Manuscript submitted to POM Journal

Balancing Revenues and Repair Costs under Partial Information about Product Reliability

Chao Ding, Paat Rusmevichientong, Huseyin Topaloglu

We consider the problem faced by a company selling a product with warranty and under partial information about the product reliability. The product can fail from multiple failure types, each of which is associated with an inherently different repair cost. If the product fails within the warranty duration, then the company is required to pay the repair cost. The company does not know the probabilities associated with different failure types, but it learns the failure probabilities as sales occur and failure information is accumulated. If the failure probabilities turn out to be too high and it becomes costly to fulfill the warranty coverage, then the company may decide to stop selling the product, possibly replacing it with a more reliable alternative. The objective is to decide if and when to stop. By formulating the problem as a dynamic program with Bayesian learning, we establish structural properties of the optimal policy. Since computing the optimal policy is intractable due to the high dimensional state space, we propose two approximation methods. The first method is based on decomposing the problem by failure types and it provides upper bounds on the value functions. The second method provides lower bounds on the value functions and it is based on a deterministic approximation. Computational experiments indicate that the policy from the first method provides noticeable benefits, especially when it is difficult to form good estimates of the failure probabilities quickly.

Key words: revenue management; optimal stopping; Bayesian learning; stochastic orders

1. Introduction

We consider the problem faced by a company selling a product with warranty, while the reliability information of the product is only partially available to the company. The product can be a physical product or a repair service agreement. Whenever a sale occurs, the company receives a one time or monthly payment from its customers. In return, the company is required to cover the repair cost of the product during a certain warranty duration. Typically, a product can fail due to multiple failure types, each of which is associated with an inherently different repair cost and an unknown failure probability. Initially, the company only has rough estimates of these failure probabilities, usually from experience with similar products or a short test marketing stage that involves a small number of units. If the true failure probabilities turn out to be too high and it becomes costly to fulfill the warranty coverage, then the company may want to stop selling the product, possibly replacing it with a more reliable alternative. As shown in the following examples, problems of this flavor naturally arise in a variety of industries.

Example 1: Printer Technical Support. To minimize the downtime of printing equipment, customers, such as schools and business organizations, often buy a technical support agreement

from either the manufacturer or a certified service provider for their printers. Companies providing such services include Xerox, MIDCOM Service Group and CPLUS Critical Hardware Support. Customers pay either a one time or monthly fee to the service provider. The service provider is responsible for on site or depot repairs of the printer during the agreed time period, ranging from one to five years; see Xerox Corporation (2011) for a sample service agreement. The printer can break down due to fuser, motor, network failures and so on. The labor and part costs associated with different failures can vary widely. If the failure probabilities prove to be too high, then the company may choose to stop selling this service or increase its fee for future contracts. Similar technical support services for other types of equipment widely exist.

Example 2: Extended Vehicle Warranty. Most new vehicles come with a manufacturer warranty, which typically covers the repair of components for a limited duration, for example, three years or 36,000 miles, whichever comes first. When the initial manufacturer warranty expires, the customer can buy an extended vehicle warranty, mostly from a third party service provider such as Warranty Direct and CARCHEX; see, for example, Top Extended Vehicle Warranty Providers (2010). The warranty contract works essentially like an insurance. The service provider charges a monthly fee to the customer for a certain warranty duration, ranging from several months to ten years. If anything covered by the warranty breaks down within this time period, then the service provider pays the repair cost, including material and labor. Depending on the type of warranty contract and the failed components, the repair costs can vary from a few hundred dollars to several thousands. If the failure probabilities turn out to be too high, then the service provider may choose to increase the payment rate of such warranty service for future demands.

Example 3: Cell Phone Warranty. Cell phone service providers routinely introduce new cell phone models into the market. These companies make a certain revenue with the sale of each cell phone, while each sold cell phone comes with a warranty that covers failures for a certain duration of time. Generally, a cell phone can fail due to failures in five to fifteen different categories, while the repair costs for failures in different categories can vary substantially. For instance, according to a nationwide repair service provider Mission Repair (2012), the screen repair for iPhone 4S costs $99, the home button repair costs $79, while the battery repair costs $39. If the failure probabilities turn out to be too high and the warranty coverage cost offsets a large portion of the sales revenue, then the company may decide to stop selling a particular cell phone model and turn its attention to other possible alternative models.

The above motivating applications share common features. First, the product covered by the warranty is fairly complex and it can fail from multiple failure types, while the repair costs

associated with these failure types can vary substantially. Second, the company has a priori knowledge of the types of failures that can occur, while it has limited information about the failure probabilities. Third, the company has the option to stop selling the current product, possibly with some follow up actions such as increasing the payment rate, terminating the product or switching to another more reliable alternative. In case a stopping decision is made, the already sold warranty contracts usually need to be honored by the company. The fundamental tradeoff is that if the company waits too long to get a better understanding of the reliability of the product, then it may introduce too many units into the market and incur excessive repair costs due to them. On the other hand, if the company does not spend enough time collecting data, then it may base its decisions on unreliable estimates of the failure probabilities.

In this paper, motivated by the applications above, we analyze an optimal stopping model that balances revenues and repair costs under partial information about product reliability. We formulate the problem as a dynamic program with a high dimensional state variable that keeps track of our beliefs about the probabilities of different failure types. As new failure information is accumulated at each time period, we adjust our beliefs about the probabilities of different failure types according to a Bayesian updating scheme. At each time period, we decide whether to continue or stop selling the product. The objective is to maximize the total expected profit, which is given by the difference between the revenue from the sales and the total expected repair cost.

We make the following contributions in this paper. We give a characterization of the optimal policy by showing that the value functions in our dynamic programming formulation are decreasing and convex in our estimates of the failure probabilities (Proposition 1). By using this result, we establish that the optimal policy is of boundary type (Propositions 2) and show that stopping will not be optimal at the end of a period that contains no failures (Proposition 3). To deal with the high dimensional state variable, we give two tractable approximation methods. The first approximation method decomposes the problem into a collection of one dimensional dynamic programs, one for each failure type. By combining the value functions obtained from each one dimensional dynamic program, we obtain upper bounds on the original value functions (Proposition 4). The second approximation method is based on a deterministic formulation that ignores the benefits from future learning. Complementing the first method, we show that the second approximation method provides lower bounds (Proposition 5). By using the second approximation method, we demonstrate that the relative value of learning the failure probabilities is larger when the demand quantities in the different time periods are smaller (Proposition 6). Finally, we establish that the lower bounds from the second approximation method are asymptotically tight as the demand in each

time period scales up linearly with the same rate (Proposition 7). Our numerical experiments compare the performance of the policies from the two approximation methods with a variety of other heuristic policies. Our computational results indicate that the policy from the first approximation method provides noticeable improvements with reasonable computational effort, especially when it is difficult to form good estimates of the failure probabilities quickly.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the related literature. We give a dynamic programming formulation of the problem in Section 3, followed by the structural properties of the optimal policy in Section 4. In Section 5, we develop an upper bound approximation based on a dynamic programming decomposition idea. In Section 6, we provide a lower bound approximation based on a deterministic formulation, along with an analysis of the benefits of learning. Computational experiments comparing the policies obtained by the two approximation methods with various heuristics appear in Section 7. In Section 8, we provide concluding remarks and discuss possible extensions of our model.

2. Literature Review

Our paper is related to several streams of literature. The first stream of work involves operations management models that have embedded optimal stopping or learning features. Feng and Gallego (1995) consider the optimal timing of a single price change between two fixed prices. They show that the optimal policy is characterized by sequences of time thresholds that depend on the number of unsold units. Aviv and Pazgal (2005) use partially observed Markov decision processes to analyze a dynamic pricing problem while learning an unknown demand parameter. They develop upper bounds on the expected revenue and propose heuristics based on modifications of the available information structure. Bertsimas and Mersereau (2007) give a model for adaptively learning the effectiveness of ads in interactive marketing. The authors formulate the problem as a dynamic program with Bayesian learning and propose a decomposition based approximation approach. Our decomposition idea resembles theirs. Caro and Gallien (2007) study a dynamic retail assortment problem, where they use the multi armed bandit framework to learn the demand. They develop an index policy and make extensions to incorporate lead times, assortment change costs and substitution effects. Araman and Caldentey (2009) study a dynamic pricing problem under incomplete information for the market size. They formulate the problem as an intensity control problem, derive structural properties of the optimal policy and give approximation methods. Farias and Van Roy (2010) study a dynamic pricing problem, where the willingness to pay distribution of the customers is known, but the customer arrival rate is unknown. The authors develop a

heuristic to learn the customer arrival rate dynamically and give a performance guarantee for their heuristic. Boyaci and Ozer (2010) study a model where a manufacturer collects advance demand information to set its capacity level and characterize when it is optimal to stop collecting the demand information. Harrison et al. (2010) consider a pricing problem where one of the two demand models is known to apply, but they do not know which. They give asymptotic analyses for a family of learning policies. Arlotto et al. (2011) study an optimal staffing problem through an infinite armed bandit model. They characterize the optimal policy as a Gittins index policy.

The second stream of related work involves decomposition methods for high dimensional dynamic programs. Hawkins (2003) develops a Lagrangian relaxation method for the so called weakly coupled dynamic programs, where the original dynamic program would decompose into a collection of independent single dimensional subproblems when one relaxes a set of linking constraints on the action space. Adelman and Mersereau (2008) compare the approximate linear programming approach and the Lagrangian relaxation method for such weakly coupled dynamic programs. Topaloglu (2009) explores the Lagrangian relaxation method in the network revenue management setting to come up with tractable policies for controlling airline ticket sales. A similar approach is adopted in Topaloglu and Kunnumkal (2010) to compute bid prices in network revenue management problems, whereas Erdelyi and Topaloglu (2010) extend this approach to obtain tractable solution methods for joint capacity allocation and overbooking problems over an airline network. Brown et al. (2010) compute bounds on the value functions by relaxing the nonanticipativity constraints and provide an analysis based on the duality theory.

Since our problem involves choosing between continuing and stopping under incomplete information, a third stream of related literature is bandit problems. The one armed bandit problem considers sampling from two options, where the reward from one option is known, but the expected reward from the other option depends on an unknown parameter. The goal is to maximize the total expected reward from sampling. Bradt et al. (1956) consider a setting that allows finite number of sampling opportunities and show that the optimal policy is to stick with the known option until the end once we switch to this option. Gittins (1979) characterizes the optimal policy for the infinite horizon one armed bandit problem as an index policy. Burnetas and Katehakis (1998) generalize the results in Bradt et al. (1956) to the single parameter exponential family, whereas Burnetas and Katehakis (2003) provide asymptotic approximations when the number of time periods goes to infinity. Goldenshluger and Zeevi (2009, 2011) consider minimax formulations of the one armed bandit problem, establish lower bounds on the minimum regret under an arbitrary policy and propose a policy that achieves a matching upper bound. When there are multiple options to choose

from, the problem is referred to as a multi armed bandit problem. Berry and Fristedt (1985) and Gittins (1989) give early analyses of multi armed bandit problems. Such problems tend to be significantly more difficult than their one armed counterparts, since one needs to keep track of the beliefs about the rewards of multiple options. The problem in this paper can be viewed as a one armed bandit problem, where the unknown expected reward depends on multiple unknown parameters. In this one armed bandit problem, the option with the unknown reward corresponds to continuing to sell the product and the multiple unknown parameters correspond to the failure probabilities. Similar to our problem, Burnetas and Katehakis (1996) study bandit problems where the unknown expected reward depends on multiple unknown parameters. They use a frequentist approach with the regret criterion, but we use a Bayesian formulation. Our Bayesian formulation allows us to work with an optimality equation and our solution strategy is based on obtaining approximate solutions to the optimality equation.

There is a rich literature in optimal stopping and its applications. Chow et al. (1971) and Peskir and Shiryaev (2006) provide overviews of optimal stopping problems. One problem in this area that is related to our work is optimal stopping of a risk process. This problem models an insurance company receiving premiums and paying out claims that occur according to a renewal process. The company invests its capital and premiums at an interest rate and claims increase at a particular rate. The objective is to stop the process at a random time to maximize the net expected gain. Boshuizen and Gouweleeuw (1993) and Muciek and Szajowski (2008) study such models of various complexities. Schöttl (1998) studies a slight modification of the problem, where the goal is to find the optimal time to recalculate the premiums so as to maximize the expected net gain. Stopping a random process to maximize the net expected gain bears resemblance to our problem, but the risk process literature usually works with known problem parameters.

Finally, the work on providing insurance under incomplete information is relevant to our paper. Learning becomes crucial when an insurance company attempts to infer the risk levels of its insurance buyers. In these settings, insurance buyers may act strategically and not report accidents to manipulate how the insurance company infers their risk levels. By doing so, they try to secure low insurance premiums in the future. Cooper and Hayes (1987), Hosios and Peters (1989) and Watt and Vazquez (1997) study multi period models to address the strategic interaction between an insurance company and its clients. In these models, the price of the service or the insurance premium is an endogenous decision made by the insurance company and this feature causes the clients to try to manipulate the perception of their risk levels. In our problem, the price of the warranty is a parameter fixed exogenously. Thus, we do not consider the strategic interaction between the company and its clients.

3. Problem Formulation

We sell a product over a finite selling horizon. The product can fail from multiple failure types and we sell the product with a warranty that covers the repairs of failed units within a certain warranty duration. We incur a repair cost when a unit fails within its warranty duration. The probabilities of different failure types are unknown at the beginning of the selling horizon, but we build estimates of the failure probabilities as failure information is accumulated over time. If the failure probabilities turn out to be high and the repairs become too costly, then we may decide to stop selling the product and avoid introducing new units into circulation, while honoring the warranty coverage of existing units. On the other hand, if the failure probabilities turn out to be low, then we may continue selling the product until the end of the selling horizon. The goal is to decide if and when to stop selling the product so as to maximize the total expected profit, which is the difference between the total revenue from sales and the total expected repair cost.

There are n failure types indexed by $1, \ldots, n$, each of which has an associated unknown failure probability. We assume that the true failure probabilities do not change overtime, but they are unknown to us. The selling horizon consists of time periods $\{0, 1, \ldots, \tau\}$. At time period zero, we sell the product to obtain an initial belief for the failure probabilities. We do not make a decision at the beginning of this time period. At the beginning of the other time periods, we need to decide whether to stop or continue selling the product. Thus, we can view time period zero as a test marketing stage during which we form an initial belief about the failure probabilities. If we already have a prior belief for the failure probabilities, then we can modify our model to skip time period zero and start directly at time period one with the prior belief. At each time period, each unit may suffer from one or more of multiple failure types. We assume that the failures of different units and the failures of each unit from different failure types are independent of each other.

We generate a revenue of r from each unit sold. A sold unit is covered by warranty for K consecutive time periods. If a unit under warranty fails from failure type i, then we incur a repair cost of c_i. A repaired unit remains in warranty only for the duration left in the original warranty contract. In other words, the warranty for a particular unit does not start from scratch after each repair. For simplicity, we ignore the possible lead time when a unit is being repaired. This assumption is reasonable in the application settings we consider because companies usually have a number of spare units that they can use to immediately replace a failed unit as the failed unit is being repaired or a technician is sent immediately when a repair is requested. We assume that the repaired units have the same unknown failure probabilities as the brand new units. This assumption is reasonable when the failure types are mostly electrical or the repair is carried out by replacing the

broken parts with new ones, which is again the case for the application settings that motivate our work. Similarly, we assume that the failure probabilities do not depend on how long ago a product was sold to a customer and how long ago a product was repaired. This assumption is reasonable when the failure types are mostly electrical, rather than wear and tear related. We expand on this assumption when listing possible extensions at the end of the paper in Section 8.

We use D_t to denote the demand for the product at time period t, which is assumed to be deterministic. The deterministic demand assumption allows us to focus on the learning dynamics for the failure probabilities and to identify structural insights from our model. We discuss possible complications from relaxing this assumption at the end of the paper. Under this assumption the number of units under warranty coverage in time period t, denoted by W_t, is simply the total demand in the last K time periods. Thus, we have $W_t = \sum_{s=0}^{t} \mathbf{1}(t-s<K)\, D_s$, where $\mathbf{1}(\cdot)$ is the indicator function. In this case, W_t corresponds to the number of units that we can potentially receive as failed units for repairs at time period t. We naturally expect to receive only a fraction of these units as failed units because not all of them fail at the same time. Finally, if we let $M_t = W_0 + W_1 + \ldots + W_{t-1}$, then M_t is the maximum number of units that we could have potentially received as failed units for repairs up until time period t. Each returned unit could have n types of failure, implying that $n\, M_t$ is the maximum possible number of failures. Since the demand is a deterministic quantity, W_t and M_t are deterministic quantities, but the total number of failures will be random and governed by the probability of each type of failure.

3.1. Learning Dynamics

The learning process is based on a Bayesian update of the failure probabilities. At the beginning of each time period, our prior belief about the probability of a particular failure type has a beta distribution. After observing the number of failed units from a particular failure type, we apply the Bayes rule to obtain an updated posterior belief about the failure probability. Since each unit fails independently and the beta distribution is a conjugate prior of the binomial distribution, our posterior belief continues to have a beta distribution.

Let P_{it} denote our random prior belief at the beginning of time period t for the probability of failure type i. We recall that M_t corresponds to the maximum number of units that we could have potentially received as failed units up until time period t. Using θ_{it} to denote the proportion of the M_t units that have actually failed from failure type i, we assume that the random variable P_{it} has a beta distribution with parameters $(\theta_{it}\, M_t, (1-\theta_{it})\, M_t)$. The parameters $\theta_{it}\, M_t$ and $(1-\theta_{it})\, M_t$ are the number of units that have failed and have not failed, respectively, from failure type i up until time period t. The expected value of P_{it} is $\theta_{it}\, M_t / [\theta_{it}\, M_t + (1-\theta_{it})\, M_t] = \theta_{it}$, which agrees

with the intuition that the expected value of our prior belief for the probability of failure type i is equal to the proportion of units that have failed from failure type i up to time period t.

Recall that W_t is the number of units that is still under warranty at time period t. If we let the random variable Y_{it} denote the number of units that we actually receive as failed units from failure type i at time period t, then our prior belief implies that Y_{it} has the binomial distribution with parameters (W_t, P_{it}), where the second parameter P_{it} is itself a beta random variable with parameters $(\theta_{it}\, M_t, (1-\theta_{it})\, M_t)$. Binomial random variables whose second parameter has a beta distribution are commonly referred to as beta binomial random variables. In this case, since the beta distribution is a conjugate prior for the binomial distribution, it is well known that our posterior belief at time period t for the probability of failure type i has a beta distribution with parameters $(\theta_{it}\, M_t + Y_{it}, (1-\theta_{it})\, M_t + W_t - Y_{it})$. The two parameters of this distribution correspond to the number of units that have failed and have not failed, respectively, from failure type i up until time period $t+1$. Throughout the rest of the paper, we prefer writing the random variables P_{it} and Y_{it} as $P_{it}(\theta_{it})$ and $Y_{it}(\theta_{it})$ to explicitly emphasize the fact that the distributions of these random variables depend on θ_{it}. By conditioning on $P_{it}(\theta_{it})$, we compute the expected value of $Y_{it}(\theta_{it})$ as $\mathbb{E}\{Y_{it}(\theta_{it})\} = \mathbb{E}\{\mathbb{E}\{Y_{it}(\theta_{it}) \,|\, P_{it}(\theta_{it})\}\} = \mathbb{E}\{W_t\, P_{it}(\theta_{it})\} = W_t\, \theta_{it}$. This computation shortly becomes useful when constructing the cost function.

