MANAGEMENT SCIENCE
Vol. 53, No. 2, February 2007, pp. 308–322
ISSN 0025-1909 | EISSN 1526-5501 | 07 | 5302 | 0308

informs

DOI 10.1287/mnsc.1060.0614

Bias and Variance Approximation in Value Function Estimates

Shie Mannor
Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada H3A 2A7, shie@ece.mcgill.ca

Duncan Simester
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, simester@mit.edu

Peng Sun
Fuqua School of Business, Duke University, Durham, North Carolina 27708, psun@duke.edu

John N. Tsitsiklis
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, jnt@mit.edu

We consider a finite-state, finite-action, infinite-horizon, discounted reward Markov decision process and study the bias and variance in the value function estimates that result from empirical estimates of the model parameters. We provide closed-form approximations for the bias and variance, which can then be used to derive confidence intervals around the value function estimates. We illustrate and validate our findings using a large database describing the transaction and mailing histories for customers of a mail-order catalog firm.

Key words: value function; confidence interval; variance; bias
History: Accepted by Wallace J. Hopp, stochastic models and simulation; received July 14, 2004. This paper was with the authors $4\frac{1}{2}$ months for 4 revisions.

1. Introduction

Bellman's value function plays a central role in the optimization of dynamic decision-making models, as well as in the structural estimation of dynamic models of rational agents. For the important case of a finite-state Markov decision process (MDP), the value function depends on two types of model parameters: the transition probabilities between states and the expected one-step rewards from each state. In many applications in the social sciences and in engineering, the transition probabilities and expected rewards are not known and instead must be estimated from finite samples of data. The estimation errors for these parameters introduce bias and variance in the value function estimates.

In this paper, we present a methodology for evaluating this bias and variance. This, in turn, allows the calculation of confidence intervals around the value function estimates. The confidence intervals are themselves approximations. For analytical and computational tractability, they rely on second-order Taylor series approximations. Moreover, because the expressions for the bias and the variance approximation require the true but unknown model parameters, we replace these unknown parameters by their estimates. We evaluate the accuracy of these approximations and validate the expressions using a large sample of real data obtained from a mail-order catalog company.

1.1. Sources of Variance

We start by distinguishing between two types of variance that can arise in an MDP: internal and parametric. Internal variance reflects the stochasticity in the transitions and rewards. For example, in a marketing setting there is rarely certainty as to whether an individual customer will purchase, resulting in genuinely stochastic transitions and rewards. Parametric variance arises if the true transition probabilities and expected rewards are estimated rather than known; the potential for error in the estimates of these parameters introduces variance in the value function estimates.

The two types of variance have different sources and can be illustrated through different experiments. To illustrate internal variance, we can fix the model parameters and then generate a number of finite-length sample trajectories (with all trajectories having the same length, starting from the same state, and using a common control policy). The variation across sample trajectories in the total rewards and/or the identity of the final state reflects internal variance. In

contrast, aggregation across samples does not mitigate parametric variance. The latter can be illustrated by comparing the average outcomes from a large number of samples generated under different estimates for the model parameters. The variation in the average outcomes under different estimates reflects parametric variance.

Internal variance has already been considered in the literature. In particular, Sobel (1982) provides an expression for the internal variance in an MDP with discounted rewards, while Filar et al. (1989) and Baukal-Gursoy and Ross (1992) consider the average reward criterion. In this paper, we focus on parametric variance. Our motivation is that in many contexts the underlying objective involves averaging outcomes across a large number of samples, in which case the internal variance is averaged out. For example, in a marketing application, firm profits typically represent the aggregation of outcomes across a large number of customers. Similarly, in a labor economics setting, a firm often aggregates across a large number of employees. Of course, there are settings where internal variance is also important. For example, when allocating financial portfolios, the (internal) variance of the return on a single financial portfolio is important in its own right.

1.2. Literature

Markov decision problems and the associated methodology of dynamic programming have found a broad range of applications in numerous fields in the social sciences and in engineering. These applications can be broadly divided in two categories based on the research objectives.

The first and more traditional category of applications focuses on optimizing the operation of human or engineering systems and on providing tools for effective decision making. The application areas are vast, and include finance (Luenberger 1997, Campbell and Viceira 2002), economics (Dixit and Pindyck 1994), inventory control and supply chain management (Zipkin 2000), revenue and yield management (McGill and van Ryzin 1999), transportation (Godfrey and Powell 2002), communications, water resource management, and electric power systems. The vast majority of this literature assumes that an accurate system model is available. There is an underlying implicit assumption that the true model will be estimated using statistical methods on the basis of whatever data are available. However, the statistical ramifications of working with finite data records have received little attention. An exception is the literature dealing with online learning of optimal policies (adaptive control of Markov chains, reinforcement learning) (Sutton and Barto 1998, Bertsekas and Tsitsiklis 1996). However, this literature is concerned with asymptotic convergence as opposed to the common statistical questions of standard errors and confidence intervals.

The second category of applications focuses on explaining observed phenomena. Among the most widely cited examples is the work of Rust (1987), who develops a discrete dynamic programming model of the optimal replacement policy for bus engines. According to this approach, the researcher starts by assuming that individuals or firms behave optimally, but that the parameters of the firm or the customer decision problem are unknown. By maximizing the likelihood of the empirically observed actions of individuals or firms under the optimal policies for different sets of parameters, the researcher seeks to identify these unobserved parameters. Similar applications of discrete dynamic programming models have become increasingly common, particularly in the labor (Keane and Wolpin 1994), industrial organization (Hendel and Nevo 2002), and marketing (Gönül and Shi 1998) literatures.

While these methods use a variety of approaches to calculate or approximate the value function, the value function relies on point estimates of the model parameters. Previous attempts to consider the impact of parameter error on the calculated value function have been limited to simulation-based approaches.

We finally note that the impact of uncertainty in the model parameters on the accuracy of the value function estimates has received attention in the finance literature. For example, Xia (2001) and Barberis (2000) investigate how dynamic learning about stock return predictability affects optimal portfolio allocations. The general problem considered in these studies is similar to the one addressed in this paper. However, the sources of variance are different. In particular, the finance literature is concerned with internal variance due to the stochasticity in the underlying process and parametric variance due to nonstationarity of the model parameters, including changes in the investment horizon and/or dynamic learning. In contrast, we abstract away from the problem of internal variance, assume that the model parameters are stationary, and focus on the parametric variance that results from estimating the model parameters from a finite sample of data.

1.3. Overview

As far as we know, this is the first paper to study parametric bias and variance in MDPs. It serves two purposes: First, to illustrate the potential for error in value function estimates and to highlight the potential magnitude of these errors; second, to provide formulas and a methodology for estimating the bias and variance in value function estimates, which can then be used to construct confidence intervals around the value function estimates.

We begin with some notations and background material in §2. In §3, we illustrate the relationship between errors in the model parameters and the accuracy of value function estimates, using actual data from a catalog mailing context. In §4, we present a methodology for estimating the bias and variance in the value function estimates. In §5, we validate our methodology using the catalog mailing data. We conclude in §6 with a review of the findings and a discussion of opportunities for future research.

2. A Formal Description of the Problem

We consider a finite-state, finite-action, infinite-horizon, discounted reward MDP, where S denotes the set of states of cardinality m, A is the set of actions, $\alpha \in (0,1)$ is the discount factor, and P_{ij}^a and R_{ij}^a ($i, j \in S$, $a \in A$) denote transition probabilities and the conditional expected rewards. The scalars P_{ij}^a and R_{ij}^a are interpreted as follows: if the current state is i and action a is applied, then the next state is j with probability P_{ij}^a; furthermore, given that a transition from i to j occurs following an action equal to a, a random reward is obtained, whose conditional expectation is equal to R_{ij}^a. Note that if action a is applied at state i, the expected reward, denoted by R_i^a, is equal to $\sum_j P_{ij}^a R_{ij}^a$.

We are interested in the value function associated with a stationary, Markovian, possibly randomized, fixed policy π. The assumption that the policy is fixed allows us to initially abstract away from the control problem. As we discuss in §4.2, the impact of parameter uncertainty on the solution to the control problem raises additional issues. We use $\pi(a \mid i)$ to denote the conditional probability of applying action a when at state i. Let $P_{ij}^\pi = \sum_a \pi(a \mid i) P_{ij}^a$, which is the transition probability from i to j. The superscript π here is a slight abuse of notation. It will be clear from the context that the Greek letter π as a superscript indicates that the parameter is a function with a policy as an argument. Similarly, denote

$$R_i^\pi = \sum_a \pi(a \mid i) R_i^a = \sum_a \pi(a \mid i) \sum_j P_{ij}^a R_{ij}^a, \qquad (1)$$

which is the expected reward at state i, under the policy π. We use P^π to denote the $m \times m$ matrix with entries P_{ij}^π, and R^π to denote the m-dimensional vector with components R_i^π.

Define the value function associated with policy π to be the m-dimensional vector given by

$$Y^\pi = \sum_{k=0}^{\infty} \alpha^k (P^\pi)^k R^\pi = (I - \alpha P^\pi)^{-1} R^\pi.$$

In our setting, the true model parameters, P_{ij}^a and R_{ij}^a, are not known. Instead, we have access to a finite sample of data from which these parameters can be estimated. Specifically, assume that for every i and a, we have a record of N_i^a transitions out of state i, under action a, and the associated rewards. We treat the numbers N_i^a as fixed (not as random variables), and assume that $N_i^a > 0$ for every i and a. This last assumption restricts attention to actions that have been tried before. For at least two reasons, we anticipate that this will be a relatively weak assumption in practice. First, the inability to evaluate actions in one state does not restrict our ability to evaluate the same action in other states because we can still evaluate an action at any state where the action has been tried before. Thus, the restriction only applies to states in which there is no past information about the outcome. Second, there is a tremendous amount of variation in historical policies in many real-world applications. This variation may arise for a lot of reasons including experimentation, implementation errors, or nonstationarity in the policy. If there is interest in untried actions and there are priors available to help predict the outcome, then a Bayesian approach can be used. For completeness, we detail such an approach in Appendix D (provided in the e-companion).[1]

Furthermore, we do not assume any relation between the sampling process and the policy π of interest; in particular, the N_i^a for different a need not be proportional to $\pi(a \mid i)$, and the number $N_i = \sum_a N_i^a$ of transitions out of state i need not be related to the steady-state probability of state i under policy π.

For the N_i^a transitions out of state i under action a in the sample data, let N_{ij}^a be the number of transitions that lead to state j. Furthermore, let C_{ij}^a be the sum of the rewards associated with these N_{ij}^a transitions (for completeness, we define $C_{ij}^a = 0$ if $N_{ij}^a = 0$). We define

$$\hat{P}_{ij}^a = \frac{N_{ij}^a}{N_i^a}, \qquad \hat{R}_{ij}^a = \frac{C_{ij}^a}{N_{ij}^a},$$

which will be our estimates of P_{ij}^a and R_{ij}^a, respectively. When $N_{ij}^a = 0$, we define $\hat{R}_{ij}^a = 0$. The possibility of N_{ij}^a being zero for feasible transitions introduces some additional bias, which will not be accounted for. However, in our analysis, we will assume that any transition with $N_{ij}^a = 0$ is infeasible. In addition, we define

$$\hat{P}_{ij}^\pi = \sum_a \pi(a \mid i) \hat{P}_{ij}^a,$$

and

$$\hat{R}_i^a = \sum_j \hat{P}_{ij}^a \hat{R}_{ij}^a = \frac{\sum_j C_{ij}^a}{N_i^a}, \qquad \hat{R}_i^\pi = \sum_a \pi(a \mid i) \hat{R}_i^a, \qquad (2)$$

[1] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

which will be our estimates of P_{ij}^{π}, R_i^a, and R_i^{π}, respectively. We finally define a matrix $\hat{P}^{\pi}$ and a vector $\hat{R}^{\pi}$, with entries $\hat{P}_{ij}^{\pi}$ and $\hat{R}_i^{\pi}$, respectively, which will be our estimates of P^{π} and R^{π}. Based on these estimates, we obtain an estimated value function $\hat{Y}^{\pi}$, given by

$$\hat{Y}^{\pi} = (I - \alpha \hat{P}^{\pi})^{-1} \hat{R}^{\pi}. \tag{3}$$

We assume that the sample data reflect the true process in the following sense. The vector $(N_{i1}^a, \ldots, N_{im}^a)$ follows a multinomial distribution with parameters $(N_i^a; P_{i1}^a, \ldots, P_{im}^a)$. Let $\mathbb{E}$ denote expectation under the true model. We then have $\mathbb{E}[N_{ij}^a] = N_i^a P_{ij}^a$. A last assumption that reflects our earlier assumptions that N_i^a is fixed and that each sample reward is conditionally independent from the past, is that $\mathbb{E}[C_{ij}^a \mid N_{ij}^a] = N_{ij}^a R_{ij}^a$. Under these assumptions, it is easily verified that $\hat{P}^{\pi}$ and $\hat{R}^{\pi}$ are unbiased estimates of P and R.

