Bias and Variance in Value Function Estimation

Shie Mannor SHIE@MIT.EDU
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139

Duncan Simester SIMESTER@MIT.EDU
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139

Peng Sun PSUN@DUKE.EDU
Fuqua School of Business, Duke University, Durham, NC 27708

John N. Tsitsiklis JNT@MIT.EDU
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139

Keywords:
Markov Processes, Reinforcement Learning, Bias, Variance, Bayesian Estimation

Abstract

We consider the bias and variance of value function estimation that are caused by using an empirical model instead of the true model. We analyze these bias and variance for Markov processes from a classical (frequentist) statistical point of view, and in a Bayesian setting. Using a second order approximation, we provide explicit expressions for the bias and variance in terms of the transition counts and the reward statistics. We present supporting experiments with artificial Markov chains and with a large transactional database provided by a mail-order catalog firm.

1. Introduction

A common method when analyzing data obtained from a Markov Process (MP) is to estimate the transition probabilities and the reward function based on an empirical sample. The cost-to-go (or the profit-to-go in a maximization problem) is then estimated by plugging the empirical model instead of the true transition probability and the reward function. A fundamental question regarding such an estimate concerns its bias and variance. Knowledge of the bias and variance is essential for evaluating the quality of the cost-to-go estimate, as well as for determining the amount of data required in order to achieve a certain confidence level. Surprisingly, little attention was given to date to the bias and variance of the cost-to-go estimate which is the topic of this paper.

From a statistical point of view, an informative estimator should be accompanied by confidence bounds. The bias and the variance of an estimator naturally lead to such confidence bounds (e.g., using a Chebychev bound). A useful bias and variance estimate should be expressed as a function of the available statistics (counts of state transitions and statistics of the reward). We will show that when assuming that the empirical model is reasonable (i.e., not "too far" from the real model) such bias and variance estimates can be developed based on a second order approximation.

A common framework for decision making under uncertainty used in decision theory and machine learning is the Markov Decision Process (MDP) framework (e.g., Puterman, 1994). In this framework there are several possible actions in every state and a decision maker is required to choose the best action. In the context of estimating the value of an MDP from data, it is common to estimate the conditional transition probabilities (conditioned on the current state and the decision maker's action) and the conditional reward function. Once the empirical model is estimated the best policy is computed by optimizing over the policy space. The value of each state in the MDP is the cumulative expected reward obtained from that state on, if the decision maker follows the optimal policy. The results in this paper are developed for Markov Processes (MPs), but are valid for Markov Decision Processes (MDPs) as long as the policy is fixed.

The paper is organized as follows. We start with de-

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. Copyright 2004 by the authors.

scribing the model in rigor and defining the problem of interest in Sec. 2. In Sec. 3 we illustrate the magnitude of the variance in both an artificial MP and in real data obtained from a mail-order catalog firm. We suggest two different approaches for estimation of the bias and variance in MPs. The first approach is a "classical" statistical (frequentist) approach, and the second approach is Bayesian. We provide the essential details in Sec. 4. We demonstrate the variance estimates for both types of data in Sec. 5. Some concluding remarks are given in Sec. 6.

2. The Model

In this section we specify the problem of interest. We start with defining the problem setup. We then point to two types of variances, one that is related to uncertainty in the parameters, and another which is inherent to the stochastic nature of the problem.

2.1. Problem Setup

We consider both MPs and MDPs. Let us define the latter, as the former is a special case. An MDP is a 4-tuple (S, A, P, R), where S is a set of the states, A is a set of actions, P_{ij}^a is the transition probability from state i to state j when performing action $a \in A$ in state i, and R_{ia} is the reward received when performing action a in state i. We assume that S and A are finite sets and that R_{ia} is a random variable. We further let $|S| = m$. At time t, the current state is s_t, the decision maker chooses some action a_t. As a result of this action the next state s_{t+1} is determined and the decision maker obtains a reward r_t which is distributed according to $R_{s_t a_t}$. We will restrict our attention to discounted reward. The discount factor will be denoted by α, where it is assumed that $\alpha < 1$.

A strategy for an MDP assigns, at each time t, for each state i a probability for performing action $a \in A$, given a history which includes the states, actions and rewards observed until time $t-1$ and the state in time t. A strategy is *stationary* if it only depends on the current state. It is well known that there exists an optimal stationary strategy for discounted reward. An MP can be considered as a special case of an MDP, where a stationary strategy is fixed by the decision maker. We will denote such a strategy by π. The expectation operator under strategy π starting from state i will be denoted by $\mathbb{E}_i^\pi$.

We will consider a nominal (empirical) model of the MDP. This model is typically the result of interacting with the environment. We denote the nominal transition probability by $\hat{P}$, and the nominal reward function by $\hat{R}$. Typically, the sampling procedure provides additional statistics, such as the variance of the reward and the counts of the transitions. We now distinguish between two types of variances.