3.2. Dynamic Program

Our prior belief at time period t for the probability of failure type i has a beta distribution with parameters $(\theta_{it}\, M_t, (1-\theta_{it})\, M_t)$. Noting that M_t is a deterministic quantity, we only need to know θ_{it} in order to keep track of our prior belief for the probability of failure type i. Therefore, we can use $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{nt})$ as the state variable in our dynamic programming formulation. Since θ_{it} corresponds to the proportion of the M_t units that we have actually received as failed units from failure type i up until time period t, the dynamics of θ_{it} is given by

$$\theta_{i,t+1} = \frac{\theta_{it}\, M_t + Y_{it}(\theta_{it})}{M_{t+1}} = \frac{M_t}{M_{t+1}}\, \theta_{it} + \Big[1 - \frac{M_t}{M_{t+1}}\Big] \frac{Y_{it}(\theta_{it})}{W_t}, \tag{1}$$

where we use the fact that $M_{t+1} - M_t = W_t$. Using the vector $\boldsymbol{Y}_t(\boldsymbol{\theta}_t) = (Y_{1t}(\theta_{1t}), \ldots, Y_{nt}(\theta_{nt}))$ and defining the deterministic quantity $\lambda_t = M_t / M_{t+1}$, we write the dynamics of $\boldsymbol{\theta}_t$ in vector notation as $\boldsymbol{\theta}_{t+1} = \lambda_t\, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t)$. The quantity $\lambda_t \in [0,1]$ is called the shrinkage factor.

To develop the cost structure of our dynamic program, we begin by considering the case where we continue selling the product at time period t. In this case, the number of units that fail from failure type i at time period t is given by the random variable $Y_{it}(\theta_{it})$. Since $\mathbb{E}\{Y_{it}(\theta_{it})\} = W_t\, \theta_{it}$, we incur an expected repair cost of $C_t(\boldsymbol{\theta}_t) = \sum_{i=1}^{n} c_i\, W_t\, \theta_{it}$, if we continue selling the product. On

the other hand, if we decide to stop selling the product at time period t, then all future demands are lost, while the warranty for the already sold units needs to be honored. In this case, the number of units that remain under warranty coverage at a future time period $\ell \geq t$ is given by $\sum_{s=0}^{t-1} \mathbf{1}(\ell - s < K)\, D_s$. Therefore, the number of units that we receive as failed units from failure type i at the future time period ℓ is given by a beta binomial random variable with parameters $\big(\sum_{s=0}^{t-1} \mathbf{1}(\ell - s < K)\, D_s, P_{it}(\theta_{it})\big)$, whose expectation is given by $\sum_{s=0}^{t-1} \mathbf{1}(\ell - s < K)\, D_s\, \theta_{it}$, where we use the same conditioning argument that we use to compute the expectation of $Y_{it}(\theta_{it})$. Adding over all of the future time periods and all of the failure types, this implies that we incur an expected repair cost of $S_t(\boldsymbol{\theta}_t) = \sum_{i=1}^{n} \sum_{\ell=t}^{\infty} \sum_{s=0}^{t-1} c_i\, \mathbf{1}(\ell - s < K)\, D_s\, \theta_{it}$ from time period t onwards, given that we stop selling the product at time period t. The implicit assumption in the last cost expression is that if a unit is covered by the warranty beyond the selling horizon, then we are responsible for fulfilling the repairs for this unit until its warranty coverage expires. If stopping to sell the product also means switching to an alternative product, then we show how to incorporate the additional revenue stream from the alternative product in Section 8.

We can formulate the problem as a dynamic program by using $\boldsymbol{\theta}_t$ as the state variable at time period t. If we continue selling the product at time period t, then we generate a revenue of $r D_t$ and incur an expected repair cost of $C_t(\boldsymbol{\theta}_t)$, whereas if we stop selling the product at time period t, then we incur an expected repair cost of $S_t(\boldsymbol{\theta}_t)$. Using the learning dynamics given in (1), the value functions satisfy the optimality equation

$$\vartheta_t(\boldsymbol{\theta}_t) = \max\Big\{ r D_t - C_t(\boldsymbol{\theta}_t) + \mathbb{E}\{\vartheta_{t+1}(\lambda_t\, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}, -S_t(\boldsymbol{\theta}_t) \Big\},$$

with the boundary condition $\vartheta_{\tau+1}(\cdot) = -S_{\tau+1}(\cdot)$. If the first term in the max operator is larger than the second term, then it is optimal to continue. Otherwise, it is optimal to stop. The expectation operator involves the random variable $\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$. It turns out that we can simplify the optimality equation by using a relationship between $C_t(\cdot)$ and $S_t(\cdot)$. Adding $S_t(\boldsymbol{\theta}_t)$ to both sides of the optimality equation and letting $V_t(\boldsymbol{\theta}_t) = \vartheta_t(\boldsymbol{\theta}_t) + S_t(\boldsymbol{\theta}_t)$, we obtain

$$V_t(\boldsymbol{\theta}_t) = \max\Big\{ r D_t - C_t(\boldsymbol{\theta}_t) + S_t(\boldsymbol{\theta}_t) + \mathbb{E}\{V_{t+1}(\lambda_t\, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} - \mathbb{E}\{S_{t+1}(\lambda_t\, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}, 0 \Big\},$$

with the boundary condition $V_{\tau+1}(\cdot) = 0$. From the definition of $S_{t+1}(\cdot)$, we see that it is a linear function, hence we have $\mathbb{E}\{S_{t+1}(\lambda_t\, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} = S_{t+1}(\mathbb{E}\{\lambda_t\, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t)\}) = S_{t+1}(\boldsymbol{\theta}_t)$, where we use the fact that $\mathbb{E}\{Y_{it}(\theta_{it})\} = W_t\, \theta_{it}$. Therefore, in the above optimality equation, we can write the expression $C_t(\boldsymbol{\theta}_t) - S_t(\boldsymbol{\theta}_t) + \mathbb{E}\{S_{t+1}(\lambda_t\, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}$ as $C_t(\boldsymbol{\theta}_t) - S_t(\boldsymbol{\theta}_t) +$

$S_{t+1}(\boldsymbol{\theta}_t)$. In addition, using the definitions of $C_t(\cdot)$ and $S_t(\cdot)$, a simple algebraic manipulation given in Appendix A shows that $C_t(\boldsymbol{\theta}_t) - S_t(\boldsymbol{\theta}_t) + S_{t+1}(\boldsymbol{\theta}_t) = K \sum_{i=1}^{n} c_i \, \theta_{it} \, D_t = K \boldsymbol{c}^\top \boldsymbol{\theta}_t \, D_t$, where we let $\boldsymbol{c} = (c_1, \ldots, c_n)$. Thus, we can write the last optimality equation as

$$V_t(\boldsymbol{\theta}_t) = \max \left\{ (r - K \boldsymbol{c}^\top \boldsymbol{\theta}_t) \, D_t + \mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}, 0 \right\} . \tag{2}$$

The optimality equation above has an intuitive explanation. According to our belief at time period t, we expect a unit to fail from failure type i with probability θ_{it} and the expression $K \boldsymbol{c}^\top \boldsymbol{\theta}_t$ can be interpreted as the expected repair cost of a unit over its whole warranty coverage. Therefore, the expression $r - K \boldsymbol{c}^\top \boldsymbol{\theta}_t$ in the optimality equation above corresponds to the expected net profit contribution of a sold unit. The optimality equation in (2) indicates that for each unit sold at time period t, the total expected repair cost over the whole warranty duration can be charged immediately at time period t according to our belief at this time period. This shift in the timing of costs is possible because the repair cost is linear in $\boldsymbol{\theta}_t$ and the Bayesian updates have the martingale property satisfying $\mathbb{E}\{\lambda_t \, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t)\} = \boldsymbol{\theta}_t$; see Whittle (1983).

In the optimality equation in (2), it is optimal to continue selling the product at time period t whenever the state $\boldsymbol{\theta}_t$ satisfies $(r - K \boldsymbol{c}^\top \boldsymbol{\theta}_t) \, D_t + \mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} > 0$. Furthermore, the state at time period t satisfies the last inequality if and only if $V_t(\boldsymbol{\theta}_t) > 0$. Therefore, the set of states at which it is optimal to continue selling the product at time period t is given by $\mathcal{C}_t = \{\boldsymbol{\theta}_t \in [0,1]^n : V_t(\boldsymbol{\theta}_t) > 0\}$. We obtain an optimal policy by continuing to sell the product at time period t if and only if the state $\boldsymbol{\theta}_t$ at this time period satisfies $\boldsymbol{\theta}_t \in \mathcal{C}_t$.

Our formulation assumes that we know all types of failures that can occur. In this case, we can keep a belief about the probability of failure for each failure type. Exploiting the assumption that the failures from different types are independent, we obtain a tractable approach to update our beliefs. At the end of the paper, we discuss a possible modification of this assumption, where each unit can fail from a single failure at a time, inducing dependence across failure types. This modification results in a Dirichlet multinomial learning model that is more difficult to work with. Also, we observe that our beliefs about the failure probabilities only affect the repair costs in the optimality equation in (2). Therefore, instead of working with multiple failure types and keeping our beliefs about the associated failure probabilities, we can keep a belief about an unknown repair cost distribution. This alternative approach bypasses the necessity to explicitly work with multiple failure types. Assuming that the repair cost distribution can be characterized by a small number of parameters, this alternative approach may require learning fewer unknown parameters. Nevertheless, we obtain a rather simple scheme to update our prior beliefs when we keep track of our beliefs about individual failure probabilities separately.

4. Structural Properties

In this section, we show that the value functions are componentwise decreasing and convex. Using these properties, we establish the monotonicity of the optimal decision as a function of the state variable and the time period. To that end, following Shaked and Shanthikumar (2007), we say that a family of random variables $\{X(\gamma) : \gamma \in \Re\}$ is stochastically increasing if $\mathbb{E}\{\phi(X(\gamma))\}$ is increasing in γ for any increasing function $\phi(\cdot)$ on $\Re$. Similarly, a family of random variables $\{X(\gamma) : \gamma \in \Re\}$ is stochastically convex if $\mathbb{E}\{\phi(X(\gamma))\}$ is convex in γ for any convex function $\phi(\cdot)$ on $\Re$. Replicating the proofs of Theorems 8.A.15 and 8.A.17 in Shaked and Shanthikumar (2007), we can show the following closure properties for stochastically increasing and convex families.

LEMMA 1 **(Closure Properties)**. *We have the following properties.*
(1) *Assume that the families of random variables* $\{X(\gamma) : \gamma \in \Re\}$ *and* $\{Y(\gamma) : \gamma \in \Re\}$ *are stochastically increasing and stochastically convex. Furthermore, assume that* $X(\gamma)$ *and* $Y(\gamma)$ *are independent of each other for all* $\gamma \in \Re$. *Then, for any* $a, b \in \Re_+$, *the family of random variables* $\{a\,X(\gamma) + b\,Y(\gamma) : \gamma \in \Re\}$ *is stochastically increasing and stochastically convex.*
(2) *Let* $\{X(\gamma) : \gamma \in \Re\}$ *be a family of real valued random variables. Assume that the families of random variables* $\{X(\gamma) : \gamma \in \Re\}$ *and* $\{Y(\theta) : \theta \in \Re\}$ *are stochastically increasing and stochastically convex. Furthermore, assume that* $X(\gamma)$ *and* $Y(\theta)$ *are independent of each other for all* $\gamma \in \Re$ *and* $\theta \in \Re$. *Then, the family of random variables* $\{Y(X(\gamma)) : \gamma \in \Re\}$ *is stochastically increasing and stochastically convex.*

If we use $\mathsf{Binomial}(n, p)$ to denote a binomial random variable with parameters (n, p), then Example 8.B.3 in Shaked and Shanthikumar (2007) establishes that the family of random variables $\{\mathsf{Binomial}(n, p) : p \in [0, 1]\}$ is stochastically increasing and linear in the sample path sense, which implies that this family of random variables is stochastically increasing and stochastically convex. Similarly, if we use $\mathsf{Beta}(\theta\, m, (1-\theta)\, m)$ to denote a beta random variable with parameters $(\theta\, m, (1-\theta)\, m)$ and define $\mathsf{Beta}(0, m) = 0$ and $\mathsf{Beta}(m, 0) = 1$, then Adell et al. (1993) show that the family of random variables $\{\mathsf{Beta}(\theta\, m, (1-\theta)\, m) : \theta \in [0, 1]\}$ is stochastically increasing and stochastically convex. Using these results together with Lemma 1, we get the following monotonicity and convexity result for the value functions. We defer the proof to Appendix B.

PROPOSITION 1 **(Monotonicity and Convexity of Value Functions)**. *For all* $t = 1, \ldots, \tau$ *and* $i = 1, \ldots, n$, *the value function* $V_t(\boldsymbol{\theta}_t)$ *is componentwise decreasing and convex in* θ_{it}.

Based on the monotonicity property, if $V_t(\boldsymbol{\theta}_t) > 0$, then for any $\boldsymbol{\theta}'_t$ with $\theta'_{it} \leq \theta_{it}, \forall\, i = 1, \ldots, n$, we have $V_t(\boldsymbol{\theta}'_t) \geq V_t(\boldsymbol{\theta}_t) > 0$. In this case, we immediately obtain the following proposition, which

gives a comparison between the decisions made by the optimal policy for different values of the state variable. In the statement of the next proposition, we recall that $\mathcal{C}_t = \{\boldsymbol{\theta}_t \in [0,1]^n : V(\boldsymbol{\theta}_t) > 0\}$ denotes the set of states for which it is optimal to continue selling the product.

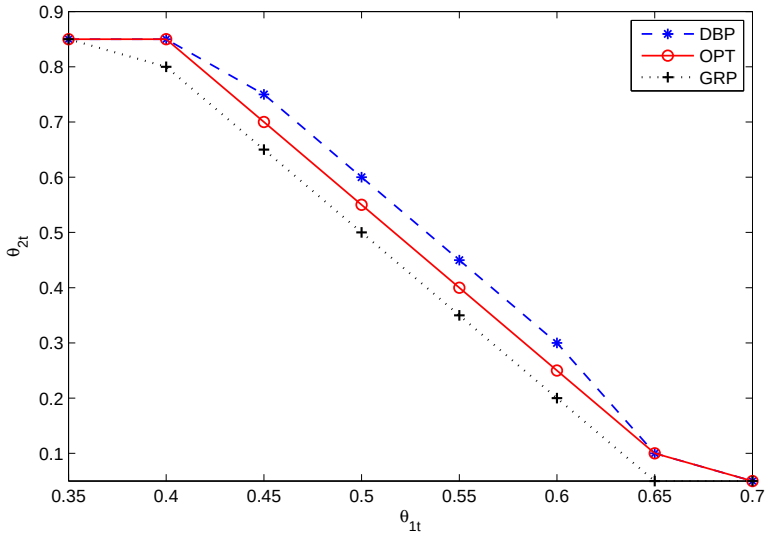
PROPOSITION 2 **(Shape of the Continuation Region)**. *For all $t = 1, ..., \tau$ and $\boldsymbol{\theta}_t, \boldsymbol{\theta}'_t \in [0,1]^n$ that satisfy $\theta'_{it} \leq \theta_{it}, \forall\, i = 1, \ldots, n$, if $\boldsymbol{\theta}_t \in \mathcal{C}_t$, then $\boldsymbol{\theta}'_t \in \mathcal{C}_t$.*

The intuition behind the proposition is that if the proportions of the units that failed in the past from different failure types are lower, then our estimates of the failure probabilities are also lower and we are more likely to continue selling the product, all else being equal. The implication of this proposition is that the optimal policy is a boundary policy. In other words, there is an optimal stopping boundary at time period t and if the proportions of the units that failed in the past from different failure types are above the optimal stopping boundary, then it is optimal to stop selling the product. Otherwise it is optimal to continue. The optimal stopping boundary at time period t is an $(n-1)$-dimensional hypersurface determined by the values of $\boldsymbol{\theta}_t$ that satisfy $(r - K\boldsymbol{c}^\top \boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{V_{t+1}(\lambda_t\, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} = 0$. For a problem instance with two failure types, the solid line in Figure 1 shows the shape of the optimal stopping boundary at a particular time period t. The horizontal and vertical axes in this figure give our expected beliefs about the probabilities of the two failure types, which we indicate by $(\theta_{1t}, \theta_{2t})$. The solid line shows the values of the expected beliefs about the probabilities of the two failure types such that we would be indifferent between stopping and continuing to sell the product. To the lower left side of the stopping boundary, the optimal decision is to continue selling the product. The dashed and dotted lines show approximations to the optimal stopping boundary that we obtain by using the methods developed in Sections 5 and 6. We dwell on these stopping boundaries later in the paper. Using the monotonicity of the value functions, the following proposition gives a comparison between the decisions made by the optimal policy at different time periods. The proof of this proposition appears in Appendix B.

PROPOSITION 3 **(No Failures Lead to Continuation)**. *Assume that the state at time period t is $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t \in \mathcal{C}_t$. If no units fail at time period t, then it is still optimal to continue selling the product at time period $t+1$.*

Proposition 3 conveys the intuition that the absence of failures at time period t strengthens our belief that failures are rare. Therefore, if it is optimal to continue selling the product at time period t and we do not observe any failures at this time period, then it is sensible to continue selling the product at time period $t+1$ as well.

Figure 1 **Optimal stopping boundary and approximations to the optimal stopping boundary at $t = 4$ for a problem instance with the following parameters: $n = 2, \tau = 5, K = 1, r = 1, c = (1.5, 0.5), D_l = 5, l = 0, ..., \tau$.**



The expectation in the optimality equation in (2) involves the random variable $Y_{it}(\theta_{it})$, which has a beta binomial distribution with parameters $(W_t, P_{it}(\theta_{it}))$, where the random variable $P_{it}(\theta_{it})$ itself is a beta random variable. Computing expectations that involve such a beta binomial random variable requires calculating beta functions, which can be computationally problematic in practice when the parameters of the beta function are large. Teerapabolarn (2008) demonstrates that a beta binomial distribution can be approximated well by a binomial distribution, especially when the expectation of the beta random variable is small. This result brings up the possibility of replacing the beta binomial random variable $Y_{it}(\theta_{it})$ in the optimality equation in (2) with a binomial random variable $Z_{it}(\theta_{it})$ with parameters (W_t, θ_{it}), which is a binomial random variable with the same expectation as $Y_{it}(\theta_{it})$. With this substitution, we can simplify the computation of the expectation in (2). We use this approximation strategy in the numerical experiments presented in Section 7. Using the fact that the value functions are componentwise convex, which is shown in Proposition 1, it is possible to establish that approximating a beta binomial random variable with the corresponding binomial one provides lower bounds on the original value functions. We provide the details of this lower bound property in Appendix C.