Based on Equation (3), we can anticipate the impact of errors in $\hat{P}^{\pi}$ and $\hat{R}^{\pi}$ on $\hat{Y}^{\pi}$. Note first that $\hat{Y}^{\pi}$ is linear in $\hat{R}^{\pi}$, so that if P^{π} were observed without error (i.e., if $\hat{P}^{\pi} = P^{\pi}$), the variance of $\hat{R}^{\pi}$ would lead to variance in $\hat{Y}^{\pi}$ but not to bias (because $\hat{R}^{\pi}$ is unbiased). In contrast, $\hat{Y}^{\pi}$ is nonlinear in $\hat{P}^{\pi}$, so that errors in $\hat{P}^{\pi}$ lead to both bias and variance in $\hat{Y}^{\pi}$. Moreover, due to the matrix inversion the nonlinearity is substantial, so that any error in $\hat{P}^{\pi}$ can translate to a large error in $\hat{Y}^{\pi}$. This is particularly true when α is close to one. Furthermore, if the errors in $\hat{P}^{\pi}$ and $\hat{R}^{\pi}$ are correlated, the nonlinearity implies that errors in $\hat{R}^{\pi}$ will also lead to bias in $\hat{Y}^{\pi}$.

3. An Illustration

To illustrate the bias and variance that can be introduced to value function estimates by errors in the model parameters, we use real data from a mail-order catalog company. While this application serves as a useful case study, our findings are not limited to this application.

Deciding who should receive a catalog is amongst the most important decisions that mail-order companies must address. Yet, identifying an optimal mailing policy is a difficult task. Customer response functions are highly stochastic, reflecting in part the relative paucity of information that firms have about each customer. Moreover, the problem is a dynamic one. Purchasing decisions are influenced not just by the firm's most recent mailing decision, but also by prior mailing decisions. As a result, the optimal mailing decision depends on past and future mailing decisions.

A typical catalog company might mail 25 catalogs per year. The number of catalogs, the dates that they are mailed, and the content of the catalogs are determined up to a year before the firm decides to whom each catalog will be mailed. For this reason, these decisions are typically treated as fixed when deciding who to mail to. Accordingly, the firm only needs to decide which customers to mail to, on each exogenously determined mailing date (a discrete infinite-horizon problem).

The firm's objective is to maximize its expected total discounted profits. Rewards (profits) in each period are calculated as the revenue earned from customer purchases (if any) less the cost of the goods sold and the mailing costs (approximately 65 cents per catalog mailed). To support their mailing decisions, catalog firms typically maintain large databases describing the individual purchase and mailing histories for each customer. We are fortunate to have access to a large database describing the transaction and mailing histories for the women's apparel division of a moderately large catalog company. This data is described in detail in Simester et al. (2006). It includes the complete transaction histories for approximately 1.72 million customers. The mailing histories are complete for the six-year period from 1996 through 2002 (the company did not maintain a record of the mailing history prior to 1996). Catalogs were mailed on 133 occasions in this six-year period, so that on average a mailing decision occurred every two to three weeks.

The catalog mailing problem can be modelled as an MDP (as in Gönül and Shi 1998), where the state is a summary of the customer's history, and the action at each period is to either mail or not mail. The construction of the state space is an interesting problem that we will not consider here. We will instead follow a standard industry approach to this problem that uses three state variables, the so-called "RFM" measures (e.g., Bult and Wansbeek 1995, Bitran and Mondschein 1996). These measures describe the recency, frequency, and monetary value of customers' prior purchases. "Recency" is measured as the number of days (in hundreds) since a customer's last purchase. "Frequency" measures the number of items that customers previously purchased. "Monetary value" measures the average price (in dollars) of the items ordered by each customer.

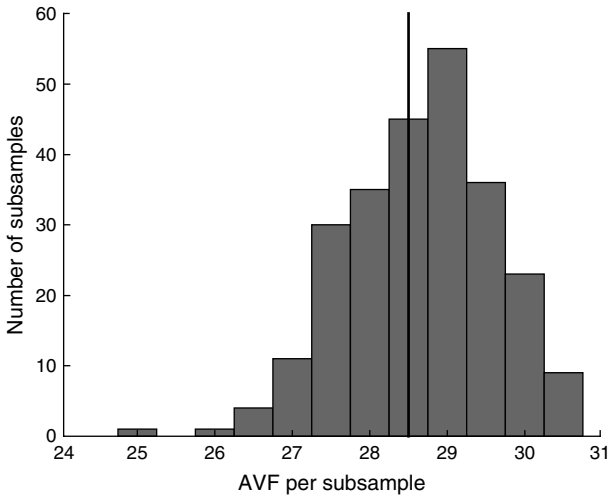
For the purposes of this illustration, we constructed a state space by quantizing each of the RFM variables to four discrete levels, yielding a state space with $|S| = 4^3 = 64$ states. At each historical mailing epoch, we evaluate the RFM variables of each customer (regardless of whether the customer received a catalog or made a purchase) and characterize him/her into one of the 64 states. We also treat the purchase amount (zero if no purchase in the epoch) less the mailing cost as a reward sample. Therefore, each customer's historical data over time serves as a sample trajectory. Following the procedure described in the previous section, we may then estimate the model parameters $\hat{P}$ and $\hat{R}$ and calculate $\hat{Y}$ for the current policy embedded in data.

Because the firm is interested in the average profit per customer rather than the profit earned from an individual customer, internal variance averages out. However, parametric variance is of interest because it affects the comparison of different policies. In particular, when evaluating a new policy, the firm would like both a prediction of the expected profits from adopting the new policy, together with confidence bounds around that prediction.

To illustrate the impact of parametric variance, we randomly divided the 1.72 million customers and 164 million observations into 250 equally-sized subsamples, each containing approximately 657,000 observations. By "observation," we mean a mailing period and an associated state transition in the history of a customer, irrespective of whether a catalog was mailed or a purchase was made during that time period. We then separately estimated the model parameters $\hat{P}^\pi$ and $\hat{R}^\pi$ following §2 using the observations from each of these subsamples. Here, we considered the policy π to be the same as the sampling policy that generated the data. Using Equation (3), we calculated 250 estimates of the value function. As a benchmark, we also estimated the model parameters using the full sample of 1.72 million customers. For the purposes of this illustration, we will interpret the model estimated using the full sample as the "true" model, which is essentially equivalent to assuming that the 1.72 million customers are the full population. Thus, within a typical subsample, the expected reward in each state $\hat{R}^\pi$ was estimated using an average of approximately 10,000 observations (N_i), while the transition matrix $\hat{P}^\pi$ was estimated using an average of 160 observations per transition. In practice, most of the transitions are infeasible; for example, a customer cannot transition from having three prior purchases to only having two prior purchases. When limiting attention to only those transitions that are feasible, the average number of observations per transition was approximately 1,400. (The average of the positive N_{ij}^as is around 1,400.)

In Figure 1, we report the empirical distribution (histogram) of the value function $\hat{Y}^\pi$ across all 250 subsamples under the historical policy used by the firm (as calculated using the whole sample). To summarize an estimated value function with a single number, we average the estimates across states for each subsample, weighing each state equally. We will refer to this measure as the *average value function* (AVF). We use equal weights to increase the clarity of illustration. By using equal weights (as with any fixed weights), we avoid potentially introducing an additional source of variance due to the weights themselves being random variables. The true AVF, computed from the parameters estimated for the full sample, is \$28.54. In comparison, the average of the 250 estimates is \$28.65, with an empirical standard deviation of \$0.97. The difference between \$28.54 and \$28.65 is not statistically significant and is of seemingly little managerial importance. However, the variance is potentially very important. The 95% confidence interval around the 250 AVF estimates ranges from \$26.59 to \$30.49, or roughly 14% of the true mean. Of course, we were able to estimate the \$0.97 standard deviation only because we had access to many subsamples. In a real-world setting, where only a single sample is available, the researcher generally relies on simulations or jack-knifing techniques to estimate the standard deviation. In this paper, we will present a procedure for deriving closed-form approximations of the standard deviation directly from the data.

Figure 1 **Mail Catalog Problem: A Histogram of the AVF of the Historical Policy for a Partition of the Customers to 250 Subsamples**



Note. The discount factor per period is $\alpha = 0.98$. The policy used is the historical (mixed) policy used by the firm, and the value function is weighted uniformly across states. The AVF obtained from the full data is \$28.54, and is plotted as a vertical line. The empirical standard deviation is \$0.97.

We can demonstrate the robustness of the above described results by varying both the size of the subsamples and the discount factor. In Table 1, we present the empirical bias and standard deviation for different discount factors (averaged over 10 repetitions). In each repetition, we divide the data set into 100 subsamples and compute the AVF for each subsample. We calculate the average absolute value of the bias and the empirical standard deviation of the AVF estimates across subsamples. It can be seen that the average bias is small for discount factors that are not too close to one. For discount factors that are close to one, the bias becomes more meaningful but still remains much smaller than the standard deviation. In another experiment, we varied the precision of the estimates by changing the size of the

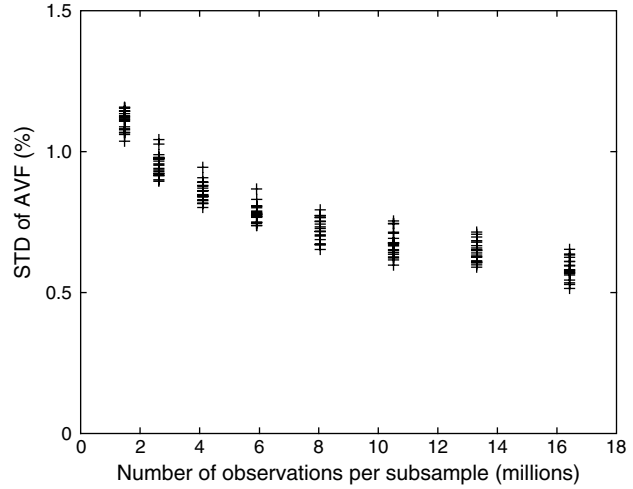
Table 1 Bias and Variance as a Function of the Discount Factor

α	Bias/AVF (%)	STD/AVF (%)
0.500	0.12	3.57
0.900	0.14	3.37
0.925	0.15	3.32
0.950	0.19	3.26
0.980	0.42	3.33
0.991	0.99	3.88
0.996	2.38	5.26

Note. For each discount factor, we partition the data 10 times, with each partition resulting in 100 subsamples (each with roughly 1.6 million observations). In the table, we present the mean absolute value of the bias and the mean empirical standard deviation each averaged across the 10 repetitions. Both of these means are standardized by dividing by the AVF associated with the historical policy (as measured on the whole data set).

subsamples and repeated the analysis using subsamples with a different number of observations. In Figure 2, we report empirical standard deviations of the AVF estimates under the different-sized subsamples. Each cross in Figure 2 represents a random assignment of the observations to subsamples (the different assignments led to variation in the subsamples between repetitions). While increasing the size of the subsamples increases the accuracy of the model parameters, and in turn reduces the variance in the AVF estimates, the rate at which the variance approaches zero slows down as the subsamples increase in size. It seems that even when estimating the model parameters with very large amounts of data, parametric variance leads to nonnegligible variance in the value function estimates.

Figure 2 Mail Catalog Problem: The Empirical Standard Deviation of the AVF as a Function of the Sample Size



Note. Each cross represents a single (random) partition of the observations into subsamples.

