OPERATIONS RESEARCH
Vol. 60, No. 5, September–October 2012, pp. 1267–1281
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1090

Performance Guarantees for Empirical Markov Decision Processes with Applications to Multiperiod Inventory Models

William L. Cooper
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455, billcoop@umn.edu

Bharath Rangarajan
Merchandising Operations, Target Corporation, Minneapolis, Minnesota 55402, bharath.rangarajan@target.com

We consider Markov decision processes with unknown transition probabilities and unknown single-period expected cost functions, and we study a method for estimating these quantities from historical or simulated data. The method requires knowledge of the system equations that govern state transitions as well as the single-period cost functions (but not the single-period expected cost functions). The estimation procedure is based upon taking expectations with respect to the empirical distribution functions of such data. Once the estimates are in place, the method computes a policy by solving the obtained "empirical" Markov decision process as if the estimates were correct. For MDPs that satisfy some conditions, we provide explicit, easily computed expressions for the probability that the procedure will produce a policy whose true expected cost is within any specified absolute distance of the actual optimal expected cost of the true Markov decision process. We also provide expressions for the number of historical or simulated data values that is sufficient for the procedure to produce a policy whose true expected cost is, with a prescribed probability, within a prescribed absolute distance of the actual optimal expected cost of the true Markov decision process. We apply our results to multiperiod inventory models. In addition, we provide a specialized analysis of such inventory models that also yields relative, rather than absolute, accuracy guarantees. We make comparisons with related results that have recently appeared, and we provide numerical examples.

Subject classifications: dynamic programming/optimal control: Markov; inventory/production; statistics; estimation.
Area of review: Stochastic Models.
History: Received September 2009; revisions received March 2011, July 2011; accepted November 2011. Published online in *Articles in Advance* October 9, 2012.

1. Introduction

In this article, we focus on Markov decision processes (MDPs) in which transition probabilities and single-period expected costs are not known, but can be estimated from data obtained from historical records or simulations. We consider finite-horizon MDPs in which state transitions are generated by system equations that map a current state, action, and realized value of an exogenous random variable to a state in the next time period. The exogenous random variables are sometimes referred to as "disturbances" in the literature (see, e.g., Puterman 1994, p. 51 or Bertsekas 2000, p. 13). Disturbances often have a natural interpretation for the particular system in question. For example, in inventory models that we consider, the disturbances are demands.

We analyze a simple, straightforward method that uses the system equations together with empirical distributions of historical or simulated disturbance data to estimate the unknown transition probabilities. To apply the method, one must know the system equations as well as the single-period cost functions (but not the single-period expected cost functions). The single-period expected costs are estimated by taking expectations with respect to these empirical distributions. Once the estimates are in place, we compute a policy by solving the obtained "empirical" Markov decision process as if the estimates were correct. This is sometimes called a model-based approach in the machine learning literature; see, e.g., Kearns and Singh (2002) or Strehl and Littman (2005). The solution method is a direct application of the usual backward induction algorithm applied to a problem with empirical estimates of transition probabilities and single-period expected costs. As such, it does not require new computational approaches and can employ the same computer programs used to solve an MDP in which all parameters are known. The only additional effort is in computing the empirical distribution functions from the disturbance data and, in the case of simulations, generating the data. The method can also be viewed as successive application of the sample average approximation method, which is reviewed in Kleywegt et al. (2001). The successive application of the sample average approximation method in stochastic programs is analyzed in Shapiro (2006) and Shapiro et al. (2009).

Given $\epsilon > 0$ and $\delta > 0$, for MDPs that satisfy some conditions (in particular, Lipschitz conditions on cost and value functions and bounded support of disturbance distributions), we provide easily computable expressions (which depend upon the Lipschitz constants) for the number of historical or simulated data values that are sufficient for the empirical procedure to yield, with probability at least $1 - \delta$, a policy for which the true expected cost-to-go from any state x at any time t is within ϵ of the actual optimal expected cost-to-go from state x at time t. In terminology common in the machine learning literature, this implies that the empirical procedure is "probably approximately correct" (PAC). For a given number of data values and given $\epsilon > 0$, we also provide a lower bound on the probability that the empirical procedure returns an ϵ-optimal policy for the true MDP. The analysis yields estimates and confidence intervals for the value function of the MDP as well. These results provide *absolute* performance guarantees, because they involve (probabilistic) assurances about the absolute difference in expected cost between the derived policy and an optimal policy. Shapiro (2006) and Shapiro et al. (2009) obtain similar absolute performance guarantees under different assumptions. Detailed discussion of their work and how it relates to ours is included later in this paper (§4) and in the online supplement. An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1090.

Our study was initially motivated by the work of Levi et al. (2007) on multiperiod inventory models with unknown demand distributions. The problem they consider is a particular example of an MDP with an unknown disturbance distribution. They devise a procedure for computing base-stock levels (which specify a policy for the MDP) from historical or simulated demand data, and they derive *relative* guarantees on performance. Such a relative guarantee specifies, for given $\epsilon > 0$ and $\delta > 0$, the number of data values so that with probability at least $1 - \delta$, the true expected cost of the derived policy is no more than $1 + \epsilon$ times the expected cost of an optimal policy. The method of Levi et al. is based on a recursive dynamic programming approach that is specifically devised for the inventory problem. At each recursive step, their algorithm and sampling procedure are designed to maintain, with high probability, certain convexity properties in what they refer to as a shadow dynamic program. The empirical approach we study does not involve such shadow dynamic programs, and instead directly solves an MDP that has empirical estimates for the expected costs and transition probabilities.

The empirical method that we analyze is not tailored to a specific problem and can be applied to a large class of MDPs. We present a specialized analysis that provides relative performance guarantees analogous to results in Levi et al. for inventory problems that include as special cases backorder as well as lost-sales models. With the additional assumption that the demand distributions have bounded support, direct application of our results for general MDPs yields an absolute performance guarantee for such inventory problems. (Here and throughout, "general" is not used to indicate a lack of assumptions, but rather to emphasize that the theory is developed for MDPs that are not necessarily inventory models. As stated above, our results rely on some assumptions regarding the underlying MDP. These assumptions are rigorously stated later.) Bounded support is not needed to apply the empirical method, but is used to obtain the absolute performance guarantee (but not the relative guarantee). As indicated earlier, Shapiro (2006) and Shapiro et al. (2009) derive absolute guarantees; they do not obtain relative guarantees.

The empirical approach we study involves application of the usual backward induction algorithm with empirical estimates in place of exact quantities. Our main results apply to MDPs with uncountable state and action spaces (that satisfy some conditions as previously mentioned). For such problems a computational difficulty may arise because the backward induction approach generally requires doing certain computations for each state and action at each time step. Absent additional structural properties of the MDP, one may need to discretize or truncate the empirical MDP to get an implementable computational approach. This issue is not unique to the empirical approach; indeed, it arises for MDPs with infinite state and/or actions spaces even when disturbance distributions are known. For inventory models with uncountable state and action spaces we provide a computational approach that solves the empirical MDP without any truncation or discretization. Even for problems with large but finite state and action spaces, the approach we study does not address the curses of dimensionality as described by, e.g., Powell (2007). The approach is intended for situations when disturbance distributions are unknown or are too complicated to work with.

The proofs of our main results (for both general MDPs and inventory problems) proceed roughly as follows. We begin by considering two separate MDPs that differ only in their disturbance distributions, and we bound the distance between the value functions of the MDPs in terms of the distance between the disturbance distributions. We also bound the expected cost obtained when one computes an optimal policy for one of the MDPs, and uses it to control the other MDP. These results are potentially of independent interest and hold outside the "empirical" context. They can be viewed as measuring the sensitivity of MDPs to changes in transition probabilities and single-period costs. (Related results have appeared in, e.g., Kearns and Singh 2002 and Strehl and Littman 2005.) To apply these sensitivity results in the analysis of the empirical MDP algorithm, we use a result of Massart (1990) to bound the probability that empirical disturbance distributions will be within a specified distance of the true disturbance distributions. Finally, this allows us to bound the probability that the solution of the empirical MDP will be "close" to the solution of the true MDP.

We also describe results of a numerical study in the multiperiod inventory setting. The study shows that even when

little data is available as input to the empirical MDP procedure, the procedure nevertheless typically returns a policy that is close to optimal. For instance, when demand is Poisson and there are 20 demand records available from each period in the time horizon, the procedure returned a policy whose true expected cost was, on average, within 6.5% above optimal. When 100 records were available from each period, the true average cost of the returned policy was found to be roughly 1% above optimal. These results are of significance when using the empirical method with historical records as input, because data may be scarce in such situations. With more-variable demand distributions, we also obtained similar results. It is important to note that the empirical procedure is nonparametric in the sense that it makes no assumption on the form of the marginal demand distributions. For example, in a problem where demand is in fact Poisson, the empirical procedure does not require the decision maker (the inventory manager) to know that demand is Poisson, nor does the method make any use of the fact that demand is Poisson. Consequently, the method is not susceptible to the types of specification errors described by Gallego et al. (2007), who show that, e.g., assuming demand to be normally distributed, when in fact it follows a gamma distribution (but with matching means and variances), can lead to very poor results. We also conducted numerical experiments in which data was used to estimate parameters of correctly and incorrectly specified parametric models. As one would expect, such parametric approaches perform quite well when correctly specified. However, the performance of incorrectly specified parametric models depends upon the particulars of the problem. In some settings, misspecified parametric models perform surprisingly well. In others, the performance was quite poor. We also found that an increase in the number of samples used to estimate parameters of an incorrectly specified model can lead to a decline in the quality of the obtained policies.

Levi et al. (2007) review the literature on inventory models with unknown demand distributions. Perakis and Roels (2008) consider an inventory model with limited information about the demand distribution, and consider robust approaches that do not use demand records as input. They also provide many references. There is also a literature on sampling-based methods for MDPs. Much of this work is reviewed in Chang et al. (2005, 2007b) and in the books by Chang et al. (2007a), Bertsekas and Tsitsiklis (1996), and Powell (2007), and the dissertation by Kakade (2003).

We next discuss work that, like ours, uses sampling to estimate parameters of an MDP, and then solves the estimated MDP to obtain a policy. As mentioned above, this is sometimes called a model-based approach. Examples of papers that use a model-based approach include Fiechter (1994), Kearns et al. (2002), Kearns and Singh (2002), Pivazyan and Shoham (2002), and Strehl and Littman (2005). Even-Dar et al. (2006) obtain PAC results for multiarmed bandits and apply the results to model-based learning for MDPs. They also propose a variation on Q-learning, which is not a model-based method. These papers do not allow infinite action spaces. (Our motivating examples of multiperiod inventory problems have infinite state and action spaces. Our results apply to such problems, whereas results in these papers do not.) Much of the prior work on model-based approaches uses some variation on the estimation scheme whereby the probability of moving from, say, state x to state x' under action q is estimated as the number of transitions to x' from x when action q was taken divided by the number of times action q was taken in state x. Such an approach can be used directly, or it can be used to build confidence intervals for transition probabilities. In either case, when the action space is infinite, there will generally be infinitely many actions that have never been tried, and hence the approach runs into problems. Moreover, the sample sizes required for PAC results in papers cited above are finite only when the action space is finite, underscoring the importance of the finiteness of the action spaces to the results in those papers. Our method of estimating the disturbance distribution avoids this difficulty. Unlike the approach we study, the methods in these papers do not require knowledge of system equations.

Mannor et al. (2007) use a model-based approach and estimate costs and transition probabilities for an MDP using data. They analyze the bias and variance of resulting value function estimates for fixed policies for infinite-horizon discounted cost MDPs. Their estimation procedure uses historical records of state transitions under different actions, and does not rely on disturbances or system equations. The main focus of the paper is estimation. It also describes how bias in value function estimates can be induced by policy optimization, and explains how this problem can be remedied through the use of separate validation samples.