2.2. Two Types of Variance

There are two different types of variance which are of interest in learning and planning. Let π be a specific stationary strategy such that $\pi(a|i)$ is the conditional probability of choosing action a in state i.

1. *Internal variance* - consider the random variable $Z = \sum_{\tau=0}^{\infty} \alpha^\tau r_\tau$. Due to random transitions and rewards, the random variable Z has some variance (i.e., $\mathrm{var}_i^\pi(Z) = \mathbb{E}_i^\pi[Z^2] - (\mathbb{E}_i^\pi[Z])^2$), even if the parameters of the model were completely specified. An expression for the variance of Z for discounted reward was given by Sobel (1982). See also Filar et al. (1989) for the average cost case. The internal variance and its reduction was studied in the context of accelerating policy gradients by, e.g., Greensmith et al. (2002).

2. *Parametric variance* - Suppose that there is some true model, P_T and R_T, and that we have an estimated model (i.e., P and R) such that there is some probabilistic law that determines the distribution of P and R. The random variable considered here is:

$$Y^\pi = (I - \alpha P^\pi)^{-1} R^\pi, \qquad (1)$$

   where the $m \times m$ matrix $P_{ij}^\pi = \sum_a P_{ij}^a \pi(a|i)$ and the vector $R_i^\pi = \sum_a R_{ia} \pi(a|i)$. Eq. (1) prescribes the cost-to-go (per state) of the estimated model. The random variable Y^π is defined with respect to a probability measure *over models*. The covariance matrix of Y^π is defined as $\mathrm{cov}(Y^\pi) = \mathbb{E}_{P_T,R_T}[Y^\pi {Y^\pi}^\top] - \mathbb{E}_{P_T,R_T}[Y^\pi] \mathbb{E}_{P_T,R_T}[Y^\pi]^\top$, where $\mathbb{E}_{P_T,R_T}$ denotes the expectation when the distribution of P and R is determined by P_T and R_T.

Each type of variance is related to a different type of uncertainty and can be associated with a different experiment. The internal variance is the variance of the cumulative discounted cost-to-go in an experiment where the decision maker starts many times from a certain state i and follows a policy π, and the model is assumed to be perfectly known. The parametric variance is the variance of the empirical cost-to-go estimate when one obtains nominal models many times. In this paper we only consider the parametric variance.

3. An illustration

We will consider artificial MPs that are randomly generated. Those MPs have $m = 10$ states, and a randomly generated transition probability according to the following rule: we sample m random numbers from a uniform distribution on $[0, 1]$. We take the two largest numbers and normalize them to sum to $1/2$, we take the rest $m - 2$ numbers and normalize them to sum to $1/2$ as well. As a result we have a transition probability that sums to one and has two states that contain 50% of the mass. The reward R_{ia} is a Normal random variable whose mean and variance were sampled from $N(0, 1)$ and $U[0, 1/4]$, respectively. To demonstrate the effect of variance we run the following experiment. We constructed a random MP using the procedure just described. We sampled this MP n times. We calculated the cost-to-go of each empirical model, and weighted the different states according to the steady-state distribution (of the true model), so that we considered $c^\top Y^\pi$ where c is the steady state vector. The reason for weighing the cost-to-go is that the vector Y^π is m dimensional and we want to consider just a one dimensional summary. We set the discount factor to $\alpha = 0.9$. Fig. 1 presents the empirical standard deviation of the weighted value function for ten randomly generated MPs as a function of number of times each state was sampled. In order to calculate the standard deviation we sampled each MP $n = 50$ times. The weighted cost-to-go was in the range $[-3, 3]$ for all MPs. It is clear from Fig. 1 that the variance in the cost-to-go estimate is significant. As expected, this variance decreases as the number of samples per state increases. In Fig. 2 we focus on a single MP that was generated in the same way. The weighted cost-to-go of the true model was $c^\top Y^\pi = 1.78$, and each state i was sampled $N_i = 200$ times, we repeated the experiment $n = 1000$ times. The histogram of $c^\top \hat{Y}$ (where $\hat{Y}$ is the empirical estimate of Y) shows that the deviation in the cost-to-go is significant.

We were fortunate to have access to a large transactional database of a mail-order catalog firm. Every time a new catalog is produced, the mail-order catalog firm has to make a decision—to mail or not to mail to each customer. The cost of producing and mailing the catalogs is not negligible, and the firm looks for a strategy that maximizes its expected revenue. The firm logs all the purchasing and mailing history for every customer, and can therefore make informed decisions. The database we used includes about 1.72 million customers and more than 160 million transactions. Following Gönül and Shi (1998), the decision problem (for every customer) can be modelled as an MDP, where the state is a summary of the customer's history, and the action at each time epoch is either to mail or not to mail. The construction of the state space is an interesting problem which we will not consider here. We have used the so-called RFM (Recency, Frequency, Monetary value) scales which is common in the mail-order catalog industry (e.g., Bult & Wansbeek, 1995; Bitran & Mondschein, 1996). In the RFM parametrization, the history of each customer is summarized by three scales: the recency of the last purchase, the frequency of purchases, and the average monetary value. We constructed an MDP model from the data by quantizing each of the RFM scales to 4 discrete levels, so that the state space has $m = 4^3 = 64$ states. Since there are many customers, the internal variance is averaged out, so the firm only cares about the parametric variance. Estimation of the parametric variance of the current policy is extremely important for the firm, so it can have confidence in projected revenues. In addition, the firm is interested in estimating the variance of new mailing policies, which might promise higher profit at the expense of larger variance.