4. Analysis

In this section, we provide closed-form approximations for the bias and variance of the estimated value function using second-order approximations. We then briefly discuss the control problem where in addition to the estimation process, we look for an optimal policy. In §4.1, we will drop the superscript π because we consider a fixed policy π.

4.1. Approximations for Bias and Variance in the Estimated Value Function

We now derive closed-form approximations for the (parametric) bias and variance of $\widehat{Y}$. The analysis follows a classical (non-Bayesian) approach, where the bias and variance are expressed in terms of the (unknown) true parameters. Because the true model parameters are unknown, we substitute the estimated parameters, which is a standard practice. However, as a result of this substitution, the values obtained for the bias and variance are themselves estimates.

For completeness, we also provide in Appendix D a Bayesian analysis. Under the Bayesian approach, P and R are treated as random variables with known prior distributions, and we deduce approximations for the conditional bias and variance, given the values of $\hat{P}$ and $\hat{R}$. The expressions obtained using the Bayesian approach are almost identical to the ones in the classical approach (unless an informative prior is available).

Our goal is to calculate $\mathbb{E}[\widehat{Y}]$ and the covariance matrix for $\widehat{Y}$, defined by

$$\operatorname{cov}(\widehat{Y}) = \mathbb{E}[\widehat{Y}\widehat{Y}^{\top}] - \mathbb{E}[\widehat{Y}]\mathbb{E}[\widehat{Y}]^{\top}.$$

We define a random $m \times m$ matrix $\tilde{P} = \hat{P} - P$ and a random m-vector $\tilde{R} = \hat{R} - R$. Note that $\tilde{P}$ and $\tilde{R}$ are zero mean random variables that represent the difference between the true model and the estimated model.

To help interpret some of the later analysis, it will be helpful to have a sense of the magnitudes of $\tilde{P}$ and $\tilde{R}$. Because the transition probabilities are bounded by zero and one, the errors in these probabilities are also bounded between zero and one. The transition probabilities themselves will tend to be smaller the larger the number of states to which transitions are feasible, while the errors in these probabilities will be smaller the more observations there are relative to the number of feasible transitions. In the example discussed in §3 and Figure 1, the maximum error in the transition probabilities in a subsample ($\max_{ij} |\tilde{P}_{ij}|$) has a mean of 0.011 and a standard deviation of 0.004. Furthermore, the average (averaged over all pairs (i, j) with a nonzero transition probability) absolute error in the transition probability estimates, $|\tilde{P}_{ij}|$, has a mean of 6.3×10^{-4} and an empirical standard deviation of

0.001. Note that in that example, the feasible transitions consist of less than 10% of the 64^2 entries in P. The expected rewards are not bounded a priori and so the errors are also unbounded. In the catalog example, the average absolute error in the reward estimates, $|\tilde{R}_{ij}|$, has a mean of \$4.25 and a standard deviation of \$1.82. The maximal error in the reward estimates, $\max_{ij}|\tilde{R}_{ij}|$, has a mean of \$56.3 and a standard deviation of \$43.2.

We now write the expectation of $\hat{Y}$ (cf. Equation (3)) as

$$\begin{aligned}\mathbb{E}[\hat{Y}] &= \mathbb{E}[(I-\alpha(P+\tilde{P}))^{-1}(R+\tilde{R})] \\ &= \mathbb{E}\left[\sum_{k=0}^{\infty}\alpha^k(P+\tilde{P})^k(R+\tilde{R})\right], \end{aligned} \tag{4}$$

where the geometric series expansion of $(I-\alpha(P+\tilde{P}))^{-1}$ was used to obtain the second equality. We use the notation $X=(I-\alpha P)^{-1}$ and $f_k(\tilde{P})=X(\tilde{P}X)^k=(X\tilde{P})^kX$. The following lemma will be useful.

Lemma 4.1. $\sum_{l=0}^{\infty}\alpha^l(P+\tilde{P})^l=\sum_{k=0}^{\infty}\alpha^k f_k(\tilde{P})$.

Proof.

$$\begin{aligned}\sum_{k=0}^{\infty}\alpha^k f_k(\tilde{P}) &= \sum_{k=0}^{\infty}\alpha^k(X\tilde{P})^kX=(I-\alpha X\tilde{P})^{-1}X \\ &= (X^{-1}-X^{-1}\alpha X\tilde{P})^{-1}=(I-\alpha P-\alpha\tilde{P})^{-1} \\ &= \sum_{l=0}^{\infty}\alpha^l(P+\tilde{P})^l,\end{aligned}$$

where we repeatedly used the definition of X and the fact that X is invertible. □

Using Lemma 4.1 in Equation (4), we obtain

$$\begin{aligned}\mathbb{E}[\hat{Y}] &= (I-\alpha P)^{-1}R+\left(\sum_{k=1}^{\infty}\alpha^k\mathbb{E}[f_k(\tilde{P})]\right)R \\ &\quad+\sum_{k=0}^{\infty}\alpha^k\mathbb{E}[f_k(\tilde{P})\tilde{R}].\end{aligned} \tag{5}$$

There are three terms on the right-hand side of Equation (5). The first term is the value function for the true model. The second term reflects the bias introduced by the uncertainty in $\hat{P}$ alone, and the third term represents the bias introduced by the correlation between the errors in $\hat{P}$ and $\hat{R}$.

Equation (5) provides a series expansion of the error in terms of high-order moments and cross moments of the errors in $\hat{P}$ and $\hat{R}$. The calculation of the bias is tedious because the term $\mathbb{E}[f_k(\tilde{P})]$ involves kth-order moments of multinomial distributions. But because $|\tilde{P}_{ij}|$ is typically small, $\tilde{P}^k$ is generally close to zero for large k. For this reason, we limit our attention to a second-order approximation and we will assume that $\mathbb{E}[f_k(\tilde{P})]\approx 0$ for $k>2$, and that $\mathbb{E}[f_k(\tilde{P})\tilde{R}]\approx 0$ for $k>1$. We use the catalog data to investigate the appropriateness of this assumption in §5. Therefore, we can write Equation (5) as

$$\begin{aligned}\mathbb{E}[\hat{Y}] &= (I-\alpha P)^{-1}R+\alpha\mathbb{E}[f_1(\tilde{P})]R+\alpha^2\mathbb{E}[f_2(\tilde{P})]R \\ &\quad+X\mathbb{E}[\tilde{R}]+\alpha\mathbb{E}[f_1(\tilde{P})\tilde{R}]+L_{\exp},\end{aligned} \tag{6}$$

where we represent all the terms of order greater than two in

$$L_{\exp}=\sum_{k=3}^{\infty}\alpha^k\mathbb{E}[f_k(\tilde{P})]R+\sum_{k=2}^{\infty}\alpha^k\mathbb{E}[f_k(\tilde{P})\tilde{R}]. \tag{7}$$

Given that we will be using second-order approximations, we expect that the mean and variance of $\hat{Y}$ can be calculated as long as we are able to compute the covariance between various entries of $\tilde{R}$ and $\tilde{P}$.

We start with $\tilde{P}$. First, we introduce some notation. We use the notation $A_{i\cdot}$ and $A_{\cdot i}$ to denote the ith row and column, respectively, of a matrix A, and $\mathrm{diag}(A_{i\cdot})$ to denote a diagonal matrix with the entries of $A_{i\cdot}$ along the diagonal. We note that $\tilde{P}_{i\cdot}$ and $\tilde{P}_{j\cdot}$ are independent when $i\neq j$. To find the covariance matrix of $\tilde{P}_{i\cdot}$, we consider the row vectors $\hat{P}^a_{i\cdot}$ and $P^a_{i\cdot}$ with the estimated and true transition probabilities, and define $\tilde{P}^a_{i\cdot}$ to be their difference. Note that

$$P_{i\cdot}=\sum_a\pi(a\mid i)P^a_{i\cdot}.$$

For each state-action pair (i,a), we define

$$M^a_i=\mathrm{diag}(P^a_{i\cdot})-(P^a_{i\cdot})^\top P^a_{i\cdot},$$

which is a symmetric positive semidefinite matrix. Recall that for each (i,a), we have $\hat{P}^a_{ij}=N^a_{ij}/N^a_i$, where the N^a_{ij} are drawn from a multinomial distribution. The covariance matrix of $\hat{P}^a_{i\cdot}$ is M^a_i/N^a_i, and the covariance matrix of $\hat{P}_{i\cdot}$ is

$$\mathrm{cov}^{(i)}=\mathbb{E}[\tilde{P}^\top_{i\cdot}\tilde{P}_{i\cdot}]=\sum_a\frac{\pi(a\mid i)^2}{N^a_i}M^a_i.$$

Now we consider $\tilde{R}$. Because C_{ij} is independent of C_{kl} whenever $i\neq k$, we have $\mathbb{E}[\tilde{R}_i\tilde{R}_k]=0$ for $i\neq k$. Furthermore,

$$\mathbb{E}[\tilde{R}_i^2]=\sum_a\pi(a\mid i)^2\mathbb{E}[(\tilde{R}^a_i)^2].$$

In the following, we use $N^a_{i\cdot}$ to represent the vector with components N^a_{ij}, $j=1,\ldots,m$, and $R^a_{i\cdot}$ to represent the vector with components R^a_{ij}, $j=1,\ldots,m$. Note that C_{ij} and C_{ik} are independent given N_{ij} and N_{ik}, so that

$$\begin{aligned}\mathbb{E}[(\tilde{R}^a_i)^2] &= \mathrm{var}\left[\frac{\sum_j C^a_{ij}}{N^a_i}\right]=\frac{1}{(N^a_i)^2}\mathrm{var}\left[\sum_j C^a_{ij}\right] \\ &= \frac{1}{(N^a_i)^2}\left\{\mathrm{var}\left(\mathbb{E}\left[\sum_j C^a_{ij}\mid N^a_{i\cdot}\right]\right)\right. \\ &\quad\left.+\mathbb{E}\left(\mathrm{var}\left[\sum_j C^a_{ij}\mid N^a_{i\cdot}\right]\right)\right\}\end{aligned}$$

$$= \frac{1}{(N_i^a)^2}\left\{\operatorname{var}\left(\sum_j R_{ij}^a N_{ij}^a\right) + \mathbb{E}\left[\sum_j V_{ij}^a N_{ij}^a\right]\right\}$$

$$= \frac{1}{(N_i^a)^2}\left\{(N_i^a)^2 \operatorname{var}(R_{i\cdot}^a(\hat{P}_{i\cdot}^a)^\top) + N_i^a \sum_j V_{ij}^a P_{ij}^a\right\}$$

$$= \frac{1}{N_i^a}(R_{i\cdot}^a M_i^a R_{i\cdot}^{a\top} + V_{i\cdot}^a P_{i\cdot}^{a\top}). \quad (8)$$

Here, V_{ij}^a is the variance of the rewards associated with a transition from i to j, under action a.

To account for the correlation between $\tilde{P}$ and $\tilde{R}_i$, we use Equation (2) to obtain

$$\hat{R}_i = \sum_a \pi(a \mid i) \sum_j \hat{R}_{ij}^a \hat{P}_{ij}^a$$

$$= \sum_a \pi(a \mid i) \sum_j (R_{ij}^a P_{ij}^a + R_{ij}^a \tilde{P}_{ij}^a + \tilde{R}_{ij}^a P_{ij}^a + \tilde{R}_{ij}^a \tilde{P}_{ij}^a), \quad (9)$$

where $\tilde{R}_{ij}^a = \hat{R}_{ij}^a - R_{ij}^a$. Comparing with Equation (1), we have

$$\tilde{R}_i = \hat{R}_i - R_i = \sum_a \pi(a \mid i) \sum_j (R_{ij}^a \tilde{P}_{ij}^a + \tilde{R}_{ij}^a P_{ij}^a + \tilde{R}_{ij}^a \tilde{P}_{ij}^a). \quad (10)$$

We use $\circ$ to denote Hadamard multiplication: for any two matrices A and B with the same dimensions, $(A \circ B)$ is a matrix (again with the same dimensions) with entries $(A \circ B)_{ij} = A_{ij} B_{ij}$. We also use e to denote the m-dimensional vector with all components equal to one, and we use π^a to denote the m-dimensional vector with the ith component being $\pi(a \mid i)$. With this notation, Equation (10) becomes

$$\tilde{R} = \left(\sum_a \pi^a \circ [(\tilde{P}^a \circ R^a + \tilde{R}^a \circ P^a + \tilde{R}^a \circ \tilde{P}^a)e]\right). \quad (11)$$

We define an $m \times m$ matrix Q with entries

$$Q_{ij} = \operatorname{cov}_{j\cdot}^{(i)} X_{\cdot i}. \quad (12)$$

(Recall the definition $X = (I - \alpha P)^{-1}$, and that $Y = XR$ is the true value function.) We define an m-dimensional vector B with its ith component defined as

$$B_i = \sum_a \frac{\pi(a \mid i)^2}{N_i^a} R_{i\cdot}^a M_i^a X_{\cdot i}.$$

The following proposition quantifies the bias under the second-order approximation assumption. The proof is given in Appendix A.