We now turn to research on non-model-based approaches. Some of this work simulates trajectories and chooses the policy (from a given family of policies) with the best empirical performance on the trajectories. Examples of papers that explore methods based upon simulating trajectories include Ng and Jordan (2000), Kearns et al. (2000), Jain and Varaiya (2006), and Bartlett and Tewari (2007). These papers use some notion of the complexity of the family of policies under consideration as measured by the VC dimension or pseudodimension. These concepts are reviewed in, e.g., Anthony and Bartlett (1999). Kearns et al. (2000) use a trajectory tree approach and require a finite action space. Antos et al. (2008) work with a single trajectory of a fixed policy as input to a procedure that combines policy iteration with value function approximation. The value functions are estimated empirically from a chosen parametric family of functions. They obtain PAC-type results for infinite-horizon discounted-cost MDPs with continuous state spaces and finite action spaces. Murphy (2005) considers finite-horizon problems with finite action spaces and proposes a variant of Q-learning in which the Q functions are empirically estimated using functions from a chosen parametric family, and derives PAC

results. The inventory problems we consider have convex piecewise-linear value functions, so one might try to adopt Murphy's approach to such problems by taking the parametric family to be the set of convex, piecewise-linear functions.

Ng and Jordan (2000), Jain and Varaiya (2006), and Bartlett and Tewari (2007) use a deterministic simulative model (analogous to the system equations in our paper), which enables them to obtain results for infinite action spaces. As pointed out by Ng and Jordan (2000), the deterministic simulative model can be thought of as a method that allows parsimonious representation of a trajectory tree. The focus in these papers is obtaining policies for time-homogeneous infinite-horizon discounted-cost problems. However, it is apparent that the methods and results, when suitably modified, can be used for finite-horizon problems with nonhomogeneous costs and transition probabilities. It is important to note, however, that finite-horizon problems typically do not have stationary optimal policies. This makes it potentially difficult to perform the optimization over multiple dimensions needed to apply the methods in these papers to finite-horizon problems. An advantage of the approaches in these papers is that they do not involve storage or computation of value function estimates for all (state, time)-pairs. For MDPs that satisfy some conditions, given an arbitrary distribution of the initial state and given $\epsilon > 0$ and $\delta > 0$, these articles provide expressions for the number of simulated trajectories that is sufficient to ensure that the obtained policy has an expected cost-to-go at the beginning of the time horizon that is, with probability at least $1-\delta$, within ϵ of the actual optimal expected cost-to-go at the beginning of the time horizon. Such results are similar to the absolute guarantees that we obtain. One difference is that the guarantees in these papers apply to the expected cost-to-go only for the given distribution of the initial state (in particular, they hold for a fixed initial state by taking the initial distribution to be a point mass). The results stated in these papers do not rule out the possibility that there may be (state, time)-pairs (including time $t = 1$) for which the derived policy will be further than ϵ from optimal, with probability more than δ. When starting from the given initial distribution, it must be unlikely to reach such (state, time)-pairs. However, conditional upon reaching such a (state, time)-pair, results in these papers do not guarantee that the policy will thereafter perform well. This issue does not arise in our approach and results, which provide a performance guarantee simultaneously across all times and states. One possible way to deal with the problem of reaching such (state, time)-pairs is to resample and reoptimize after each state transition.

Chang et al. (2005, 2007b) describe sampling-based methods for finite-horizon MDPs. The methods are based upon simulating state transitions using what are called sampled trees. Chang et al. (2007b) consider a method called the recursive automata sampling algorithm (RASA). For problems with finite action spaces, bounded single-period rewards, and unique optimal policies, they provide (a) a lower bound for the probability that RASA samples an optimal action at the beginning of the horizon, and (b) an upper bound on the probability that the difference between the RASA value function estimate and the true value function exceeds a specified error. Chang et al. (2005) consider a method called adaptive multistage sampling (AMS). For MDPs with finite state and action spaces, they prove that AMS produces asymptotically unbiased value function estimates, and they find the rate at which the bias converges to zero. Kearns et al. (2002) consider tree-based methods, but with fixed rather than adaptive sampling procedures. For problems with finite action spaces and bounded single-period rewards, they prove that the value of the randomized policy associated with their method can be made arbitrarily close to the true optimal value, and they provide expressions, independent of the size of the state space, for how many samples are needed to achieve such a guarantee. The articles by Chang et al. (2005, 2007b) and Kearns et al. (2002) use sampling and trees to approximately solve MDPs with large (or infinite) state spaces, bounded single-period costs, and finite action spaces. Our analysis, which is centered on using empirical distribution functions of the disturbances and which exploits (and relies upon) the known form of the state equation, does not involve trees, and applies to many problems with infinite state and action spaces and unbounded costs.

In summary, the main contribution of our paper is the analysis of an empirical approach to finite-horizon Markov decision processes that is applicable when disturbance distributions are not known, but can be estimated from historical or simulated data. More specifically, the contributions are the following. (1) We derive absolute performance guarantees for the empirical approach for MDPs that satisfy some conditions and apply the results to multiperiod inventory problems. (2) We provide an analysis for multiperiod inventory problems that yields relative performance guarantees. With the exception of Levi et al. (2007), which develops a specialized approach for inventory problems, none of the papers above derive relative performance guarantees (for inventory problems or general MDPs). In contrast to the work of Levi et al., our method is not specifically tailored to inventory problems. (3) For our proofs, we develop results on the sensitivity of MDPs to perturbations of their disturbance distributions. (4) We describe a numerical study that shows the empirical approach to be effective in the inventory setting, even when there is little data available to estimate the demand distributions.

The remainder of this paper is organized as follows: §2 provides some background on MDPs; §3 describes the empirical MDP approach; §4 provides our absolute performance guarantee for finite-horizon MDPs; §5 applies the result to multiperiod inventory problems and also specializes the analysis to obtain relative performance guarantees for inventory problems; §6 contains the numerical study; §7 contains some concluding comments. Sections S-1 and S-2 of the online supplement contain proofs and supporting material, including the sensitivity results for

MDPs. Section S-3 of the online supplement contains material related to the work of Shapiro (2006) and Shapiro et al. (2009). Section S-4 of the online supplement describes a finite computational approach that does not require truncation or discretization of the empirical MDP in the inventory setting. Section S-5 of the online supplement contains a numerical study of parametric estimation for inventory MDPs.

2. Markov Decision Processes

In this section, we describe the Markov decision processes under consideration. See, e.g., Puterman (1994) or Bertsekas (2000), for a complete treatment of MDPs. Let $\mathbb{R}$ denote the real numbers and $\mathbb{R}_+$ be the nonnegative real numbers. Suppose there is a finite number of time periods, indexed $t = 1, \ldots, \tau$ where $\tau \geqslant 1$ is the final period. The state space is $\mathscr{X} \subseteq \mathbb{R}$, and the action space is $\mathscr{A} = \bigcup_{x \in \mathscr{X}} \mathscr{A}(x)$ where $\mathscr{A}(x)$ is the set of allowable actions in state $x \in \mathscr{X}$. We do not require $\mathscr{X}$ or $\mathscr{A}(x)$ to be finite or countable. Let $\{X_t\colon\ t = 1, \ldots, \tau\}$ and $\{Q_t\colon\ t = 1, \ldots, \tau\}$ denote the sequence of states and actions. At each time t, the decision maker observes the value of the state X_t and selects an action. This is expressed as choosing a function $q_t\colon \mathscr{X} \to \mathscr{A}$ and setting $Q_t = q_t(X_t) \in \mathscr{A}(X_t)$ to be the action in period t. A sequence of such functions $\{q_t(\cdot)\colon t = 1, \ldots, \tau\}$ is called a policy. (There is no loss of optimality in restricting attention to such *Markovian deterministic policies* that are functions only of the current state; see Puterman 1994).

After the decision maker selects an action in period t, the state in period $t+1$ is determined by the period-t state and action, as well as by the realized value of a real-valued random disturbance Z_t. In a typical inventory context, the state will represent a period's starting inventory, the action will represent the order quantity, and the disturbance will be demand. For each period t, we assume that there is a function $\psi_t\colon \mathscr{X} \times \mathscr{A} \times \mathbb{R} \to \mathscr{X}$ that specifies the next state (say x') according to the system equation, $x' = \psi_t(x, q, z)$. The disturbances $\{Z_t\colon\ t = 1, \ldots, \tau\}$ are mutually independent, and we denote by F_t the distribution function of Z_t. Given a state x and action q in period t, the state in period $t+1$ is conditionally independent of past states and actions. The transition probabilities of the MDP are given by $P(X_{t+1} \in B \mid X_t = x, Q_t = q) = \int_{z \in S_t(x,q,B)} dF_t(z)$, where $S_t(x, q, B) = \{z \in \mathbb{R}\colon\ \psi_t(x, q, z) \in B\}$ for $x \in \mathscr{X}$, $q \in \mathscr{A}(x)$, and $B \subseteq \mathscr{X}$.

In period t, the function $c_t\colon \mathscr{X} \times \mathscr{A} \times \mathbb{R} \to \mathbb{R}$ specifies the cost $c_t(x, q, z)$ that is incurred if the state is x, the action is q, and the disturbance is z. The objective is to identify a policy that minimizes the expected total cost; that is, to minimize $\mathrm{E}\sum_{t=1}^{\tau} c_t(X_t, q_t(X_t), Z_t)$ over all $\{q_t(\cdot)\colon\ t = 1, \ldots, \tau\}$ where $X_{t+1} = \psi_t(X_t, q_t(X_t), Z_t)$ for $t = 1, \ldots, \tau - 1$ and X_1 is any initial state. A policy that yields the minimum expected total cost is called an optimal policy. A policy that yields an expected total cost that exceeds the minimum by no more than $\varepsilon > 0$ is called an ε-optimal policy.

Define the single-period expected cost function for period t to be $C_t(x, q) = \int_{\mathbb{R}} c_t(x, q, z)\, dF_t(z)$. The functions c_t and C_t may be unbounded. Let $V_t(x)$ denote the optimal expected cost over periods $t, \ldots, \tau$ (the optimal expected cost-to-go at time t), given that the state in period t is x (so the objective is to obtain $\{V_t(\cdot)\}$ and, if possible, an associated optimal policy). We may obtain the functions $\{V_t\}$ recursively by

$$V_t(x) = \inf_{q \in \mathscr{A}(x)} W_t(x, q), \tag{1}$$

where

$$W_t(x, q) = \begin{cases} C_t(x, q) + \int V_{t+1}(\psi_t(x, q, z))\, dF_t(z) \\ \qquad\qquad \text{if } t = 1, \ldots, \tau - 1 \\ C_\tau(x, q) \quad \text{if } t = \tau. \end{cases} \tag{2}$$

Carrying out the minimization in (1) may be difficult. In addition, if $\mathscr{X}$ or $\mathscr{A}$ is large, then the MDP may suffer from the well-known curse of dimensionality, rendering exact solution of (1)–(2) impractical. If $\mathscr{X}$ or $\mathscr{A}$ is infinite, then it may be necessary to truncate or discretize the problem to use (1)–(2). Such computational issues are beyond the scope of this article.

If for each x and t, a minimum is attained in (1), then a policy that uses action $q_t^*(x) \in \arg\min_{q \in \mathscr{A}(x)} W_t(x, q)$ when the state is x at time t is an optimal policy, and $V_t(x) = W_t(x, q_t^*(x))$. For simplicity, we sometimes use $V_{\tau+1}(\cdot) = 0$.

If a minimum is not attained in (1), then no optimal policy exists and we must be content with finding an ε-optimal policy, where $\varepsilon > 0$ is arbitrary. This can be done by selecting for each t and x an action $q_t^\varepsilon(x) \in \mathscr{A}(x)$ that satisfies

$$W_t(x, q_t^\varepsilon(x)) \leqslant V_t(x) + \varepsilon/\tau\,. \tag{3}$$

The expected total cost $V_1^\varepsilon(x)$ from following such a policy can be computed recursively by $V_t^\varepsilon(x) = C_t(x, q_t^\varepsilon(x)) + \int V_{t+1}^\varepsilon(\psi_t(x, q_t^\varepsilon(x), z))\, dF_t(z)$ for $t = 1, \ldots, \tau - 1$ and $V_\tau^\varepsilon(x) = C_\tau(x, q_\tau^\varepsilon(x))$. It can be checked using backward induction that $V_t^\varepsilon(x) \leqslant V_t(x) + \varepsilon(\tau - t + 1)/\tau$, and hence the policy that specifies actions $\{q_t^\varepsilon(x)\}$ is indeed ε-optimal; see (Puterman 1994, Theorem 4.3.4).