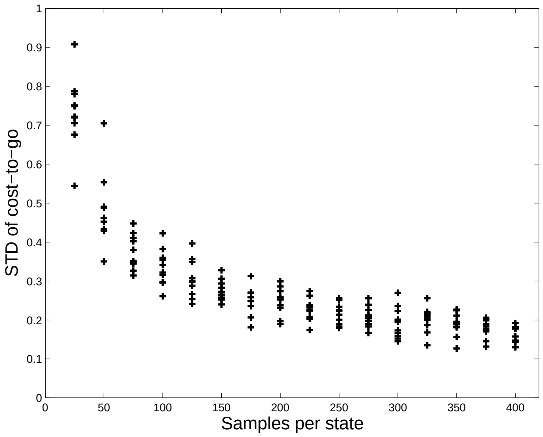

Figure 1. Artificial data: Empirical standard deviation of the weighted cost-to-go as a function of sample size. Each cross represents one empirical standard deviation that was computed based on $n = 50$ runs of a single artificially generated MP.

In Tab. 1 we present the empirical standard deviation of the profit-to-go for the real data, weighted uniformly over the states, for the policy used by the firm. The empirical standard deviation was calculated by dividing the data to roughly equal segments (since the history of every customer is integral, we did not split a customer between segments). The empirical profit-to-go of the whole data is \$5.88. It can be seen that the standard deviation is rather significant, and accounts for as much as 5% of the profit-to-go for as many as 1.5 million transactions.

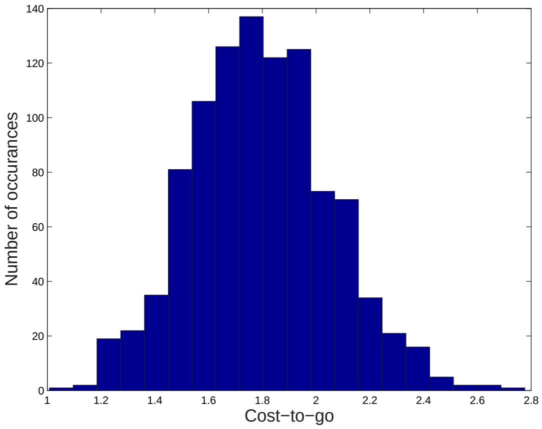

Figure 2. Artificial data: A histogram of the cost-to-go for a single MP, as predicted from empirical samples. Each state i was sampled $N_i = 200$ times. The MP was sampled $n = 1000$ times. The true $c^\top Y = 1.78$.

Number of transactions per segment (millions)	STD of value	Relative error
0.66	0.3548	6.4%
1.48	0.2821	5.1%
2.62	0.2455	4.4%
4.09	0.2293	4.1%
5.91	0.2126	3.8%
16.26	0.1629	2.9%

Table 1. Mail catalog data: The empirical standard deviation of the profit-to-go as a function of the segment size. The relative error is 1 empirical standard deviation divided by the value as calculated using all the data.

4. Two Approaches

Suppose that we have access to the counts that generate $\hat{P}$ and to the statistics of $\hat{R}$ (i.e., its empirical variance and mean). Given those statistics we consider two approaches for estimating the parametric bias and variance of the cost-to-go. In this section we assume that a fixed stationary policy π is used so we therefore drop the superscript π.

1. The "classical" (or frequentist) approach. According to the classical approach we assume that (P, R) are given, and treat $(\hat{P}, \hat{R})$ as random variables. For a given pair of P and R we calculate the bias and variance of the cost-to-go estimator (in terms of the unknown P and R). Since we only have access to the empirical estimates of P and R we substitute those estimates instead, and obtain estimates of the bias and variance of the cost-to-go. We further assume that the total number of transitions out of each state is provided as well (rather than being a random variable depending on P and R). This approach is motivated by the assumption that there is a fair amount of data and that the empirical estimates of P and R are pretty close to the true P and R.

2. The Bayesian approach. Since $(\hat{P}, \hat{R})$ are given, we can treat the true (P, R) as random variables in a Bayesian framework. Using Bayes law we have that: $\mathbf{P}(P|\hat{P}) = \mathbf{P}(\hat{P}|P)\mathbf{P}(P)/C$, where C is a normalizing constant. We need to come up with "reasonable" priors for P and R so that the posterior calculation would be feasible. In Section 4.2 we assume that P has a Dirichlet prior and R has a normal prior, and that P and R are not correlated between states.