Proposition 4.1. *The expectation of the estimated value function $\hat{Y}$ satisfies*

$$\mathbb{E}[\hat{Y}] = Y + \alpha^2 XQY + \alpha XB + L_{\exp},$$

where $L_{\exp}$ is defined in Equation (7) *and*

$$L_{\exp} = o\left(\frac{1}{N_{i^*}^{a^*}}\right),$$

where $N_{i^*}^{a^*} = \min_{(i,a):\,\pi(a\mid i)>0} N_i^a$ and the term $o(\cdot)$ satisfies $\lim_{N\to\infty} o(1/N)\cdot N = 0$.

In the above proposition, (i^*, a^*) represents the least sampled state-action pair that is used by the policy. The term $L_{\exp}$ decreases to zero faster that $1/N_{i^*}^{a^*}$, whereas Q and B can be shown to decrease like $1/N_{i^*}^{a^*}$. Therefore, our approximation for the bias in the value function estimates will be

$$\alpha^2 XQY + \alpha XB.$$

For the purposes of the next proposition, we introduce more notation. We define the diagonal matrix W, whose diagonal entries are given by

$$W_{ii} = \sum_a \frac{\pi(a \mid i)^2}{N_i^a}[(\alpha Y^\top + R_{i\cdot}^a) \cdot M_i^a(\alpha Y + R_{i\cdot}^{a\top}) + V_{i\cdot}^{a\top} P_{i\cdot}^a]. \quad (13)$$

The next proposition provides an expression for the second moment, $\mathbb{E}[Y\hat{Y}^\top]$. Together with the expression for $\mathbb{E}[\hat{Y}]$ in the preceding proposition, it leads to an approximation for the covariance matrix of $\hat{Y}$. The proof is given in Appendix B.

Proposition 4.2. *The second moment of $\hat{Y}$ satisfies*

$$\mathbb{E}[\hat{Y}\hat{Y}^\top] = YY^\top + X\{\alpha^2(QYR^\top + RY^\top Q^\top) + \alpha(BR^\top + RB^\top) + W\}X^\top + L_{\text{var}},$$

where L_{var} is given by

$$L_{\text{var}} = \sum_{k,l:\,k+l>2} \alpha^{k+l}\mathbb{E}[f_k(\tilde{P})(RR^\top + (\tilde{R})(\tilde{R})^\top) f_l(\tilde{P})^\top] + \alpha\mathbb{E}[X(\tilde{R})(\tilde{R})^\top f_1(\tilde{P})^\top] + \alpha\mathbb{E}[f_1(\tilde{P})(\tilde{R})(\tilde{R})^\top X^\top] = o\left(\frac{1}{N_{i^*}^{a^*}}\right).$$

By taking the difference between $\mathbb{E}[\hat{Y}\hat{Y}^\top]$, as given by Proposition 4.2, and $\mathbb{E}[\hat{Y}]\mathbb{E}[\hat{Y}^\top]$, as prescribed by Proposition 4.1, the following corollary is easily derived.

Corollary 4.1. *The covariance matrix of the estimated value function satisfies*

$$\operatorname{cov}(\hat{Y}) = XWX^\top + o\left(\frac{1}{N_{i^*}^{a^*}}\right).$$

The expressions in Propositions 4.1 and 4.2 and Corollary 4.1 yield several insights. First, as the counts N_i^a increase to infinity, $\operatorname{cov}^{(i)}$ approaches zero, and thus all the terms involving the matrices Q, B, and W converge to zero. As expected, this implies that as the sample size increases and the accuracy of the estimated parameters improves, both the bias and the variance decrease to zero. Second, the expressions for the bias and variance rely on the true model parameters, which are unknown. As discussed in the introduction, to obtain computable approximations of the

bias and variance, we will use instead $\hat{P}$, $\hat{R}$, and the empirical variance of each R^a_{ik}. In principle, we could also estimate the bias and variance due to this approximation, but this is tedious and, as suggested by the experimental results in the next section, generally unnecessary. Third, when $\min_{i,a} N^a_i$ is large, it follows that the nonzero entries of B, W, and Q decrease to zero like $1/N^{a^*}_{i^*}$. Therefore, the standard deviation decreases to zero like $1/\sqrt{N^{a^*}_{i^*}}$, which is the usual behavior of empirical estimates.

The expressions in Proposition 4.1 and Corollary 4.1 allow us to qualitatively compare the magnitude of the bias and variance. According to Corollary 4.1, the standard deviation of $\widehat{Y}_i$ can be approximately estimated as

$$\sigma(\widehat{Y}_i) = \sqrt{X_{i\cdot} W X_{i\cdot}^\top}. \tag{14}$$

The next proposition, proved in Appendix C, quantifies the ratio between the standard deviation and the bias. Recall that for two positive functions f and g (defined on the real numbers), we write $f(n) = \Omega(g(n))$ if there exist constants N_0 and $C > 0$ such that $f(n) \geq Cg(n)$ for $n \geq N_0$.

Proposition 4.3. *Suppose that* $\sigma(\widehat{Y}_i) > 0$ *and* $N^{a^*}_{i^*}/N^a_i > c > 0$ *for all* a *and* i. *Then,*

$$\frac{\sigma(\widehat{Y}_i)}{|\mathbb{E}[\widehat{Y}_i] - Y_i|} = \Omega\left(\sqrt{N^{a^*}_{i^*}}\right) \qquad \textit{for all } i.$$

Proposition 4.3 implies that the errors introduced by the parametric variance will generally be much larger than the bias. Note that because W is a positive semidefinite matrix, $\sigma(\widehat{Y}_i) > 0$ is a very weak nondegeneracy assumption. The condition $N^{a^*}_{i^*}/N^a_i > c > 0$ requires that sample sizes increase "uniformly."

While the expression in Corollary 4.1 allows us to approximate the covariance matrix of the estimated value function, the findings on their own do not allow us to calculate confidence intervals around these estimates. Calculating a confidence interval requires that we know the distribution of the value function estimates. A central limit theorem (Serfling 1980, p. 122, Theorem A) speaks to this issue.

Theorem 4.1 (Serfling 1980). *Suppose that a sequence of random vectors* $\{\mathbf{X}_n := (X_{n1}, \ldots, X_{nk})\}$ *is* $\mathcal{AN}(\boldsymbol{\mu}, b_n^2\boldsymbol{\Sigma})$ (*asymptotically normal, that is,* $(\mathbf{X}_n - \boldsymbol{\mu})/b_n \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$); *and the sequence of scalars* b_n *converges to* 0. *Let* $g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))$, $\mathbf{x} = (x_1, \ldots, x_k)$, *be a vector-valued function for which each component function* $g_i(\mathbf{x})$ *is a real-valued function and has a nonzero gradient at* $\mathbf{x} = \boldsymbol{\mu}$. *Let*

$$\mathbf{D} = \left[\left.\frac{\partial g_i}{\partial x_j}\right|_{x=\mu}\right]_{m\times k}.$$

Then, $g(\mathbf{X}_n)$ *is* $\mathcal{AN}(g(\boldsymbol{\mu}), b_n^2\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top)$.

Because $\hat{P}^a_{ij}$ and $\hat{R}^a_{ij}$ are all estimators that asymptotically follow normal distributions, we may consider $\widehat{Y}$ as the function g in the above theorem and conclude that $\widehat{Y}$ is asymptotically normal. We further investigate this issue using catalog mailing data in §5, where we report that a Kolmogorov-Smirnov test cannot reject the hypothesis that $\widehat{Y}$ is normally distributed.

Readers may wonder whether we could have used Serfling's (1980) result to derive our earlier findings. It is technically possible to do so. Indeed, under the assumption that all of the N^a_is are identical, we were able to show that the two approaches yield the same result, and observed that the two derivations were of comparable length and complexity. However, if sampling occurs at different rates in different states, the rate at which the N^a_is approach infinity will generally vary. In this case, use of the Serfling theorem, or any related central limit theorem, requires extensive additional derivation. Moreover, these theorems do not address the issue of bias.

The same approach can be used for infinite-horizon average reward MDPs. Under mild assumptions on the structure of the Markov chain, we get similar approximations to the bias and variance for the average reward. The idea can also be extended to semi-Markov processes, where the transition times between time epochs are random and estimated from sampled data.

4.2. The Control Problem

To this point, we have focused on the value function under a fixed policy. In many applications, we are interested in comparing an existing policy with an alternative policy, possibly derived through a policy optimization process. We know from the MDP theory that there exists an optimal policy $\boldsymbol{\pi}^*$ such that $Y_i^{\pi^*} \geq Y_i^{\pi}$ for all admissible policies $\boldsymbol{\pi}$ and all states $i \in S$. An optimal policy may be obtained using value iteration, policy iteration, or linear programming algorithms. See, for example, Bertsekas (2000).

Because we do not have access to the true model parameters P and R, optimization based on the estimated parameters $\hat{P}$ and $\hat{R}$ produces an "optimal" policy $\hat{\boldsymbol{\pi}}$ such that $\widehat{Y}^{\hat{\pi}} \geq \widehat{Y}^{\pi}$ for all admissible policies $\boldsymbol{\pi}$. In general, policy $\hat{\boldsymbol{\pi}}$ is different from $\boldsymbol{\pi}^*$. Moreover, because the policy $\hat{\boldsymbol{\pi}}$ is obtained through an optimization process, the estimates of the model parameters for that policy ($\hat{P}^{\hat{\pi}}$ and $\hat{R}^{\hat{\pi}}$) will no longer be unbiased estimates of the true model parameters ($P^{\hat{\pi}}$ and $R^{\hat{\pi}}$). Therefore, we cannot use the approximation derived in Proposition 4.1 (for a fixed policy) to evaluate the bias in the optimal value function—nor can we use the approximations in Proposition 4.2 and Corollary 4.1 to estimate the covariance matrix.

We can illustrate the problem through a simple example. Consider a single-state MDP with two actions, that is, $S = \{1\}$ and $A = \{0, 1\}$. Both actions yield identical zero-mean random rewards. Clearly, in such a problem π^* could be either action 0 or 1, with value functions

$$Y^{\pi^*} = Y^{\hat{\pi}} = 0.$$

Now assume that we have n samples to estimate the expected reward $\hat{R}^a$ for either action. Indeed, both $\hat{R}^a$ follow (approximately) a normal distribution $\mathcal{N}(0, 1/n)$. The policy optimization procedure chooses the action with the largest $\hat{R}^a$. If we use $\hat{R}^*$ to denote the maximum of $\hat{R}^0$ and $\hat{R}^1$, we know from Jensen's inequality that $\mathbb{E}[\hat{R}^*] > 0$, and so the value function estimated for the chosen policy will on average be positively biased:

$$\begin{aligned}\mathbb{E}[\hat{Y}^{\hat{\pi}}] = \mathbb{E}[\hat{R}^*] &= \mathbb{E}[\max\{\hat{R}^0, \hat{R}^1\}] \\ &> \max\{\mathbb{E}[\hat{R}^0], \mathbb{E}[\hat{R}^1]\} = 0.\end{aligned}$$

The magnitude of $\mathbb{E}[\hat{Y}^{\hat{\pi}}]$, and therefore the bias in this example, is studied in the order statistics literature (Leadbetter et al. 1983). We also refer readers to Clark (1961), which presents a procedure to approximate moments of the maximum of a finite number of correlated Gaussian random variables.