The difficulty in obtaining the optimal policy for our problem is due to the fact that the optimality equation in (2) has a high dimensional state vector. This high dimensionality makes it difficult to compute the value functions exactly when the number of failure types exceeds two or three. In the next two sections, we develop computationally tractable methods to construct approximations to the value functions. These methods scale gracefully with the number of failure types.

5. Upper Bounds and Decomposition Based Policy

In this section, we develop a tractable method for approximating the value functions by decomposing the problem into a collection of one dimensional dynamic programs, each involving a single failure type. To begin with, we observe that the expected repair cost $K\boldsymbol{c}^{\top}\boldsymbol{\theta}_t\, D_t = K\sum_{i=1}^{n} c_i\,\theta_{it}\, D_t$ in the optimality equation in (2) decomposes by the failure types. Furthermore, the dynamics of $\boldsymbol{\theta}_t$ given in (1) implies that our beliefs about the different failure probabilities evolve independently of each other. These observations motivate writing the revenue expression in the optimality equation in (2) as $rD_t = \sum_{i=1}^{n} \rho_i\, D_t$, where we assume that the vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$ satisfies $\sum_{i=1}^{n} \rho_i = r$, but we leave it unspecified otherwise for the time being. In this case, for each $i = 1, \ldots, n$, we propose solving the optimality equation

$$V_{it}^{U}(\theta_{it} \,|\, \rho_i) = \max\Big\{(\rho_i - Kc_i\,\theta_{it})\, D_t + \mathbb{E}\{V_{i,t+1}^{U}(\lambda_t\,\theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\}, 0\Big\}, \tag{3}$$

with the boundary condition $V_{i,\tau+1}^{U}(\cdot \,|\, \rho_i) = 0$. The optimality equation in (3) finds the optimal policy for the case where the failures are only of type i and the revenue that we generate from each sold unit is given by ρ_i. We use the subscript i in the value functions to emphasize that the optimality equation in (3) focuses only on a single failure type i. The argument ρ_i in the value functions emphasizes that the solution to the optimality equation depends on the choice of ρ_i. As shown in the following proposition, the optimality equation in (3) can be used to construct upper bounds on the original value functions and the superscript U in the value functions emphasizes this upper bound property. Since the state variable in the optimality equation in (3) is a scalar, we can solve this optimality equation in a tractable manner.

PROPOSITION 4 **(Upper Bound)**. *For any vector $\boldsymbol{\rho}$ satisfying $\sum_{i=1}^{n} \rho_i = r$, we have $V_t(\boldsymbol{\theta}_t) \leq \sum_{i=1}^{n} V_{it}^{U}(\theta_{it} \,|\, \rho_i)$ for all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$.*

Proof. We show the result by using induction over the time periods. The result trivially holds at time period $\tau + 1$. Assuming that the result holds at time period $t + 1$, we have

$$\begin{aligned}
\sum_{i=1}^{n} V_{it}^{U}(\theta_{it} \,|\, \rho_i) &= \sum_{i=1}^{n} \max\Big\{(\rho_i - Kc_i\,\theta_{it})\, D_t + \mathbb{E}\{V_{i,t+1}^{U}(\lambda_t\,\theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\}, 0\Big\} \\
&\geq \max\left\{\sum_{i=1}^{n}(\rho_i - Kc_i\,\theta_{it})\, D_t + \sum_{i=1}^{n}\mathbb{E}\{V_{i,t+1}^{U}(\lambda_t\,\theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\}, 0\right\} \\
&= \max\Big\{(r - K\boldsymbol{c}^{\top}\boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{\textstyle\sum_{i=1}^{n} V_{i,t+1}^{U}(\lambda_t\,\theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\}, 0\Big\} \\
&\geq \max\Big\{(r - K\boldsymbol{c}^{\top}\boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}, 0\Big\} = V_t(\boldsymbol{\theta}_t),
\end{aligned}$$

where the first inequality follows by noting that $\max\{a, 0\} + \max\{b, 0\} \geq \max\{a + b, 0\}$, the second equality uses the fact that $\sum_{i=1}^{n} \rho_i = r$, the second inequality follows from the induction assumption

and the fact that the components of $\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$ are independent of each other and the last equality follows from the optimality equation in (2). $\square$

From the proof of the proposition above, we observe that the sufficient conditions for the upper bound property to hold are the following. First, the immediate expected profit should be of the form a constant plus a separable function of the state. In our problem setting, this expected profit is of the form $r\,D_t - K\,D_t \sum_{i=1}^n c_i\,\theta_{it}$. Second and perhaps more importantly, our beliefs about the probabilities of the different failure types should evolve independently of each other. As long as these two conditions are satisfied, the decomposition approach described above provides upper bounds on the original value functions, irrespective of the learning model that we use.

It is natural to ask how we can use Proposition 4 to choose a value for $\boldsymbol{\rho}$ and how we can use the optimality equation in (3) to decide whether we should continue or stop selling the product at a particular time period. By Proposition 4, the optimal objective value of the problem

$$V_t^U(\boldsymbol{\theta}_t) = \min_{\boldsymbol{\rho}} \left\{ \sum_{i=1}^n V_{it}^U(\theta_{it} \,|\, \rho_i) : \sum_{i=1}^n \rho_i = r \right\} \tag{4}$$

provides the tightest possible upper bound on $V_t(\boldsymbol{\theta}_t)$. Indeed, the problem above finds the tightest possible upper bound on $V_t(\boldsymbol{\theta}_t)$ among all upper bounds of the form $\sum_{i=1}^n V_{it}^U(\theta_{it} \,|\, \rho_i)$. In this case, we can mimic the optimal policy by defining the set of states $\mathcal{C}_t^U = \{\boldsymbol{\theta}_t \in [0,1]^n : V_t^U(\boldsymbol{\theta}_t) > 0\}$ and continuing to sell the product at time period t if and only if the state $\boldsymbol{\theta}_t$ at this time period satisfies $\boldsymbol{\theta}_t \in \mathcal{C}_t^U$. We refer to this policy as the decomposition based policy. Since we have $V_t^U(\boldsymbol{\theta}_t) \geq V_t(\boldsymbol{\theta}_t)$, we obtain $\mathcal{C}_t^U \supseteq \mathcal{C}_t$. Therefore, the decomposition based policy is more likely to continue selling the product when compared with the optimal policy. The dashed line in Figure 1 shows the approximation to the optimal stopping boundary that we obtain by using the decomposition based policy. We observe from this figure that we indeed have $\mathcal{C}_t^U \supseteq \mathcal{C}_t$.

In Appendix D, we use induction over the time periods to show that $V_{it}^U(\theta_{it} \,|\, \rho_i)$ is a convex function of ρ_i. Furthermore, we establish that if we start at time period t in state θ_{it}, then the total expected demand that we observe until we stop selling the product according to the optimality equation in (3) gives a subgradient of $V_{it}^U(\theta_{it} \,|\, \cdot)$ at ρ_i. In other words, if we want to compute a subgradient of $V_{it}^U(\theta_{it} \,|\, \cdot)$ at ρ_i, then we can solve the optimality equation in (3) with this value of ρ_i. In this way, we obtain the value functions $\{V_{is}^U(\cdot \,|\, \rho_i) : s = 0, 1, \ldots, \tau\}$. According to the optimality equation in (3), it is optimal to continue selling the product as long as the state variable θ_{is} at a time period s satisfies $V_{is}^U(\theta_{is} \,|\, \rho_i) > 0$. In this case, starting at time period t in state θ_{it}, the total expected demand that we observe until we stop selling the product provides a subgradient of $V_{it}^U(\theta_{it} \,|\, \cdot)$ at ρ_i. Since we can use Monte Carlo simulation to estimate the total expected demand

that we observe until we stop selling the product, it is straightforward to compute a subgradient of $V_{it}^U(\theta_{it} \,|\, \rho_i)$ with respect to ρ_i. Alternatively, we can compute this total expected demand exactly by exploiting the fact that $Y_{it}(\theta_{it})$ has a finite number of realizations. Once we have the subgradient of $V_{it}^U(\theta_{it} \,|\, \rho_i)$ with respect to ρ_i, we can solve problem (4) by using standard subgradient search for minimizing a convex function subject to linear constraints; see Ruszczynski (2011).

6. Lower Bounds and Greedy Policy

The goal of this section is to complement the approach in the previous section by providing lower bounds on the value functions. We begin this section by motivating our lower bounds through Jensen's inequality. Following this motivation, we show that our lower bounds correspond to the expected profits obtained by a greedy policy that makes its decisions based only on the current beliefs, ignoring future learning. Using the lower bounds, we construct a policy to decide if and when to stop selling the product. Finally, we show that our lower bounds become asymptotically tight as we scale up the demand at each time period linearly with the same rate.

6.1. A Deterministic Approximation

Our approach is based on exchanging the order in which we compute the expectation and the value function on the right side of the optimality equation in (2). In particular, we observe that $\mathbb{E}\{\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t)\} = \boldsymbol{\theta}_t$, which follows by the fact that $\mathbb{E}\{\boldsymbol{Y}_t(\theta_t)\} = W_t \boldsymbol{\theta}_t$ or simply by the martingale property of Bayesian updates. In this case, replacing $\mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}$ on the right side of the optimality equation in (2) with $V_{t+1}(\mathbb{E}\{\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t)\}) = V_{t+1}(\boldsymbol{\theta}_t)$, we obtain the optimality equation

$$V_t^L(\boldsymbol{\theta}_t) = \max\Big\{(r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t)\, D_t + V_{t+1}^L(\boldsymbol{\theta}_t), 0\Big\}, \tag{5}$$

with the boundary condition $V_{\tau+1}^L(\cdot) = 0$. The optimality equation above does not involve any uncertainty and the state does not change as long as we continue selling the product. In this case, letting $[\cdot]^+ = \max\{\cdot, 0\}$ and using induction over the time periods, it is possible to show that the value functions computed through the optimality equation in (5) are explicitly given by

$$V_t^L(\boldsymbol{\theta}_t) = [r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \sum_{s=t}^{\tau} D_s \tag{6}$$

for all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$. We defer the details of this simple induction argument to Appendix E. The following proposition shows that the value function $V_t^L(\cdot)$ provides a lower bound on the original value function. The superscript L emphasizes this lower bound property.

PROPOSITION 5 **(Lower Bound)**. *For all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$, we have $V_t(\boldsymbol{\theta}_t) \geq V_t^L(\boldsymbol{\theta}_t)$.*

Proof. We show the result by using induction over the time periods. The result trivially holds at time period $\tau+1$. Assuming that the result holds at time period $t+1$, we have

$$\begin{aligned}\mathbb{E}\{V_{t+1}^L(\lambda_t\,\boldsymbol{\theta}_t+\tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} &= \mathbb{E}\{[r-K\boldsymbol{c}^\top(\lambda_t\,\boldsymbol{\theta}_t+\tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t))]^+\}\sum_{s=t+1}^{\tau}D_s\\ &\geq [r-K\boldsymbol{c}^\top\mathbb{E}\{\lambda_t\,\boldsymbol{\theta}_t+\tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t)\}]^+\sum_{s=t+1}^{\tau}D_s\\ &= (r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t)^+\sum_{s=t+1}^{\tau}D_s = V_{t+1}^L(\boldsymbol{\theta}_t), \qquad (7)\end{aligned}$$

where the equalities use the closed form expression for $V_t^L(\cdot)$ given in (6) and the inequality follows by noting that $[\cdot]^+$ is a convex function and using Jensen's inequality. Thus, we obtain

$$\begin{aligned}V_t(\boldsymbol{\theta}_t) &= \max\Big\{(r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\,D_t+\mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t+\tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t))\},0\Big\}\\ &\geq \max\Big\{(r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\,D_t+\mathbb{E}\{V_{t+1}^L(\lambda_t\,\boldsymbol{\theta}_t+\tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t))\},0\Big\}\\ &\geq \max\Big\{(r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\,D_t+V_{t+1}^L(\boldsymbol{\theta}_t),0\Big\} = V_t^L(\boldsymbol{\theta}_t),\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second one is by (7). □

From the discussion at the end of Section 3, we can interpret $r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t$ as the expected net profit contribution of a sold unit. Therefore, the expression for $V_t^L(\boldsymbol{\theta}_t)$ given in (6) corresponds to the total expected profit obtained by a policy that continues selling the product at all future time periods whenever the expected net profit contribution of a sold unit is positive. This policy does not consider the benefits from learning the failure probabilities at future time periods.

We can use the optimality equation in (5) to come up with a policy to decide whether we should continue or stop selling the product at a particular time period. In particular, if we use $V_t^L(\cdot)$ as an approximation to $V_t(\cdot)$, then we can mimic the optimal policy by defining the set of states $\mathcal{C}_t^L=\{\boldsymbol{\theta}_t\in[0,1]^n : V_t^L(\boldsymbol{\theta})>0\}$ and continuing to sell the product at time period t if and only if the state $\boldsymbol{\theta}_t$ at this time period satisfies $\boldsymbol{\theta}_t\in\mathcal{C}_t^L$. We refer to this policy as the greedy policy. Since we have $V_t^L(\boldsymbol{\theta}_t)\leq V_t(\boldsymbol{\theta}_t)$, we obtain $\mathcal{C}_t^L\subseteq\mathcal{C}_t$, which implies that the greedy policy is more likely to stop selling the product when compared with the optimal policy. Furthermore, noting (6), we can write $\mathcal{C}_t^L$ as $\mathcal{C}_t^L=\{\boldsymbol{\theta}_t\in[0,1]^n : r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t>0\}$ and the stopping boundary from the greedy policy is an $(n-1)$-dimensional hyperplane determined by $r-K\boldsymbol{c}^\top\boldsymbol{\theta}_t=0$. Therefore, we do not even need to solve an optimality equation to find the decisions made by the greedy policy. The dotted line in Figure 1 shows the approximation to the optimal stopping boundary that we obtain by using the greedy policy and it demonstrates that we indeed have $\mathcal{C}_t^L\subseteq\mathcal{C}_t$.

6.2. Asymptotic Analysis

Although the greedy policy is simple to compute, it does not consider the benefits from learning the failure probabilities and a natural question is when we can expect the greedy policy to perform reasonably well. In this section, we consider an asymptotic regime where we scale up the demand at each time period linearly with the same rate. We show that the performance of the greedy policy becomes optimal under this regime. The asymptotic regime we consider is interesting in the following sense. On the one hand, if the demand at each time period is scaled up, then we collect a large amount of information right after the first time period and our estimates of the failure probabilities immediately become accurate. Thus, it may not be a huge problem to make decisions without considering the benefits from learning the failure probabilities and the greedy policy is expected to perform well. On the other hand, since the demand quantities at the future time periods are also large, we have the potential to collect a large amount of information about failure probabilities in the future, which may change our current belief about failure probabilities dramatically. Furthermore, since the demand quantities are large, small errors in estimating the failure probabilities may have serious consequences in absolute terms. Thus, ignoring the benefits from learning may cause serious problems and the greedy policy may perform poorly. From these two conflicting perspectives, it is not clear a priori whether the greedy policy is expected to perform well or not. The rest of this section dwells on this question by showing that the greedy policy is optimal under our asymptotic scaling regime.

We consider a family of problems $\{\mathcal{P}^m : m = 1, 2, \ldots\}$ indexed by the parameter $m \in \mathbb{Z}_+$. In problem $\mathcal{P}^m$, the demand at time period t is mD_t, but all other problem parameters are the same as those in Section 3. Thus, the parameter m scales up the demand at each time period. Accordingly, in problem $\mathcal{P}^m$, we have mM_t units that we could potentially have received as failed units up until time period t. We continue using θ_{it} to denote the proportion of the mM_t units that we have actually received as failed units from failure type i. In this case, if we let $P_{it}^m(\theta_{it})$ be our prior belief at time period t for the probability of failure type i, then $P_{it}^m(\theta_{it})$ has a beta distribution with parameters $(\theta_{it}\, mM_t, (1-\theta_{it})\, mM_t)$. Similarly, in problem $\mathcal{P}^m$, we have mW_t units that we can potentially receive as failed units for repairs at time period t. We use $Y_{it}^m(\theta_{it})$ to denote the number of units that we actually receive as failed units from failure type i at time period t. In this case, $Y_{it}^m(\theta_{it})$ has a beta binomial distribution with parameters $(mW_t, P_{it}^m(\theta_{it}))$.

We use $\{V_t(\cdot \,|\, m) : t = 1, \ldots, \tau\}$ to denote the value functions that we obtain by solving the optimality equation in (2) for problem $\mathcal{P}^m$. In other words, these value functions are obtained by replacing D_t with mD_t and $\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$ with $\boldsymbol{Y}_t^m(\boldsymbol{\theta}_t) = (Y_{1t}^m(\theta_{1t}), \ldots, Y_{nt}^m(\theta_{nt}))$ in the optimality equation

in (2) and solving this optimality equation. We note that λ_t in problem $\mathcal{P}^m$ does not depend on m since we have $\lambda_t = mM_t/mM_{t+1} = M_t/M_{t+1}$. Similarly, we use $\{V_t^L(\cdot\,|\,m) : t = 1,\ldots,\tau\}$ to denote the value functions that we obtain by solving the optimality equation in (5) for problem $\mathcal{P}^m$. Noting (6), we have $V_t^L(\boldsymbol{\theta}_t\,|\,m) = m\,V_t^L(\boldsymbol{\theta}_t\,|\,1)$ for all $t = 1,\ldots,\tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$.

For problem $\mathcal{P}^m$, it is optimal to continue selling the product at time period t whenever the state $\boldsymbol{\theta}_t$ at this time period is in the set $\mathcal{C}_t(m) = \{\boldsymbol{\theta}_t \in [0,1]^n : V_t(\boldsymbol{\theta}_t\,|\,m) > 0\}$. Replacing $V_t(\boldsymbol{\theta}_t\,|\,m)$ with the lower bound $V_t^L(\boldsymbol{\theta}_t\,|\,m)$, we can obtain an approximate policy for problem $\mathcal{P}^m$ by continuing to sell the product at time period t whenever the state $\boldsymbol{\theta}_t$ at this time period is in the set $\mathcal{C}_t^L(m) = \{\boldsymbol{\theta}_t \in [0,1]^n : V_t^L(\boldsymbol{\theta}_t\,|\,m) > 0\}$. Since we have $V_t^L(\boldsymbol{\theta}_t\,|\,m) \leq V_t(\boldsymbol{\theta}_t\,|\,m)$ for all $\boldsymbol{\theta}_t \in [0,1]^n$, we naturally obtain $\mathcal{C}_t^L(m) \subseteq \mathcal{C}_t(m)$. Furthermore, since $V_t^L(\boldsymbol{\theta}_t\,|\,m) = m\,V_t^L(\boldsymbol{\theta}_t\,|\,1)$, we have $\mathcal{C}_t^L(m) = \mathcal{C}_t^L(1)$ by the definition of $\mathcal{C}_t^L(m)$. Therefore, we have $\mathcal{C}_t^L(1) \subseteq \mathcal{C}_t(m)$ for all $m \in \mathbb{Z}_+$. The following proposition gives an ordering for $\{\mathcal{C}_t(m) : m \in \mathbb{Z}_+\}$, showing that $\mathcal{C}_t(m)$ shrinks as m increases. Its proof is in Appendix F and uses the componentwise convexity of the value function shown in Proposition 1.