4.1. The classical approach

In the classical approach we assume the existence of true P and R that generate data (a collection of sample trajectories). Using these data we generate the nominal model, i.e., $\hat{P}$ and $\hat{R}$. Both, $\hat{P}$ and $\hat{R}$, are random variables that depend on the true P and R. We will provide expressions for the bias and variance of the cost-to-go estimate if P and R were known, and later suggest to replace P with $\hat{P}$ and R with $\hat{R}$ in those estimates.

We assume that the number of transitions out of each state, N_i, is given. The number of transitions from state i to all states j, N_{ij}'s, thus follows a multinomial distribution. We use the estimate $\hat{P}_{ij} = \frac{N_{ij}}{N_i}$, and we have that $\mathbb{E}[\hat{P}] = P$. We use zero mean random variables ΔP and ΔR to represent the difference between the true model and nominal model, i.e., $\hat{P} = P + \Delta P$ and $\hat{R} = R + \Delta R$. Note that the random variables ΔP and ΔR may be correlated. We therefore write the expectation of $\hat{Y} := (I - \alpha\hat{P})^{-1}\hat{R}$ as:

$$\begin{aligned}\mathbb{E}\left[\hat{Y}\right] &= \mathbb{E}\left[(I - \alpha(P + \Delta P))^{-1}(R + \Delta R)\right] \\ &= \mathbb{E}\left[\sum_{i=0}^{\infty} \alpha^i (P + \Delta P)^i (R + \Delta R)\right] . \quad (2)\end{aligned}$$

We use notation $X \triangleq (I - \alpha P)^{-1}$ and let $f_k(\Delta P) \triangleq X\left(\Delta P X\right)^k = \left(X \Delta P\right)^k X$. The following technical lemma will be useful:

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

Proof.

$$\sum_{k=0}^{\infty} \alpha^k f_k(\Delta P) = \sum_{k=0}^{\infty} \alpha^k \left(X \Delta P\right)^k X$$

$$
\begin{aligned}
&= (I-\alpha X\Delta P)^{-1}X = (X^{-1}-X^{-1}\alpha X\Delta P)^{-1} \\
&= (I-\alpha P-\alpha\Delta P)^{-1} = \sum_{i=0}^{\infty}\alpha^i(P+\Delta P)^i,
\end{aligned}
$$

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

Substituting Lemma 4.1 in Eq. (2), and separating the first term in the sum ($k=0$) from the rest of the terms, we obtain:

$$
\begin{aligned}
\mathbb{E}[\hat{Y}] &= (I-\alpha P)^{-1}R + \left(\sum_{k=1}^{\infty}\alpha^k\mathbb{E}[f_k(\Delta P)]\right)R + \\
&\qquad \sum_{k=0}^{\infty}\alpha^k\mathbb{E}[f_k(\Delta P)\Delta R]. \qquad (3)
\end{aligned}
$$

There are three terms in Eq. (3). The first term is the cost-to-go of the true model. The second term reflects the uncertainty in P and the third term represents the correlation in errors between the estimates of R and P. The immediate implication of Eq. (2) is that using the nominal model induces bias.

The estimation of $\mathbb{E}[f_k(\Delta P)]$ involves kth order moments of multinomial distributions. This can be conducted however is rather tedious. We will consider a second order approximation and assume that $\mathbb{E}[f_k(\Delta P)] \approx 0$ for $k > 2$. As a justification, notice that as long as $\|\Delta P\| < (1-\alpha)/\alpha$ (where $\|\cdot\|$ is any matrix norm) we have that $q := \alpha\|\Delta P\|\left\|(I-\alpha P)^{-1}\right\| < 1$ thus $\alpha^k\|\mathbb{E}[f_k(\Delta P)]\| \leq \alpha^k\|\Delta P\|^k\left\|(I-\alpha P)^{-1}\right\|^{k+1} < q^k/(1-\alpha)$ decays exponentially with increase of k.

In many cases, the correlation between ΔP and ΔR can be modelled as the result having a true model whose rewards come from an $m\times m$ matrix R^m while R is just the aggregated values from R^m in the following way

$$
\hat{R}_i = \sum_j\left(R^m_{ij}P_{ij} + R^m_{ij}\Delta P_{ij} + \Delta R^m_{ij}P_{ij} + \Delta R^m_{ij}\Delta P_{ij}\right). \qquad (4)
$$

Here ΔR^m_{ij}'s and ΔP are independent. Under this modelling assumption, we have $\Delta R_i = \sum_j\left(R^m_{ij}\Delta P_{ij} + \Delta R^m_{ij}P_{ij} + \Delta R^m_{ij}\Delta P_{ij}\right)$, or in matrix notations:

$$
\Delta R = (R^m\circ\Delta P + \Delta R^m\circ P + \Delta R^m\circ\Delta P)e, \qquad (5)
$$

where $\circ$ is the Hadamard multiplication, e is an $m\times 1$ vector of ones, and ΔR^m is the $m\times m$ difference matrix between the true R^m and the empirical estimate $\hat{R}^m$.