This problem raises two issues. First, how can we de-bias the estimates of $\hat{P}^{\hat{\pi}}$ and $\hat{R}^{\hat{\pi}}$ so that we can use our earlier results to estimate the bias and covariance matrix of a value function when the policy is derived from an optimization procedure? Second, because the optimization procedures themselves rely on estimates $\hat{P}^{\pi}$ and $\hat{R}^{\pi}$, the policies derived from standard dynamic programming algorithms will generally not be truly optimal ($\hat{\pi} \neq \pi^*$). In the remainder of this section, we propose a cross-validation approach that can help to address the first issue. Unfortunately, we do not have a solution to the second issue. Indeed, it seems unlikely that a general procedure can be found that resolves the second issue as the suboptimality reflects the absence of complete information in the training data.

The bias in the estimates of $\hat{P}^{\hat{\pi}}$ and $\hat{R}^{\hat{\pi}}$ arises because optimization methods tend to favor actions for which the estimation errors in $\hat{P}^{\pi}$ and $\hat{R}^{\pi}$ lead to inflated estimates of the value function. As long as the errors in $\hat{P}$ and $\hat{R}$ are independent across samples, we can derive unbiased estimates of P and R if we use a different sample of data to evaluate the policy $\hat{\pi}$ than the sample we used to design the policy. In particular, consider the following approach: Start by dividing the training data into two subsamples—a calibration sample and a validation sample. Use the calibration sample to estimate the model parameters $\hat{P}_{\text{cal}}$ and $\hat{R}_{\text{cal}}$ and obtain the "optimal" policy

$$\hat{\pi}_{\text{cal}} = \arg\max_{\pi}(I - \alpha\hat{P}^{\pi}_{\text{cal}})^{-1}\hat{R}^{\pi}_{\text{cal}}.$$

Then, estimate model parameters $\hat{P}_{\text{val}}$ and $\hat{R}_{\text{val}}$ from the validation sample and (following Equation (3)) evaluate the policy using these new parameters:

$$\hat{Y}^{\hat{\pi}_{\text{cal}}}_{\text{val}} = \left(I - \alpha\hat{P}^{\hat{\pi}_{\text{cal}}}_{\text{val}}\right)^{-1}\hat{R}^{\hat{\pi}_{\text{cal}}}_{\text{val}}.$$

Through this procedure, we can de-bias the value function estimates by reporting $\hat{Y}^{\hat{\pi}_{\text{cal}}}_{\text{val}}$ instead of $\hat{Y}^{\hat{\pi}_{\text{cal}}}_{\text{cal}}$, where $\hat{Y}^{\hat{\pi}_{\text{cal}}}_{\text{cal}} = (I - \alpha\hat{P}^{\hat{\pi}_{\text{cal}}}_{\text{cal}})^{-1}\hat{R}^{\hat{\pi}_{\text{cal}}}_{\text{cal}}$. Accordingly, we may also approximate the bias and variance and therefore the confidence bounds of $\hat{Y}^{\hat{\pi}_{\text{cal}}}_{\text{val}}$ following Proposition 4.1 and Corollary 4.1.

The assumption that the estimation errors in $\hat{P}$ and $\hat{R}$ are independent across the calibration and validation subsamples is obviously critical. In this paper, we have assumed that estimates $\hat{P}$ and $\hat{R}$ are derived from straightforward nonparametric aggregates of the available data. Under this approach, the estimation errors are independent across the subsamples as long as any measurement errors are independent across observations. However, in some settings, it is common to estimate the model parameters from maximum likelihood estimates that require functional form and distribution assumptions (this is particularly common in the economics literature). Under this alternative approach, any errors introduced by the functional form and distribution assumptions will be correlated across the subsamples. As a result, the cross-validation procedure that we have proposed will not de-bias the estimates of $\hat{P}^{\hat{\pi}}$ and $\hat{R}^{\hat{\pi}}$, even if the measurement errors are independent across the observations.

5. Experiments

The reliance on a second-order expansion in deriving the approximations for the bias and variance presumes that higher-order terms are relatively unimportant. We now examine this assumption in further detail by using the catalog mailing data to validate the findings. These data also enable us to investigate the impact (if any) of using estimates of the model parameters in these expressions (in the absence of the true model parameters).

If the value function estimates follow a normal distribution, the variance and bias expressions derived in the previous section facilitate calculation of confidence intervals around the de-biased value function estimates. We can investigate the accuracy of these confidence intervals by comparing how frequently the "true" value function falls within the confidence intervals. We would expect that on average the true value

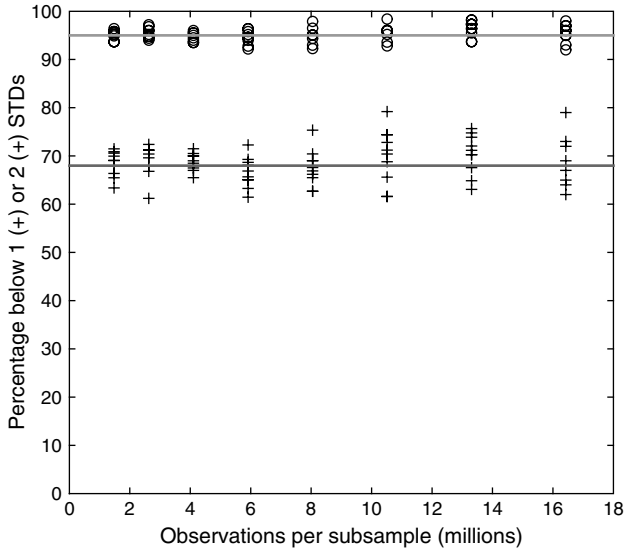
will fall within one standard deviation of the unbiased mean 68% of the time and within two standard deviations 95% of the time.

We begin by investigating whether the value function estimates follow a normal distribution. We do so by using a Kolmogorov-Smirnov test on each of the data points reported in §3. The hypothesis that the reward is a two-sided Gaussian could not be rejected with confidence 0.05 at any instance. The average p-value was 0.612 with a minimum of 0.061 and a maximum of 0.991. This indicates that it cannot be determined that the data do not follow a Gaussian rule.

We use the same partitions of the data as in §2. In Figure 3, the percentage of times that the true value function was within one standard deviation is denoted by a "+" and within two standard deviations by an "o." For example, for the 250 subsamples (with about 657,000 observations each), we report the percentage of the 250 estimates in which the true AVF (as estimated on the full sample) was within the estimated confidence interval. By redrawing the 250 subsamples 10 times, we report 10 instances of this percentage. An analogous process was used with other choices of the subsample size. The findings in Figure 3 confirm that the percentage of estimates that fall within one and two standard deviations of the true AVF are close to the targets of 68% and 95%, respectively.

We next consider the importance of the second-order approximations. We do so by taking advantage of the role played by the discount factor α. The importance of higher-order terms in the series expansions increases as the discount factor approaches one. In Table 2, we repeat the analysis for 250 subsamples of a fixed size, but for different discount factors (same settings as in Table 1). As expected, as α approaches one, the accuracy of the confidence intervals degrades. We attribute this to the error introduced by the second-order approximation.

Figure 3 **The Percentage of the AVF Estimates that Fall Within One ("+") and Two ("o") Standard Deviations from the Value Calculated Based on the Full Data Set**



Note. Each "+" and "o" represents a random partition of the full data to subsamples. The discount factor is $\alpha = 0.98$.

Table 2 **Percentage of the AVF Estimates that Fall Within One and Two Standard Deviations**

α	Samples with one STD (%)	Samples with two STDs (%)
0.500	67.68 (63.2-73.6)	95.44 (93.2-98.0)
0.900	69.12 (64.8-72.0)	94.84 (93.6-96.0)
0.925	68.12 (60.8-73.6)	95.08 (93.2-96.8)
0.950	67.88 (64.0-70.4)	94.76 (92.0-96.8)
0.980	68.84 (61.2-72.4)	95.52 (94.0-97.2)
0.991	66.60 (64.0-70.0)	94.92 (92.0-97.6)
0.996	63.04 (58.8-68.4)	92.20 (89.6-93.2)

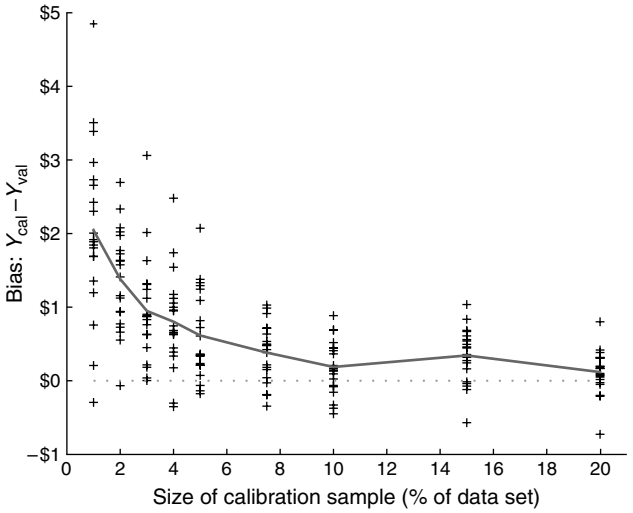
Note. We randomly partitioned the data while varying the discount factor. For each discount factor, we performed the partition 10 times; each partition was to 250 subsamples (each with roughly 657,000 million observations). We present the percentage of samples in which the estimated AVF is within one standard deviation (as predicted by Proposition 4.2) of the value as measured on all the data; the minimum and maximum percentages over the 10 runs are provided in parentheses. The same statistics are presented for two standard deviations.

5.1. The Control Problem

As discussed in §4.2, an obvious application of our analysis is the comparison of a current policy with a new policy generated through some optimization process. We cautioned that before applying the expressions for the bias and the variance to a policy derived from such a process, we should first obtain unbiased estimates of the model parameters, using an independent validation sample. We will use the catalog mailing data to illustrate the importance of this first step.

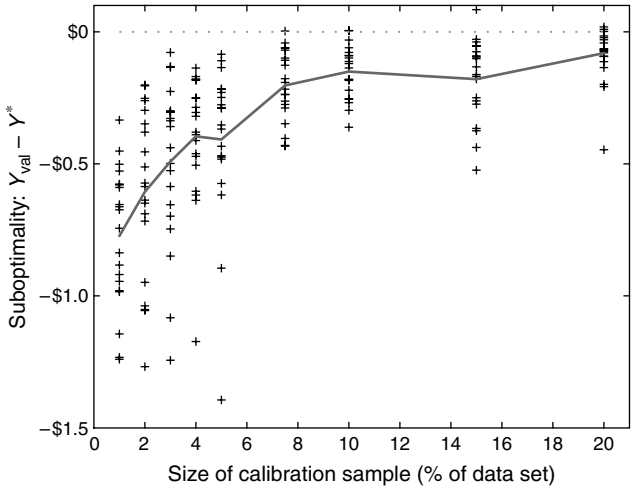
We begin by randomly selecting a portion of the available data to be used as a calibration sample, and retain the remaining data as a validation sample. To demonstrate how the size of the calibration sample affects the findings, we repeat this process for calibration samples of different sizes. The calibration sample is used to estimate model parameters $\hat{P}_{\text{cal}}$ and $\hat{R}_{\text{cal}}$. Then, we run a policy iteration algorithm to identify an "optimal" policy $\hat{\pi}_{\text{cal}}$ from $\hat{P}_{\text{cal}}$ and $\hat{R}_{\text{cal}}$. We will compare two AVF estimates for this policy: the AVF calculated on the basis of the model estimated using the calibration sample (denoted by Y_{cal}), and the AVF of that policy as estimated using the validation sample (denoted by Y_{val}). The difference between the two estimates represents the bias introduced by the error in the model parameters (the errors no longer have zero expectation due to the optimization process).

Figure 4 **The Differences (Marked by "+") Between the AVF Estimates (in Dollars, and Averaged Over All States) Based on the Calibration Sample and the Validation Sample for the Policy Identified Through an Optimization Process**



Note. Each "+" was generated by randomly partitioning the data to a calibration and a validation sample. The horizontal axis corresponds to the size of the calibration sample, as a percentage of the full data sample. Here, $\alpha = 0.98$ for which the true optimal AVF is approximately \$33.59.