3. An Empirical Approach

We now describe an empirical approach that is applicable when the distributions $\{F_t\}$ are not known. The input to the algorithm is a set of historical or simulated data $\{Z_t^i\colon\ t = 1, \ldots, \tau, i = 1, \ldots, n_t\}$, where Z_t^i is the ith value of the disturbance in period t and n_t is the number of observations of the period-t disturbance. The random variables $\{Z_t^i\}$ are defined on probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and we assume that for each t, the sequence $\{Z_t^i\colon\ i = 1, \ldots, n_t\}$ is i.i.d.

with common distribution function F_t. If $\{Z_t^i\}$ are historical data, then it is natural for the values also to be independent across time periods (this will be the case because the problem is an MDP). If, however, $\{Z_t^i\}$ are generated by simulations, then it may possibly be useful to introduce dependence between (say) samples in period t and samples in period t'. In any event, our analysis assumes that for each t, the sequence $\{Z_t^i\colon i = 1, \ldots, n_t\}$ is i.i.d., but does not require independence across time periods.

For each t, let $\hat{F}_t$ be the empirical distribution function of $\{Z_t^i\colon i = 1, \ldots, n_t\}$; i.e., $\hat{F}_t(z) = n_t^{-1} \sum_{i=1}^{n_t} 1\{Z_t^i \leqslant z\}$ for $z \in \mathbb{R}$. The empirical approach mimics (1)–(2), recursively defining $\{\hat{V}_t\}$ as

$$\hat{V}_t(x) = \inf_{q \in \mathscr{A}(x)} \hat{W}_t(x, q) \tag{4}$$

where

$$\hat{W}_t(x, q) = \begin{cases} \hat{C}_t(x, q) + \int \hat{V}_{t+1}(\psi_t(x, q, z))\, d\hat{F}_t(z) \\ \qquad\qquad \text{if } t = 1, \ldots, \tau - 1 \\ \hat{C}_\tau(x, q) \quad \text{if } t = \tau \end{cases} \tag{5}$$

and $\hat{C}_t(x, q) = \int c_t(x, q, z)\, d\hat{F}_t(z)$. [For clarification, for any $f\colon \mathbb{R} \to \mathbb{R}$, we have $\int f(z)\, d\hat{F}_t(z) = n_t^{-1} \sum_{i=1}^{n_t} f(Z_t^i)$.] For each x and t, let $\hat{q}_t^\varepsilon(x) \in \mathscr{A}(x)$ satisfy

$$\hat{W}_t(x, \hat{q}_t^\varepsilon(x)) \leqslant \hat{V}_t(x) + \varepsilon/\tau. \tag{6}$$

Note that for the above to be useful for computations, it is necessary for one to know the functions $\{c_t\}$ and $\{\psi_t\}$, but not the distributions $\{F_t\}$.

One may view the above as estimating C_t with its sample-average approximation $\hat{C}_t$ and estimating the time-t state transition probabilities by

$$\hat{P}_t(B \mid x, q) = n_t^{-1} \sum_{i=1}^{n_t} 1\{\psi_t(x, q, Z_t^i) \in B\} = \int_{z \in S_t(x, q, B)} d\hat{F}_t(z).$$

With these estimates in hand, the quantities in relations (4)–(6) are computed via the usual backwards induction algorithm of MDP theory, applied with estimates of the single-period expected costs and transition probabilities in place of the true (but unknown) values.

Once we have $\{\hat{F}_t\}$, the computation and storage requirements with (4)–(5) are roughly the same as with (1)–(2). When such requirements render exact solution of (4)–(5) impractical, approximation or specialized sampling schemes may be necessary. Such topics for general empirical MDPs are beyond the scope of this article. (Note that exact solution of (4)–(5) does not mean exact solution of the true MDP.) For problems in which (4)–(5) are not computationally practical, there is still insight to be gained from studying (4)–(5) because they may allow us to identify structural properties of the empirical MDP, which may aid in its solution. See Puterman (1994, p. 93) for similar comments regarding (nonempirical) MDPs. Our analysis of (4)–(5) separates the issue of how the empirical MDP approximates a true MDP with infinite state and action spaces and unknown disturbance distributions from the issue of how one would implement a computational procedure to solve (exactly or approximately) the empirical MDP. In the online supplement, we describe a computational procedure for some inventory problems that solves (4)–(5) exactly, even though the true MDP for such problems has uncountable state and action spaces.

Here we should also point out that the data $\{Z_t^i\}$ can be used to simulate state transitions needed for methods based upon trees. In particular, for a given state x, time t, and action q, a value (say) Z_t^i generates a transition from x to $\psi_t(x, q, Z_t^i)$. As indicated earlier, the method (4)–(5) does not involve trees.

In case that the infimum in (4) is attained, then we may take $\varepsilon = 0$ above, in which case the actions $\{\hat{q}_t^0(x)\}$ specify a policy that is optimal with respect to the estimated disturbance distributions $\{\hat{F}_t\}$. Otherwise the policy is ε-optimal with respect to the estimated disturbance distributions. As a practical matter, ε may depend upon the data; i.e., ε may be random. Intuitively, this means that for different realizations of $\{Z_t^i\}$, someone implementing the algorithm (4)–(6) may select different ε. However, in our analysis below, we will suppose that ε is a constant.

Below, we analyze the performance (with respect to the true rather than estimated distributions) of the policy specified by the empirical MDP algorithm (4)–(6). That is, we study how close the true expected total cost (where expectation is taken with respect to the distributions $\{F_t\}$) of the policy determined by (4)–(6) is to the true minimum expected total cost as determined by (1)–(3). It is intuitive that if $\{n_t\}$ are large, then the empirical disturbance distributions will be close to the true disturbance distributions, and hence the solution to (4)–(6) should be close to that of (1)–(3) under suitable conditions.

Let $\tilde{V}_t(x)$ be the true expected total cost from time t onward from following the policy specified by the algorithm (4)–(6), given that the state is x at the start of period t. (Note we are suppressing the dependence upon ε from the notation.) We have $\tilde{V}_\tau(x) = C_\tau(x, \hat{q}_\tau^\varepsilon(x))$ and

$$\tilde{V}_t(x) = C_t(x, \hat{q}_t^\varepsilon(x)) + \int \tilde{V}_{t+1}(\psi_t(x, \hat{q}_t^\varepsilon(x), z))\, dF_t(z)$$

for $t = 1, \ldots, \tau - 1$. How close is $\tilde{V}_1$ to V_1? The main difficulty in answering this question is that computation of $\tilde{V}_1$ and V_1 requires $\{F_t\}$, which are unknown. Moreover, one should only hope for probabilistic performance guarantees, because $\tilde{V}_1$ is a (function-valued) random element on Ω. Note that for each fixed $x \in \mathscr{X}$, the value $\tilde{V}_1(x)$ depends upon the actions $\{\hat{q}_t^\varepsilon(x)\}$, which depend upon the random variables $\{Z_t^i\}$, which are functions of $\omega \in \Omega$. So, for each fixed $x \in \mathscr{X}$, $\tilde{V}_1(x)$ is a real-valued random variable defined

on Ω. As is customary, we will usually suppress the dependence upon $\omega \in \Omega$ from the notation, rather than writing $\tilde{V}_t(x, \omega)$, $\hat{q}_t^{\varepsilon}(x, \omega)$, $Z_t^i(\omega)$, and so forth.

It is apparent that $\mathbb{P}[V_1(x) \leqslant \tilde{V}_1(x) \text{ for all } x \in \mathscr{X}] = 1$; i.e., the true expected total cost of the policy specified by the algorithm is no less than the true minimum expected total cost. For additional clarification, the measure $\mathbb{P}$ is on the probability space Ω upon which $\{Z_t^i\}$ are defined. The expectations in the "true expected total cost(s)" are *not* computed by integrating on Ω with respect to $\mathbb{P}$, but rather by integrating on $\mathbb{R}^\tau$ with respect to $dF_1 \times \cdots \times dF_\tau$. The measure $\mathbb{P}$ is related to $\{F_t\}$ by $\mathbb{P}[\{\omega \in \Omega\colon Z_t^i(\omega) \leqslant z_i \text{ for } i = 1, \ldots, n_t\}] = \Pi_{i=1}^{n_t} F_t(z_i)$ for any $\{z_i\}$.

4. A Performance Guarantee

In this section we provide a performance guarantee for the method introduced in the previous section. For $f\colon \mathscr{X} \to \mathbb{R}$, let $\|f\| = \sup_{x \in \mathscr{X}} |f(x)|$, and for $f\colon \mathscr{X} \times \mathscr{A} \to \mathbb{R}$, then $\|f\| = \sup_{x \in \mathscr{X},\, q \in \mathscr{A}} |f(x, q)|$. For $\mathscr{S} \subseteq \mathbb{R}$ we say that a function $f\colon \mathscr{S} \to \mathbb{R}$ is Lipschitz with constant L if $|f(x_2) - f(x_1)| \leqslant L|x_2 - x_1|$ for all $x_1, x_2 \in \mathscr{S}$. The first part of the upcoming theorem gives, for fixed numbers of disturbance observations, the probability of obtaining a policy that is "nearly optimal." The second part uses the first part to provide expressions for the number of samples that suffice to obtain, with a specified probability, a policy whose true expected cost deviates by at most a specified amount from optimality. The third part applies to situations where there is $\mathbb{P}$-almost surely an optimal policy for the empirical MDP with the estimates in place of the true values. In preparation for the theorem, for $\epsilon > 0$ and $\varepsilon \in [0, \epsilon)$ define

$$\xi_t(\epsilon, \varepsilon) = (\tau^2 + 3\tau - 2\tau t - 3t + t^2 + 2)(\epsilon - \varepsilon)/(\tau^2 + \tau) + (\tau - t + 1)\varepsilon/\tau.$$

The theorem assumes each $V_{t+1}(\psi_t(x, q, \cdot))$ to be Lipschitz with constant H_{t+1}. Proposition S-3 in §S-1 of the online supplement shows how to obtain $\{H_{t+1}\}$ from properties of the cost functions $\{c_t\}$ and state transition functions $\{\psi_t\}$—without knowledge of the distributions $\{F_t\}$. See also the text preceding Proposition S-3 for comments on the applicability of the theorem to finite-state and action MDPs. Lemma 2 in §5 identifies values of $\{H_{t+1}\}$ for some inventory models, also without using $\{F_t\}$.

THEOREM 1. *Suppose there exist constants* $\{(\alpha_t, \beta_t)\colon t = 1, \ldots, \tau\}$ *so that* $F_t(\alpha_t-) = \lim_{z \uparrow \alpha_t} F_t(z) = 0$ *and* $F_t(\beta_t) = 1$ *for each* t. *Suppose also that* $V_{t+1}(\psi_t(x, q, \cdot))$ *is Lipschitz with constant* H_{t+1} *and* $c_t(x, q, \cdot)$ *is Lipschitz with constant* ρ_t *for all* x, q *for each* $t = 1, \ldots, \tau$. *Let* $\lambda_\tau = (\beta_\tau - \alpha_\tau)\rho_\tau$ *and* $\lambda_t = (\beta_t - \alpha_t)(\rho_t + H_{t+1})$ *for* $t < \tau$. *Suppose* $\{\hat{q}_t^{\varepsilon}(x)\}$ *are given by* (6).