Let Q be the $m\times m$ matrix satisfying:

$$
Q_{ij} = \frac{P_{ij}}{N_i}\left(X_{ji} - \sum_k P_{ik}X_{ki}\right). \qquad (6)
$$

When ΔP is "small" one can use a second order approximation and estimate the bias and variance of the cost-to-go. The following proposition prescribes the bias. The proof of the proposition is technical and lengthy. This proof and other proofs are deferred to the full version of this paper.

Proposition 4.1 *Assume that ΔP and ΔR are correlated according to Eq. (5). Then the expectation of $\hat{Y}$ satisfies:*

$$
\mathbb{E}[\hat{Y}] \approx Y + \alpha^2 XQXR + \alpha X(Q\circ R^m)e\ ,
$$

where $X := (I-\alpha P)^{-1}$, $Y = (I-\alpha P)^{-1}R$ is the true cost-to-go, and Q is computed according to Eq. (6).

Since we can calculate $\mathbb{E}[\hat{Y}]$, it suffices to calculate $\mathbb{E}[\hat{Y}\hat{Y}^\top]$ in order to obtain the covariance matrix. The following proposition provides this estimation.

Proposition 4.2 *Using the same notations and under the same assumptions of Prop. 4.1, the second moment of $\hat{Y}$ is approximately*

$$
\begin{aligned}
\mathbb{E}[\hat{Y}\hat{Y}^\top] \approx YY^\top + X\Big\{&\alpha^2(Q^{(1)} + QYR^\top + RY^\top Q^\top) + \\
&\alpha\left[((R^m\circ Q)eR^\top + Q^{(2)}) + ((R^m\circ Q)eR^\top + Q^{(2)})^\top\right] \\
&+Q^{(3)} + Q^{(4)}\Big\}X^\top,
\end{aligned}
$$

where $Q^{(1)}$, $Q^{(2)}$, $Q^{(3)}$ and $Q^{(4)}$ are all diagonal matrices such that

$$
\begin{aligned}
Q^{(1)}_{ii} &= \frac{1}{N_i}\left(\sum_k Y_k^2P_{ik} - \sum_k\sum_l Y_kY_lP_{ik}P_{il}\right) \\
Q^{(2)}_{ii} &= \frac{1}{N_i}\left(\sum_k Y_kR^m_{ik}P_{ik} - \sum_k Y_k\sum_l R^m_{il}P_{ik}P_{il}\right) \\
Q^{(3)}_{ii} &= \frac{1}{N_i}\left(\sum_k (R^m_{ik})^2P_{ik} - \sum_k\sum_l R^m_{ik}R^m_{il}P_{ik}P_{il}\right) \\
Q^{(4)}_{ii} &= \sum_k\frac{1}{N_{ik}}P_{ik}^2\,\mathrm{var}(R^m_{ik})\ .
\end{aligned}
$$

Note that as the N_i's increase to ∞ all the terms involving Q decrease to 0, so that both the bias and the variance decrease to 0. The true model (P and R) is used in the above estimates. According to the classical approach we plug in $\hat{P}$ and $\hat{R}$ in place of P and R (and the empirical variance of R_{ik} instead of $\mathrm{var}(R_{ik})$ for $Q^{(4)}$). Simple algebra shows that the variance and bias are roughly of the same order of magnitude. Since both are typically a number much smaller than 1, this implies that the standard deviation (which is the square root of the variance) will be typically *much larger* than the bias. The conclusion is that de-biasing the cost-to-go estimate is not useful since the noise caused by the variance is typically more significant.

4.2. The Bayesian Approach

In this section we describe a Bayesian approach. As before, we assume that the data is the number of transitions out of each state N_i, the number of transitions from state i to all states j, N_{ij}'s. We also observe the sample of the rewards obtained when moving between the states. We assume that for every pair of states i, j the reward moving from state i to state j, R_{ij} is a random variable with a Normal prior. We further assume that the probability P is a random variable with a Dirichlet prior (as in Strens, 2000). See Dearden et al. (1998) for a somewhat different Bayesian formulation in the context of Q-learning. An additional assumption is that the priors of P and R is not correlated between states.

We first recall the following properties of a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_m$ (here we define $\alpha_0 := \sum_k \alpha_k$). We refer the reader to Gelman et al. (1995) for further details. For a vector $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ the probability of $\mathbf{p}$ is $\mathbf{P}(\mathbf{p}) = (1/Z(\alpha)) \prod_{i=1}^{m} p_i^{\alpha_i - 1}$, where $Z(\alpha)$ is a normalizing constant. Some useful properties of the Dirichlet distribution are:

1. Mean of the k^{th} component: α_k / α_0.
2. Variance of the k^{th} component: $\mathrm{var}(P_k) = \alpha_k(\alpha_0 - \alpha_k)/(\alpha_0^2(\alpha_0 + 1))$.
3. Covariance between the k^{th} and ℓ^{th} components: $\mathrm{cov}(P_k, P_\ell) = -(\alpha_k \alpha_\ell)/(\alpha_0^2(\alpha_0 + 1))$.