PROPOSITION 6 **(Learning is More Beneficial for Smaller Demand)**. *For all $t = 1,\ldots,\tau$, $\boldsymbol{\theta}_t \in [0,1]^n$ and $m \in \mathbb{Z}_+$, we have $\mathcal{C}_t^L(1) \subseteq \mathcal{C}_t(m+1) \subseteq \mathcal{C}_t(m)$.*

Proposition 6 indicates that the optimal policy is more likely to continue selling the product when m is small. Noting that the demand quantities are smaller when m is smaller, this result builds the intuition that if the demand quantities are smaller, then we should be more willing to learn the failure probabilities by continuing to sell the product. In other words, we intuitively expect the value of learning to be large when the demand quantities are small. In addition, Proposition 6 shows that the set of states at which we continue selling the product under our deterministic approximation is always smaller than the set of states at which we continue selling the product under the optimal policy. Thus, our deterministic approximation is at the extreme end in the sense that no matter how large the demand quantities are, our deterministic approximation is more likely to stop selling the product when compared with the optimal policy. In the following proposition, we show that $V_t(\boldsymbol{\theta}_t\,|\,m)$ deviates from $V_t^L(\boldsymbol{\theta}_t\,|\,m)$ by a term that grows in the order of $\sqrt{m}$. We defer the proof of this result to Appendix G.

PROPOSITION 7 **(Vanishing Relative Gap with Lower Bound)**. *For all $t = 1,\ldots,\tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$, we have $V_t^L(\boldsymbol{\theta}_t\,|\,m) \leq V_t(\boldsymbol{\theta}_t\,|\,m) \leq V_t^L(\boldsymbol{\theta}_t\,|\,m) + \bar{G}_t\sqrt{m}$, where $\bar{G}_t$ is a constant that depends on $(D_0,\ldots,D_\tau)$, $(c_1,\ldots,c_n)$ and K, but it is independent of m.*

Since we have $V_t^L(\boldsymbol{\theta}_t\,|\,m) = m\,V_t^L(\boldsymbol{\theta}_t\,|\,1)$ for all $t = 1,\ldots,\tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$, Proposition 7 implies that $m\,V_t^L(\boldsymbol{\theta}_t\,|\,1) \leq V_t(\boldsymbol{\theta}_t\,|\,m) \leq m\,V_t^L(\boldsymbol{\theta}_t\,|\,1) + \bar{G}_t\sqrt{m}$. Therefore, as long as $V_t^L(\boldsymbol{\theta}_t\,|\,1)$ is strictly

positive, both $V_t^L(\boldsymbol{\theta}_t \,|\, m)$ and $V_t(\boldsymbol{\theta}_t \,|\, m)$ grow linearly with m, but the difference between $V_t^L(\boldsymbol{\theta}_t \,|\, m)$ and $V_t(\boldsymbol{\theta}_t \,|\, m)$ grows only in the order of $\sqrt{m}$. In other words, we have $\lim_{m\to\infty} \frac{V_t(\boldsymbol{\theta}_t \,|\, m)}{V_t^L(\boldsymbol{\theta}_t \,|\, m)} = 1$ as long as $V_t^L(\boldsymbol{\theta}_t \,|\, 1)$ is strictly positive. Intuitively speaking, when the demand quantities get large, the value of learning vanishes in a relative sense and the problem becomes reasonably easy. We emphasize that although the relative gap between $V_t(\boldsymbol{\theta}_t \,|\, m)$ and $V_t^L(\boldsymbol{\theta}_t \,|\, m)$ vanishes as m gets large, the absolute gap $V_t(\boldsymbol{\theta}_t \,|\, m) - V_t^L(\boldsymbol{\theta}_t \,|\, m)$ can still grow in the order of $\sqrt{m}$. Heyman and Sobel (2003) and Muckstadt and Sapra (2010) give examples of stochastic optimization problems in inventory control, marketing and capacity expansion settings where a greedy policy provides an optimal policy. Our results suggest that the relative performance of the greedy policy is expected to be good when m is large, but this policy should not be regarded as a consistently strong policy. In particular, the greedy policy may suffer from a large absolute performance loss even when m is large. Furthermore, our computational experiments indicate that when m is small so that there is limited amount of demand at each time period, the greedy policy may also have a large relative performance loss, as it ignores the benefits from future learning.

7. Computational Experiments

In this section, we provide computational experiments to test the performance of the policies developed in Sections 5 and 6. We begin by describing our benchmark policies and experimental setup. Following this description, we provide our computational results.

7.1. Benchmark Policies

We compare the following five benchmark policies.

Ideal Policy (IDE). This benchmark policy corresponds to an idealized decision rule computed under the assumption that the true failure probabilities are known. If we know that the true failure probabilities are given by $\boldsymbol{p} = (p_1, \ldots, p_n)$, then we generate a revenue of r and incur an expected repair cost of $K\boldsymbol{c}^\intercal\boldsymbol{p}$ for each sold unit. Therefore, if $r > K\boldsymbol{c}^\intercal\boldsymbol{p}$, then it is optimal to sell the product until the end of the selling horizon, whereas if $r \leq K\boldsymbol{c}^\intercal\boldsymbol{p}$, then it is optimal not to sell the product at all. Since there is no decision to be made at the beginning of time period zero, the expected profit obtained by IDE is simply $D_0\,[r - K\boldsymbol{c}^\intercal\boldsymbol{p}] + \sum_{t=1}^{\tau} D_t\,[r - K\boldsymbol{c}^\intercal\boldsymbol{p}]^+$. This expected profit corresponds to the idealized scenario of knowing the true failure probabilities and it provides an unattainable upper bound on the expected profit from any policy that tries to learn the failure probabilities. Nevertheless, by comparing the performance of a particular policy with this upper bound, we can assess how difficult the problem is and how well the policy performs.

Decomposition Based Policy (DBP). This benchmark policy corresponds to the one described in Section 5. In particular, if our belief for the failure probabilities at time period t is captured by

the vector $\boldsymbol{\theta}_t$, then DBP solves problem (4) to compute $V_t^U(\boldsymbol{\theta}_t)$ and continues selling the product if and only if $\boldsymbol{\theta}_t$ is in the set $\mathcal{C}_t^U = \{\boldsymbol{\theta}'_t \in [0,1]^n : V_t^U(\boldsymbol{\theta}'_t) > 0\}$.

Greedy Policy (GRP). This benchmark policy corresponds to the one described in Section 6. If our belief for the failure probabilities at time period t is captured by the vector $\boldsymbol{\theta}_t$, then GRP continues selling the product if and only if $\boldsymbol{\theta}_t$ is in the set $\mathcal{C}_t^L = \{\boldsymbol{\theta}'_t \in [0,1]^n : r - K\boldsymbol{c}^\top \boldsymbol{\theta}'_t > 0\}$. We note that implementing GRP does not require solving an optimality equation at all. Furthermore, the set $\mathcal{C}_t^L$ does not depend on the time period.

One Step Look Ahead Policy (OSL). This benchmark policy tries to learn the failure probabilities only for the current time period and evaluates the future expected profit after the current time period by using the value function of a heuristic policy. As for the heuristic policy, we use GRP that we describe above. Thus, if our belief for the failure probabilities at time period t is captured by the vector $\boldsymbol{\theta}_t$, then OSL solves the problem

$$V_t^O(\boldsymbol{\theta}_t) = \max\Big\{(r - K\boldsymbol{c}^\top \boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{V_{t+1}^L(\lambda_t\, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}, 0\Big\} \tag{8}$$

and continues selling the product if and only if $\boldsymbol{\theta}_t$ is in the set $\mathcal{C}_t^O = \{\boldsymbol{\theta}'_t \in [0,1]^n : V_t^O(\boldsymbol{\theta}'_t) > 0\}$. We observe that the value function $V_{t+1}^L(\cdot)$ has an explicit form given by (6). Therefore, problem (8) is not difficult to solve. While GRP ignores the benefits from future learning, OSL can be viewed as an improvement of GRP, since it attempts to learn the failure probabilities for one time period and switches to the decision rule followed by GRP after this time period.

Single Failure Policy (SFP). This benchmark policy is motivated by the observation that our problem is not difficult to solve when there is one failure type. Thus, SFP approximates our belief for the failure probabilities by using a single weighted average of these probabilities. The weights that we put on the different failure probabilities are the repair costs so that more costly failures tend to get more weight. In particular, SFP solves the optimality equation

$$V_t^S(\Theta_t) = \max\Big\{(r - K\, c^S\, \Theta_t)\, D_t + \mathbb{E}\{V_{t+1}^S(\lambda_t\, \Theta_t + \tfrac{1-\lambda_t}{W_t}\, Y_t^S(\Theta_t))\}, 0\Big\},$$

where the state variable at time period t is the scalar Θ_t and we let $c^S = \sum_{i=1}^n c_i$ and $Y_t^S(\Theta_t) = \mathsf{Binomial}(W_t, \mathsf{Beta}(\Theta_t\, M_t, (1-\Theta_t)\, M_t))$. We define $\mathcal{C}_t^S = \{\Theta_t \in [0,1] : V_t^S(\Theta_t) > 0\}$ to represent the set of states for which we continue selling the product. If our belief for the failure probabilities at time period t is captured by the vector $\boldsymbol{\theta}_t$, then SFP continues selling the product if and only if $\frac{\boldsymbol{c}^\top \boldsymbol{\theta}_t}{c^S}$ is in the set $\mathcal{C}_t^S$. We note that $\frac{\boldsymbol{c}^\top \boldsymbol{\theta}_t}{c^S}$ is a weighted average of the components of $\boldsymbol{\theta}_t$.

7.2. Experimental Setup

We use simulation to test the performance of our benchmark policies. For each test problem, we simulate the performance for 500 sample paths. At the beginning of each sample path, we sample the true failure probabilities $\boldsymbol{p} = (p_1, \ldots, p_n)$ such that p_i has a beta distribution with mean μ_i and standard deviation σ_i and the components of the vector $\boldsymbol{p}$ are independent of each other. Once we fix the true failure probabilities, we use these probabilities to generate the numbers of failed units throughout the selling horizon. In particular, letting $\tilde{Y}_{it}$ be the number of units that fail from failure type i at time period t, we generate $\tilde{Y}_{it}$ from the binomial distribution with parameters (W_t, p_i). At each time period, we update the state of the system by using the dynamics $\boldsymbol{\theta}_{t+1} = \lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \tilde{\boldsymbol{Y}}_t$, where $\tilde{\boldsymbol{Y}}_t = (\tilde{Y}_{1t}, \ldots, \tilde{Y}_{nt})$. The initial value of the state variable is given by $\theta_{i1} = \frac{1}{W_0} \tilde{Y}_{i0}$ for all $i = 1, \ldots, n$, which corresponds to the fraction of units under warranty that fail from failure type i at time period zero. If we are testing DBP, then we continue selling the product at time period t if and only if the state $\boldsymbol{\theta}_t$ at this time period satisfies $\boldsymbol{\theta}_t \in \mathcal{C}_t^U$. Similarly, if we are testing GRP, then we continue selling the product at time period t if and only if we have $\boldsymbol{\theta}_t \in \mathcal{C}_t^L$. If we are testing OSL, then we continue selling the product at time period t if and only if $\boldsymbol{\theta}_t \in \mathcal{C}_t^O$, whereas if we are testing SFP, then we continue selling the product at time period t if and only if $\frac{\boldsymbol{c}^\intercal \boldsymbol{\theta}_t}{c^S} \in \mathcal{C}_t^S$. Finally, if we are testing IDE, then we continue selling the product until the end of the selling horizon if and only if $r > K \boldsymbol{c}^\intercal \boldsymbol{p}$. Otherwise, we stop selling the product as early as possible, which is the beginning of time period 1. We note that IDE has access to the true failure probabilities to make its decisions, whereas DBP, GRP, OSL and SFP make their decisions only by using the samples of the failed units given by $\{\tilde{\boldsymbol{Y}}_t : t = 0, 1, \ldots, \tau\}$.

By simulating the decisions of each benchmark policy as described above, we accumulate the profit obtained over the selling horizon. Averaging the accumulated profits on 500 sample paths, we estimate the expected profit obtained by a particular benchmark policy. We note that this way of testing the performance of the benchmark policies corresponds to a frequentist framework, where the true failure probabilities are fixed at the beginning of a sample path and all failures occur according to these fixed true failure probabilities, in which case, the immediate expected profit obtained at each time period is driven by these fixed true failure probabilities. In contrast, the dynamic programming formulation in (2) is under a Bayesian framework, where the failure probabilities are assumed to evolve over the time periods according to the Bayes rule. In particular, as a function of our beliefs $\boldsymbol{\theta}_t$ about the probabilities of different failure types, the immediate expected profit at time period t in (2) is given by $(r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t) D_t$ and our beliefs about the failure probabilities evolve from time period t to time period $t+1$ according to $\boldsymbol{\theta}_{t+1} = \lambda_t \boldsymbol{\theta}_t +$

$\frac{1-\lambda_t}{W_t}\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$. Therefore, the immediate expected profit at each time period in the optimality equation in (2) is not computed according to fixed failure probabilities. As a result, although one can use the optimality equation in (2) to construct a policy to make stopping and continuing decisions over time, if the performance of this policy is tested in a frequentist framework, then the total expected profit obtained by the policy is not necessarily equal to the total expected profit predicted by the value function $V_1(\cdot)$ evaluated at the initial belief. Similarly, although we obtain tractable upper bounds $\{V_t^U(\cdot): t=1,\ldots,\tau\}$ on $\{V_t(\cdot): t=1,\ldots,\tau\}$ by using the decomposition approach in Section 5, the total expected profits collected by our benchmark policies can be larger or smaller than the upper bound $V_1^U(\cdot)$ evaluated at the initial belief.

A few setup runs demonstrated that if r is substantially larger than $K\boldsymbol{c}^\top\boldsymbol{p}$, then it is clearly optimal to continue selling the product and DBP, GRP, OSL and SFP are all quick to realize this fact. In this case, the performance of the four benchmark policies is comparable. Similarly, if r is substantially smaller than $K\boldsymbol{c}^\top\boldsymbol{p}$, then it is clearly optimal to stop selling the product, in which case, DBP, GRP, OSL and SFP end up performing comparably as well. Therefore, test problems tend to be more difficult when r is roughly equal to $K\boldsymbol{c}^\top\boldsymbol{p}$ so that it is not easy to detect immediately whether it is optimal to continue or stop selling the product. To generate test problems with this nature, we set the mean vector $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)$ of $\boldsymbol{p}$ to satisfy $r=K\boldsymbol{c}^\top\boldsymbol{\mu}$. In this case, if the standard deviation $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_n)$ of $\boldsymbol{p}$ is small, then the sampled value of $\boldsymbol{p}$ in many sample paths roughly satisfies $r=K\boldsymbol{c}^\top\boldsymbol{p}$ and we obtain test problems for which it is difficult to detect whether it is optimal to continue or stop selling the product. Therefore, we generally expect the test problems to be more difficult as $\boldsymbol{\sigma}$ gets smaller.

In all of our test problems, the number of time periods in the selling horizon is 12 and the warranty coverage is for six time periods so that $\tau=12$ and $K=6$. We work with test problems with two or five failure types, corresponding to $n=2$ or $n=5$. For the demand at each time period, one set of test problems assume that $D_0=50$ and $D_1=\ldots=D_\tau=\bar{D}$, where we vary $\bar{D}$ over $50, 100, 200, 500$ and 1000. Another set of test problems assume that $D_0=D_1=\ldots=D_\tau=\bar{D}$, varying $\bar{D}$ over $50, 100, 200, 500$ and 1000. Therefore, the test marketing stage in the first set of test problems always involves a small number of units, irrespective of the magnitude of the demand at the subsequent time periods. The second set of test problems correspond to the case where the demand in the test marketing stage is equal to the demand at the subsequent time periods and the demand at each time period is scaled up as $\bar{D}$ gets larger. This way of scaling the demand matches the asymptotic regime in Section 6. The first set of test problems arguably correspond to a more realistic situation, where firms can go through only limited test marketing effort.

In all of our test problems, we normalize the unit revenue to $r = 1$. For the test problems with two failure types, we set the repair costs as $\boldsymbol{c} = (0.4, 0.8)$, whereas we use $\boldsymbol{c} = (0.4, 0.5, 0.6, 0.7, 0.8)$ for the test problems with five failure types. To choose the vector $\boldsymbol{\mu}$, we assume that $\mu_1 = \mu_2 = \ldots = \mu_n$ and pick the common value by solving $r = K\boldsymbol{c}^{\mathsf{T}}\boldsymbol{\mu}$ for $\boldsymbol{\mu}$. Finally, we choose the value of $\boldsymbol{\sigma}$ by letting $\sigma_i = CV\,\mu_i$ for all $i = 1, \ldots, n$ so that CV gives the coefficient of variation for each component of the vector $\boldsymbol{p}$. We vary CV over $0.6, 0.6^2, 0.6^3$ and 0.6^4. We label our test problems by using the tuple $(D_0, \bar{D}, n, CV)$. In our first set of test problems, we have $D_0 = 50$, whereas in the second set of test problems, we have $D_0 = \bar{D}$. We vary $\bar{D}$ over $\{50, 100, 200, 500, 1000\}$, n over $\{2, 5\}$ and CV over $\{0.6, 0.6^2, 0.6^3, 0.6^4\}$, yielding 80 test problems for our computational experiments.

7.3. Computational Results

We give our main computational results in Tables 1 and 2. Table 1 focuses on the test problems where the demand at the test marketing stage is always 50, whereas Table 2 focuses on the test problems where the demand at the test marketing stage is as large as the demand at the subsequent time periods. The layouts of the two tables are the same. The left and right portions of the tables respectively give the results for the test problems with two and five failure types. In each portion, the first column shows the characteristics of the test problems by using the tuple $(D_0, \bar{D}, n, CV)$. The second column shows the percent gaps between the expected profits obtained by IDE and DBP, giving an indication of how much the expected profit from DBP lags behind the unattainable upper bound provided by IDE. Our computational experiments indicate that DBP generally provides improvements over GRP, OSL and SFP. We design the presentation of our results to demonstrate the additional benefits from using DBP. So, the third column in Tables 1 and 2 shows the percent gaps between the expected profits obtained by DBP and GRP, whereas the fourth column shows the percent gaps between the expected profits obtained by DBP and OSL. Finally, the fifth column shows the percent gaps between the expected profits obtained by DBP and SFP. Positive gaps in the last three columns favor DBP, whereas negative gaps favor GRP, OSL or SFP.