1. *Fix* $\epsilon_1', \ldots, \epsilon_\tau' > 0$ *and suppose* $\varepsilon > 0$. *Define* $\epsilon_t = \sum_{k=t}^{\tau} \epsilon_k'$ *and* $\gamma_t = \epsilon_t'/\lambda_t$ *for* $t = 1, \ldots, \tau$. *Then* $\mathbb{P}[A] \geqslant 1 - \sum_{t=1}^{\tau} 2\exp(-2\gamma_t^2 n_t)$, *where*

$$A = \bigcap_{t=1,\ldots,\tau} \Big\{ \|V_t - \hat{V}_t\| \leqslant \|W_t - \hat{W}_t\| \leqslant \epsilon_t;\ 0 \leqslant \tilde{V}_t(x) - V_t(x) \leqslant 2\sum_{j=t}^{\tau} \epsilon_j + \frac{(\tau - t + 1)\varepsilon}{\tau} \ \forall x \Big\}. \quad (7)$$

2. *Fix* $\epsilon > 0$ *and* $\delta > 0$, *and suppose* $\varepsilon \in (0, \epsilon)$. *If* $n_t \geqslant n_t^* = [2(\epsilon - \varepsilon)^2]^{-1}(\tau^2 + \tau)^2 \lambda_t^2 \log(2\tau/\delta)$ *for* $t = 1, \ldots, \tau$, *then* $\mathbb{P}[B] \geqslant 1 - \delta$, *where*

$$B = \bigcap_{t=1,\ldots,\tau} \Big\{ \|V_t - \hat{V}_t\| \leqslant \|W_t - \hat{W}_t\| \leqslant \frac{(\tau - t + 1)(\epsilon - \varepsilon)}{(\tau^2 + \tau)};\ 0 \leqslant \tilde{V}_t(x) - V_t(x) \leqslant \xi_t(\epsilon, \varepsilon) \ \forall x \Big\}. \quad (8)$$

3. *If* $\mathbb{P}[\min_{q \in \mathscr{A}(x)} \hat{W}_t(x, q) \text{ exists for all } x \in \mathscr{X} \text{ and } t = 1, \ldots, \tau] = 1$, *then we may also take* $\varepsilon = 0$ *in parts* 1 *and* 2 *above.*

Occurrence of the event A in part 1 implies that the value function estimates are close to the true value function in the sense that $\|\hat{V}_t - V_t\| \leqslant \epsilon_t$ for all t, because $\{\|\hat{V}_t - V_t\| \leqslant \epsilon_t \text{ for all } t = 1, \ldots, \tau\} \supseteq A$. Occurrence of A also implies that the policy determined by (4)–(6) yields a true expected cost that is near the true optimal expected cost; more precisely, $0 \leqslant \tilde{V}_1(x) - V_1(x) \leqslant 2\sum_{j=1}^{\tau} \epsilon_j + \varepsilon$ for all x, because $\{0 \leqslant \tilde{V}_1(x) - V_1(x) \leqslant 2\sum_{j=1}^{\tau} \epsilon_j + \varepsilon \text{ for all } x\} \supseteq A$. Part 1 provides a lower bound for the probability of the event A. As the number of data points grows, it can be seen that $\mathbb{P}[A]$ approaches 1 at a rate that is exponential in $\min_t n_t$. Implicit above is the assumption that the event A is measurable. The proof of the theorem identifies a measurable event A' such that $A' \subseteq A$ and $\mathbb{P}[A'] \geqslant 1 - \sum_{t=1}^{\tau} 2\exp(-2\gamma_t^2 n_t)$.

In part 2, note that $\xi_1(\epsilon, \varepsilon) = \epsilon$ for any $\varepsilon \in (0, \epsilon)$, so $B \subseteq \{0 \leqslant \tilde{V}_1(x) - V_1(x) \leqslant \epsilon \text{ for all } x\}$. Hence, it is immediate from part 2 of the theorem that if $n_t \geqslant n_t^*$ for all t, then we have

$$\mathbb{P}\big[0 \leqslant \tilde{V}_1(x) - V_1(x) \leqslant \epsilon \text{ for all } x \in \mathscr{X}\big] \geqslant 1 - \delta. \quad (9)$$

This means that if each n_t is large enough, then with probability at least $1 - \delta$ the algorithm will yield a policy that is ϵ-optimal. Hence, if we have enough samples from each F_t, then the expected cost from following the policy specified by the algorithm (4)–(6) is, with probability at least $1 - \delta$, within ϵ of the true optimal expected cost. Note also that, not surprisingly, n_t^* is decreasing in ϵ and δ, and increasing in ε and $(\beta_t - \alpha_t)$. In particular, n_t^* is proportional to $(\epsilon - \varepsilon)^{-2}$ and affine in $\log(1/\delta)$.

Part 2 also implies that if $n_t \geqslant n_t^*$ for all t, then we have

$$\mathbb{P}\Big[\hat{V}_t(x) - \frac{(\tau - t + 1)(\epsilon - \varepsilon)}{\tau^2 + \tau} \leqslant V_t(x) \leqslant \hat{V}_t(x) + \frac{(\tau - t + 1)(\epsilon - \varepsilon)}{\tau^2 + \tau} \text{ for all } x \text{ and } t\Big] \geqslant 1 - \delta. \quad (10)$$

The relation (10) states that

$$\hat{V}_t(x) \pm (\tau - t + 1)(\epsilon - \varepsilon)/(\tau^2 + \tau)$$

are simultaneous $100 \times (1 - \delta)\%$ confidence intervals for $\{V_t(x)\}$.

When $\{(\alpha_t, \beta_t)\}$, $\{H_t\}$, and $\{\rho_t\}$ are known, the values $\{n_t^*\}$ can be readily calculated, and so the theorem allows easy determination of how many samples are needed to obtain any desired level of accuracy ($\epsilon > 0$) with any desired probability ($1 - \delta \in [0, 1)$). When the functions $\{c_t\}$ are known, then a decision maker can determine the values $\{\rho_t\}$, provided that they exist. Assuming knowledge of $\{(\alpha_t, \beta_t)\}$ is not unreasonable. In an inventory context where the disturbance represents demand, it is natural that $\alpha_t = 0$. A manager may use prior experience or knowledge of the system to obtain values $\{\beta_t\}$. In settings where $\{(\alpha_t, \beta_t)\}$, $\{H_t\}$, or $\{\rho_t\}$ are not known, but all exist, the theorem is still valid but less useful from a practical standpoint.

We should point out that by using the usual inverse transform method, an MDP (even one whose disturbance distributions have unbounded support) can be transformed to an equivalent MDP in which the disturbance distributions are all uniform distributions on $[0, 1]$, in which case $\alpha_t = 0$ and $\beta_t = 1$. However, the transformation requires knowledge of F_t—which often would not be available when using the empirical approach. We also must appropriately redefine the functions $\{c_t\}$ and $\{\psi_t\}$ so that their third arguments are the uniform-$[0, 1]$ realizations rather than the original disturbances from the "untransformed" problem. This will generally alter the Lipschitz properties of these functions, thereby possibly violating the assumptions of the theorem. See Jain and Varaiya (2006) for an example that shows that convergence properties of sampling-based approaches to MDPs depend not just on the true MDP, but also on the choice of simulative model.

A slightly different idea, which may be useful in a simulation context when $\{F_t\}$ are known, is to generate a single sequence of i.i.d. uniform-$[0, 1]$ random variables $\{\xi^i\colon i = 1, \ldots, \max_t n_t\}$ and to let $Z_t^i = F_t^{-1}(\xi^i)$ for $t = 1, \ldots, \tau$, $i = 1, \ldots, n_t$ (this is the inverse transform). Each individual uniform sample generates (dependent) realizations of the disturbance in all time periods. In this way, we can apply the empirical method with nonuniform disturbance realizations $\{Z_t^i = F_t^{-1}(\xi^i)\}$ and apply the theorem with nonuniform disturbance distributions $\{F_t\}$. In this case, there is no need to redefine the functions $\{c_t\}$ and $\{\psi_t\}$, but to apply the theorem we still require the bounded support of the $\{F_t\}$. Here, $\max_t n_t$ independent uniform-$[0, 1]$ random variables are generated, which in turn generate a total of $\sum_t n_t$ disturbances.

The theorem also allows us to obtain relative, rather than absolute, performance guarantees in some cases. For example, if there is a $\underline{c}$ such that $c_t(x, q, z) \geqslant \underline{c} > 0$ for all x, q, z, t (i.e., costs are bounded away from 0), then $V_1(x) \geqslant \tau \underline{c}$ for all x. In this case, if $\tilde{V}_1(x) - V_1(x) \leqslant \epsilon$, then $\tilde{V}_1(x)/V_1(x) \leqslant 1 + \epsilon/V_1(x) \leqslant 1 + \epsilon/(\tau \underline{c})$ and hence (9) implies $\mathbb{P}[\tilde{V}_1(x) \leqslant V_1(x)(1 + \epsilon/(\underline{c}\tau))$ for all $x \in \mathscr{X}] \geqslant 1 - \delta$. No such $\underline{c}$ exists for the inventory models discussed in the next section. In the next section, we obtain relative performance guarantees for inventory models where costs are *not* bounded away from 0 and with no restriction on support of the demand distributions.

Part 2 of the theorem provides a performance guarantee for *all* state-time pairs, that is,

$$\mathbb{P}[0 \leqslant \tilde{V}_t(x) - V_t(x) \leqslant \xi_t(\epsilon, \varepsilon) \text{ for all } t = 1, \ldots, \tau \text{ and } x \in \mathscr{X}] \geqslant 1 - \delta. \tag{11}$$

In contrast, results in other papers that consider infinite action spaces provide guarantees on the expected cost-to-go for a fixed distribution [say $H(\cdot)$] of the initial state (in particular, they hold for a fixed initial state). To make this more precise, suppose that $\bar{V}_t(x)$ denotes the true expected cost-to-go of the obtained policy in Ng and Jordan (2000) when the state is x and the time is t. The main results in Ng and Jordan (2000) state that

$$\mathbb{P}\left[\int_x \bar{V}_1(x)\, dH(x) - \int_x V_1(x)\, dH(x) \leqslant \epsilon\right] > 1 - \delta$$

provided that the number of sampled trajectories is large enough. The results do not rule out the possibility that $\mathbb{P}[\bar{V}_t(x) - V_t(x) > \epsilon] \geqslant \delta$ for some values of x and t. This same distinction applies when comparing our work to that of Kearns et al. (2000), Jain and Varaiya (2006), and Bartlett and Tewari (2007) as well. One possible way to deal with the problem of reaching such (x, t)-pairs when using, e.g., the Ng and Jordan approach, is to sample new trajectories after each state transition with each new trajectory starting at the just-reached (state, time)-pair. Once the new trajectories are sampled, then a new policy for the remainder of the horizon could be computed by performing the appropriate optimization. Only the action prescribed for the current period would be used, however, as resampling and reoptimization would be done again in subsequent periods. This modification would be useful in settings in which there is adequate time to do the resampling and reoptimization at each step.

In the inventory context, the "real-time" optimization mentioned in the previous paragraph would involve choosing base-stock levels for each future period that minimize the empirical cost averaged over the sampled trajectories. Hence, there would be one decision variable (a base-stock level) in the optimization for each future period. The optimization problem will generally be nonconvex, but it would likely be computationally tractable absent an inordinately long horizon. (Choosing base-stock levels that minimize the empirical cost averaged over the trajectories is different from the model-based approach that is our focus.)

Given any $\epsilon_* > 0$, if we take $\epsilon = c\epsilon_* + (\tau+1)\epsilon_*/3$ and $\varepsilon = c\epsilon_*$ for any $c > 0$ in part 2 of the theorem, then $\mathbb{P}[\|W_1 - \hat{W}_1\| > \epsilon_*/3] \leqslant \delta$ if $n_k \geqslant (9/2)\epsilon_*^{-2}\tau^2\lambda_k^2 \log(2\tau/\delta)$ for $k = 1, \ldots, \tau$. If for a fixed initial state x we select an action q' for period 1 that satisfies $\hat{W}_1(x, q') \leqslant \inf_{q\in\mathcal{A}(x)} \hat{W}_1(x, q) + \epsilon_*/3$, then it follows that $W_1(x, q') \leqslant \hat{W}_1(x, q') + \sup_{q\in\mathcal{A}(x)} |W_1(x, q) - \hat{W}_1(x, q)| \leqslant \inf_{q\in\mathcal{A}(x)} W_1(x, q) + 2\sup_{q\in\mathcal{A}(x)} |W_1(x, q) - \hat{W}_1(x, q)| + \epsilon_*/3$. Recall that $V_1(x) = \inf_{q\in\mathcal{A}(x)} W_1(x, q)$, and note that $\|W_1 - \hat{W}_1\| \geqslant \sup_{q\in\mathcal{A}(x)} |W_1(x, q) - \hat{W}_1(x, q)|$. Therefore, Theorem 1 implies that

$$\mathbb{P}[W_1(x, q') \leqslant V_1(x) + \epsilon_*] \geqslant 1 - \delta \tag{12}$$

for the state x if the numbers of samples satisfy $n_k \geqslant (9/2)\epsilon_*^{-2}\tau^2\lambda_k^2 \log(2\tau/\delta)$ and $q'(\omega)$—which depends on the disturbance observations—is selected so that $\hat{W}_1(x, q') \leqslant \inf_{q\in\mathcal{A}(x)} \hat{W}_1(x, q) + \epsilon_*/3$. These lower bounds on the numbers of samples are proportional to $\lambda_k^2\tau^2 \log\tau$. We will see in the next section that λ_k^2 is proportional to τ^2 for the inventory problems with time-homogeneous cost rates that we consider. Hence, the numbers of samples sufficient for (12) are proportional to $\tau^4 \log\tau$ for such problems.

