*[g(y,f(y,X),D)]\big\} \\
&\quad - \inf_{f\in\mathcal{F}} \mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}} \big\{cf(y,X) + q[g(y,f(y,X),D)] + \gamma V^*[g(y,f(y,X),D)]\big\}\Big| \\
\leq& \sup_{f\in\mathcal{F}} \Big|\mathbb{E}_{(X,D)\sim\hat{\mathbb{P}}_{joint}} \big\{cf(y,X) + q[g(y,f(y,X),D)] + \gamma\hat{V}^*[g(y,f(y,X),D)]\big\} \\
&\quad - \mathbb{E}_{(X,D)\sim\mathbb{P}_{joint}} \big\{cf(y,X) + q[g(y,f(y,X),D)] + \gamma V^*[g(y,f(y,X),D)]\big\}\Big| \\
\leq& \sup_{f\in\mathcal{F}} \Big| \sum_{X,D} [\hat{\mathbb{P}}_{joint}(X,D) - \mathbb{P}_{joint}(X,D)]\big[q(y + f(y,X) - D) + \gamma\big\{V^*(y + f(y,X) - D)\big\}\big] \\
&\quad + \gamma \sum_{X,D} \hat{\mathbb{P}}_{joint}(X,D)[\hat{V}^*(y + f(y,X) - D) - V^*(y + f(y,X) - D)] \\
&\quad + c\sum_{X} [\hat{\mathbb{P}}_1(X) - \mathbb{P}_1(X)] f(y,X)\Big|
\end{aligned}
$$

Therefore, by taking maximum of the above inequality with respect to $y \in \mathcal{Y}$, we obtain that

$$
\begin{aligned}
\|\hat{V}^* - V^*\|_\infty &\leq \Big(q_{\max} + \frac{\gamma(ca_{\max} + q_{\max})}{1-\gamma}\Big)\|\hat{\mathbb{P}}_{joint} - \mathbb{P}_{joint}\|_1 + \gamma\|\hat{V}^* - V^*\|_\infty + ca_{\max}\|\hat{\mathbb{P}}_1 - \mathbb{P}_1\|_1 \\
&\overset{(i)}{\leq} \gamma\|\hat{V}^* - V^*\|_\infty + \frac{ca_{\max} + q_{\max}}{1-\gamma} \cdot \sqrt{\frac{2}{n}\log(|\mathcal{P}_{joint}|/\delta)}
\end{aligned}
$$

where (i) uses Theorem 21 of (Agarwal et al. 2020) and $\mathcal{P}_{joint}$ denotes function class from which the MLE $\hat{\mathbb{P}}_{joint}$ is selected. By rearranging the above inequality, we obtain that

$$
\|\hat{V}^* - V^*\|_\infty \leq \frac{2(ca_{\max} + q_{\max})}{n(1-\gamma)^2}\log(|\mathcal{P}_{joint}|/\delta). \tag{10}
$$

Step 4: Obtaining convergence results

By adding up eqs. (9) & (10), we prove eq. (7) as follows.

$$
\begin{aligned}
\|V^K - V^*\|_\infty \leq& \|\hat{V}^* - V^*\|_\infty + \|V^K - \hat{V}^*\|_\infty \\
\leq& \frac{2\gamma^K(ca_{\max} + q_{\max})}{1-\gamma} + \frac{2(ca_{\max} + q_{\max})}{n(1-\gamma)^2}\log(|\mathcal{P}_{joint}|/\delta) \\
\leq& \frac{2(ca_{\max} + q_{\max})}{1-\gamma}\Big(\gamma^K + \frac{\log(|\mathcal{P}_{joint}|/\delta)}{n(1-\gamma)}\Big)
\end{aligned}
$$

When $K \geq \frac{1}{\ln(\gamma^{-1})}\ln\Big(\frac{4(ca_{\max}+q_{\max})}{\epsilon(1-\gamma)}\Big) = \widetilde{O}\big((1-\gamma)^{-1}\big)$ and $n \geq \frac{4(ca_{\max}+q_{\max})}{\epsilon(1-\gamma)^2}\log(|\mathcal{P}_{joint}|/\delta) = \widetilde{O}\big((1-\gamma)^{-2}\epsilon^{-1}\big)$, the above inequality yields that $\|V^K - V^*\|_\infty \leq \epsilon$. The corresponding sample complexity is $KT \geq \widetilde{O}\big((1-\gamma)^{-3}\epsilon^{-1}\big)$.