For the test problems where the demand at the test marketing stage is always 50, the results in Table 1 indicate that DBP provides significantly better performance than GRP. The average performance gap between DBP and GRP is 22.26% and there are test problems where the performance gap between the two benchmark policies can exceed 50%. DBP generally provides improvements over OSL as well and the performance gap between DBP and OSL becomes particularly noticeable for the test problems with five failure types, reaching 20.95%. DBP and SFP are comparable for the test problems with two failure types. This observation is sensible because SFP makes its decisions by aggregating all failure types and if the number of failure types

Table 1 **Computational result for the test problems with $D_0 = 50$ and $D_\ell = \bar{D}$ for $\ell = 1, \ldots, \tau$. In the last three columns, cells shaded in gray are statistically significant at 5% level.**

Test Problem				Percent Gaps			
D_0	$\bar{D}$	n	CV	IDE vs. DBP	DBP vs. GRP	DBP vs. OSL	DBP vs. SFP
50	50	2	0.6	3.67%	2.70%	-0.18%	1.15%
50	50	2	0.6^2	8.95%	8.63%	2.79%	1.05%
50	50	2	0.6^3	23.29%	17.32%	-0.21%	-1.78%
50	50	2	0.6^4	27.44%	28.92%	5.53%	-5.00%
50	100	2	0.6	3.23%	3.08%	-0.17%	0.71%
50	100	2	0.6^2	8.18%	8.01%	1.56%	0.85%
50	100	2	0.6^3	19.70%	19.30%	1.56%	0.06%
50	100	2	0.6^4	22.05%	35.82%	7.94%	-3.03%
50	200	2	0.6	2.57%	3.75%	0.40%	0.91%
50	200	2	0.6^2	7.40%	7.58%	2.18%	0.78%
50	200	2	0.6^3	17.39%	21.93%	3.32%	-1.71%
50	200	2	0.6^4	20.24%	34.31%	12.98%	-5.01%
50	500	2	0.6	2.29%	3.72%	0.53%	1.08%
50	500	2	0.6^2	6.35%	7.68%	3.00%	1.18%
50	500	2	0.6^3	15.77%	17.75%	1.94%	-1.38%
50	500	2	0.6^4	17.79%	30.42%	10.48%	-3.70%
50	1000	2	0.6	2.17%	3.81%	1.31%	0.82%
50	1000	2	0.6^2	5.99%	7.95%	3.94%	0.61%
50	1000	2	0.6^3	14.69%	16.10%	7.10%	-2.32%
50	1000	2	0.6^4	16.05%	29.47%	13.12%	-4.46%
Average				12.26%	15.41%	3.96%	-0.96%

Test Problem				Percent Gaps			
D_0	$\bar{D}$	n	CV	IDE vs. DBP	DBP vs. GRP	DBP vs. OSL	DBP vs. SFP
50	50	5	0.6	10.05%	8.87%	3.44%	1.89%
50	50	5	0.6^2	19.00%	25.03%	7.04%	-0.71%
50	50	5	0.6^3	32.11%	43.08%	15.55%	-0.46%
50	50	5	0.6^4	37.98%	54.99%	20.95%	5.21%
50	100	5	0.6	9.85%	8.51%	2.17%	0.88%
50	100	5	0.6^2	14.11%	24.28%	7.77%	4.16%
50	100	5	0.6^3	25.23%	39.71%	17.15%	2.86%
50	100	5	0.6^4	37.74%	51.75%	5.92%	10.29%
50	200	5	0.6	7.88%	8.80%	2.72%	3.17%
50	200	5	0.6^2	12.69%	26.36%	9.86%	4.52%
50	200	5	0.6^3	22.21%	41.02%	17.73%	5.19%
50	200	5	0.6^4	31.20%	34.29%	13.23%	5.96%
50	500	5	0.6	6.97%	9.40%	4.04%	2.23%
50	500	5	0.6^2	10.84%	26.20%	11.74%	3.79%
50	500	5	0.6^3	19.95%	35.56%	15.21%	4.43%
50	500	5	0.6^4	26.68%	42.79%	17.36%	8.61%
50	1000	5	0.6	6.67%	8.65%	3.51%	3.84%
50	1000	5	0.6^2	10.25%	24.66%	11.82%	7.28%
50	1000	5	0.6^3	17.72%	37.49%	20.93%	9.93%
50	1000	5	0.6^4	24.90%	30.97%	16.12%	11.34%
Average				19.20%	29.12%	11.21%	4.72%

Table 2 **Computational result for the test problems with $D_0 = \bar{D}$ and $D_\ell = \bar{D}$ for $\ell = 1, \ldots, \tau$. In the last three columns, cells shaded in gray are statistically significant at 5% level.**

Test Problem				Percent Gaps			
D_0	$\bar{D}$	n	CV	IDE vs. DBP	DBP vs. GRP	DBP vs. OSL	DBP vs. SFP
50	50	2	0.6	3.67%	2.70%	-0.18%	1.15%
50	50	2	0.6^2	8.95%	8.63%	2.79%	1.05%
50	50	2	0.6^3	23.29%	17.32%	-0.21%	-1.78%
50	50	2	0.6^4	27.44%	28.92%	5.53%	-5.00%
100	100	2	0.6	2.49%	0.63%	-0.17%	0.40%
100	100	2	0.6^2	4.51%	3.98%	0.34%	0.79%
100	100	2	0.6^3	11.97%	9.48%	0.66%	1.02%
100	100	2	0.6^4	17.95%	28.82%	5.98%	-1.20%
200	200	2	0.6	1.27%	0.84%	0.13%	0.33%
200	200	2	0.6^2	3.66%	2.68%	0.42%	0.03%
200	200	2	0.6^3	8.01%	5.86%	0.44%	1.22%
200	200	2	0.6^4	14.65%	12.47%	1.66%	0.24%
500	500	2	0.6	0.52%	0.36%	-0.07%	0.09%
500	500	2	0.6^2	1.75%	0.75%	-0.04%	0.13%
500	500	2	0.6^3	4.39%	2.04%	0.50%	0.27%
500	500	2	0.6^4	7.33%	4.34%	2.06%	0.50%
1000	1000	2	0.6	0.52%	0.34%	0.15%	0.01%
1000	1000	2	0.6^2	0.75%	0.48%	-0.18%	0.18%
1000	1000	2	0.6^3	2.77%	1.86%	-0.21%	0.81%
1000	1000	2	0.6^4	4.51%	6.09%	2.46%	0.80%
Average				7.52%	6.93%	1.10%	0.05%

Test Problem				Percent Gaps			
D_0	$\bar{D}$	n	CV	IDE vs. DBP	DBP vs. GRP	DBP vs. OSL	DBP vs. SFP
50	50	5	0.6	10.05%	8.87%	3.44%	1.89%
50	50	5	0.6^2	19.00%	25.03%	7.04%	-0.71%
50	50	5	0.6^3	32.11%	43.08%	15.55%	-0.46%
50	50	5	0.6^4	37.98%	54.99%	20.95%	5.21%
100	100	5	0.6	6.38%	3.80%	1.27%	3.74%
100	100	5	0.6^2	12.09%	19.70%	5.75%	4.13%
100	100	5	0.6^3	24.12%	32.26%	17.03%	6.19%
100	100	5	0.6^4	33.40%	24.87%	4.42%	6.10%
200	200	5	0.6	4.89%	2.57%	-0.58%	1.55%
200	200	5	0.6^2	9.21%	5.35%	-0.73%	3.56%
200	200	5	0.6^3	17.76%	20.26%	6.53%	4.72%
200	200	5	0.6^4	24.99%	25.62%	5.12%	6.13%
500	500	5	0.6	3.44%	-0.45%	-1.58%	1.77%
500	500	5	0.6^2	6.00%	1.96%	-2.63%	3.00%
500	500	5	0.6^3	13.14%	6.20%	-3.90%	5.90%
500	500	5	0.6^4	20.32%	8.79%	-2.64%	5.50%
1000	1000	5	0.6	2.32%	-0.60%	-1.35%	0.97%
1000	1000	5	0.6^2	5.55%	-2.80%	-4.07%	2.11%
1000	1000	5	0.6^3	10.74%	-0.50%	-5.48%	4.48%
1000	1000	5	0.6^4	17.16%	0.16%	-6.25%	5.59%
Average				15.53%	13.96%	2.89%	3.57%

is smaller, then the error induced by the aggregation is likely to be less significant. However, DBP provides significant benefits over SFP when we have five failure types. For the test problems with five failure types, the performance gap between DBP and SFP can exceed 10%.

We observe some interesting trends in Table 1. To begin with, the performance gap between

DBP and GRP increases as CV decreases. As mentioned above, as CV gets smaller, it becomes more likely that the true failure probabilities $\boldsymbol{p}$ roughly satisfy $r = K\boldsymbol{c}^\top\boldsymbol{p}$. In this case, it is more difficult to detect whether it is profitable to continue or stop selling the product. It is encouraging that DBP, which captures the learning process by using a dynamic program, performs better than GRP as the underlying problem gets more difficult. A similar trend holds when we compare the performance of DBP with that of OSL and SFP, especially for the test problems with five failure types. On the other hand, comparing the left and right portions of Table 1, we observe that the gap between IDE and DBP gets larger when we have a larger number of failure types, indicating that the problem tends to be more difficult when there are more failure types. Furthermore, the performance gaps between DBP and the remaining three benchmark policies GRP, OSL and SFP tend to increase as the number of failure types increases. Finally, the performance gap between IDE and DBP decreases as the demand at each time period increases. This trend is sensible because we collect more information on the failure probabilities when the demand at each time period is larger, in which case, DBP is able to assess the failure probabilities quickly and make a sound decision on whether to continue or stop selling the product.

Table 2 gives our results for the test problems where the demand at the test marketing stage is equal to the demand at the subsequent time periods. For these test problems, DBP continues to provide improvements over GRP, but the performance gap between these two benchmark policies decreases as $\bar{D}$ increases and we observe more demand at each time period. As we observe more demand at each time period, GRP can quickly assess the failure probabilities. This observation also agrees with the results in Section 6, which show that the lower bounds obtained by GRP become asymptotically tight as the demand at each time period is scaled up linearly with the same rate. When both the demand at each time period and the value of CV are small, DBP provides improvements over OSL. For example, considering the test problems where the demand at each time period is 50, 100 or 200 and CV is 0.6^4, the performance gap between DBP and OSL ranges roughly around 5 to 20%. When the demand at each time period is small, it becomes difficult to learn the failure probabilities quickly. Furthermore, as CV gets smaller, it becomes more difficult to detect whether it is profitable to continue or stop selling the product. In this type of situation, DBP can provide improvements over OSL. However, for the test problems with larger demands or with larger values of CV, OSL can perform noticeably better than DBP. For the test problems with two failure types, DBP and SFP perform comparably, but DBP performs significantly better than SFP when there are five failure types.

To sum up, when we have smaller demand quantities at each time period, it is more difficult to learn the failure probabilities. Furthermore, if the value of CV is small, then it is difficult to detect whether it is profitable to continue or stop selling the product and it becomes important to have good estimates of the failure probabilities. In these cases, DBP generally performs better than GRP, OSL and SFP. We observe that DBP bases its decisions on a dynamic programming formulation. In addition, this benchmark policy attempts to capture the interactions between the different failure types to a certain extent through solving problem (4). These features help DBP perform particularly well when it is difficult to form good estimates for the failure probabilities quickly and it is important to have such good estimates. As we are able to observe more demand at each time period, the performance gap between DBP and GRP gets smaller. OSL, which is obtained by applying a one step look ahead on the decision rule used by GRP, can improve the performance of GRP to such an extent that it can provide better expected profits than DBP, but such improvements tend to occur only when there is a large amount of demand at the test marketing stage so that we can quickly assess the failure probabilities. Since SFP is based on aggregating our beliefs about the failure probabilities of different failure types, it is effective when we have a small number of failure types, but this benchmark policy becomes unreliable as the number of failure types gets larger. With the goal of understanding how the behavior of different benchmark policies is affected by various problem parameters, we give additional computational results in Appendix H that compare when we stop selling the product under different benchmark policies.

Considering the computational effort for the different benchmark policies, the computation time per decision for GRP is negligible. OSL requires computing the expectation of a piecewise linear function of a beta binomial random variable and it still requires a fraction of a second per decision. SFP requires solving a dynamic program with a scalar state variable. This dynamic program is solved only once, after which the computation time per decision is also negligible. The computation time for DBP grows roughly linearly with the number of failure types. For the test problems with five failure types, DBP takes about one minute per decision on a Pentium IV Desktop PC with 2.4 GHz CPU and 4 GB RAM running Windows XP. Majority of this computation time is spent on solving problem (4) by using standard subgradient search.

8. Conclusions and Extensions

We studied the problem faced by a company selling a product under limited information about the probabilities of different failure types. The goal is to learn the failure probabilities as sales take place and dynamically decide whether it is profitable to continue or stop selling the product. Our

approach builds on a dynamic programming formulation with embedded optimal stopping and learning features. To deal with the high dimensional state variable in the dynamic program, we proposed approximation methods that provide upper and lower bounds on the value functions. Our computational experiments indicated that the approximation method based on upper bounds is more computationally intensive, but it can provide significant improvements.

There are a number of possible extensions to our model. Our model assumes that there is no other revenue stream once we stop selling the product, but it is not difficult to extend our model to deal with the case where there is a standard substitute product that brings a known expected net profit contribution and stopping to sell the current product means switching to this substitute product. Similarly, our model assumes that the warranty coverage is for K consecutive time periods, but we can work with other forms of warranty coverage. We can also incorporate repair costs that depend on how long a unit has been with the owner. Although our model assumes that each unit can fail from multiple failure types at a time period, we can use a Dirichlet multinomial learning model to capture the situation where each unit can fail from at most one failure at a time period. Finally, our model assumes that the failure probabilities are constant, but we can work with failure probabilities that depend on how long ago the product was sold to a customer. In Appendix I, we describe in detail how to incorporate all of these extensions into our model.

The extensions described above are not too difficult to incorporate into our model, but some other extensions need significant changes in our approach, requiring additional research. In this paper, we assume that the demand quantities are deterministic. There are some settings with accurate forecasts or advance demand information that may make this assumption reasonable, but it is certainly desirable to incorporate demand uncertainty into our model. The advantage of working with deterministic demand quantities is that when this assumption holds, it is straightforward to compute the number of units that are under warranty coverage at any time period. In contrast, if the demand quantities are random, then we need to extend the state variable of our dynamic programming formulation to keep track of the numbers of units that were sold at different time periods. This additional component in the state variable brings complications. For example, when the state variable has this additional component, it is not possible to shift the timing of the costs, as it was done in Section 3 to obtain the optimality equation in (2). Also, if the state variable keeps track of the numbers of units sold at different time periods, then the update of our belief about a particular failure probability depends on the numbers of units sold at different time periods. In this case, the different components of the state variable do not evolve independently and the decomposition idea that we use in Section 5 may become ineffective.

Also, our beta binomial learning model assumes that different types of failures occur independently of each other, while the Dirichlet multinomial learning model described above ensures that each unit can fail from at most one failure at a time, inducing negative correlations between numbers of failures of different types. In reality, there may be general correlations among the different failure types and one may adopt other learning models to allow more general correlations. For example, we may use generalized Dirichlet distribution which allows general correlations and is still a conjugate prior for the multinomial distribution. Finally, it is worthwhile to investigate the possibility that it may be better for the company not to serve all of the demand in a time period. By rationing the supply in the early time periods, the company may be able to learn about the reliability of the product while controlling the risk of facing too many returns.

Acknowledgements

The authors would like to thank the department editor, the senior editor and two anonymous referees for their valuable comments that improved the paper in many ways.

References

Adell, J.A., F.G. Badia, J. de la Cal. 1993. Beta-type operators preserve shape properties. *Stochastic Processes and their Applications* **48**(1) 1–8.

Adell, J.A., F.G. Badia, J. de la Cal, Fernando Plo. 1996. On the property of monotonic convergence for beta operators. *Journal of Approximation Theory* **84**(1) 61–73.

Adelman, D., A.J. Mersereau. 2008. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research* **56**(3) 712–727.

Araman, V.F., R. Caldentey. 2009. Dynamic pricing for non-perishable products with demand learning. *Operations Research* **57**(5) 1169–1188.

Arlotto, A., S.E. Chick, N. Gans. 2011. Optimal hiring and retention policies for heterogeneous workers who learn. Tech. rep., INSEAD.

Aviv, Y., A. Pazgal. 2005. A partially observed Markov decision process for dynamic pricing. *Management Science* **51**(9) 1400–1416.

Berry, D.A., B. Fristedt. 1985. *Bandit problems: Sequential allocation of experiments*. Springer, New York, NY.

Bertsimas, D., A.J. Mersereau. 2007. A learning approach for interactive marketing to a customer segment. *Operations Research* **55**(6) 1120–1135.

Boshuizen, F.A., J.M. Gouweleeuw. 1993. General optimal stopping theorems for semi-Markov processes. *Advances in Applied Probability* 825–846.

Boyaci, T., O. Ozer. 2010. Information acquisition for capacity planning via pricing and advance selling: When to stop and act? *Operations Research* **58**(5) 1328–1349.

Bradt, R.N., S.M. Johnson, S. Karlin. 1956. On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics* **27**(4) 1060–1074.

Brown, D.B., J.E. Smith, P. Sun. 2010. Information relaxations and duality in stochastic dynamic programs. *Operations Research* **58**(4) 785–801.

Burnetas, A.N., M.N. Katehakis. 1996. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* **17** 122–142.

Burnetas, A.N., M.N. Katehakis. 1998. Dynamic allocation policies for the finite horizon one armed bandit problem. *Stochastic Analysis and Applications* **16**(5) 811–824.

Burnetas, A.N., M.N. Katehakis. 2003. Asymptotic Bayes Analysis for the Finite-Horizon One-Armed-Bandit Problem. *Probability in the Engineering and Informational Sciences* **17**(1) 53–82.

Caro, F., J. Gallien. 2007. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science* **53**(2) 276.

Chow, Yuan Shih, Herbert Robbins, David Siegmund. 1971. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin Boston.

Cooper, R., B. Hayes. 1987. Multi-period insurance contracts. *International Journal of Industrial Organization* **5**(2) 211–231.

Erdelyi, A., H. Topaloglu. 2010. A dynamic programming decomposition method for making overbooking decisions over an airline network. *INFORMS Journal on Computing* **22**(3) 443–456.

Farias, V.F., B. Van Roy. 2010. Dynamic pricing with a prior on market response. *Operations Research* **58**(1) 16–29.

Feng, Y., G. Gallego. 1995. Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science* **41**(8) 1371–1391.

Gallego, G. 1992. A minmax distribution free procedure for the (q,r) inventory model. *Operations Research Letters* **11**(1) 55–60.

Gittins, J.C. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**(2) 148–177.