Shapiro (2006) and Shapiro et al. (2009, §5.8.2) (hereafter SDR) study the iterative application of the sample average approximation method to Markovian stochastic programs with three time periods. (In their setup, problems with three time periods have two disturbances that must be sampled.) Their main results are guarantees of the form (12) that hold under some assumptions. Similar to our work, they allow infinite state and action spaces. They indicate that their results can be extended to an arbitrary number of periods. To facilitate comparisons, we have worked through the details of such an extension in §S-3 of the online supplement; in particular, see (S-20) and (S-21). Among the assumptions used in our proof of the τ-period extension of the results of Shapiro and SDR are finite-diameter action spaces, Lipschitz conditions on the rough analogs of $\{V_{t+1}(\psi_t(x, \cdot, z))\colon t = 1, \ldots, \tau\}$, and conditions on the moment-generating functions of the rough analogs of $\{V_{t+1}(\psi_t(x, q, Z_t^i)) - \mathbb{E}[V_{t+1}(\psi_t(x, q, Z_t^i))]\colon t = 1, \ldots, \tau\}$. (We use the term "rough" because the setup of Shapiro and SDR differs from ours in some ways.) They allow multidimensional actions and disturbances, whereas we do not. They do not require bounded support of disturbance distributions. We do not impose conditions on the diameter of the action spaces, nor do we make Lipschitz assumptions on $\{V_{t+1}(\psi_t(x, \cdot, z))\colon t = 1, \ldots, \tau\}$ when viewed as functions of the actions. Our Lipschitz assumption on $\{V_{t+1}(\psi_t(x, q, \cdot))\colon t = 1, \ldots, \tau\}$ together with our assumption of bounded support together imply the "conditions on the moment-generating functions" mentioned above.

The extension of the Shapiro and SDR results yields the guarantee (12) if the numbers of samples each period k exceed lower bounds that are proportional to $\sigma_k^2\tau^2 \log\tau$, where σ_k^2 is a value (which one must identify to use the results) that must satisfy some conditions involving the moment-generating functions mentioned above. It is desirable to have a low number of samples, so one should choose the values $\{\sigma_k^2\}$ as small as possible. For inventory problems under the assumptions of Corollary 1 below, we show in §S-3 of the online supplement that σ_k^2 can be chosen to be of order τ^2, but not smaller. Hence, for such inventory problems, the lower bounds on the number of samples in each period from the extension of the Shapiro and SDR approach are of order $\tau^4 \log\tau$. Please see §S-3 for further discussion.

We close this section by noting that we may view the empirical procedure as generating a randomized policy, where the randomization is done, only prior to time $t = 1$, by realizing $\{Z_t^i\}$. This is similar to the viewpoint taken by Kearns et al. (2002). To clarify informally, the empirical procedure can be viewed as an algorithm for obtaining a randomized policy, which immediately prior to time $t = 1$ selects Markovian deterministic policy (say) $\{q_t(\cdot)\colon t = 1, \ldots, \tau\}$ with probability $\mathbb{P}[\{\omega \in \Omega\colon \{\hat{q}_t^{\varepsilon}(\cdot, \omega)\} = \{q_t(\cdot)\}\}]$. The true value of the randomized policy is $v_1(x) := \mathbb{E}\tilde{V}_1(x) := \int_{\omega\in\Omega} \tilde{V}_1(x, \omega)\, d\mathbb{P}(\omega)$. Using part (a) of Theorem 1, we may compute the suboptimality of the randomized policy as

$$\begin{aligned} v_1(x) - V_1(x) &= \int_A [\tilde{V}_1(x, \omega) - V_1(x)]\, d\mathbb{P}(\omega) \\ &\quad + \int_{\Omega\setminus A} [\tilde{V}_1(x, \omega) - V_1(x)]\, d\mathbb{P}(\omega) \\ &\leqslant 2\sum_{t=1}^{\tau} \epsilon_t + \varepsilon + \sup_{\omega\in\Omega\setminus A} \{\tilde{V}_1(x, \omega) - V_1(x)\} \\ &\quad \cdot \sum_{t=1}^{\tau} 2\exp(-2\gamma_t^2 n_t). \end{aligned}$$

In general, $\sup_{\omega\in\Omega\setminus A}\{\tilde{V}_1(x, \omega) - V_1(x)\} = \infty$. However, under the additional assumption that $0 \leqslant c_t(x, q, z) \leqslant \check{c} < \infty$ (bounded costs), it is easy to see that $\tilde{V}_1(x, \omega) - V_1(x) \leqslant \tau\check{c}$, and hence $v_1(x) - V_1(x) \leqslant 2\sum_{t=1}^{\tau} \epsilon_t + \varepsilon + \tau\check{c}\sum_{t=1}^{\tau} 2\exp(-2\gamma_t^2 n_t)$. Therefore, under the additional assumption, we can make $v_1(x) - V_1(x)$ as small as we like through appropriate choices of ε, $\{\epsilon_t'\}$ and $\{n_t\}$.

5. Application to Multiperiod Inventory Models

In this section we consider an MDP formulation of a multiperiod inventory model. The objective is to determine an ordering policy that minimizes the expected total cost over a finite number of time periods, $t = 1, \ldots, \tau$. Demands in different periods are independent, but not necessarily identically distributed. The per-unit holding cost in period t is denoted by $h_t > 0$. We consider a general model that includes settings with backorders and with lost sales as special cases. In the backorder model, $b_t > 0$ is the per-unit backorder cost in period t. In the lost-sales model, $b_t > 0$

is the per-unit lost-sales cost in period t. The distribution of the random demand Z_t in period t is F_t. We assume that F_t has finite mean and $F_t(0-) = 0$ so that demand is nonnegative.

In each time period (say t) the sequence of events is as follows: the inventory manager observes the inventory level (say $x \in \mathbb{R}$) at the beginning of the period and places an order (of say $q \geqslant 0$) that is delivered immediately (the lead time is zero) bringing the inventory level to $x+q$; demand Z_t realizes; costs are incurred on any backorders/lost sales $(Z_t - x - q)^+$ and on any leftover inventory $(x + q - Z_t)^+$. Here $u^+ = \max\{0, u\}$. The inventory level at the beginning of period $t+1$ is $\psi_t(x, q, Z_t) = \phi(x + q - Z_t)$, where the function ϕ is assumed to be convex, nondecreasing, and Lipschitz with constant 1 (for simplicity, we assume ϕ is independent of t). We can take $\phi(u) = u$ to get a system with backorders and $\phi(u) = u^+$ to get a system with lost sales. (The theory can be extended with very minor modifications to settings where the Lipschitz constant of ϕ is an arbitrary $\ell \geqslant 0$.) In the backorders model, a negative value of x above means there are backorders at the beginning of period t.

The period-t cost is $c_t(x, q, z) = k_t(x + q - z)$ where $k_t(u) = b_t(-u)^+ + h_t u^+$ is the cost incurred in period t if the net inventory is u after the order and demand have arrived. Note that c_t depends upon x and q only through their sum $x+q$. It is easy to see that $c_t(x, q, \cdot)$ is Lipschitz with constant $\rho_t = \max\{b_t, h_t\}$. The single-period expected cost function for period t is $C_t(x, q) = K_t(x + q)$, where $K_t(y) := \int k_t(y - z)\, dF_t(z)$.

Let $V_t(x)$ denote the optimal expected cost over periods $t, \ldots, \tau$, given that the inventory level before ordering in period t is x (so the objective is to obtain $V_1(x)$ and an associated optimal ordering policy). As described in, e.g., Zipkin (2000), the optimality Equations (1)–(2) simplify to

$$V_t(x) = \min_{q:\, q \geqslant 0} U_t(x + q) \tag{13}$$

where

$$U_t(y) = \begin{cases} K_t(y) + \int V_{t+1}(\phi(y - z))\, dF_t(z) \\ \qquad\qquad \text{if } t = 1, \ldots, \tau - 1 \\ K_\tau(y) \quad \text{if } t = \tau. \end{cases} \tag{14}$$

To clarify the connection between (13)–(14) and (1)–(2), let $W_t(x, q)$ be defined as in (2). Using $C_t(x, q) = K_t(x + q)$ and $\psi_t(x, q, z) = \phi(x + q - z)$, we see that $W_t(x, q) = U_t(x + q)$.

The following lemma (presented without proof) summarizes some well-known properties that we will use frequently later. A proof can be found in, e.g., Zipkin (2000).

Lemma 1. *For each $t = 1, \ldots, \tau$,* (a) *the function U_t is convex and attains a finite minimum, i.e., there exists $y_t^* \in \mathbb{R}$ such that $U_t(y) \geqslant U_t(y_t^*) \in \mathbb{R}$ for all $y \in \mathbb{R}$;* (b) *the function V_t is nondecreasing and convex*; (c) *there is an optimal policy that specifies order $q_t^*(x) = (y_t^* - x)1\{x \leqslant y_t^*\}$ in state x at time t*; *and* (d) *$V_t(x) = U_t(\max\{x, y_t^*\})$.*

If at the start of period t we have an inventory level of x, it is optimal to order $y_t^* - x$ if $x \leqslant y_t^*$ and to order nothing if $x > y_t^*$. The quantities $\{y_t^*\}$ are called optimal base-stock levels, and the policy that orders $q_t^*(x)$ if the inventory is x at the beginning of period t is optimal in the sense that it minimizes, over all possible policies, the expected total cost accrued in $t = 1, \ldots, \tau$.

As in the previous section, suppose now that $\{F_t\}$ are not known and that $\{Z_t^i\colon\ t = 1, \ldots, \tau;\ i = 1, \ldots, n_t\}$ is a set of historical or simulated demand data. Let $\{\hat{F}_t\}$ be the empirical distribution functions of the demand data. The empirical version of (13)–(14) is as follows.

$$\hat{V}_t(x) = \min_{q:\, q \geqslant 0} \hat{W}_t(x, q) = \min_{q:\, q \geqslant 0} \hat{U}_t(x + q) \tag{15}$$

where

$$\hat{U}_t(y) = \begin{cases} \hat{K}_t(y) + \int \hat{V}_{t+1}(\phi(y - z))\, d\hat{F}_t(z) \\ \qquad\qquad \text{if } t = 1, \ldots, \tau - 1 \\ \hat{K}_\tau(y) \quad \text{if } t = \tau, \end{cases} \tag{16}$$

where $\hat{K}_t(y) = \int k_t(y - z)\, d\hat{F}_t(z)$.

Note that (15)–(16) have the same simple form as do (13)–(14). Consequently, the structural properties described in Lemma 1 hold for (15)–(16) as well. In particular, (15)–(16) yield the following base-stock policy: order $\hat{q}_t(x) = (\hat{y}_t - x)1\{x \leqslant \hat{y}_t\}$ if inventory is x at the start of period t where $\hat{y}_t$ is a minimizer of the convex function $\hat{U}_t$. In §S-4 of the online supplement we provide additional discussion regarding the computation of the functions $\{\hat{U}_t\}$ and $\{\hat{V}_t\}$ and the base-stock levels $\{\hat{y}_t\}$. There we explain how to compute these quantities without truncating or discretizing the state or action spaces. This is notable because the state and action spaces of the true MDP are both uncountably infinite.