Figure 5 **The Differences (Marked by "+") Between the AVF Estimates (in Dollars) of the Optimal Policy Based on the Calibration Sample and the AVF of the Optimal Policy Found by Optimizing on the Validation Sample**



Note. Each "+" was generated by randomly partitioning the data to a calibration and a validation sample. The horizontal axis corresponds to the size of the calibration sample, as a percentage of the full data sample. Here, $\alpha = 0.98$ for which the true optimal AVF is approximately \$33.59.

This bias is illustrated in Figure 4 for calibration samples of varying sizes. It can be seen that value function estimates from the calibration sample are almost uniformly greater than the estimates from the validation sample. This bias is statistically significant. It is also managerially relevant, averaging around 6.3% of the true optimal AVF (\$33.59) for a calibration sample that consists of approximately 1.6 million observations (1% of the data). In addition, the \$33.59 AVF for the optimal policy can be compared with the \$28.54 AVF for the historical policy (reported in Figure 1). These results indicate that the optimal policy offers a potential profit improvement of approximately 17%.

We can also use the catalog data to investigate the extent to which parametric variance leads to suboptimal policies. To do so, we compared the "optimal" policy derived using each subsample, with the true optimal policy derived using the entire data set. Both policies are evaluated on the validation sample. We use Y^* to denote the AVF for the optimal policy found by optimizing on the entire data set. The findings are reported in Figure 5. As expected, the optimal policy always outperforms the policy derived from the calibration subsample. The differences are again statistically significant. Note that the computation of Y^* and Y_{val} uses the same data, which may introduce correlation between the two quantities. This will tend to diminish our estimates of the "suboptimality." We also computed Y^*_{val}, the optimal AVF over the validation set, in place of Y^* for Table 3 and Figures 4 and 5. The results are similar.

To demonstrate the robustness of the findings, we performed an experiment similar to the one reported in Table 2. In Table 3, we present the bias and suboptimality introduced by the optimization process for different values of α. Specifically, the bias was calculated as $(Y_{cal} - Y_{val})/Y^*$; the suboptimality was calculated as $(Y_{val} - Y^*)/Y^*$. From Table 3, we can easily obtain the mean standard errors as the sample standard deviations divided by 10 (the square root of the sample size, 100). It is clear that both the bias and the suboptimality are generally significantly greater than zero, with the bias averaging around 2% of the AVF and the suboptimality averaging around 1%.

Table 3 **Optimization Bias for Different Values of α**

	Bias in percent		Suboptimality in percent	
α	Mean	STD	Mean	STD
0.500	1.19	1.45	−0.64	0.58
0.900	1.66	1.25	−0.84	0.61
0.925	1.59	1.45	−0.77	0.63
0.950	1.83	1.44	−0.96	0.70
0.980	1.59	1.42	−0.87	0.54
0.991	1.14	1.66	−0.69	0.63
0.996	0.42	1.85	−0.38	0.41

Note. For each discount factor, we performed a random sampling of the data 100 times. Each time we use a random calibration sample of 20% of the entire data set (each with roughly eight million observations) and the other 80% as a validation sample. We found the optimal policy in each such MDP and present in the table the bias, $Y_{cal} - Y_{val}$, normalized by Y^*. We also present the suboptimality, $Y_{val} - Y^*$, normalized similarly. The means of the biases are significantly greater than zero.

We conclude that parametric variance introduces two issues in policy optimization. First, the estimates of the transition probabilities and the rewards for the "optimal" policy are biased, leading to positive bias in the value function estimates. This problem can be remedied relatively easily by evaluating the policy on a separate validation sample. The second problem is more difficult to resolve: errors in the model parameters also lead to suboptimal policies. As we discussed in §4.2, this second problem is at least to some extent inevitable in the absence of the true model parameters.

There is an interesting question raised by a referee: Given a fixed amount of data, how do we divide it into the calibration and validation samples? Including more data in the calibration sample potentially leads to a better policy, while more data in the validation sample means a tighter confidence interval when we evaluate the policy. This trade-off can often be resolved empirically.

6. Concluding Remarks

We have provided closed-form approximations for the bias and variance of estimated value functions caused by uncertainty in the true model parameters. For small and mid-sized MDP models, the expressions can be easily calculated and used to evaluate existing policies or to compare new polices with existing ones. For the case where a new policy is derived through a policy optimization process, we also demonstrated how to remove the additional bias introduced by the optimization process, by using a validation sample.

The expressions are based on second-order approximations. Moreover, in the absence of the true model parameters, the expressions are evaluated by relying on estimates of the model parameters and are therefore themselves estimates, subject to parametric variance. We used a large sample of data from a catalog mailing company to investigate the impact of these approximations. The findings indicate that the confidence intervals obtained on the basis of the bias and variance expressions are reassuringly accurate.

Both the catalog mailing data and our theoretical analysis provide a comparison of the relative magnitude of the variance and the different biases. The variance introduced by parametric uncertainty is considerable, suggesting both practical and statistical importance. Of the two biases, only the bias in "optimal" policies introduced by the optimization process is significant. For a fixed policy, the bias introduced by parametric uncertainty will generally be negligible when compared to the variance.

While we report the average of the value functions (averaged over all states), the disaggregate results may also be of interest. In particular, the variance of the value functions for alternative policies could be used to guide future experimentation. Future experimentation may favor actions that might have a large effect on the variance of the value function estimate. In this manner, the findings may contribute to our understanding of the trade-off between exploration and exploitation. The findings may also help to improve the policy optimization process. The policy improvement portions of standard algorithms focus on point estimates of the value functions and overlook the variance around these estimates.

Finally, we caution that, as with all analyses of MDPs, our findings rely on an assumption that the data are sampled from a Markov process. In our experiments, we ensured satisfaction of this condition by sampling observations rather than trajectories (a trajectory here would be the complete history of a customer).

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

This research was partially supported by NSF Grant DMI-0322823. This paper has benefited from comments by workshop participants at Duke University, University of Pennsylvania, Washington University at St. Louis, the 2004 International Conference on Machine Learning, and the INFORMS Annual Meeting 2004. The authors are thankful to Yann Le Tallec for finding an error in a previous version and to the referees for constructive comments. The authors are especially thankful to the department editor for a detailed review and many constructive suggestions.

Appendix A

Proof of Proposition 4.1.

Lemma A.1. *For any action a, we have $\mathbb{E}[\tilde{P}^a] = 0$, $\mathbb{E}[\tilde{R}^a] = 0$, and furthermore, $\tilde{R}^a$ is uncorrelated with any function of $\tilde{P}^1, \ldots, \tilde{P}^{|A|}$, in which $|A|$ is the cardinality of the action space A.*

Proof. The property $\mathbb{E}[\tilde{P}^a] = 0$ is obvious. Let $\mathbf{N}$ stand for the collection of random variables N^b_{jk} for every j, k, and b. We have $\mathbb{E}[\hat{R}^a_{ij} - R^a_{ij} \mid \mathbf{N}] = 0$. The fact $\mathbb{E}[\tilde{R}^a] = 0$ follows by taking the unconditional expectation. Furthermore, because any $\tilde{P}^{a'}$ is completely determined by $\mathbf{N}$, it follows that for any function g, we have

$$\mathbb{E}[g(\tilde{P}^1, \ldots, \tilde{P}^{|A|})\tilde{R}^a_{ij} \mid \mathbf{N}] = g(\tilde{P}^1, \ldots, \tilde{P}^{|A|})\mathbb{E}[\tilde{R}^a_{ij} \mid \mathbf{N}] = 0.$$

By taking unconditional expectations, we obtain the last part of the lemma. □

For the proof of Proposition 4.1, we start from Equation (5) and substitute the expression from Equation (11) for $\tilde{R}$, to obtain

$$\mathbb{E}[\hat{Y}] = (I - \alpha P)^{-1}R + \left(\sum_{k=1}^{\infty} \alpha^k \mathbb{E}[f_k(\tilde{P})]\right)R$$

$$+\sum_{k=0}^{\infty}\alpha^k\mathbb{E}\left[f_k(\tilde{P})\left(\sum_a \pi^a\circ((\tilde{R}^a\circ\tilde{P}^a)e)\right)\right]$$
$$+\sum_{k=0}^{\infty}\alpha^k\mathbb{E}\left[f_k(\tilde{P})\left(\sum_a \pi^a\circ((\tilde{P}^a\circ R^a)e)\right)\right]$$
$$+\sum_{k=0}^{\infty}\alpha^k\mathbb{E}\left[f_k(\tilde{P})\left(\sum_a \pi^a\circ((\tilde{R}^a\circ P^a)e)\right)\right].$$

From Lemma A.1, terms that are linear in $\tilde{P}^a$ or $\tilde{R}^a$, as well as terms that involve products of entries of $\tilde{P}^a$ and $\tilde{R}^{a'}$, vanish. That is,

$$\alpha\mathbb{E}[f_1(\tilde{P})]R+\mathbb{E}\left[X\left(\sum_a \pi^a\circ((\tilde{P}^a\circ R^a)e)\right)\right]$$
$$+\sum_{k=0}^{\infty}\alpha^k\mathbb{E}\left[f_k(\tilde{P})\left(\sum_a \pi^a\circ((\tilde{R}^a\circ P^a)e)\right)\right]$$
$$+\sum_{k=0}^{\infty}\alpha^k\mathbb{E}\left[f_k(\tilde{P})\left(\sum_a \pi^a\circ((\tilde{R}^a\circ \tilde{P}^a)e)\right)\right]=0.$$

We then consider a second-order approximation. This leaves us with

$$\mathbb{E}[\hat{Y}] = (I-\alpha P)^{-1}R+\alpha^2\mathbb{E}[f_2(\tilde{P})]R$$
$$+\alpha X\mathbb{E}\left[\tilde{P}X\left(\sum_a \pi^a\circ((R^a\circ\tilde{P}^a)e)\right)\right]+L_{\exp}. \quad \text{(A1)}$$

The proof is completed by using the definition of $f_2(\tilde{P})$, which yields $\mathbb{E}[f_2(\tilde{P})]R = X\mathbb{E}[\tilde{P}X\tilde{P}]XR = X\mathbb{E}[\tilde{P}X\tilde{P}]Y$, and the lemma follows. □

Lemma A.2. *We have* $\mathbb{E}[\tilde{P}X\tilde{P}]=Q$ *and*

$$\mathbb{E}\left[\tilde{P}X\left(\sum_a \pi^a\circ((R^a\circ\tilde{P}^a)e)\right)\right]=B.$$

Proof. We first observe that the errors in the transition probabilities from two different states ($\tilde{P}_{i\cdot}$ and $\tilde{P}_{j\cdot}$) are independent. Thus, $\mathbb{E}[\tilde{P}_{ik}\tilde{P}_{lj}]=\mathbb{E}[\tilde{P}_{ik}]\mathbb{E}[\tilde{P}_{lj}]=0$ for $i\neq l$.

For the first assertion, we note that the ijth entry of $\mathbb{E}[\tilde{P}X\tilde{P}]$ is equal to

$$\mathbb{E}\left[\sum_{k,l}X_{kl}\tilde{P}_{ik}\tilde{P}_{lj}\right]=\sum_k X_{ki}\mathbb{E}[\tilde{P}_{ik}\tilde{P}_{ij}]=\sum_k X_{ki}\mathrm{cov}^{(i)}_{jk}=\mathrm{cov}^{(i)}_{j\cdot}X_{\cdot i},$$

which is the same as the ijth entry of Q (cf. Equation (12)).