Gittins, J.C. 1989. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons Inc.

Goldenshluger, A., A. Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *Annals of Applied Probability* **19**(4) 1603–1633.

Goldenshluger, A., A. Zeevi. 2011. Performance limitations in bandit problems with side observations. *To appear in IEEE Transactions on Information Theory* .

Harrison, J., N. Keskin, A. Zeevi. 2010. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. Tech. rep., Columbia and Stanford University.

Hawkins, J.T. 2003. A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications. Ph.D. thesis, MIT.

Heyman, D.P., M.J. Sobel. 2003. *Stochastic Models in Operations Research: Stochastic Optimization*. Dover Books on Computer Science Series, Dover Publications Inc.

Hosios, A.J., M. Peters. 1989. Repeated insurance contracts with adverse selection and limited commitment. *The Quarterly Journal of Economics* **104**(2) 229–253.

Mission Repair. 2012. Mission repair: Iphone 4s. URL `http://www.missionrepair.com/iPhone_4S_Repair_and_Diagnosis_s/762.htm`.

Muciek, Bogdan K, Krzysztof J Szajowski. 2008. Optimal stopping of a risk process when claims are covered immediately. *arXiv preprint arXiv:0812.3925* .

Muckstadt, J.A., A. Sapra. 2010. *Principles of Inventory Management: When You Are Down to Four, Order More*. Springer Series in Operations Research and Financial Engineering, Springer.

Peskir, G., A. Shiryaev. 2006. *Optimal Stopping and Free-Boundary Problems*. Birkhauser.

Ruszczynski, A. 2011. *Nonlinear Optimization*. Princeton University Press.

Schöttl, A. 1998. Optimal stopping of a risk reserve process with interest and cost rates. *Journal of Applied Probability* **35**(1) 115–123.

Shaked, M., J.G. Shanthikumar. 2007. *Stochastic Orders*. Springer Series in Statistics, Springer.

Teerapabolarn, K. 2008. A bound on the binomial approximation to the beta binomial distribution. *International Mathematical Forum* **3**(28) 1355–1358.

Top Extended Vehicle Warranty Providers. 2010. A top ten list of automotive extended warranty companies. URL `http://ezinearticles.com/?A-Top-Ten-List-of-Automotive-Extended-Warranty-Companies&id=3728026`.

Topaloglu, H. 2009. Using Lagrangian relaxation to compute capacity-dependent bid prices in network revenue management. *Operations Research* **57**(3) 637–649.

Topaloglu, H., S. Kunnumkal. 2010. Computing time-dependent bid-prices in network revenue management problems. *Transportation Science* **44**(1) 38–62.

Watt, R., F.J. Vazquez. 1997. Full insurance, Bayesian updated premiums, and adverse selection. *The Geneva Risk and Insurance Review* **22**(2) 135–150.

Whittle, P. 1983. *Optimization Over Time: Dynamic Programming and Stochastic Control*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley.

Xerox Corporation. 2011. U.S. and Canada on-site service agreement. URL `http://www.office.xerox.com/latest/SERTC-16.PDF`.

Online Appendix to "Balancing Revenues and Repair Costs under Partial Information about Product Reliability "

Appendix A: Simplification of the Cost Function

In this section, we show that $C_t(\boldsymbol{\theta}_t) - S_t(\boldsymbol{\theta}_t) + S_{t+1}(\boldsymbol{\theta}_t) = K \sum_{i=1}^n c_i \, \theta_{it} \, D_t$. To see that this identity holds, we use the definitions of $C_t(\cdot)$ and $S_t(\cdot)$ to obtain

$$\begin{aligned}
C_t(\boldsymbol{\theta}_t) - S_t(\boldsymbol{\theta}_t) + S_{t+1}(\boldsymbol{\theta}_t) &= \sum_{i=1}^n \sum_{s=0}^t c_i \, \mathbf{1}(t-s<K) \, D_s \, \theta_{it} - \sum_{i=1}^n \sum_{\ell=t}^\infty \sum_{s=0}^{t-1} c_i \, \mathbf{1}(\ell-s<K) \, D_s \, \theta_{it} \\
&\qquad + \sum_{i=1}^n \sum_{\ell=t+1}^\infty \sum_{s=0}^t c_i \, \mathbf{1}(\ell-s<K) \, D_s \, \theta_{it} \\
&= \sum_{i=1}^n \sum_{\ell=t}^\infty \sum_{s=0}^t c_i \, \mathbf{1}(\ell-s<K) \, D_s \, \theta_{it} - \sum_{i=1}^n \sum_{\ell=t}^\infty \sum_{s=0}^{t-1} c_i \, \mathbf{1}(\ell-s<K) \, D_s \, \theta_{it} \\
&= \sum_{i=1}^n \sum_{\ell=t}^\infty c_i \, \mathbf{1}(\ell-t<K) \, D_t \, \theta_{it} = K \sum_{i=1}^n c_i \, \theta_{it} \, D_t,
\end{aligned}$$

where the first equality uses the fact that $W_t = \sum_{s=0}^t \mathbf{1}(t-s<K) \, D_s$.

Appendix B: Proofs of Structural Properties

Proof of Proposition 1. Using the notation defined after Lemma 1, we observe that the random variable $Y_{it}(\theta_{it})$ can be written as $\mathsf{Binomial}(W_t, \mathsf{Beta}(\theta_{it} \, M_t, (1-\theta_{it}) \, M_t))$. In this case, using the second part of Lemma 1 and the discussion that follows this lemma, we observe that the family of random variables $\{Y_{it}(\theta_{it}) : \theta_{it} \in [0,1]\}$ is stochastically increasing and stochastically convex. Thus, the first part of Lemma 1 implies that $\{\lambda_t \, \theta_{it} + \frac{1-\lambda_t}{W_t} \, Y_{it}(\theta_{it}) : \theta_{it} \in [0,1]\}$ is a stochastically increasing and stochastically convex family of random variables. Based on these observations, we proceed to showing the desired result by using induction over the time periods. The result trivially holds at time period $\tau+1$. Assuming that the result holds at time period $t+1$, we write the expectation on the right side of (2) as

$$\mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} = \mathbb{E}\{\mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t)) \, | \, Y_{jt}(\theta_{jt}) \text{ for all } j \in \{1,\ldots,n\} \setminus \{i\}\}\}.$$

We note that $V_{t+1}(\boldsymbol{\theta}_{t+1})$ is decreasing and convex in $\theta_{i,t+1}$ by the induction argument. Furthermore, the family of random variables $\{\lambda_t \, \theta_{it} + \frac{1-\lambda_t}{W_t} \, Y_{it}(\theta_{it}) : \theta_{it} \in [0,1]\}$ is stochastically increasing and stochastically convex. Thus, it follows that

$$\mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t)) \, | \, Y_{jt}(\theta_{jt}) \text{ for all } j \in \{1,\ldots,n\} \setminus \{i\}\}$$

is a decreasing and convex function of θ_{it} for any realization of the random variables $Y_{jt}(\theta_{jt})$ for $j \in \{1,\ldots,n\} \setminus \{i\}$. Noting that $\{Y_{it}(\theta_{it}) : i = 1,\ldots,n\}$ are independent of each other, taking the expectation of the expression above, we get that $\mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}$ is decreasing and convex in θ_{it}. Thus, since $(r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t) \, D_t$ is a decreasing and linear function of θ_{it}, $(r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t) \, D_t + \mathbb{E}\{V_{t+1}(\lambda_t \, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \, \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\}$ is decreasing and convex in θ_{it} as well. Since the pointwise maximum of two decreasing and convex functions is decreasing and convex, the optimality equation in (2) implies that $V_t(\boldsymbol{\theta}_t)$ is decreasing and convex in θ_{it}. $\square$

Proof of Proposition 3. Since it is optimal to continue at time period t when the state of the system is $\boldsymbol{\theta}_t$, it must be the case that $(r - K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} > 0$. If no units fail at time period t, then the state at time period $t+1$ is $\boldsymbol{\theta}_{t+1} = \lambda_t\,\boldsymbol{\theta}_t$. To obtain a contradiction, we assume that $(r - K\boldsymbol{c}^\top\boldsymbol{\theta}_{t+1})\, D_{t+1} + \mathbb{E}\{V_{t+2}(\lambda_{t+1}\,\boldsymbol{\theta}_{t+1} + \frac{1-\lambda_{t+1}}{W_{t+1}}\,\boldsymbol{Y}_{t+1}(\boldsymbol{\theta}_{t+1}))\} \leq 0$ so that it is optimal to stop at time period $t+1$ when the state is $\boldsymbol{\theta}_{t+1} = \lambda_t\,\boldsymbol{\theta}_t$. The last inequality implies that $(r - K\boldsymbol{c}^\top\boldsymbol{\theta}_{t+1})\, D_{t+1} \leq 0$ since we have $V_t(\cdot) \geq 0$ for all $t = 1, \ldots, \tau$ by the optimality equation in (2). In this case, we have

$$\begin{aligned} 0 < (r - K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} &\leq (r - K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t)\} \\ &= (r - K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\, D_t \leq (r - K\lambda_t\,\boldsymbol{c}^\top\boldsymbol{\theta}_t)\, D_t = (r - K\boldsymbol{c}^\top\boldsymbol{\theta}_{t+1})\, D_t \leq 0, \end{aligned}$$

which is a contradiction and this completes the proof. The second inequality in the chain of inequalities above follows by noting that $\lambda_t\,\boldsymbol{\theta}_t \leq \lambda_t\,\boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\,\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$ and using the fact that the value functions are decreasing by Proposition 1. The first equality follows from the fact that $V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t) = 0$ since we assume that it is optimal to stop at time period $t+1$ when the state is $\lambda_t\,\boldsymbol{\theta}_t$. □

Appendix C: Simplifying the Computation of Expectations Involving Beta Binomial Random Variables

In this section, we show that by replacing the beta binomial random variable $Y_{it}(\theta_{it})$ in the optimality equation (2) with a binomial random variable $Z_{it}(\theta_{it})$ with parameters (W_t, θ_{it}), we obtain lower bounds on the value functions. Using the vector $\boldsymbol{Z}_t(\boldsymbol{\theta}_t) = (Z_{1t}(\theta_{1t}), \ldots, Z_{nt}(\theta_{nt}))$, consider the optimality equation

$$\bar{V}_t(\boldsymbol{\theta}_t) = \max\Big\{(r - K\boldsymbol{c}^\top\boldsymbol{\theta}_t)\, D_t + \mathbb{E}\{\bar{V}_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t}\,\boldsymbol{Z}_t(\boldsymbol{\theta}_t))\}, 0\Big\}, \tag{C.1}$$

with the boundary condition $\bar{V}_{\tau+1}(\cdot) = 0$. The expectation above is computed with respect to the vector of binomial random variables $\boldsymbol{Z}_t(\boldsymbol{\theta}_t)$. The main result of this section is the following proposition, which shows that the value function $\bar{V}_t(\cdot)$ provides a lower bound on the original value function.

Proposition C1. *For all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$, we have $V_t(\boldsymbol{\theta}_t) \geq \bar{V}_t(\boldsymbol{\theta}_t)$.*

The proof of the above result makes use of the following general form of Jensen's inequality for componentwise convex functions.

Lemma C1. *If $g(\cdot) : \Re^n \to \Re$ is a componentwise convex function and $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random variable taking values in $\Re^n$ with independent components, then $\mathbb{E}\{g(\boldsymbol{X})\} \geq g(\mathbb{E}\{\boldsymbol{X}\})$.*

Proof. We prove the result by induction on n. When $n = 1$, we have a scalar convex function and the result trivially holds by the standard form of Jensen's inequality. Assuming that the result holds when we deal with functions that map $\Re^{n-1}$ to $\Re$, we have

$$\mathbb{E}\{g((X_1, \ldots, X_n))\} = \mathbb{E}\{\mathbb{E}\{g((X_1, \ldots, X_{n-1}, X_n)) \,|\, X_1, \ldots, X_{n-1}\}\}.$$

Since $g(\cdot)$ is componentwise convex, $g((X_1, \ldots, X_{n-1}, X_n))$ is a convex function of X_n and applying Jensen's inequality on the scalar convex function $g(X_1, \ldots, X_{n-1}, \cdot)$, we obtain

$$\mathbb{E}\{g((X_1, \ldots, X_{n-1}, X_n)) \,|\, X_1, \ldots, X_{n-1}\} \geq g((X_1, \ldots, X_{n-1}, \mathbb{E}\{X_n\})),$$

where we use the fact that the distribution of X_n conditional on $X_1, \ldots, X_{n-1}$ is the same as the unconditional distribution of X_n. Viewing $g(\cdot, \ldots, \cdot, \mathbb{E}\{X_n\})$ on the right side of the inequality above as a function that maps $\Re^{n-1}$ to $\Re$, by the induction assumption, we have

$$\mathbb{E}\{g((X_1, \ldots, X_{n-1}, \mathbb{E}\{X_n\}))\} \geq g((\mathbb{E}\{X_1\}, \ldots, \mathbb{E}\{X_{n-1}\}, \mathbb{E}\{X_n\})),$$

so that we obtain

$$\begin{aligned}\mathbb{E}\{g(\boldsymbol{X})\} &= \mathbb{E}\{\mathbb{E}\{g((X_1, \ldots, X_{n-1}, X_n)) \,|\, X_1, \ldots, X_{n-1}\}\} \\ &\geq \mathbb{E}\{g((X_1, \ldots, X_{n-1}, \mathbb{E}\{X_n\}))\} \geq g((\mathbb{E}\{X_1\}, \ldots, \mathbb{E}\{X_{n-1}\}, \mathbb{E}\{X_n\})) = g(\mathbb{E}\{\boldsymbol{X}\}). \quad \square\end{aligned}$$

Proof of Proposition C1. We show the result by using induction over the time periods. The result trivially holds at time period $\tau + 1$. Assuming that the result holds at time period $t + 1$, we have

$$\mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} = \mathbb{E}\{\mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t)) \,|\, \boldsymbol{P}_t(\boldsymbol{\theta}_t)\}\} = \mathbb{E}\{G_{t+1}(\boldsymbol{P}_t(\boldsymbol{\theta}_t))\}, \tag{C.2}$$

where we use the vector $\boldsymbol{P}_t(\boldsymbol{\theta}_t) = (P_{1t}(\theta_{1t}), \ldots, P_{nt}(\theta_{nt}))$ and define the function $G_{t+1}(\cdot) : [0,1]^n \to \Re$ as $G_{t+1}(\boldsymbol{p}) = \mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t)) \,|\, \boldsymbol{P}_t(\boldsymbol{\theta}_t) = \boldsymbol{p}\}$. Conditional on $P_{it}(\theta_{it}) = p_i$, the random variable $Y_{it}(\theta_{it})$ has a binomial distribution with parameters (W_t, p_i). Therefore, noting that the family of random variables $\{\mathsf{Binomial}(W_t, p_i) : p_i \in [0,1]\}$ is stochastically convex and $V_{t+1}(\cdot)$ is componentwise convex by Proposition 1, $G_{t+1}(\boldsymbol{p}) = \mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t)) \,|\, \boldsymbol{P}_t(\boldsymbol{\theta}_t) = \boldsymbol{p}\}$ is componentwise convex in $\boldsymbol{p}$. In this case, Jensen's inequality in Lemma C1 implies that $\mathbb{E}\{G_{t+1}(\boldsymbol{P}_t(\boldsymbol{\theta}_t))\} \geq G_{t+1}(\mathbb{E}\{\boldsymbol{P}_t(\boldsymbol{\theta}_t)\}) = G_{t+1}(\boldsymbol{\theta}_t)$ and we continue the chain of equalities in (C.2) as

$$\begin{aligned}\mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} \geq G_{t+1}(\boldsymbol{\theta}_t) &= \mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t)) \,|\, \boldsymbol{P}_t(\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t\} \\ &= \mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \boldsymbol{Z}_t(\boldsymbol{\theta}_t))\} \geq \mathbb{E}\{\bar{V}_{t+1}(\lambda_t \boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{W_t} \boldsymbol{Z}_t(\boldsymbol{\theta}_t))\},\end{aligned}$$

where the second equality follows from the fact that conditional on $\boldsymbol{P}_t(\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t$, the random variable $\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$ has the same distribution as $\boldsymbol{Z}_t(\boldsymbol{\theta}_t)$ and the second inequality follows from the induction assumption. Therefore, we have $\mathbb{E}\{V_{t+1}(\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Y}_t(\boldsymbol{\theta}_t))\} \geq \mathbb{E}\{\bar{V}_{t+1}(\lambda_t \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t} \boldsymbol{Z}_t(\boldsymbol{\theta}_t))\}$. In this case, noting the optimality equations in (2) and (C.1), it follows that $V_t(\boldsymbol{\theta}_t) \geq \bar{V}_t(\boldsymbol{\theta}_t)$. $\square$

Appendix D: Upper Bound on the Value Functions

In this section, we establish that $V_{it}^U(\theta_{it} \,|\, \rho_i)$ is a convex function of ρ_i and show how to compute a subgradient of $V_{it}^U(\theta_{it} \,|\, \rho_i)$ with respect to ρ_i. We begin by using induction over the time periods to show that $V_{it}^U(\theta_{it} \,|\, \rho_i)$ is a convex function of ρ_i. The result trivially holds at time period $\tau + 1$. Assuming that the result holds at time period $t+1$, it follows that $V_{i,t+1}^U(\lambda_t \theta_{it} + \frac{1-\lambda_t}{W_t} Y_{it}(\theta_{it}) \,|\, \rho_i)$ is a convex function of ρ_i for every realization of $Y_{it}(\theta_{it})$. Therefore, $(\rho_i - K c_i \theta_{it}) D_t + \mathbb{E}\{V_{i,t+1}^U(\lambda_t \theta_{it} + \frac{1-\lambda_t}{W_t} Y_{it}(\theta_{it}) \,|\, \rho_i)\}$ is a convex function of ρ_i. Since the pointwise maximum of two convex functions is also convex, the optimality equation in (3) implies that $V_{it}^U(\theta_{it} \,|\, \rho_i)$ is a convex function of ρ_i.