As in the previous sections, we are interested in evaluating the true performance of the policy that uses actions $\{\hat{q}_t(x)\}$; that is, we are interested in evaluating $\tilde{V}_\tau(x) = K_\tau(x + \hat{q}_t(x)) = K_\tau(\max\{x, \hat{y}_\tau\})$ and

$$\tilde{V}_t(x) = K_t(\max\{x, \hat{y}_t\}) + \int \tilde{V}_{t+1}(\phi(\max\{x, \hat{y}_t\} - z))\, dF_t(z) \tag{17}$$

for $t = 1, \ldots, \tau - 1$. Note that $\tilde{V}_t$ is generally not convex.

To apply the theorem of the previous section, we need the following result about the true value function of the inventory model.

Lemma 2. *For the inventory model, V_t is Lipschitz with constant $H_t = h_t + \cdots + h_\tau$; i.e., $|V_t(x_2) - V_t(x_1)| \leqslant H_t|x_2 - x_1|$ for all $x_1, x_2 \in \mathbb{R}$. Moreover, $V_t(\psi_{t-1}(x, q, \cdot)) = V_t((\phi(x + q - \cdot))$ is also Lipschitz with constant H_t.*

We may compare the Lipschitz constant provided by the preceding lemma with that provided by Proposition S-3

in §S-1 of the online supplement (which applies to general MDPs). In the present inventory setting, Proposition S-3 gives a Lipschitz constant of $\max\{h_t, b_t\} + \cdots + \max\{h_\tau, b_\tau\}$ for $V_t(\psi_{t-1}(x, q, \cdot))$, which is "worse" than the constant in the above lemma. Hence, appealing to the specific structure of the inventory problem allows us to obtain a stronger result in Lemma 2.

Theorem 1, Lemma 2, and the existence of a minimizer of $\hat{U}_t$ (by Lemma 1 and the discussion that follows (16)) together immediately yield the following.

THEOREM 2. *Suppose there exist constants* $\{\beta_t : t=1, \ldots, \tau\}$ *so that* $F_t(\beta_t) = 1$ *for each* t. *Let* $\rho_t = \max\{b_t, h_t\}$ *and* $H_t = h_t + \cdots + h_\tau$ *for* $t = 1, \ldots, \tau$, *and let* $\lambda_\tau = \beta_\tau \rho_\tau$ *and* $\lambda_t = \beta_t(\rho_t + H_{t+1})$ *for* $t = 1, \ldots, \tau - 1$.

1. *Fix* $\epsilon'_1, \ldots, \epsilon'_\tau > 0$ *and suppose* $\varepsilon \geqslant 0$. *Define* $\epsilon_t = \sum_{k=t}^{\tau} \epsilon'_k$ *and* $\gamma_t = \epsilon'_t/\lambda_t$ *for* $t = 1, \ldots, \tau$. *Then*

$$\mathbb{P}\Bigg[\|V_t - \hat{V}_t\| \leqslant \|U_t - \hat{U}_t\| \leqslant \epsilon_t,\ 0 \leqslant \tilde{V}_t(x) - V_t(x) \leqslant 2\sum_{j=t}^{\tau} \epsilon_j + \frac{(\tau - t + 1)\varepsilon}{\tau}\ \forall x \text{ for } t = 1, \ldots, \tau \Bigg] \geqslant 1 - \sum_{t=1}^{\tau} 2\exp(-2\gamma_t^2 n_t).$$

2. *Fix* $\epsilon > 0$ *and* $\delta > 0$, *and* $\varepsilon \in (0, \epsilon)$. *Suppose* $n_t \geqslant n_t^* = [2(\epsilon - \varepsilon)^2]^{-1}(\tau^2 + \tau)^2 \lambda_t^2 \log(2\tau/\delta)$ *for* $t = 1, \ldots, \tau$. *Then*

$$P\Bigg[\|V_t - \hat{V}_t\| \leqslant \|U_t - \hat{U}_t\| \leqslant \frac{(\tau - t + 1)(\epsilon - \varepsilon)}{(\tau^2 + \tau)},\ 0 \leqslant \tilde{V}_t(x) - V_t(x) \leqslant \xi_t(\epsilon, \varepsilon)\ \forall x \text{ for } t = 1, \ldots, \tau \Bigg] \geqslant 1 - \delta.$$

It is of interest to consider a setting in which $b_t \equiv b$, $h_t \equiv h$, and $\beta_t \equiv \beta$. The first two of these conditions mean that costs are time homogeneous and the third condition can be made to hold by setting $\beta_t \equiv \beta = \max_k \beta_k$. We now have the following corollary. The second part of the corollary follows from the argument leading up to (12). An extension of the results of Shapiro (2006) and Shapiro et al. (2009) to an arbitrary number of periods also yields guarantees of the form in part 2 of the corollary; see §S-3 of the online supplement.

COROLLARY 1. *Fix* $\delta > 0$. *Suppose* $b_t = b$, $h_t = h$, $\rho = \max\{h, b\}$, *and* $F_t(\beta) = 1$ *for all* t.

1. *Suppose* $\epsilon > 0$ *and* $\varepsilon \in (0, \epsilon)$. *If*

$$n_t \geqslant n_t^*(\epsilon, \delta, \varepsilon) := \frac{(\tau^2 + \tau)^2 \beta^2 (\rho + (\tau - t)h)^2}{2(\epsilon - \varepsilon)^2} \log(2\tau/\delta) = O(\tau^6 \log \tau), \tag{18}$$

for all $t = 1, \ldots, \tau$, *then* $\mathbb{P}\big[0 \leqslant \tilde{V}_1(x) - V_1(x) \leqslant \epsilon \text{ for all } x \in \mathbb{R}\big] \geqslant 1 - \delta$.

2. *Consider an initial inventory level* x *and suppose* $\epsilon_* > 0$. *If*

$$n_t \geqslant (9/2)\epsilon_*^{-2}\tau^2(\rho + (\tau - t)h)^2 \log(2\tau/\delta) = O(\tau^4 \log \tau)$$

for all $t = 1, \ldots, \tau$, *then* $\mathbb{P}[U_1(x + q') \leqslant V_1(x) + \epsilon_*] \geqslant 1 - \delta$ *for any order quantity* $q'(\omega)$ *for period* 1 *that satisfies* $\hat{U}_1(x + q') \leqslant \inf_{q \geqslant 0} \hat{U}_1(x + q) + \epsilon_*/3$.

Next we derive a relative, rather than absolute, performance guarantee for policy from the empirical MDP (15)–(16). [The approach outlined in §4 will not work here, because $c_t(x, q, x + q) = 0$.] For the following theorem, we assume that we use particular minimizers of $\{\hat{U}_t\}$ to determine the policy $\{\hat{q}_t(x) = (\hat{y}_t - x)1\{x \leqslant \hat{y}_t\}\}$. Specifically, we take

$$\hat{y}_t := \min\{y :\ \hat{U}_t^r(y) \geqslant 0\}, \tag{19}$$

where $\hat{U}_t^r$ is the right-derivative of $\hat{U}_t$. In addition, we hereafter assume that $\tilde{V}_t$ in (17) is defined using these specific choices of $\hat{y}_t$. In preparation for the theorem, for $\epsilon > 0$ define $\Xi_k(\epsilon) = 1 + ((\tau - k + 1)(\tau - k + 2))\epsilon/(\tau^2 + \tau)$. Observe that $\Xi_1(\epsilon) = 1 + \epsilon$.

THEOREM 3. *Let* $\zeta_t = b_t + H_t = b_t + h_t + \cdots + h_\tau$ *for* $t = 1, \ldots, \tau$, *and* $c = \min_t\{\min\{b_t, h_t\}\}$.

1. *Fix* $\epsilon'_1, \ldots, \epsilon'_\tau > 0$ *such that* $\sum_{i=1}^{\tau} \epsilon'_i \leqslant c/3$. *Define* $\epsilon_t = (3\sum_{i=t}^{\tau} \epsilon'_i)/\min\{b_t, h_t\}$ *and* $\gamma_t = \epsilon'_t/\zeta_t$ *for* $t = 1, \ldots, \tau$. *Then*

$$\mathbb{P}\Bigg[\tilde{V}_k(x) \leqslant V_k(x) \prod_{j=k}^{\tau} (1 + \epsilon_j) \text{ for all } x \in \mathbb{R};\ k = 1, \ldots, \tau \Bigg] \geqslant 1 - \sum_{t=1}^{\tau} 2\exp(-2\gamma_t^2 n_t).$$

2. *Fix* $\epsilon \in (0, 2\log 2]$ *and* $\delta > 0$. *If*

$$n_t \geqslant n_t^{\ddagger} := (2c^2\epsilon^2)^{-1} 9(\tau^2 + \tau)^2 \zeta_t^2 \log(2\tau/\delta)$$

for $t = 1, \ldots, \tau$, *then*

$$\mathbb{P}\big[\tilde{V}_k(x) \leqslant V_k(x)\Xi_k(\epsilon) \text{ for all } x \in \mathbb{R};\ k = 1, \ldots, \tau\big] \geqslant 1 - \delta.$$

Observe that, unlike our previous theorems, we do not need to assume the existence of $\{\beta_t\}$ for the above result. Part 2 provides conditions under which

$$\mathbb{P}\Big[\sup_x (\tilde{V}_1(x) - V_1(x))/V_1(x) \leqslant \epsilon\Big] \geqslant 1 - \delta. \tag{20}$$

With time-homogeneous parameters, we have $\zeta_t = b + (\tau - t + 1)h$ for every t. Substituting these values in part 2 of Theorem 3, we obtain the following corollary. Below, $n_t^{\ddagger}(\epsilon, \delta)$ is proportional to ϵ^{-2} and affine in $\log(1/\delta)$.

COROLLARY 2. *Fix* $\epsilon \in (0, 2\log 2]$ *and* $\delta > 0$ *and suppose* $b_t = b$, $h_t = h$. *If*

$$n_t \geqslant n_t^{\ddagger}(\epsilon, \delta) := \frac{9(\tau^2 + \tau)^2 (b + (\tau - t + 1)h)^2}{2(\epsilon \min\{b, h\})^2} \log(2\tau/\delta) = O(\tau^6 \log \tau), \tag{21}$$

then $\mathbb{P}[\tilde{V}_1(x) \leqslant V_1(x)(1 + \epsilon) \text{ for all } x \in \mathbb{R}] \geqslant 1 - \delta$.

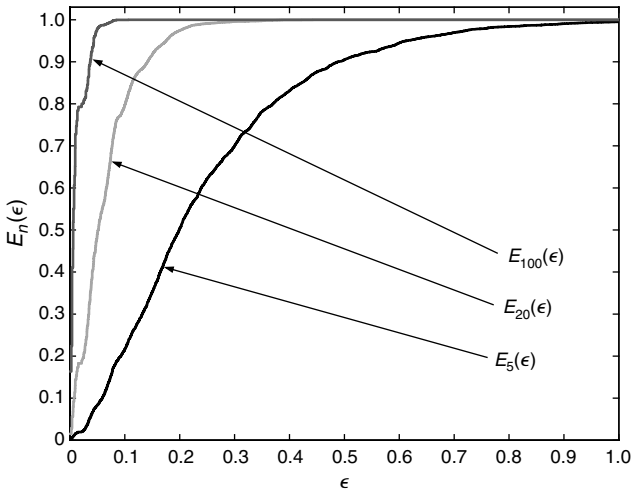
6. Numerical Examples

We close with a brief numerical study of the performance of algorithm (15)–(16) with base-stock levels (19) (which is (4)–(6) with $\varepsilon = 0$ applied to the inventory setting). We provide comparisons with parametric approaches in §S-5 of the online supplement. Suppose $\tau = 5$, with time-homogeneous cost rates $h = 1$ and $b = 10$. Excess demand is backordered; that is, $\phi(u) = u$. Demands in periods $t = 1, \ldots, 5$ are Poisson with respective means of $1, 2, 6, 10, 1$. To clarify, this means, for instance, that F_3 is the distribution function of a Poisson with mean 6. We used (13)–(14) to compute a true optimal policy and the true value functions $\{V_t\}$. At this point, we should emphasize that the computations of $\{y_t^*\}$ and $\{V_t\}$ were done for benchmarking purposes. Someone using the empirical approach studied in this article would typically not know the true demand distributions, and hence would not be able to compute these quantities.

