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.