We define some new notation to compute the subgradient of $V_{it}^U(\theta_{it} \,|\, \rho_i)$ with respect to ρ_i. We let $\mathbf{1}_{it}^*(\cdot \,|\, \rho_i) : [0,1] \to \{0,1\}$ be the decision function at time period t that we obtain by solving the optimality equation

in (3). In other words, we have $\mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) = 1$ when $(\rho_i - K c_i\, \theta_{it})\, D_t + \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \frac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\} > 0$ and $\mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) = 0$ otherwise. In the optimality equation in (3), if we start with state θ_{it} at time period t and use $\Delta_{it}(\theta_{it} \,|\, \rho_i)$ to denote the total expected demand that we observe until we stop selling the product, then $\Delta_{it}(\theta_{it} \,|\, \rho_i)$ satisfies the recursion

$$\Delta_{it}(\theta_{it} \,|\, \rho_i) = \mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) \Big\{ D_t + \mathbb{E}\{\Delta_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\} \Big\}, \tag{D.1}$$

with the boundary condition $\Delta_{i,\tau+1}(\cdot \,|\, \rho_i) = 0$. In the rest of this section, we use induction over the time periods to show that $V^U_{it}(\theta_{it} \,|\, \cdot)$ satisfies the subgradient inequality

$$V^U_{it}(\theta_{it} \,|\, \hat{\rho}_i) \geq V^U_{it}(\theta_{it} \,|\, \rho_i) + \Delta_{it}(\theta_{it} \,|\, \rho_i)\, [\hat{\rho}_i - \rho_i], \tag{D.2}$$

which implies that $\Delta_{it}(\theta_{it} \,|\, \rho_i)$ is a subgradient of $V^U_{it}(\theta_{it} \,|\, \rho_i)$ with respect to ρ_i. If we start with state θ_{it} at time period t and follow the decision function $\mathbf{1}^*_{is}(\cdot \,|\, \rho_i)$ at time periods $s = t, t+1, \ldots, \tau$, then $\Delta_{it}(\theta_{it} \,|\, \rho_i)$ corresponds to the total expected demand that we observe until we stop selling the product. Therefore, we can use Monte Carlo simulation to estimate $\Delta_{it}(\theta_{it} \,|\, \rho_i)$. Alternatively, noting that the random variable $Y_{it}(\theta_{it})$ has a finite number of realizations, we can compute $\Delta_{it}(\theta_{it} \,|\, \rho_i)$ by computing $\Delta_{is}(\cdot \,|\, \rho_i)$ for a finite number of values at time periods $s = t+1, t+2, \ldots, \tau$.

The subgradient inequality in (D.2) trivially holds at time period $\tau + 1$. Assuming that this subgradient inequality holds at time period $t+1$, we write the optimality equation in (3) as

$$V^U_{it}(\theta_{it} \,|\, \rho_i) = \mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) \Big\{ (\rho_i - K c_i\, \theta_{it})\, D_t + \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\} \Big\}.$$

On the other hand, since we have $\mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) \in \{0, 1\}$, if we solve the optimality equation in (3) after replacing ρ_i with $\hat{\rho}_i$, then the value function $V^U_{it}(\theta_{it} \,|\, \hat{\rho}_i)$ satisfies

$$V^U_{it}(\theta_{it} \,|\, \hat{\rho}_i) \geq \mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) \Big\{ (\hat{\rho}_i - K c_i\, \theta_{it})\, D_t + \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \hat{\rho}_i)\} \Big\}.$$

Subtracting the last equality from the last inequality side by side, we obtain

$$\begin{aligned} V^U_{it}(\theta_{it} \,|\, \hat{\rho}_i) \geq V^U_{it}(\theta_{it} \,|\, \rho_i) + \mathbf{1}^*_{it}(\theta_{it} \,|\, \rho_i) \Big\{ & D_t\, [\hat{\rho}_i - \rho_i] \\ & + \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \hat{\rho}_i)\} - \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\} \Big\}. \end{aligned} \tag{D.3}$$

The induction assumption implies that

$$\begin{aligned} \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \hat{\rho}_i)\} - \mathbb{E}\{V^U_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\} & \\ \geq \mathbb{E}\{\Delta_{i,t+1}(\lambda_t\, \theta_{it} + \tfrac{1-\lambda_t}{W_t}\, Y_{it}(\theta_{it}) \,|\, \rho_i)\}\, [\hat{\rho}_i - \rho_i], & \end{aligned} \tag{D.4}$$

in which case, the subgradient inequality in (D.2) follows by using (D.4) and (D.1) in (D.3).

Appendix E: Closed Form Expression for the Lower Bound

In this section, we use induction over the time periods to show that the value functions computed through the optimality equation in (5) are given by the expression in (6). The result trivially holds at time period $\tau + 1$. Assuming that the result holds at time period $t+1$, the optimality equation in (5) implies that

$$V^L_t(\boldsymbol{\theta}_t) = \max \Big\{ (r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t)\, D_t + [r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \sum_{s=t+1}^{\tau} D_s, 0 \Big\} = [r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \sum_{s=t}^{\tau} D_s,$$

where the second equality follows by noting that the fact that the max operator above returns a nonzero value if and only if we have $r - K \boldsymbol{c}^\intercal \boldsymbol{\theta}_t > 0$.

Appendix F: Shrinking Property of $\mathcal{C}_t(m)$ as a Function of m

In this section, our ultimate goal is to give a proof for Proposition 6. This proof requires establishing results for stochastic convex orders that do not appear in the earlier literature. We begin this section with two new results for stochastic convex orders. For two random variables X and Y, we say that X is greater than or equal to Y in stochastic convex order whenever $\mathbb{E}\{\phi(Y)\} \leq \mathbb{E}\{\phi(X)\}$ for any convex function $\phi(\cdot): \Re \to \Re$. We denote this stochastic convex order between X and Y by using $Y \leq_{cx} X$. The following two lemmas show convex orders for the families of random variables $\{\mathsf{Binomial}(W,\theta)/W : W = 1,2,\ldots\}$ and $\{\mathsf{Binomial}(mW, \mathsf{Beta}(\theta\, mM, (1-\theta)\, mM))/mW : m = 1,2,\ldots\}$ for any fixed $\theta \in [0,1]$. These lemmas are useful for proving Proposition 6, but they may also have independent interest.

LEMMA F2. *For any fixed $\theta \in [0,1]$, letting $X(W) = \mathsf{Binomial}(W,\theta)/W$, we have $X(W+1) \leq_{cx} X(W)$.*

Proof. Throughout the proof, we use $=_{st}$ to denote equality in distribution. By Theorem 3.A.4 in Shaked and Shanthikumar (2007), it suffices to construct a pair of random variables $\hat{X}(W+1)$ and $\hat{X}(W)$ such that $\hat{X}(W+1) =_{st} X(W+1), \hat{X}(W) =_{st} X(W)$ and $\{\hat{X}(W+1), \hat{X}(W)\}$ is a martingale.

We let $\hat{X}(W+1) = X(W+1) = \mathsf{Binomial}(W+1,\theta)/(W+1)$, in which case, $\hat{X}(W+1)$ takes values in the set $\{0, 1/(W+1), \ldots, W/(W+1), 1\}$. We construct $\hat{X}(W)$ in the following way. If $\hat{X}(W+1) = 0$ or 1, then we set $\hat{X}(W) = \hat{X}(W+1)$. If, on the other hand, $\hat{X}(W+1) = j/(W+1)$ for some $j = 1,\ldots,W$, then we set $\hat{X}(W) = (j-1)/W$ with probability $\frac{j/W - j/(W+1)}{1/W}$ and $\hat{X}(W) = j/W$ with probability $\frac{j/(W+1) - (j-1)/W}{1/W}$. It is straightforward to check that these probabilities add up to one.

We proceed to showing that $\hat{X}(W) =_{st} X(W)$. To see this equivalence in distribution, we note that

$$\begin{aligned}
\mathbb{P}\{\hat{X}(W) = 0\} &= \mathbb{P}\{\hat{X}(W) = 0 \,|\, \hat{X}(W+1) = 0\}\, \mathbb{P}\{\hat{X}(W+1) = 0\} \\
&\quad + \mathbb{P}\{\hat{X}(W) = 0 \,|\, \hat{X}(W+1) = 1/(W+1)\}\, \mathbb{P}\{\hat{X}(W+1) = 1/(W+1)\} \\
&= 1 \binom{W+1}{0} (1-\theta)^{W+1} + \frac{1/W - 1/(W+1)}{1/W} \binom{W+1}{1} \theta\, (1-\theta)^{W} \\
&= (1-\theta)^{W+1} + \frac{1}{W+1}\, (W+1)\, \theta\, (1-\theta)^{W} = (1-\theta)^{W}.
\end{aligned}$$

Similarly, we can show that $\mathbb{P}\{\hat{X}(W) = 1\} = \theta^W$. On the other hand, for $j = 1,\ldots,W-1$, the definition of $\hat{X}(W)$ implies that

$$\begin{aligned}
\mathbb{P}\{\hat{X}(W) = \tfrac{j}{W}\} &= \mathbb{P}\{\hat{X}(W) = \tfrac{j}{W} \,|\, \hat{X}(W+1) = \tfrac{j}{W+1}\}\, \mathbb{P}\{\hat{X}(W+1) = \tfrac{j}{W+1}\} \\
&\quad + \mathbb{P}\{\hat{X}(W) = \tfrac{j}{W} \,|\, \hat{X}(W+1) = \tfrac{j+1}{W+1}\}\, \mathbb{P}\{\hat{X}(W+1) = \tfrac{j+1}{W+1}\} \\
&= \frac{j/(W+1) - (j-1)/W}{1/W} \binom{W+1}{j} \theta^j\, (1-\theta)^{W+1-j} \\
&\quad + \frac{(j+1)/W - (j+1)/(W+1)}{1/W} \binom{W+1}{j+1} \theta^{j+1}\, (1-\theta)^{W-j} \\
&= \frac{W-j+1}{W+1} \left[\frac{(W+1)W\ldots(W-j+2)}{j!}\right] \theta^j\, (1-\theta)^{W+1-j} \\
&\quad + \frac{j+1}{W+1} \left[\frac{(W+1)W\ldots(W-j+1)}{(j+1)j!}\right] \theta^{j+1}\, (1-\theta)^{W-j} \\
&= \binom{W}{j} \theta^j\, (1-\theta)^{W-j}\, (1-\theta+\theta) = \binom{W}{j} \theta^j\, (1-\theta)^{W-j}.
\end{aligned}$$

Therefore, we have $\hat{X}(W) =_{st} \mathsf{Binomial}(W,\theta)/W =_{st} X(W)$.

It remains to show that $\{\hat{X}(W+1), \hat{X}(W)\}$ is a martingale. Correspondingly, if $\hat{X}(W+1) = 0$ or 1, then we naturally have $\mathbb{E}\{\hat{X}(W)|\hat{X}(W+1)\} = \hat{X}(W+1)$ and the martingale equality holds. If, on the other hand, $\hat{X}(W+1) = j/(W+1)$ for some $j = 1, \ldots, W$, then we have

$$\mathbb{E}\{\hat{X}(W)|\hat{X}(W+1)\} = \frac{j-1}{W}\left[\frac{j/W - j/(W+1)}{1/W}\right] + \frac{j}{W}\left[\frac{j/(W+1) - (j-1)/W}{1/W}\right] = \frac{j}{W+1} = \hat{X}(W+1).$$

Therefore, $\{\hat{X}(W+1), \hat{X}(W)\}$ is indeed a martingale. □

LEMMA F3. *For any fixed $\theta \in [0,1]$, let $X(m) = \frac{Y^m(\theta)}{mW}$, where $Y^m(\theta) = \mathsf{Binomial}(mW, P^m(\theta))$ and $P^m(\theta) = \mathsf{Beta}(\theta\, mM, (1-\theta)\, mM)$. Then, it holds that $X(m+1) \leq_{cx} X(m)$ for all $m = 1, 2, \ldots$.*

Proof. The result follows by noting that for any convex function $\phi(\cdot) : \Re \to \Re$, we have

$$\begin{aligned}\mathbb{E}\Big\{\phi\Big(\frac{Y^m(\theta)}{mW}\Big)\Big\} &= \mathbb{E}\Big\{\phi\Big(\frac{\mathsf{Binomial}(mW, P^m(\theta))}{mW}\Big)\Big\} \geq \mathbb{E}\Big\{\phi\Big(\frac{\mathsf{Binomial}(mW, P^{m+1}(\theta))}{mW}\Big)\Big\} \\ &\geq \mathbb{E}\Big\{\phi\Big(\frac{\mathsf{Binomial}((m+1)W, P^{m+1}(\theta))}{(m+1)W}\Big)\Big\} = \mathbb{E}\Big\{\phi\Big(\frac{Y^{m+1}(\theta)}{(m+1)W}\Big)\Big\}. \qquad \text{(F.1)}\end{aligned}$$

To see that the first inequality holds, for $p \in [0,1]$, we let $g(p) = \mathbb{E}\left\{\phi\left(\frac{Y^m(\theta)}{mW}\right) \mid P^m(\theta) = p\right\}$ so that the expression on the left side of the first inequality can be written as $\mathbb{E}\{g(P^m(\theta))\}$. Therefore, the first inequality can be written as $\mathbb{E}\{g(P^m(\theta))\} \geq \mathbb{E}\{g(P^{m+1}(\theta))\}$. To see that this last inequality holds, conditional on $P^m(\theta) = p$, $Y^m(\theta)$ has a binomial distribution with parameters (mW, p). Due to stochastic convexity of the binomial family $\{\mathsf{Binomial}(mW,p) : p \in [0,1]\}$, we get that $g(\cdot)$ is also a convex function. In this case, the first inequality in (F.1) follows from the monotone convergence property of the beta operator under convex functions shown by Adell et al. (1996). The second inequality in (F.1) follows by conditioning on $P^{m+1}(\theta)$ and successively applying Lemma F2 above W times. □

Proof of Proposition 6. The proposition states that if $\boldsymbol{\theta}_t \in \mathcal{C}_t(m+1)$, then we have $\boldsymbol{\theta}_t \in \mathcal{C}_t(m)$. Noting the definition of $\mathcal{C}_t(m)$, it suffices to show that

$$\frac{1}{m} V_t(\boldsymbol{\theta}_t \,|\, m) \geq \frac{1}{m+1} V_t(\boldsymbol{\theta}_t \,|\, m+1)$$

for all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$. We show this inequality by using induction over the time periods. The result trivially holds at time period $\tau+1$. Assuming that the result holds at time period $t+1$, we have

$$\begin{aligned}\frac{1}{m}\,\mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{mW_t}\,\boldsymbol{Y}_t^m(\boldsymbol{\theta}_t) \,|\, m)\} &\geq \frac{1}{m+1}\,\mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{mW_t}\,\boldsymbol{Y}_t^m(\boldsymbol{\theta}_t) \,|\, m+1)\} \\ &\geq \frac{1}{m+1}\,\mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{(m+1)W_t}\,\boldsymbol{Y}_t^{m+1}(\boldsymbol{\theta}_t) \,|\, m+1)\}. \qquad \text{(F.2)}\end{aligned}$$

The first inequality above follows from the induction assumption. To see that the second inequality holds, we note that $V_{t+1}(\cdot \,|\, m+1)$ is a componentwise convex function by Proposition 1, in which case, we can apply Lemma F3 component by component n times. To obtain the desired result, we observe that

$$\begin{aligned}\frac{1}{m}\,V_t(\boldsymbol{\theta}_t \,|\, m) &= \max\Big\{(r - K\boldsymbol{c}^\intercal\boldsymbol{\theta}_t)D_t + \frac{1}{m}\,\mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{mW_t}\,\boldsymbol{Y}_t^m(\boldsymbol{\theta}_t) \,|\, m)\}, 0\Big\} \\ &\geq \max\Big\{(r - K\boldsymbol{c}^\intercal\boldsymbol{\theta}_t)D_t + \frac{1}{m+1}\,\mathbb{E}\{V_{t+1}(\lambda_t\,\boldsymbol{\theta}_t + \tfrac{1-\lambda_t}{(m+1)W_t}\,\boldsymbol{Y}_t^{m+1}(\boldsymbol{\theta}_t) \,|\, m+1)\}, 0\Big\} \\ &= \frac{1}{m+1} V_t(\boldsymbol{\theta}_t \,|\, m+1),\end{aligned}$$

where the two equalities follow from the optimality equation in (2) and the inequality uses (F.2). □

Appendix G: Proof of Proposition 7

In this section, we give a proof for Proposition 7, showing that $V_t(\boldsymbol{\theta}_t \,|\, m)$ deviates from $V_t^L(\boldsymbol{\theta}_t \,|\, m)$ by a term that grows in the order of $\sqrt{m}$.

Proof. Since Proposition 5 shows the first inequality, we only focus on the second inequality. The proof uses induction over the time periods to show that $V_t(\boldsymbol{\theta}_t \,|\, m) \leq V_t^L(\boldsymbol{\theta}_t \,|\, m) + G_t(\boldsymbol{\theta}_t)\sqrt{m}$, where $G_t(\boldsymbol{\theta}_t)$ is a componentwise concave function of $\boldsymbol{\theta}_t$, which does not depend on m. The result trivially holds at time period $\tau+1$. Assuming that the result holds at time period $t+1$, we define the random variable $\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)$ as $\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t) = \lambda_t\,\boldsymbol{\theta}_t + \frac{1-\lambda_t}{mW_t}\,\boldsymbol{Y}_t^m(\boldsymbol{\theta}_t)$. In this case, we obtain

$$\begin{aligned}
\mathbb{E}\{V_{t+1}(\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t) \,|\, m)\} &\leq \mathbb{E}\{V_{t+1}^L(\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t) \,|\, m)\} + \mathbb{E}\{G_{t+1}(\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t))\}\,\sqrt{m} \\
&\leq \mathbb{E}\{V_{t+1}^L(\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t) \,|\, m)\} + G_{t+1}(\mathbb{E}\{\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)\})\,\sqrt{m} \\
&= \mathbb{E}\{[r - K\boldsymbol{c}^\intercal\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)]^+\} \sum_{s=t+1}^{\tau} mD_s + G_{t+1}(\boldsymbol{\theta}_t)\,\sqrt{m}, \qquad \text{(G.1)}
\end{aligned}$$

where the first inequality follows from the induction assumption, the second inequality follows by noting that $G_{t+1}(\cdot)$ is a componentwise concave function and using the general form of Jensen's inequality for componentwise convex functions that we derive in Appendix C and the last equality follows by using the closed form expression for $V_{t+1}^L(\cdot \,|\, m)$ given in (6) and noting that $\mathbb{E}\{\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)\} = \boldsymbol{\theta}_t$.

