A Control Variable Perspective for the Optimal Combination of Truncated Corrected Returns

Article · February 1995

Source: CiteSeer

Michael Duff
UConn Health Center

A Control Variable Perspective for the Optimal Combination of Truncated Corrected Returns

Michael O. Duff
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
duff@cs.umass.edu

Abstract

This paper details further development of an idea first suggested in (Barto & Duff, 1994)—that of bringing variance reduction techniques to bear upon the problem of optimally combining corrected truncated returns.

1 INTRODUCTION

Consider a system whose dynamics are described by a finite state Markov chain with transition matrix P, and suppose that at each time step, in addition to making a transition from state $x_t = i$ to $x_{t+1} = j$ with probability p_{ij}, the system produces a randomly determined reward, r_{t+1}, whose expected value is $R(i)$. The *evaluation function*, V, maps states to their expected, infinite-horizon discounted returns: $V(i) = E\{\sum_{t=0}^{\infty} \gamma^t r_{t+1} | x_0 = i\}$, $0 < \gamma < 1$. It is well known that V uniquely satifies a linear system of equations decribing local consistency: $V = R + \gamma P V$.

One way (Watkins, 1989) of viewing the TD(λ) algorithm (Sutton, 1988) is that it updates its current value function estimate $\tilde{V}(x_t)$ in the direction of a weighted combination of the following (infinite) family of estimators:

$$X^{[k]} = r_t + \gamma r_{t+1} + \cdots \gamma^{k-1} r_{t+k-1} + \gamma^k \tilde{V}(x_{t+k}) \qquad k = 1, 2, ... \qquad (1)$$

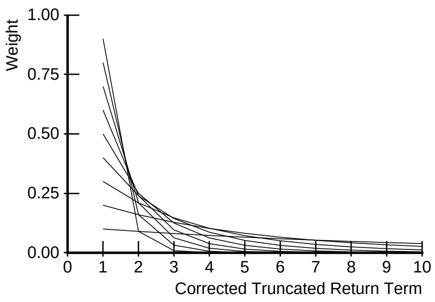

Figure 1: TD(λ) weight sets.

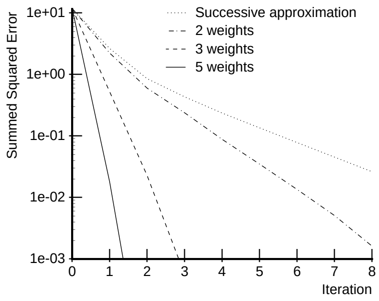

Figure 2: Convergence results.

For large values of k (such that γ^k is small), the truncated corrected return, $X^{[k]}$, is approxmately equal to the sampled discounted return, which is an unbiased estimate for $V(x_t)$, and has variance equal to the variance of accumulated discounted rewards encountered along sample paths. For smaller values of k, the portion of total variance due to the variance of summed, discounted rewards beyond time horizon k is replaced by the variance of $\tilde{V}(x_{t_k})$. If the approximate value function is equal to the true value function, then the variance of this "tail" will be less than the variance of the tail of sampled returns; if the approximate value function is in error, then in general $X^{[k]}$ will be a biased estimator.

In TD(λ), the estimators $X^{[k]}$ are combined via $X_w = \sum_{k=1}^{\infty} w_k X^{[k]}$, where $w_k = \lambda^{(k-1)}(1-\lambda)$, and $\tilde{V}(x_t)$ is updated via $\tilde{V}^{(new)}(x_t) = (1-\alpha)\tilde{V}^{(old)}(x_t) + \alpha X_w$. The weighting function, w_k, for the combination of truncated corrected returns is illustrated in Figure 1. For small values of λ, the resulting estimate X_w is more heavily weighted toward $X^{[k]}$'s with small values of k, estimators with potentially small variance but high bias. Watkins has suggested that a reasonable strategy might be to apply TD(λ) with λ near one at the outset of learning, so as to avoid the ill-effects of using biased $\tilde{V}$'s, then as confidence in the $\tilde{V}$'s grows, to slowly reduce λ to reduce variance. The decaying exponential form of the weighting function gives rise to an update rule that can be implemented incrementally, but there is no reason to presume that the "best" combination of estimators would be weighted in this way for *any* value of λ, in fact one might well suspect that the best weighting scheme would (at least) vary with state and time.