We considered $n_t \equiv n = 20$ and generated 10,000 independent sets of historical demand data. For each set, we computed a policy using the empirical approach, and we also computed the true expected cost $\tilde{V}_1$ of the policy. (Again, note that $\tilde{V}_1$ would typically not be computable by someone using the empirical method, because doing so requires the true distributions $\{F_t\}$.) We refer to one set of demand data and the associated computations as a replication. More precisely, a *single* replication is comprised of $n = 20$ realizations of demand in each period $t = 1, \ldots, 5$; or put differently, it is comprised of one realization of the set of data $\{Z_t^i\colon t = 1, \ldots, 5,\ i = 1, \ldots, 20\}$. From the realized $\{Z_t^i\colon t = 1, \ldots, 5,\ i = 1, \ldots, 20\}$, we formed the empirical distribution functions $\{\hat{F}_t(\cdot) = (1/20)\sum_{i=1}^{20} 1\{Z_t^i \leqslant \cdot\}\colon t = 1, \ldots, 5\}$, and then ran the algorithm (15)–(16) and (19) to obtain base-stock levels $\{\hat{y}_t\colon t = 1, \ldots, 5\}$. We then computed the functions $\{\tilde{V}_t\colon t = 1, \ldots, 5\}$ using (17), and subsequently we computed $R = \max_x(\tilde{V}_1(x) - V_1(x))/V_1(x)$ to measure the relative suboptimality of the policy obtained from the empirical procedure. By computing the appropriate statistics for R we can estimate the left side of (20) for different values of ϵ to get an idea of the performance of the empirical method as of the beginning of the horizon.

The average value of R over the 10,000 independent replications was 0.0652 (the standard deviation was 0.0545). That is, the empirical procedure, using $n = 20$ demand records from each time period, generated a policy with an expected cost 6.52% above the true optimal, on average. One interpretation is that 0.0652 is an (unbiased) estimate of $\mathbb{E}[\max_x(\tilde{V}_1(x) - V_1(x))/V_1(x)]$, where $\mathbb{E}$ is expectation with respect to $\mathbb{P}$. Likewise, a 95% confidence interval for $\mathbb{E}[\max_x(\tilde{V}_1(x) - V_1(x))/V_1(x)]$ is $0.0652 \pm 1.96 \times 0.0545/\sqrt{10{,}000} = (0.0641, 0.0663)$. The curve labeled $E_{20}(\epsilon)$ in Figure 1 shows the empirical distribution function of the 10,000 values of R. Here, we may interpret $E_{20}(\epsilon)$ as an estimate of $\mathbb{P}[\max_x(\tilde{V}_1(x) - V_1(x))/V_1(x) \leqslant \epsilon] = \mathbb{P}[\tilde{V}_1(x) \leqslant V_1(x)(1+\epsilon)$ for all $x]$. For a given value of ϵ on the horizontal axis, the curve $E_{20}(\epsilon)$ shows the fraction of the 10,000 values of R that did not exceed ϵ. For example, with $n = 20$, the empirical method produced a policy that was within 10% of optimal in 79.76% of the replications; i.e., $E_{20}(0.1) = 0.7976$. In addition, 90% of the 10,000 replications yielded an R no greater than 0.1411; $E_{20}(0.1411) = 0.90$.

Figure 1. Empirical distribution $E_n(\epsilon)$ of $\max_x(\tilde{V}_1(x) - V_1(x))/V_1(x)$ for $n = 5, 20, 100$.



We also repeated the above procedure with $n = 5$ and $n = 100$ (again with 10,000 replications each) in order to see how performance is affected by the amount of data that is used as input. The empirical distributions of the relative cost increase R are labeled $E_5(\epsilon)$ and $E_{100}(\epsilon)$ on Figure 1. Not surprisingly, the algorithm did better with more data; this is to be expected based upon part 1 of Theorems 1–3 (and intuition). The average value of R was 0.2458 for $n = 5$ and was 0.0122 for $n = 100$. The respective standard deviations were 0.1881 and 0.0151. The empirical method produced solutions within 10% of the optimum for 21.68% and 99.97% of the replications for $n = 5$ and $n = 100$, respectively [from $E_5(0.1) = 0.2168$ and $E_{100}(0.1) = 0.9997$]. Also, 90% of the replications gave R no greater than 0.4886 and 0.0369 for $n = 5$ and $n = 100$, respectively [from $E_5(0.4886) = 0.90$ and $E_{100}(0.0369) = 0.90$]. With more data ($n = 100$), the empirical method returned an optimal policy 16.35% of the time; $E_{100}(0) = 0.1635$.

To compare the performance we observed in the numerical study with the theoretical guarantees, Table 1 shows values of $n^{\ddagger}(\epsilon, \delta) = \sum_{t=1}^{\tau} n_t^{\ddagger}(\epsilon, \delta)$, as obtained from Corollary 2; recall that $\mathbb{P}[\max_x(\tilde{V}_1(x) - V_1(x))/V_1(x) \leqslant \epsilon] = \mathbb{P}[\tilde{V}_1(x) \leqslant V_1(x)(1+\epsilon)$ for all $x] \geqslant 1 - \delta$ when each $n_t \geqslant n_t^{\ddagger}(\epsilon, \delta)$. We show the total number of samples (across all time periods $t = 1, \ldots, \tau$), because it is indicative of the general relationship between theoretical and observed performance. For different values of n and ϵ, the table reports $n^{\ddagger}(\epsilon, 1 - E_n(\epsilon))$, which is the number of samples such that $\mathbb{P}[\tilde{V}_1(x) \leqslant V_1(x)(1+\epsilon)$ for all $x]$ is at least $E_n(\epsilon)$, where $E_n(\epsilon)$ represents the fraction of replications in the numerical study for which $\tilde{V}_1(x) \leqslant V_1(x)(1+\epsilon)$ for all x.

Table 1. Approximate values of $n^{\ddagger} \times 10^{-9}$ and $n^{\circ} \times 10^{-9}$, where $n^{\ddagger} = n^{\ddagger}(\epsilon, 1 - E_n(\epsilon))$ and $n^{\circ} = n^{\circ}(\epsilon, 1 - E_n(\epsilon))$.

	$\epsilon = 0.05$			$\epsilon = 0.10$		
n	$E_n(\epsilon)$	$n^{\ddagger} \times 10^{-9}$	$n^{\circ} \times 10^{-9}$	$E_n(\epsilon)$	$n^{\ddagger} \times 10^{-9}$	$n^{\circ} \times 10^{-9}$
5	0.0839	3.31	46.85	0.2168	0.88	12.48
20	0.4917	4.13	58.40	0.7976	1.35	19.11
100	0.9784	8.50	120.31	0.9997	3.60	51.04

Notes. It is simple to compute $n^{\ddagger}$ and n° exactly. We do not show all digits because doing so would not be very informative here. Above, we have $n^{\circ}(\cdot) > n^{\ddagger}(\cdot)$. However, for large enough τ we have $n^{\circ}(\epsilon, \delta) < n^{\ddagger}(\epsilon, \delta)$ because $n^{\ddagger}(\epsilon, \delta) = O(\tau^7 \log \tau)$ and $n^{\circ}(\epsilon, \delta) = O(\tau^6 \log \tau)$.

It can be seen that the derived performance guarantees are extremely conservative. For example, in our experiments, when taking just 20 samples from each period ($100 = 20\tau = 20 \cdot 5$ samples in total), the empirical procedure obtained a policy with a cost within 10% of optimal in 79.76% of the replications; $E_{20}(0.1) = 0.7976$. The table shows that Corollary 2 asks for $n^{\ddagger}(0.1, 0.2024) \approx 1.79 \times 10^9$ demand samples to be guaranteed to be within 10% of optimal with probability at least 0.7976. Hence, in this particular example, Corollary 2 appears to severely overestimate how many samples are needed. This suggests the possibility of improving upon the results in the corollary and Theorems 1–3. This overestimation may be partly attributed to the fact that no assumptions are placed on the demand distributions in Theorem 3 (although finite mean demands are needed for the value function of the MDP to be finite). It might be possible to improve the bounds if there were additional information on demand distributions. It also may be that there are distributions for which the bounds in Theorem 3 are not very conservative; however, at this point, we do not have examples for which this is the case. Another possible source of the apparent overestimation is our use of Lemma S-2 of §S-1 in our proofs. The lemma (which is used with the value function in place of the generic function f) does not use information about the value function except Lipschitz continuity. Finally, it is also worth noting that sample sizes needed to obtain performance guarantees for sampling approaches to other types of stochastic optimization problems have been found to be quite conservative as well; see, e.g., Kleywegt et al. (2001, §4.2). (We draw these comparisons not to criticize earlier works, but rather to emphasize the difficulty of obtaining tight performance guarantees.)

To get the same performance guarantee (with the same ϵ and δ) for a policy computed using the approach in Levi et al. (2007), Corollary 3.3 of Levi et al. requires

$$n_t^{\circ}(\epsilon, \delta) = \frac{72\tau^2(h+b)^2}{(\epsilon \min\{b, h\})^2} \log(2\tau/\delta) \sum_{k=1}^{t} (\tau - k + 1)^2$$

samples in each period $t = 1, \ldots, \tau$ and $n^{\circ}(\epsilon, \delta) = \sum_{t=1}^{\tau} n_t^{\circ}(\epsilon, \delta)$ samples in total. For the particular problem at hand, values of $n^{\circ}(\epsilon, 1 - E_n(\epsilon))$ are shown in Table 1. The table shows $n^{\circ}(\epsilon, \delta) > n^{\ddagger}(\epsilon, \delta)$ for this specific example with $\tau = 5$; however, the result of Levi et al. requires asymptotically fewer samples as the length of the horizon τ grows (see the comment in the caption of the table). This may be attributable to the fact that their approach is specifically tailored to inventory problems.

To test the effect of demand variability on the performance of the empirical method, we conducted experiments with negative binomial demand distributions. A random variable X has the negative binomial distribution with parameters $a > 0$ and $m > 0$ when

$$P(X = x) = \frac{\Gamma(x + 1/a)}{(x+1)!\,\Gamma(1/a)} \left(\frac{am}{1+am}\right)^x (1+am)^{-1/a} \quad \text{for } x = 0, 1, 2, \ldots, \quad (22)$$

in which case $E(X) = m$ and $\mathrm{Var}(X) = m + am^2$. (Here, $\Gamma(u) = \int_0^{\infty} e^{-y} y^{u-1}\, dy$.) For discussion and references on inventory models with negative binomial demand distributions, refer to Gallego et al. (2007).

In the experiments, we considered inventory problems parameterized by a value k, which was used to scale demand variability. In each problem, the parameters were identical to those described above for the problem with Poisson demand, except that the demand was instead assumed to be negative binomial with mean m_t and variance km_t in each period $t = 1, 2, \ldots, 5$. The coefficient of variation (c.o.v.) of demand in period t was then $\sqrt{km_t}/m_t = \sqrt{k/m_t}$. This corresponds to parameters $a = (k-1)/m_t$ and $m = m_t$ in period t. We used $m_1 = 1$, $m_2 = 2$, $m_3 = 6$, $m_4 = 10$, $m_5 = 1$ so m_t is the true mean and variance of demand in period t in the problem described above with Poisson demand. For instance, for $k = 16$, the mean demands in periods $t = 1, \ldots, 5$ were 1, 2, 6, 10, 1, the variances were 16, 32, 96, 160, 16, and c.o.v.s were 4.0, 2.8, 1.6, 1.3, 4.0.