Assume that $P_{i\cdot}$, the prior transition probability distribution out of state i is Dirichlet with initial parameters $\alpha_1^i, \ldots, \alpha_m^i$. After observing sample trajectories, summarized by N_{ij} transitions out of state i to state j and $N_i = \sum_j N_{ij}$, the posterior distribution of $P_{i\cdot}$ is again Dirichlet with parameters $\alpha_1^i + N_{i1}, \ldots, \alpha_m^i + N_{im}$. It then follows that the posterior distribution for P_i has mean $\mathbb{E}_{post}[P_{ij}] = (\alpha_j^i + N_{ij})/(\alpha_0^i + N_i)$, where $\alpha_0^i := \sum_j \alpha_j^i$ and $\mathbb{E}_{post}$ is expectation w.r.t. the posterior. This motivates us to define the nominal model, which is also an unbiased estimator for the posterior of P, to be $\hat{P}_{ij} = (\alpha_j^i + N_{ij})/(\alpha_0^i + N_i)$. The difference between the nominal and the true model is then a zero mean random matrix $\Delta P := P - \hat{P}$. The following lemma is a result of the useful facts regarding the properties of the Dirichlet distribution.

Lemma 4.2 *Under the assumption of a Dirichlet prior we have that:*
i. $\mathbb{E}_{post}\Big[P\Big] = \hat{P}$.
ii. $\mathbb{E}_{post}\left[\Delta P_{ik} \Delta P_{ij}\right] = -\frac{(\alpha_k^i + N_{ik})(\alpha_j^i + N_{ij})}{(\alpha_0^i + N_i)^2(\alpha_0^i + N_i + 1)}$.
iii. $\mathbb{E}_{post}\left[(\Delta P_{ij})^2\right] = \frac{(\alpha_j^i + N_{ij})(\alpha_0^i + N_i - \alpha_j^i - N_{ij})}{(\alpha_0^i + N_i)^2(\alpha_0^i + N_i + 1)}$.

We note that if $\alpha_j^i = 0$ (for $j = 0, \ldots, m$) then we get the same estimates 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 R^m. Notice that R^m could be drawn from any family of distributions that has a close form Bayesian updates. As a special case, here we assume the prior distribution for R^m is Normal with parameters μ_{ij}, ρ_{ij} and denote the sample variance by s_{ij}.

If for each component R_{ij}^m we observe a series of N_{ij} observations $\hat{x}_1^{ij}, \ldots, \hat{x}_{N_{ij}}^{ij}$, the posterior distribution for R_{ij}^m is also Normal with expectation: $\mu_{ij}^{post} = \left(\mu_{ij}/\rho_{ij}^2 + \sum_{k=1}^{N_{ij}} \hat{x}_k^{ij}/s_{ij}^2\right) / \left((1/\rho_{ij}^2) + (N_{ij}/s_{ij}^2)\right)$, and variance: $\rho_{ij}^{post} = 1/\left((1/\rho_{ij}^2) + (N_{ij}/s_{ij}^2)\right)$, e.g., Gelman et al., 1995. So we define the nominal model $\hat{R}^m$ to be the $m \times m$ matrix whose ij-th entry is $\hat{R}_{ij}^m = \mu_{ij}^{post}$, and accordingly:

$$\begin{aligned} \hat{R} &= (\hat{R}^m \circ \hat{P})e\,, \qquad R^m = \hat{R}^m + \Delta R^m\ , \\ \Delta R &= (R^m \circ \Delta P + \Delta R^m \circ P + \Delta R^m \circ \Delta P)e\ . \end{aligned}$$

Using a second order approximation and following similar derivation as in Section 4.1, we have the following results.

Proposition 4.3 *Assume that random matrices ΔR^m and ΔP are independent, the expectation (w.r.t. the posterior) of $Y := (I - \alpha P)^{-1} R$ satisfies:*

$$\mathbb{E}_{post}[Y] \approx \hat{Y} + \alpha^2 \hat{X}\hat{Q}\hat{Y} + \alpha \hat{X}(\hat{Q} \circ \hat{R})e\ ,$$

where $\hat{X} := (I - \alpha \hat{P})^{-1}$, $\hat{Y} = \hat{X}\hat{R}$ and $\hat{Q}$ is computed according to

$$\hat{Q}_{ij} = \sum_{k:k \neq j} \hat{X}_{ki}\,\mathrm{cov}\left(P_{ik}, P_{ij}\right) + \hat{X}_{ji}\,\mathrm{var}\left(P_{ij}\right), \qquad (7)$$

where $\mathrm{cov}\left(P_{ik}, P_{ij}\right)$ *and* $\mathrm{var}\left(P_{ij}\right)$ *are computed according to Lemma 4.2.*