For the second assertion, let

$$\bar{B} = \mathbb{E}\left[\tilde{P}X\left(\sum_a \pi^a\circ((R^a\circ\tilde{P}^a)e)\right)\right]$$
$$=\mathbb{E}\left[\left(\sum_a \pi^a e^\top\circ\tilde{P}^a\right)X\left(\sum_a \pi^a\circ((R^a\circ\tilde{P}^a)e)\right)\right]$$
$$=\sum_a\mathbb{E}[(\pi^a e^\top\circ\tilde{P}^a)X(\pi^a\circ((R^a\circ\tilde{P}^a)e))].$$

Then,

$$\bar{B}_i=\sum_a\mathbb{E}\left[\sum_{k,l}\pi(a\mid i)\tilde{P}^a_{ik}X_{kl}\pi(a\mid l)\sum_j\tilde{P}^a_{lj}R^a_{lj}\right]$$
$$=\sum_a\mathbb{E}\left[\sum_{k}\pi(a\mid i)\tilde{P}^a_{ik}X_{ki}\pi(a\mid i)\sum_j\tilde{P}^a_{ij}R^a_{ij}\right]$$
$$=\sum_{a,k,j}\pi(a\mid i)^2R^a_{ij}X_{ki}\mathbb{E}[\tilde{P}^a_{ik}\tilde{P}^a_{ij}]=\sum_{a,k,j}\pi(a\mid i)^2R^a_{ij}X_{ki}\frac{(M^a_i)_{jk}}{N^a_i}$$
$$=\sum_{a,j}\frac{\pi(a\mid i)^2}{N^a_i}R^a_{ij}(M^a_i)_{j\cdot}X_{\cdot i}=\sum_a\frac{\pi(a\mid i)^2}{N^a_i}R^a_{i\cdot}M^a_iX_{\cdot i}=B_i. \quad \square$$

Finally, we outline the idea that validates $L_{\exp}=o(1/N^{a^*}_{i^*})$. From the expression of $L_{\exp}$ in the proposition, it is clear that only third and higher moments of $\tilde{P}$ are involved. The above claim can be seen from the moment expressions of the corresponding multinomial distributions. We omit a detailed proof.

Appendix B

Proof of Proposition 4.2. The second moment of $\hat{Y}$ is

$$\mathbb{E}[\hat{Y}\hat{Y}^\top]=\mathbb{E}\left[\left(\sum_{i=0}^{\infty}\alpha^i(P+\tilde{P})^i\right)(R+\tilde{R})(R+\tilde{R})^\top\right.$$
$$\left.\cdot\left(\sum_{i=0}^{\infty}\alpha^i((P+\tilde{P})^i)\right)^\top\right].$$

Using Lemma 4.1, we have

$$\mathbb{E}[\hat{Y}\hat{Y}^\top]=\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\alpha^{k+l}\mathbb{E}[f_k(\tilde{P})(RR^\top+\tilde{R}R^\top+R\tilde{R}^\top+\tilde{R}\tilde{R}^\top)f_l(\tilde{P})^\top]. \quad \text{(B1)}$$

Following Lemma A.1, we may drop the zero terms

$$\mathbb{E}\big[X(R\tilde{R}^\top+\tilde{R}R^\top)X^\top\big]+\alpha\mathbb{E}[XRR^\top f_1(\tilde{P})^\top]+\alpha\mathbb{E}\big[f_1(\tilde{P})RR^\top X^\top\big]$$
$$+\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\alpha^{k+l}\mathbb{E}\big[f_k(\tilde{P})(\tilde{R}R^\top+R\tilde{R}^\top)f_l(\tilde{P})^\top\big]=0.$$

Taking the second-order approximation, we obtain

$$\mathbb{E}[\hat{Y}\hat{Y}^\top] = (I-\alpha P)^{-1}RR^\top((I-\alpha P)^{-1})^\top$$
$$+X\big(\alpha^2\mathbb{E}[\tilde{P}XRR^\top X^\top\tilde{P}^\top]+\alpha^2\mathbb{E}[\tilde{P}X\tilde{P}]XRR^\top$$
$$+\alpha^2RR^\top X^\top\mathbb{E}[\tilde{P}X\tilde{P}]^\top+\alpha\mathbb{E}\big[\tilde{P}X(\tilde{R}R^\top+R\tilde{R}^\top)\big]$$
$$+\alpha\mathbb{E}\big[(\tilde{R}R^\top+R\tilde{R}^\top)X^\top\tilde{P}^\top\big]+\mathbb{E}[\tilde{R}\tilde{R}^\top]\big)X^\top+L_{\mathrm{var}}. \quad \text{(B2)}$$

Expanding the above terms, and keeping in mind Lemma A.1 and Equation (11), we just need to calculate the following terms:

$$VY_1 := \mathbb{E}[\tilde{P}XRR^\top X^\top\tilde{P}^\top]+QXRR^\top+RR^\top X^\top Q^\top,$$
$$VY_2 := \mathbb{E}[\tilde{P}X(\tilde{R}R^\top+R\tilde{R}^\top)]$$
$$=\mathbb{E}\left[\tilde{P}X\left(\left(\sum_a \pi^a\circ((R^a\circ\tilde{P}^a)e)\right)R^\top\right.\right.$$
$$\left.\left.+R\left(\sum_a \pi^a\circ((R^a\circ\tilde{P}^a)e)\right)^\top\right)\right],$$
$$VY_3 := VY_2^\top,$$
$$VY_4 := \mathbb{E}[\tilde{R}\tilde{R}^\top].$$

To summarize, Equation (B2) can be written in terms of VY_1, VY_2, VY_3, and VY_4 as

$$\mathbb{E}[\hat{Y}\hat{Y}^\top]=YY^\top+X(\alpha^2VY_1+\alpha VY_2+\alpha VY_3+VY_4)X^\top+L_{\mathrm{var}}.$$

We now provide expressions for VY_1, VY_2, VY_3, and VY_4.

VY_1: We have $VY_1 = \mathbb{E}[\tilde{P}YY^\top\tilde{P}^\top] + QYR^\top + RY^\top Q^\top$. If we define $Q^{(1)} := \mathbb{E}[\tilde{P}YY^\top\tilde{P}^\top]$, then $Q^{(1)}_{ij} = \mathbb{E}[\tilde{P}_{i\cdot}YY^\top\tilde{P}_{j\cdot}^\top]$. Thus, $Q^{(1)}_{ij} = 0$ for $i \neq j$ and

$$Q^{(1)}_{ii} = \mathbb{E}[(\tilde{P}_{i\cdot}Y)^2] = Y^\top\mathbb{E}[\tilde{P}_{i\cdot}^\top\tilde{P}_{i\cdot}]Y = Y^\top \mathrm{cov}^{(i)}Y.$$

VY_2: We have

$$VY_2 = \mathbb{E}\left[\tilde{P}X\left(\sum_a \pi^a \circ ((R^a \circ \tilde{P}^a)e)\right)\right]R^\top + \mathbb{E}\left[\tilde{P}Y\left(\sum_a \pi^a \circ ((R^a \circ \tilde{P}^a)e)\right)^\top\right].$$

Following Lemma A.2, $\mathbb{E}[\tilde{P}X(\sum_a \pi^a \circ ((R^a \circ \tilde{P}^a)e))] = B$. If we define $Q^{(2)} := \mathbb{E}[\tilde{P}Y(\sum_a \pi^a \circ ((R^a \circ \tilde{P}^a)e))^\top]$, then

$$Q^{(2)}_{ij} = \mathbb{E}\left[(\tilde{P}_{i\cdot}Y)\left(\sum_a \pi^a_j((R^a_{j\cdot} \circ \tilde{P}^a_{j\cdot})e)\right)^\top\right] = \sum_a \pi(a \mid i)\pi^a_j R^a_{j\cdot}\mathbb{E}[(\tilde{P}^a_{j\cdot})^\top\tilde{P}^a_{i\cdot}]Y.$$

Thus, $Q^{(2)}_{ij} = 0$ for $i \neq j$ and

$$Q^{(2)}_{ii} = \sum_a (\pi(a \mid i)^2/N^a_i)Y^\top M^a_i R^a_{i\cdot}.$$

VY_4: We have $(VY_4)_{ij} = \mathbb{E}[\tilde{R}_i\tilde{R}_j] = 0$ for $i \neq j$ and with Equation (8),

$$(VY_4)_{ii} = \mathbb{E}[\tilde{R}_i^2] = \sum_a \pi(a \mid i)^2\mathbb{E}[(\tilde{R}^a_i)^2] = \sum_a \frac{\pi(a \mid i)^2}{N^a_i}(R^a_{i\cdot}M^a_i R^{a\top}_{i\cdot} + V^{a\top}_{i\cdot}P^a_{i\cdot}).$$

Define W as in Equation (13). We have $W = \alpha^2 Q^{(1)} + \alpha(Q^{(2)} + (Q^{(2)})^\top) + VY_4$. The result follows by collecting the different terms. □

Appendix C

Proof of Proposition 4.3. For a fixed state i, we define

$$f_{r,a} = X^2_{ir}\pi(a \mid r)^2\left[(\alpha Y^\top + R^a_{r\cdot})M^a_r(\alpha Y + (R^a_{r\cdot})^\top) + \sum_k P^a_{rk}V^a_{rk}\right].$$

Also, we define $F = \sum_{r,a}(N^{a^*}_{i^*}/N^a_r)f_{r,a}$. Using Equation (14), it can be easily verified that $\sigma(\widehat{Y}_i) = \sqrt{F/N^{a^*}_{i^*}}$. Because we assume that $\sigma(Y_i) > 0$, there exists some (r, a) such that $f_{r,a} > 0$; $f_{r,a}$ does not depend on N^a_{ir}. Then, the assumption $N^{a^*}_{i^*}/N^a_r > c > 0$ guarantees that F is bounded from below.

Similarly, $\mathbb{E}[\widehat{Y}_i] - Y_i = G/N^{a^*}_{i^*}$, where

$$G = \sum_{r,a}(N^{a^*}_{i^*}/N^a_r)\alpha\pi(a \mid r)^2 X_{ir}(\alpha M^a_r X_{\cdot r}Y + R^a_{r\cdot}M^a_r X_{\cdot i})$$

is bounded from above because $N^{a^*}_{i^*} \leq N^a_r$ and terms M^a_r, X, Y, and $\pi(a \mid r)$ do not depend on N^a_r.

Thus, the bias decreases like $N^{a^*}_{i^*}$, whereas the standard deviation decreases no faster than $\sqrt{N^{a^*}_{i^*}}$, which yields the desired result. □

References

Barberis, N. 2000. Investing for the long-run when returns are predictable. *J. Finance* **55** 225–264.

Baukal-Gursoy, M., K. Ross. 1992. Variability sensitive Markov decision processes. *Math. Oper. Res.* **17**(3) 558–571.

Bertsekas, D. P. 2000. *Dynamic Programming and Optimal Control*, 2nd ed., Vol. 1. Athena Scientific, Belmont, MA.

Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Bitran, G. R., S. V. Mondschein. 1996. Mailing decisions in the catalog sales industry. *Management Sci.* **42**(9) 1364–1381.

Bult, J., T. Wansbeek. 1995. Optimal selection for direct mail. *Marketing Sci.* **14**(4) 378–394.

Campbell, J. Y., L. M. Viceira. 2002. *Strategic Asset Allocation*: *Portfolio Choice for Long-Term Investors*. Oxford University Press, Oxford, UK.

Clark, C. 1961. The greatest of a finite set of random variables. *Oper. Res.* **9** 145–162.

Dixit, A. K., R. S. Pindyck. 1994. *Investment under Uncertainty*. Princeton University Press, Princeton, NJ.

Filar, J. A., L. Kallenberg, H. Lee. 1989. Variance-penalized Markov decision processes. *Math. Oper. Res.* **14** 147–161.

Godfrey, G., W. Powell. 2002. An adaptive dynamic programming algorithm for dynamic fleet management I: Single period travel time. *Transportation Sci.* **36**(1) 21–39.

Gönül, F., M. Shi. 1998. Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Sci.* **44**(9) 1249–1262.

Hendel, I., A. Nevo. 2002. Measuring the implications of sales and consumer stockpiling behavior. *Econometrica*. Forthcoming.

Keane, M., K. Wolpin. 1994. The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence. *Rev. Econom. Statist.* **76**(4) 648–672.

Leadbetter, M. R., G. Lindgren, H. Rootzén. 1983. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York.

Luenberger, D. G. 1997. *Investment Science*. Oxford University Press, New York.

McGill, J., G. van Ryzin. 1999. Revenue management: Research overview and prospects. *Transportation Sci.* **33** 233–256.

Rust, J. 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* **55**(5) 999–1033.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Simester, D., P. Sun, J. N. Tsitsiklis. 2006. Dynamic catalog mailing policies. *Management Sci.* **52**(5) 683–696.

Sobel, M. J. 1982. The variance of a discounted Markov decision process. *J. Appl. Probab.* **19** 794–802.