For each $k = 2, 4, 8, 16, 32, 64$, we again generated 10,000 independent replications for $n = 5, 20, 100$ as described above for the Poisson example. Results are shown in Table 2. Each cell in the table contains four numbers. The top number is the sample mean (over 10,000 replications) of R, the second number is the sample standard deviation of R, the third number is the fraction of replications for which R was no greater than 0.1 [which is given by $E_n(0.1)$]; and the fourth is the 90% quantile of $E_n(\cdot)$, which we denote by $E_n^{-1}(0.9) = \min\{\epsilon\colon E_n(\epsilon) \geqslant 0.9\}$. For instance, for $k = 16$ and $n = 20$, the sample mean of R was 0.0698, the sample standard deviation was 0.0547, $E_{20}(0.1) = 0.7909$ [so 79.09% of the 10,000 replications with $n = 20$ yielded a policy within 10% of optimal], and $E_{20}^{-1}(0.9) = 0.1362$ [so 90% of the replications yielded a policy within 13.62% of optimal]. For sake of comparison, the results of the example with Poisson demand are shown as well. Overall, there does not appear to be a discernible pattern to how the variability parameter k affects the statistical properties of R, and in general,

Table 2. Sample means of R, sample standard deviations of R, fractions of replications that produced a policy within 10% of optimal, and 90% quantiles of $E_n(\cdot)$.

Actual demand distributions		$n=5$	$n=20$	$n=100$
Poisson	mean	0.2458	0.0652	0.0122
c.o.v.'s:	std. dev.	0.1881	0.0545	0.0151
(1.0, 0.7, 0.4, 0.3, 1.0)	$E_n(0.1)$	0.2168	0.7976	0.9997
	$E_n^{-1}(0.9)$	0.4886	0.1411	0.0369
NB, $k=2$	mean	0.2547	0.0749	0.0180
c.o.v.'s:	std. dev.	0.1790	0.0515	0.0133
(1.4, 1.0, 0.6, 0.5, 1.4)	$E_n(0.1)$	0.1745	0.7503	0.9997
	$E_n^{-1}(0.9)$	0.4911	0.1440	0.0365
NB, $k=4$	mean	0.2473	0.0737	0.0162
c.o.v.'s:	std. dev.	0.1762	0.0503	0.0128
(2.0, 1.4, 0.8, 0.6, 2.0)	$E_n(0.1)$	0.1734	0.7609	0.9998
	$E_n^{-1}(0.9)$	0.4709	0.1398	0.0329
NB, $k=8$	mean	0.2382	0.0715	0.0169
c.o.v.'s:	std. dev.	0.1917	0.0505	0.0124
(2.8, 2.0, 1.2, 0.9, 2.8)	$E_n(0.1)$	0.2050	0.7795	0.9999
	$E_n^{-1}(0.9)$	0.4713	0.1372	0.0333
NB, $k=16$	mean	0.2320	0.0698	0.0166
c.o.v.'s:	std. dev.	0.2233	0.0547	0.0127
(4.0, 2.8, 1.6, 1.3, 4.0)	$E_n(0.1)$	0.2856	0.7909	0.9995
	$E_n^{-1}(0.9)$	0.4933	0.1362	0.0332
NB, $k=32$	mean	0.2501	0.0682	0.0164
c.o.v.'s:	std. dev.	0.3019	0.0636	0.0141
(5.7, 4.0, 2.3, 1.8, 5.7)	$E_n(0.1)$	0.3385	0.7993	0.9985
	$E_n^{-1}(0.9)$	0.5719	0.1418	0.0342
NB, $k=64$	mean	0.2757	0.0635	0.0139
c.o.v.'s:	std. dev.	0.4215	0.0773	0.0140
(8.0, 5.7, 3.3, 2.5, 8.0)	$E_n(0.1)$	0.4264	0.8386	0.9979
	$E_n^{-1}(0.9)$	0.7036	0.1330	0.0309

Note. True demand distributions are negative binomial (except for the row labeled "Poisson") with means 1, 2, 6, 10, 1 in periods $t=1$, 2, 3, 4, 5.

the data suggest that for moderate and large n ($n=20$ and 100), the empirical procedure typically produced effective ordering policies.

To examine the effect of the magnitude of demand, we also considered a problem with Poisson demand with means $5, 10, 30, 50, 5$ in periods $t=1,\ldots,5$ and other parameters as above. The results were not markedly different from those detailed above. For example, with $n=20$, we found that the sample mean of R, sample standard deviation of R, $E_{20}(0.1)$, and $E_{20}^{-1}(0.9)$ over 10,000 replications were 0.0722, 0.0524, 0.7681, and 0.1405, respectively.

Finally, we considered situations in which demand was "lumpy," i.e., it could take only two possible values in each period. In each period $t=1,\ldots,5$ the actual demand was a scaled Bernoulli random variable; $\mathbb{P}(Z_t^i=0)=1-1/k$ and $\mathbb{P}(Z_t^i=km_t)=1/k$ so $\mathbb{E}Z_t^i=m_t$ and $\mathrm{Var}(Z_t^i)=m_t^2(k-1)$ and the c.o.v. of demand in period t was $\sqrt{k-1}$ for each t. We again used $m_1=1$, $m_2=2$, $m_3=6$, $m_4=10$, and $m_5=1$. Table 3 shows the results of 10,000 simulation replications for $k=2,3,4$. Observe that for $k=2$ the empirical method worked extremely well, and in fact it obtained the actual optimal solution in each of the 10,000 replications when $n=100$; this is reflected by the 0 mean and standard deviation of R in the $n=100$ column of the $k=2$ row. The method performs well when $k=2$ because in that case the only two probabilities it needs to "learn" are $\mathbb{P}(Z_t^i=0)=\mathbb{P}(Z_t^i=2m_t)=1/2$, neither of which is small. For $k=3$ and 4, $\mathbb{P}(Z_t^i=0)$ is smaller, making the estimation problem more difficult. In these cases, performance of the empirical approach is comparable to that for the problems with Poisson or negative binomial demand.

Table 3. Sample means of R, sample standard deviations of R, fractions of replications that produced a policy within 10% of optimal, and 90% quantiles of $E_n(\cdot)$.

Actual demand distributions		$n=5$	$n=20$	$n=100$
SB, $k=2$	mean	0.1425	0.0018	0
c.o.v.'s: 1.0	std. dev.	0.4999	0.0567	0
	$E_n(0.1)$	0.8568	0.9990	1
	$E_n^{-1}(0.9)$	0.3103	0	0
SB, $k=3$	mean	0.2700	0.0395	0.0001
c.o.v.'s: 1.4	std. dev.	0.4505	0.1555	0.0083
	$E_n(0.1)$	0.5688	0.9327	0.9998
	$E_n^{-1}(0.9)$	0.7701	0	0
SB, $k=4$	mean	0.2729	0.0856	0.0053
c.o.v.'s: 1.7	std. dev.	0.3181	0.1675	0.0353
	$E_n(0.1)$	0.4179	0.7329	0.9781
	$E_n^{-1}(0.9)$	0.7852	0.2414	0

Note. True demand distributions are scaled Bernoulli with means 1, 2, 6, 10, 1 in periods $t=1$, 2, 3, 4, 5.

7. Conclusion

In this paper we analyzed an empirical approach to Markov decision processes applicable in situations where disturbance distributions are not known, but can be estimated from historical or simulated disturbance data. The approach requires knowledge of the MDP system equations and single-period cost functions. We derived absolute performance guarantees for the approach for finite-horizon MDPs that satisfy some conditions and applied the results to multi-period inventory problems with unknown demand distributions. We also provided a specialized analysis of such inventory problems that subsequently yielded relative performance guarantees. For our proofs, we developed results on the sensitivity of MDPs to changes in their disturbance distributions. A numerical study revealed the empirical approach to be effective in the inventory setting, even when little data was available to estimate the demand distributions.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1090.

Acknowledgments

The authors thank the referees for their constructive comments that helped improve this article. The second author acknowledges the support of Radha and Akshara while working on this article. The second author conducted research for this paper while employed at the University of Minnesota and would like to acknowledge the help received from the Graduate Program in Industrial and Systems Engineering, University of Minnesota.

References

Anthony M, Bartlett PL (1999) *Neural Network Learning*: *Theoretical Foundations* (Cambridge University Press, Cambridge, UK).

Antos A, Szepesvári C, Munos R (2008) Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning* 71(1):89–129.

Bartlett PL, Tewari A (2007) Sample complexity of policy search with known dynamics. Schölkopf B, Platt J, Hoffman T, eds. *Advances in Neural Information Processing Systems*, Vol. 19 (MIT Press, Cambridge, MA), 97–104.

Bertsekas DP (2000) *Dynamic Programming and Optimal Control*, Vol. I (Athena Scientific, Belmont, MA).

Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming* (Athena Scientific, Belmont, MA).

Chang HS, Fu MC, Hu J, Marcus SI (2005) An adaptive sampling algorithm for solving Markov decision processes. *Oper. Res.* 53(1):126–139.

Chang HS, Fu MC, Hu J, Marcus SI (2007a) *Simulation-based Algorithms for Markov Decision Processes* (Springer-Verlag, London).

Chang HS, Fu MC, Hu J, Marcus SI (2007b) Recursive learning automata approach to Markov decision processes. *IEEE Trans. Automatic Control* 52(7):1349–1355.

Even-Dar E, Mannor S, Mansour Y (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Machine Learning Res.* 7(June):1079–1105.

Fiechter, C-N (1994) Efficient reinforcement learning. *Proc. Seventh Ann. Conf. Comput. Learn. Theory* (ACM Press, New York), 88–97.

Gallego G, Katircioglu K, Ramachandran B (2007) Inventory management under highly uncertain demand. *Oper. Res. Lett.* 35(3):281–289.

Jain R, Varaiya PP (2006) Simulation-based uniform value function estimates of Markov decision processes. *SIAM J. Control Optim.* 45(5):1633–1656.

Kakade SM (2003) On the sample complexity of reinforcement learning. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, London.

Kearns M, Singh S (2002) Near-optimal reinforcement learning in polynomial time. *Machine Learning* 49(2–3):209–232.

Kearns M, Mansour Y, Ng AY (2000) Approximate planning in large POMDPS via reusable trajectories. Solla SA, Leen TK, Müller K-R, eds. *Advances in Neural Information Processing Systems*, Vol. 12 (MIT Press, Cambridge, MA), 1001–1007.

Kearns M, Mansour Y, Ng AY (2002) A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning* 49(2–3):193–208.

Kleywegt AJ, Shapiro A, Homem-de-Mello T (2001) The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* 12(2):479–502.

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* 32(4):821–839.

Mannor S, Simester D, Sun P, Tsitsiklis JN (2007) Bias and variance approximation in value function estimates. *Management Sci.* 53(2):308–322.

Massart P (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* 18(3):1269–1283.

Murphy SA (2005) A generalization error for Q-learning. *J. Machine Learning Res.* 6(December):1073–1097.

Ng AY, Jordan M (2000) PEGASUS: A policy search method for large MDPs and POMDPs. Boutilier C, Goldszmidt M, eds. *Proc. Sixteenth Conf. Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco), 406–415.

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Oper. Res.* 56(1):188–203.

Pivazyan K, Shoham Y (2002) Polynomial-time reinforcement learning of near-optimal policies. Dechter R, Kearns M, Sutton R, eds. *Eighteenth Natl. Conf. Artificial Intelligence* (American Association for Artificial Intelligence, Palo Alto, CA).

Powell WB (2007) *Approximate Dynamic Programming* (John Wiley & Sons, Hoboken, NJ).

Puterman ML (1994) *Markov Decision Processes*: *Discrete Stochastic Dynamic Programming* (John Wiley & Sons, New York).

Shapiro A (2006) On complexity of multistage stochastic programs. *Oper. Res. Lett.* 34(1):1–8.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming*: *Modeling and Theory* (SIAM and MPS, Philadelphia).

Strehl AL, Littman ML (2005) A theoretical analysis of model-based interval estimation. De Raedt L, Wrobel S, eds. *Proc. 22nd Internat. Conf. Machine Learning* (ACM Press, New York), 856–863.

Zipkin PH (2000) *Foundations of Inventory Management* (McGraw-Hill, New York).

William L. Cooper is an associate professor in the Department of Industrial and Systems Engineering at the University of Minnesota. His research interests include stochastic models and revenue management.

Bharath Rangarajan is lead optimization analyst at the Target Corporation. He wrote the article in this issue while he was employed as the Benjamin Mayhugh Assistant Professor in the Graduate Program in Industrial and Systems Engineering at the University of Minnesota from 2005–2009. He completed his Ph.D. in operations research at Cornell University in 2004. His research interests include optimization algorithms, numerical analysis, convex programming, and stochastic processes with applications to finance, economics, and engineering.