Proposition 4.4 *Using the same notations and under the same assumptions of Prop. 4.3, the second moment of $Y := (I - \alpha P)^{-1} R$ is approximately*

$$\mathbb{E}_{post}[YY^\top] \approx \hat{Y}\hat{Y}^\top + \hat{X}\Big\{\alpha^2(\hat{Q}^{(1)} + \hat{Q}\hat{Y}\hat{R}^\top + \hat{R}\hat{Y}^\top\hat{Q}^\top)$$
$$+\alpha\left[(\hat{R}^m \circ \hat{Q})e\hat{R}^\top + \hat{Q}^{(2)}) + ((\hat{R}^m \circ \hat{Q})eR^\top + \hat{Q}^{(2)})^\top\right]$$
$$+\hat{Q}^{(3)} + \hat{Q}^{(4)}\Big\}\hat{X}^\top,$$

where $\hat{X} := (I - \alpha\hat{P})^{-1}$, $\hat{Y} := \hat{X}\hat{R}$, and $\hat{Q}^{(1)}$, $\hat{Q}^{(2)}$, $\hat{Q}^{(3)}$ and $\hat{Q}^{(4)}$ are all diagonal matrices such that

$$\hat{Q}^{(1)}_{ii} = \textstyle\sum_{k,\ell:k\neq\ell} \hat{Y}_k\hat{Y}_\ell \operatorname{cov}(P_{ik}, P_{i\ell}) + \sum_k \hat{Y}_k^2 \operatorname{var}(P_{ik}),$$
$$\hat{Q}^{(2)}_{ii} = \textstyle\sum_{k,\ell:k\neq\ell} \hat{Y}_k\hat{R}^m_{i\ell} \operatorname{cov}(P_{ik}, P_{il}) + \sum_k \hat{Y}_k\hat{R}^m_{ik} \operatorname{var}(P_{ik}),$$
$$\hat{Q}^{(3)}_{ii} = \textstyle\sum_{k,\ell:k\neq\ell} \hat{R}^m_{ik}\hat{R}^m_{i\ell} \operatorname{cov}(P_{ik}, P_{il}) + \sum_k \hat{R}^{m^2}_{ik} \operatorname{var}(P_{ik}),$$
$$\hat{Q}^{(4)}_{ii} = \textstyle\sum_k \hat{P}^2_{ik} \operatorname{var}(R^m_{ik}) = \sum_k \hat{P}^2_{ik}\left(\frac{1}{\rho^2_{ik}} + \frac{N_{ik}}{s^2_{ik}}\right)^{-1},$$

and $\hat{Q}$ is calculated according to Eq. (7).

5. Experiments

To validate the variance estimation, and to show that we can compute confidence bounds with reasonable accuracy, we performed the following experiment. We generated random MPs as in Sec. 3. For each such random MP we sampled the process 1000 times. We then compared the difference between the empirical cost-to-go and the true cost-to-go (weighted by the steady-state frequency of the true model) and divided the difference by the estimated standard deviation (based on the empirical sample and the classical approach). In Fig. 3 we show the percentage of experiments that were within 1 standard deviation (marked by '+') and within 2 standard deviations (marked by 'x') from the true weighted cost-to-go, as a function of the number of samples per state. Under a Gaussian distribution assumption these percentages would ideally equal 68% and 95%, respectively. It can be seen that the variance estimation is rather accurate.

We performed a similar experiment using the mail-order catalog data. We divided the data to 1000 groups. We then randomly chose different segments, and based on the empirical model for each segment estimated the variance of a certain fixed policy (the policy used by the firm). We then compared the difference between the empirical profit-to-go of the segment and the profit-to-go of the model that uses *all* of the data (weighted equally across all the 64 states) and normalized by one standard deviation using the variance estimate (based on the classical approach). In Fig. 4 we show the percentage of customer segments that were within 1 standard deviation (marked by '+') and within 2 standard deviations (marked by 'x') from the profit-to-go as calculated based on all the data, as a function of the number of transactions per segment. In Fig. 5 we present a histogram of the normalized difference between the estimated profit-to-go and the profit-to-go based on all of the data, across different partitions of the data. The difference appears to be Gaussian with high confidence (as validated by a Kolmogorov-Smirnof test). The experiments presented validate the accuracy of the variance estimate. We note that the variance estimate appears less tight for the mail-catalog data than for the simulated MPs. We attribute this lack of tightness to the fact the data was not sampled from a "real" Markov process.