Sutton, R., A. Barto. 1998. *Reinforcement Learning*. MIT Press, Cambridge, MA.

Xia, Y. 2001. Learning about predictability: The effects of parameter uncertainty on dynamic asset allocation. *J. Finance* **56** 205–246.

Zipkin, P. 2000. *Foundations of Inventory Management*. McGraw-Hill/Irwin, Boston, MA.

MANAGEMENT SCIENCE
DOI 10.1287/mnsc.1060.0614ec
pp. ec1–ec3

e-companion
ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Bias and Variance Approximation in Value Function Estimates" by Shie Mannor, Duncan Simester, Peng Sun, and John N. Tsitsiklis, *Management Science* 2007, 53(2) 308–322.

Appendix D. The Bayesian Approach

In this appendix we describe a Bayesian approach to variance and bias approximation. The expressions for the mean and variance in the Bayesian setting are very similar to the ones for the classical setting. The only difference is that certain expectations in the classical setting are replaced by conditional (posterior) expectations in the Bayesian setting. However, for these expressions to be useful, one should be able to compute the conditional expectations in a tractable manner. This will be the case for Dirichlet priors on the transition probabilities and normal priors on the rewards, which is the case that we consider in the sequel.

As before, we assume that the sample data consist of the number of transitions out of each state for every action (N_i^a) and the number of transitions from each state i to any other state j for every action a (N_{ij}^a). We also observe the rewards associated with each transition in the sample data. We assume that the expected reward R_{ij}^a associated with a transition from i to j under action a is a random variable with a normal prior. We further assume that each transition probability P_{ij}^a is a random variable with a Dirichlet prior (as in Strens 2000) and that the priors of P_{ij}^a and R_i^a are independent for different i or a.

We first recall some properties of a Dirichlet distribution. We refer the reader to Gelman et al. (1995) for further details. Let $\alpha_0 = \sum_k \alpha_k$. We say that a vector $(p_1, p_2, \ldots, p_m)$ has a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_m$, if its distribution is described by a joint probability density function of the form $(1/Z(\alpha)) \prod_{i=1}^m p_i^{\alpha_i - 1}$, where $Z(\alpha)$ is a normalizing constant. Some useful properties of the Dirichlet distribution are:

1. The mean of p_k is α_k/α_0.
2. The variance of p_k is $\alpha_k(\alpha_0 - \alpha_k)/(\alpha_0^2(\alpha_0 + 1))$.
3. The covariance between p_k and p_ℓ, for $k \neq \ell$, is $-(\alpha_k \alpha_\ell)(\alpha_0^2(\alpha_0 + 1))$.

Assume that the prior distribution of $P_{i\cdot}^a$, the vector of transition probabilities out of state i under action a, is Dirichlet with parameters $\alpha_{i1}^a, \ldots, \alpha_{im}^a$. If in N_i^a observed transitions, and for every j, exactly N_{ij}^a transitions lead to state j, then the posterior distribution of P_i^a is again Dirichlet with parameters $\alpha_{i1}^a + N_{i1}^a, \ldots, \alpha_{im}^a + N_{im}^a$. It then follows that the posterior distribution for P_i has mean $\mathbb{E}_{\text{post}}[P_{ij}^a] = (\alpha_{ij}^a + N_{ij}^a)/(\alpha_{i0}^a + N_i^a)$, where $\alpha_{i0}^a := \sum_j \alpha_{ij}^a$ and $\mathbb{E}_{\text{post}}$ is expectation w.r.t. the posterior. This motivates us to define the estimated transition probabilities by

$$\hat{P}_{ij}^a = \mathbb{E}_{\text{post}}[P_{ij}^a] = (\alpha_{ij}^a + N_{ij}^a)/(\alpha_{i0}^a + N_i^a).[-3pt]$$

The difference between the estimated and the true model is then a zero mean random matrix $\tilde{P} := P - \hat{P}$. The following lemma is an immediate consequence of the properties of the Dirichlet distribution given earlier. Here, var_{post} and cov_{post} are used to denote the posterior variance and covariance.

Lemma D.1. *Under the assumption of a Dirichlet prior we have that*:

(i) $\mathbb{E}_{\text{post}}[P_{ij}^a] = \hat{P}_{ij}^a = (\alpha_{ij}^a + N_{ij}^a)/(\alpha_{i0}^a + N_i^a)$.

(ii) $\mathbb{E}_{\text{post}}[\tilde{P}_{ik}^a \tilde{P}_{ij}^a] = \text{cov}_{\text{post}}(P_{ik}^a, P_{ij}^a) = -((\alpha_{ik}^a + N_{ik}^a)(\alpha_{ij}^a + N_{ij}))/((\alpha_{i0}^a + N_i^a)^2(\alpha_{i0}^a + N_i^a + 1))$.

(iii) $\mathbb{E}_{\text{post}}[(\tilde{P}_{ij}^a)^2] = \text{var}_{\text{post}}(P_{ij}^a) = ((\alpha_{ij}^a + N_{ij}^a)(\alpha_{i0}^a + N_i^a - \alpha_{ij}^a - N_{ij}^a))/((\alpha_{i0}^a + N_i^a)^2(\alpha_{i0}^a + N_i^a + 1))$.

We note that if $\alpha_{ij}^a = 0$ (for $j = 0, \ldots, m$), then we get the same approximations as in the classical approach (up to the $+1$ in the denominator of the variance and the covariance).

Similarly, we define the prior distribution for the immediate reward when moving from state i to state j when using action a. Notice that this reward can be drawn from any family of distributions for which Bayesian updates can be carried out in closed form. As a special case, we assume the reward distribution is normal with mean R_{ij}^a and variance τ_{ij}^a. We assume independence of the priors, i.e., that the prior distribution of R_{ij}^a given τ_{ij}^a does not depend on τ_{ij}^a, and that the prior distribution of R_{ij}^a is normal with mean μ_{ij}^a and standard deviation σ_{ij}^a.

For each i, j, and a, we observe N_{ij}^a sample rewards $\hat{x}_1^{ij,a}, \ldots, \hat{x}_{N_{ij}}^{ij,a}$. We denote the sample variance by s_{ij}^a. Following the analysis of normal data with a semiconjugate prior distribution (see, e.g., Gelman et al. 1995), the posterior distribution (given τ_{ij}^a) for the mean reward is then normal with mean

$$(\mu_{ij}^a)^{\text{post}} = \frac{\mu_{ij}^a/(\sigma_{ij}^a)^2 + \sum_{k=1}^{N_{ij}} \hat{x}_k^{ij,a}/(\tau_{ij}^a)^2}{1/(\sigma_{ij}^a)^2 + N_{ij}^a/(\tau_{ij}^a)^2},$$

and standard deviation: $(\sigma_{ij}^a)^{\text{post}} = 1/(1/(\sigma_{ij}^a)^2 + (N_{ij}^a/(\tau_{ij}^a)^2))^{1/2}$. We may further assume priors for τ_{ij}^a and derive its posterior. For simplicity, we can approximate $(\mu_{ij}^a)^{\text{post}}$ and $(\sigma_{ij}^a)^{\text{post}}$ by substituting s_{ij}^a for τ_{ij}^a. This leads us to define $\hat{R}_{ij}^a$ as the approximation for $(\mu_{ij}^a)^{\text{post}}$ that results from this substitution.

As in the classical case, we consider a fixed (possibly randomized) stationary policy π, and define the following quantities:

1. An (unknown) $m \times m$ matrix P representing the transition probabilities, whose ith row is $P_{i\cdot} = \sum_a \pi(a \mid i) P_{i\cdot}^a$, its estimate $\hat{P}_{i\cdot} = \sum_a \pi(a \mid i) \hat{P}_{i\cdot}^a$, and the difference matrix $\tilde{P} = P - \hat{P}$.
2. An m-dimensional vector representing the immediate reward whose i component is $R_i = \sum_a \pi(a \mid i) \sum_j P_{ij}^a R_{ij}^a$, and its estimate $\hat{R}_i = \sum_a \pi(a \mid i) \sum_j \hat{P}_{ij}^a \hat{R}_{ij}^a$.

Using a second-order approximation and applying Lemma D.1, we obtain expressions for the posterior bias and variance of the estimated value function estimate under the posterior. The proofs are almost identical to those of Propositions 4.1 and 4.2 and are omitted.

Proposition D.1. *The expectation (under the posterior) of* $Y := (I - \alpha P)^{-1} R$ *satisfies*:

$$\mathbb{E}_{\text{post}}[Y] = \hat{Y} + \alpha^2 \hat{X} \hat{Q} \hat{Y} + \alpha \hat{B} + L_{\text{exp}}^b,$$

where $\hat{X} := (I - \alpha \hat{P})^{-1}$; $\hat{Y} = \hat{X}\hat{R}$; *vector* $\hat{B}$ *and matrix* $\hat{Q}$ *are computed according to*

$$\hat{B}_i = \sum_a \pi(a \mid i)^2 \hat{R}_{i\cdot}^a \widehat{M}_i^a \hat{X}_{\cdot i} \tag{EC1}$$

and

$$\hat{Q}_{ij} = \widehat{\text{cov}}_{j\cdot}^{(i)} \hat{X}_{\cdot i} \quad \textit{in which } \widehat{\text{cov}}^{(i)} = \sum_a \pi(a \mid i)^2 \widehat{M}_i^a, \tag{EC2}$$

where matrix $\widehat{M}_i^a$ *is the posterior covariance matrix of* $P_{i\cdot}^a$ *as specified by parts* ii *and* iii *of Lemma* D.1, *and higher order terms*

$$L_{\text{exp}}^b = \sum_{k=3}^{\infty} \alpha^k \mathbb{E}[f_k^b(\tilde{P})] \hat{R} + \sum_{k=2}^{\infty} \alpha^k \mathbb{E}[f_k^b(\tilde{P}) \tilde{R}],$$

in which $\tilde{P} = P - \hat{P}$, $\tilde{R} = R - \hat{R}$ *and* $f_k^b(\tilde{P}) = \hat{X}(\tilde{P}\hat{X})^k$.

Proposition D.2. *Using the same notation as in Proposition* D.1, *the second moment of* $Y := (I - \alpha P)^{-1} R$ *is approximately*

$$\mathbb{E}_{\text{post}}[YY^\top] = \hat{Y}\hat{Y}^\top + \hat{X}\{\alpha^2(\hat{Q}\hat{Y}\hat{R}^\top + \hat{R}\hat{Y}^\top\hat{Q}^\top) + \alpha[\hat{B}\hat{R}^\top + \hat{R}\hat{B}^\top] + \widehat{W}\}\hat{X}^\top + L_{\text{var}}^b,$$

where $\hat{X} := (I - \alpha\hat{P})^{-1}$, $\hat{Y} := \hat{X}\hat{R}$, $\widehat{W}$ *is a diagonal matrix such that*

$$\widehat{W}_{ii} = \sum_a \pi(a \mid i)^2 \left\{ (\alpha\hat{Y}^\top + \hat{R}_{i\cdot}^a)\widehat{M}_i^a(\alpha\hat{Y} + (\hat{R}_{i\cdot}^a)^\top) + \sum_k \hat{P}_{ik}^a \left(\frac{1}{(\sigma_{ik}^a)^2} + \frac{N_{ik}^a}{(s_{ik}^a)^2} \right)^{-1} \right\}$$

and $\hat{Q}$ *and* $\hat{B}$ *are calculated according to Equations* (EC2) *and* (EC1); *and higher-order terms*

$$L_{\text{var}} = \sum_{k,l:\,k+l>2} \alpha^{k+\ell} \mathbb{E}[f_k^b(\tilde{P})(\hat{R}\hat{R}^\top + (\tilde{R})\hat{R}^\top + \hat{R}(\tilde{R})^\top + (\tilde{R})(\tilde{R})^\top) f_\ell^b(\tilde{P})^\top].$$

References

Gelman, A., J. Carlin, H. Stern, D. B. Rubin. 1995. *Bayesian Data Analysis*. Chapman and Hall, London.

Strens, M. 2000. A Bayesian framework for reinforcement learning. *Proc. 17th Internat. Conf. Machine Learn.* Morgan Kaufmann, San Francisco, CA, 943–950.