We proceed to bound the expectation $\mathbb{E}\{[r - K\boldsymbol{c}^\intercal\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)]^+\}$ on the right side of (G.1). The variance of the random variable $K\boldsymbol{c}^\intercal\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)$ can be bounded by noting that

$$\begin{aligned}
\mathrm{Var}(K\boldsymbol{c}^\intercal\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)) &= K^2 \sum_{i=1}^n c_i^2\,\mathrm{Var}(\boldsymbol{\Theta}_{i,t+1}^m(\theta_{it})) = K^2\,\frac{(1-\lambda_t)^2}{m^2W_t^2}\sum_{i=1}^n c_i^2\,\mathrm{Var}(Y_{it}^m(\theta_{it})) \qquad \text{(G.2)} \\
&= K^2\,\frac{(1-\lambda_t)^2}{m^2W_t^2}\sum_{i=1}^n c_i^2\,\frac{mW_t\,(mM_t\,\theta_{it})\,mM_t\,(1-\theta_{it})(mM_t+mW_t)}{(mM_t)^2\,(mM_t+1)} \\
&\leq K^2\,\frac{(1-\lambda_t)^2(M_t+W_t)}{mW_t\,M_t}\sum_{i=1}^n c_i^2\,\theta_{it}\,(1-\theta_{it}),
\end{aligned}$$

where the first equality follows because components of $\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)$ are independent, the third equality uses the fact that $Y_{it}^m(\theta_{it})$ is a beta binomial random variable with parameters $(mW_t, P_{it}^m(\theta_{it}))$ and $P_{it}^m(\theta_{it})$ is a beta random variable with parameters $(\theta_{it}\,mM_t, (1-\theta_{it})\,mM_t)$, and the inequality follows simply by noting that $mM_t + 1 \geq mM_t$. For a deterministic scalar z and a real valued random variable Z with finite mean μ and finite variance σ^2, Gallego (1992) shows that $\mathbb{E}\{[z-Z]^+\} \leq [\sqrt{\sigma^2 + (z-\mu)^2} + (z-\mu)]/2 \leq [z-\mu]^+ + \sigma/2$. Thus, we can bound the expectation $\mathbb{E}\{[r - K\boldsymbol{c}^\intercal\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)]^+\}$ by

$$\begin{aligned}
\mathbb{E}\{[r - K\boldsymbol{c}^\intercal\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t)]^+\} &\leq [r - K\boldsymbol{c}^\intercal\boldsymbol{\theta}_t]^+ + \frac{1}{2}\sqrt{K^2\,\frac{(1-\lambda_t)^2(M_t+W_t)}{mW_t\,M_t}\sum_{i=1}^n c_i^2\,\theta_{it}\,(1-\theta_{it})} \\
&= [r - K\boldsymbol{c}^\intercal\boldsymbol{\theta}_t]^+ + \frac{L_t(\boldsymbol{\theta}_t)}{\sqrt{m}},
\end{aligned}$$

where we let $L_t(\boldsymbol{\theta}_t) = \frac{K\,(1-\lambda_t)}{2}\sqrt{\frac{(M_t+W_t)}{W_tM_t}\sum_{i=1}^n c_i^2\,\theta_{it}\,(1-\theta_{it})}$. In this case, we can continue the chain of inequalities in (G.1) as

$$\mathbb{E}\{V_{t+1}(\boldsymbol{\Theta}_{t+1}^m(\boldsymbol{\theta}_t) \,|\, m)\} \leq \Big[[r - K\,\boldsymbol{c}^\intercal\boldsymbol{\theta}_t]^+ + \frac{L_t(\boldsymbol{\theta}_t)}{\sqrt{m}}\Big] \sum_{s=t+1}^{\tau} mD_s + G_{t+1}(\boldsymbol{\theta}_t)\,\sqrt{m}$$

$$= [r - K\,\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \sum_{s=t+1}^{\tau} mD_s + \sqrt{m}\, L_t(\boldsymbol{\theta}_t) \sum_{s=t+1}^{\tau} D_s + G_{t+1}(\boldsymbol{\theta}_t)\sqrt{m}.$$
$$= [r - K\,\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \sum_{s=t+1}^{\tau} mD_s + G_t(\boldsymbol{\theta}_t)\sqrt{m},$$

where we let $G_t(\boldsymbol{\theta}_t) = L_t(\boldsymbol{\theta}_t) \sum_{s=t+1}^{\tau} D_s + G_{t+1}(\boldsymbol{\theta}_t)$. We note that $L_t(\cdot)$ is componentwise concave and since $G_{t+1}(\cdot)$ is componentwise concave by the induction assumption, it follows that $G_t(\cdot)$ is also componentwise concave. To finish the proof, we write the optimality equation in (2) as

$$\begin{aligned} V_t(\boldsymbol{\theta}_t \,|\, m) &= \max\Big\{(r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t)\, mD_t + \mathbb{E}\{V_{t+1}(\boldsymbol{\Theta}^m_{t+1}(\boldsymbol{\theta}_t) \,|\, m)\}, 0\Big\} \\ &\leq [r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ mD_t + \mathbb{E}\{V_{t+1}(\boldsymbol{\Theta}^m_{t+1}(\boldsymbol{\theta}_t) \,|\, m)\} \\ &\leq [r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ mD_t + [r - K\,\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \sum_{s=t+1}^{\tau} mD_s + G_t(\boldsymbol{\theta}_t)\sqrt{m} = V^L_t(\boldsymbol{\theta}_t \,|\, m) + G_t(\boldsymbol{\theta}_t)\sqrt{m}, \end{aligned}$$

where the first inequality follows by noting that $[r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t]^+ \geq r - K\boldsymbol{c}^\intercal \boldsymbol{\theta}_t$ and $V_{t+1}(\cdot \,|\, m) \geq 0$ and the second equality follows by noting (6). This completes the induction argument and we have $V_t(\boldsymbol{\theta}_t \,|\, m) \leq V^L_t(\boldsymbol{\theta}_t \,|\, m) + G_t(\boldsymbol{\theta}_t)\sqrt{m}$ for all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$.

To obtain the result in the statement of the proposition, we let $\bar{\frac{1}{2}} = (\frac{1}{2}, \ldots, \frac{1}{2}) \in \Re^n$ and note that $L_t(\boldsymbol{\theta}_t) \leq L_t(\bar{\frac{1}{2}})$ for all $t = 1, \ldots, \tau$ and $\boldsymbol{\theta}_t \in [0,1]^n$. In this case, we obtain $G_t(\boldsymbol{\theta}_t) \leq G_t(\bar{\frac{1}{2}})$ and letting $\bar{G}_t = G_t(\bar{\frac{1}{2}})$ in the statement of the proposition suffices. $\square$

Appendix H: Stopping Times for Different Benchmark Policies

In this section, we provide additional experimental results that give a comparison of when we stop selling the product under different benchmark policies. Our goal is to understand how the behavior of different benchmark policies is affected by various problem parameters and how each benchmark policy compares to others in terms of its stopping decisions. Throughout this section, for economy of space, we consider a subset of our test problems.

Table 3 shows the average stopping times of different benchmark policies, averaged over 500 sample paths. This table focuses on test problems with five failure types. Recalling that there are 12 time periods in the selling horizon, if a benchmark policy does not stop selling the product over the whole selling horizon, then we define its stopping time to be time period 13 by convention. The left portion of the table focuses on the test problems where the demand at the test marketing stage is always 50, whereas the right portion focuses on the test problems where the demand at the test marketing stage is equal to the demand at the subsequent time periods. In both portions of the table, the first column shows the characteristics of the test problems by using the tuple $(D_0, \bar{D}, n, CV)$. Each one of the remaining five columns shows the average stopping times for IDE, DBP, GRP, OSL and SFP. We note that the stopping decision for IDE is based on comparing r with $K\boldsymbol{c}^\intercal \boldsymbol{p}$. The problem parameters r, K and $\boldsymbol{c}$ are constants and the sampled values of the true failure probabilities $\boldsymbol{p}$ depend on CV. Therefore, the stopping times for IDE depend on CV, but they do not depend on $\bar{D}$.

The results in Table 3 indicate that the stopping times for DBP tend to be larger than those for GRP, which is in agreement with Propositions 4 and 5. In particular, the discussion that follows Proposition 4

Table 3 Average stopping times for the test problems with five failure types.

Test Problem				Average Stopping Times				
D_0	$\bar{D}$	n	CV	IDE	DBP	GRP	OSL	SFP
50	50	5	0.6	7.43	7.84	5.65	6.52	8.33
50	50	5	0.6^2	7.43	8.17	5.16	6.30	8.83
50	50	5	0.6^3	7.38	8.20	4.25	5.77	8.92
50	50	5	0.6^4	7.29	8.36	4.16	5.83	9.24
50	100	5	0.6	7.43	7.90	5.65	6.38	8.42
50	100	5	0.6^2	7.43	8.28	5.39	6.46	8.87
50	100	5	0.6^3	7.38	8.61	4.28	5.71	9.26
50	100	5	0.6^4	7.29	8.48	3.92	5.52	9.20
50	200	5	0.6	7.43	8.02	5.77	6.50	8.57
50	200	5	0.6^2	7.43	8.43	5.31	6.34	8.95
50	200	5	0.6^3	7.38	8.61	4.48	5.83	9.29
50	200	5	0.6^4	7.29	8.68	4.39	5.67	9.66
50	500	5	0.6	7.43	8.12	5.75	6.36	8.50
50	500	5	0.6^2	7.43	8.40	5.43	6.28	8.83
50	500	5	0.6^3	7.38	8.60	4.60	5.67	9.32
50	500	5	0.6^4	7.29	8.91	4.22	5.37	9.60
50	1000	5	0.6	7.43	7.98	5.84	6.35	8.52
50	1000	5	0.6^2	7.43	8.29	5.53	6.16	8.89
50	1000	5	0.6^3	7.38	8.64	4.45	5.29	9.36
50	1000	5	0.6^4	7.29	8.94	4.51	5.31	9.67

Test Problem				Average Stopping Times				
D_0	$\bar{D}$	n	CV	IDE	DBP	GRP	OSL	SFP
50	50	5	0.6	7.43	7.84	5.65	6.52	8.33
50	50	5	0.6^2	7.43	8.17	5.16	6.30	8.83
50	50	5	0.6^3	7.38	8.20	4.25	5.77	8.92
50	50	5	0.6^4	7.29	8.36	4.16	5.83	9.24
100	100	5	0.6	7.43	8.01	6.14	6.85	8.38
100	100	5	0.6^2	7.43	8.44	5.48	6.50	8.98
100	100	5	0.6^3	7.38	8.47	4.59	6.06	9.15
100	100	5	0.6^4	7.29	8.70	4.44	5.94	9.54
200	200	5	0.6	7.43	8.33	6.33	6.93	8.26
200	200	5	0.6^2	7.43	8.51	6.08	6.90	8.81
200	200	5	0.6^3	7.38	8.91	4.97	6.30	9.39
200	200	5	0.6^4	7.29	9.03	4.67	6.08	9.66
500	500	5	0.6	7.43	7.97	6.56	7.05	8.21
500	500	5	0.6^2	7.43	8.43	6.38	7.00	8.70
500	500	5	0.6^3	7.38	8.87	5.71	6.75	9.30
500	500	5	0.6^4	7.29	9.19	5.25	6.49	9.64
1000	1000	5	0.6	7.43	7.96	6.72	7.16	8.11
1000	1000	5	0.6^2	7.43	8.32	6.88	7.26	8.59
1000	1000	5	0.6^3	7.38	8.77	5.99	6.82	9.11
1000	1000	5	0.6^4	7.29	9.22	5.67	6.66	9.66

indicates that DBP is more likely to continue selling the product when compared with the optimal policy, whereas the discussion that follows Proposition 5 indicates that GRP is more likely to stop selling the product. The stopping times for OSL are consistently larger than those for GRP. GRP ignores the benefits from future learning altogether. In contrast, by applying a one step look ahead on the decision rule used by GRP, OSL tries to take the benefits from future learning into consideration, spending more time to learn the failure probabilities, which results in larger stopping times. The stopping times for SFP are generally larger than those for the other benchmark policies, but larger amount of time spent on learning the failure probabilities does not necessarily translate into larger expected profits, as indicated by the inferior expected profit performance of SFP for the test problems with five failure types.

Our findings indicate that GRP tends to stop selling the product early. To make up for this shortcoming of GRP, we check the performance of an augmented version of this benchmark policy, where we continue selling the product for at least T time periods irrespective of our belief about the failure probabilities and switch to the decision rule of GRP after these T time periods. For each test problem, we try every possible value of T and report the results corresponding to the best value. Table 4 gives our findings. To conserve space, we focus on the test problems where the demand in the test marketing stage is always 50. For the other test problems, the augmented version of GRP does not improve the original version noticeably. The left and right portions of the table respectively give the results for the test problems with two and five failure types. In each portion, the first column shows the characteristics of the test problems by using the tuple $(D_0, \bar{D}, n, CV)$ and the second column shows the percent gaps between the expected profits obtained by DBP and the augmented version of GRP. Comparing the results in Table 4 with those in the third column of Table 1 indicates that the augmented version of GRP significantly improves the earlier version of GRP. Nevertheless, the augmented version of GRP generally lags behind DBP unless the demand quantities are large and it is possible to learn the failure probabilities quickly.

Table 4 Performance of GRP when we continue selling the product for at least T time periods.

Test Problem				DBP vs.
D_0	$\bar{D}$	n	CV	GRP
50	50	2	0.6	2.70%
50	50	2	0.6^2	7.61%
50	50	2	0.6^3	7.61%
50	50	2	0.6^4	10.61%
50	100	2	0.6	3.08%
50	100	2	0.6^2	7.82%
50	100	2	0.6^3	9.85%
50	100	2	0.6^4	14.21%
50	200	2	0.6	3.75%
50	200	2	0.6^2	6.54%
50	200	2	0.6^3	5.14%
50	200	2	0.6^4	6.48%
50	500	2	0.6	3.72%
50	500	2	0.6^2	4.68%
50	500	2	0.6^3	1.06%
50	500	2	0.6^4	0.12%
50	1000	2	0.6	3.81%
50	1000	2	0.6^2	4.59%
50	1000	2	0.6^3	-1.96%
50	1000	2	0.6^4	-1.90%
Average				4.98%

Test Problem				DBP vs.
D_0	$\bar{D}$	n	CV	GRP
50	50	5	0.6	8.87%
50	50	5	0.6^2	12.86%
50	50	5	0.6^3	18.11%
50	50	5	0.6^4	37.06%
50	100	5	0.6	8.51%
50	100	5	0.6^2	6.44%
50	100	5	0.6^3	12.80%
50	100	5	0.6^4	31.23%
50	200	5	0.6	6.50%
50	200	5	0.6^2	6.66%
50	200	5	0.6^3	17.41%
50	200	5	0.6^4	10.44%
50	500	5	0.6	5.34%
50	500	5	0.6^2	3.24%
50	500	5	0.6^3	0.69%
50	500	5	0.6^4	6.35%
50	1000	5	0.6	4.27%
50	1000	5	0.6^2	1.20%
50	1000	5	0.6^3	-0.64%
50	1000	5	0.6^4	-4.05%
Average				9.66%

Appendix I: Possible Extensions

In this section, we discuss possible extensions of our model and point out which results in the paper continue to hold under these extensions.

Substitute Product. Our model assumes that there is no other revenue stream for the company once we stop selling the product. We can extend our model to deal with the case where there is a standard substitute product that brings a known expected net profit contribution of r_0 per sold unit and stopping to sell the current product means switching to this substitute product. To incorporate this extension, all we need to do is to replace r with $r - r_0$ in the optimality equation in (2). In this case, all of our results in the paper go through without any modifications.

Alternative Warranty Coverage. In this paper, we assume that the warranty coverage is for K time periods, but it is possible to work with other forms of warranty coverage. For example, units may be covered until the end of the selling horizon irrespective of when they were sold or the warranty duration may depend on when the unit was actually sold to a customer. The key observation is that no matter which form of warranty coverage we use, since the demand quantities at different time periods are deterministic, it is simple to compute the number of units that are under warranty coverage at any time period. All of our results hold as long as we can compute the number of units that are under warranty coverage at any time period.

Alternative Costs. The cost structure in our model assumes that the cost of repairing a unit that fails from a particular failure type is constant, irrespective of how long the unit has been with the customer. It is not difficult to work with repair costs that depend on how long the customer has used the unit. In particular, if we use c_{ik} to denote the repair cost of a unit that fails from failure type i after having been with the customer for k time periods, then a close inspection of the dynamic programming formulation in Section 3 shows that we can capture such age dependent repair costs by replacing $K\boldsymbol{c}^\intercal\boldsymbol{\theta}_t$ in the optimality equation

in (2) with $\sum_{i=1}^{n} \sum_{k=0}^{K-1} c_{ik}\, \theta_{it}$. All of our results continue to hold under this modification. More generally, all of our results go through when we replace $K \boldsymbol{c}^{\top} \boldsymbol{\theta}_t$ in the optimality equation in (2) with a separable concave increasing function, allowing us to model possible nonlinear costs associated with product failures.

Dirichlet Multinomial Learning. By using the Dirichlet multinomial learning model, we can capture the situation where each product can fail from at most one failure at a time. Under the Dirichlet multinomial learning model, the learning dynamics is similar to that under the beta binomial model that we use throughout the paper. In particular, at time period t, the learning dynamics can still be summarized by $\boldsymbol{\theta}_{t+1} = \lambda_t\, \boldsymbol{\theta}_t + \frac{1-\lambda_t}{W_t}\, \boldsymbol{Y}_t(\boldsymbol{\theta}_t)$, but the difference is that $\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$ has to be a Dirichlet multinomial random variable instead of a random vector consisting of independent beta binomial random variables. With this new interpretation of the learning dynamics, we reach the same optimality equation as in (2). It is worthwhile to observe that if each product can fail from at most one failure at a time, then we end up introducing correlations in our beliefs about the probabilities of different failure types.

Under the Dirichlet multinomial learning model, the structural properties in Propositions 1, 2 and 3 continue to hold. In particular, the value functions are componentwise decreasing, an optimal stopping boundary exists and the optimality of continuing decision is preserved at the next time period when there are no failures at the current time period. The proof of Proposition 4 does not necessarily hold anymore, since the components of $\boldsymbol{Y}_t(\boldsymbol{\theta}_t)$ are not independent. We need more advanced analysis in multi dimensional stochastic orders to show an analogue of the upper bound property given in Proposition 4. The lower bound property given in Proposition 5 holds, which implies that we still obtain a lower bound from the deterministic approximation. The proof of Proposition 6 uses properties of beta binomial random variables and it does not necessarily hold under the Dirichlet multinomial learning model. Proposition 7 still holds so that the lower bound is asymptotically tight when we scale up the demand at each time period.

Age Dependent Failures. In our model, we assume that a product fails from each failure type with a fixed probability. In particular, the failure probabilities do not depend on how long ago the product was sold to a customer or how long ago the product was last repaired. This assumption is reasonable when the failure types are mostly electrical, rather than wear and tear related. It is possible to extend our model to address the case where the probability of failure from a certain failure type depends on when the product was sold to a customer. To make this extension, we use p_{ik} to denote the true probability that a product that was sold to a customer k time periods ago fails from failure type i. Since the warranty duration is K time periods, we do not need to know the failure probabilities for the products that were sold more than K time periods ago. Therefore, we need to learn nK parameters. Furthermore, since the demand quantities are deterministic, it is straightforward to compute the number of products that were sold to a customer a certain number of time periods ago. In this case, we can use essentially the same learning dynamics that we use in the paper to update our beliefs about the nK unknown failure probabilities.

This discussion indicates that it is not difficult to deal with failure probabilities that depend on how long ago the product was sold to a customer. In contrast, we need to augment our modeling approach when the

failure probabilities depend on how long ago the product was last repaired. Although the demand quantities are deterministic, the product failures are uncertain and there is no straightforward way to compute the number of products that were repaired a certain number of time periods ago. This complication requires us to extend the state variable of our dynamic programming formulation to keep track of the number of products that were repaired a certain number of time periods ago. The dynamic programming formulation becomes significantly more difficult to analyze with this extra component in the state variable.