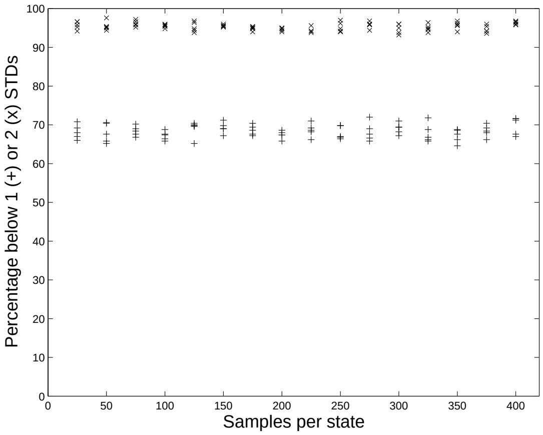

Figure 3. Simulated data: The percentage of value estimates that were within 1 (marked by '+') and 2 (marked by 'x') estimated standard deviations of the true value function. Each point in the plot was calculated by sampling a single MP 1000 times.

6. Concluding Remarks

In this paper we provided explicit expressions for the bias and variance of the cost-to-go in MPs using both classical and Bayesian approaches. We assumed that we have access to a rather accurate estimate of the model, which allowed us to use a second order approximation. It is not clear how to go about estimating the variance when a second order approximation is not valid. A natural question is how to assess the validity of the second order approximation given data that are suspected to be generated from a Markov process.

In this paper we did not address the maximization problem encountered in MDPs, where an additional maximization over the space of policies is performed.

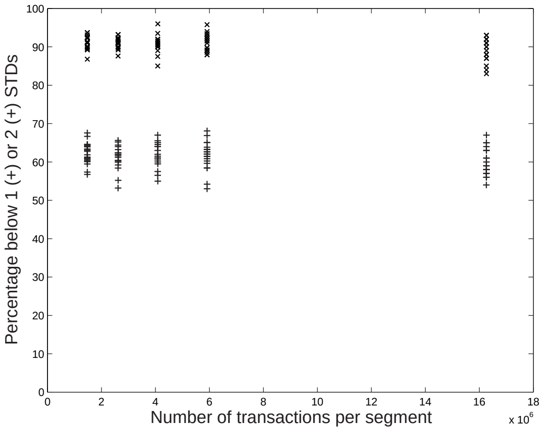

Figure 4. Mail catalog data: The percentage of value estimates that were within 1 and 2 standard deviations from the profit-to-go calculated based on the whole data set.

This maximization may introduce an additional bias to the value function estimation, since actions that are not sampled enough may appear better than they really are. Estimating this bias and accounting for it as part of the optimization process are important research questions.

The statistical setup of this paper assumes that the sample trajectories are provided. A learning setup, where an agent can actively sample trajectories, is a natural extension. In such a setup, a learning agent may guide the exploration to minimize the parametric variance. We note that in model-based reinforcement learning (e.g., Kearns & Singh, 2002) one typically assumes that the current estimation of the model is accurate enough to allow accurate policy evaluations. By estimating the parametric variance we may, perhaps, allow the agent to focus on sampling critical areas of the state space where the variance can be reduced significantly.

Acknowledgements

We are grateful for the helpful suggestions made by three anonymous reviewers. This work was partially supported by the National Science Foundation under grant 0322823.

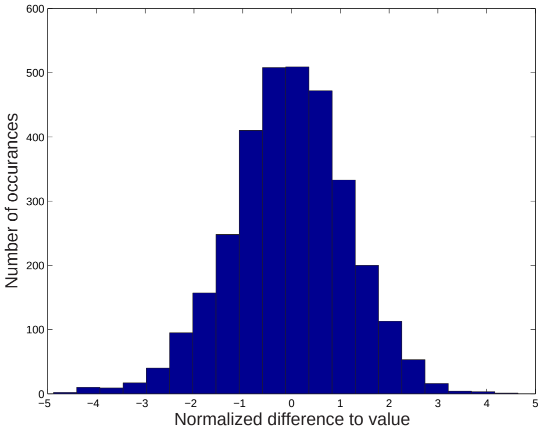

Figure 5. Mail catalog data: a histogram for the difference between the estimated profit-to-go and the profit-to-go calculated using all the data. To normalize the difference it is divided by the estimated standard deviation. Each sample included 10,000-20,000 customers. The F-score of the data is 1.15, indicating that the normalization works well. A Kolmogorov-Smirnof test indicates that the data appears Gaussian with high confidence.

References

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

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

Dearden, R., Friedman, N., & Russell, S. J. (1998). Bayesian Q-learning. *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 761–768). AAAI Press / The MIT Press.

Filar, J. A., Kallenberg, L., & Lee, H. (1989). Variance-penalized Markov decision processes. *Mathematics of operations Research*, *14*, 147–161.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall.

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

Greensmith, E., Bartlett, P. L., & Baxter, J. (2002). Variance reduction techniques for gradient estimates in reinforcement learning. *Advances in Neural Information Processing Systems 14*.

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

Puterman, M. (1994). *Markov decision processes*. Wiley-Interscience.

Sobel, M. J. (1982). The variance of discounted Markov decision process. *Journal of Applied Probability*, *19*, 794–802.

Strens, M. (2000). A Bayesian framework for reinforcement learning. *In Proceedings of the 17th International Conference on Machine Learning* (pp. 943–950).