In (Barto & Duff, 1994), it was suggested that the estimators $X^{[k]}$ could be interpreted as what are known as "control variables" in the literature of Monte Carlo variance reduction techniques. A control variable (Lavenberg & Welch, 1981) for a random variable Y, whose expected value, μ, we are trying to estimate, is another random variable, C, that is correlated with Y and whose expected value is known or known to be the same as $Y's$. A new estimator for μ can be constructed that is a linear combination of Y and C, and if the combining is done in the right way, then the new estimator will be unbiased and will have variance less than the variance of Y alone.

This paper investigates the consequences of applying a multivariable form of this variance reduction technique to the problem of estimating a value function. In this setting, the control variables are the current estimate, $\tilde{V}$, together with the truncated corrected returns, $X^{[k]}$. Admittedly, the algorithms suggested by the full-blown control-variable approach, even when realizable, are computationally intensive, but the real goal of this analysis is to gain increased insight into temporal difference methods and, in particular, to improve upon heuristic schemes like the one offered by Watkins for introducing or "scheduling" bias—the efforts summarized in this paper seek more sophisticated methods based upon a rigorous theory.

2 A MINIMUM VARIANCE APPROACH

Consider the (infinite) family of estimators, (1), for $V(x_t)$, together with $X^{[0]} = \tilde{V}(x_t)$. If the $\tilde{V}(\cdot)$ are unbiased, then so are the $X^{[k]}$, and the unbiased linearly weighted combination of the $X^{[k]}$, $\sum_{k=0}^{\infty} w_k X^{[k]}$, having minimum variance is obtained by choosing [1]

$$\boldsymbol{w}^* = \frac{\Sigma_X^{-1}\mathbf{1}}{\mathbf{1}'\Sigma_X^{-1}\mathbf{1}} \tag{2}$$

—where Σ_X denotes the covariance matrix of the estimators $X^{[k]}$ and $\mathbf{1}$ is a column vector of 1's. The resulting minimum variance estimator has

$$Var\left\{\sum_{k=0}^{\infty} w_k^* X^{[k]}\right\} = \boldsymbol{w}^{*\prime}\Sigma_X \boldsymbol{w}^* = \frac{1}{\mathbf{1}'\Sigma_X^{-1}\mathbf{1}}. \tag{3}$$

With regard to the covariance matrix Σ_X, consider a generic entry:

$$[\Sigma_X]_{m,n} = Cov(X^{[m]}, X^{[n]}) = Cov\left(\sum_{k=0}^{m-1}\gamma^k r_{t+k} + \gamma^m \tilde{V}(x_{t+m}), \sum_{k=0}^{n-1}\gamma^k r_{t+k} + \gamma^n \tilde{V}(x_{t+n})\right).$$

Without loss of generality, suppose that $m \leq n$, and let

$$Z' \stackrel{def}{=} \left[\begin{array}{cccccc} r_t & r_{t+1} & \cdots & r_{t+n-1} & \tilde{V}(x_{t+m}) & \tilde{V}(x_{t+n}) \end{array}\right].$$

Then

$$\begin{aligned} X^{[m]} &= \left[\begin{array}{cccccccccc} 1 & \gamma & \gamma^2 & \cdots & \gamma^{m-1} & 0 & \cdots & 0 & \gamma^m & 0 \end{array}\right] Z = C_m' Z \\ X^{[n]} &= \left[\begin{array}{cccccccccc} 1 & \gamma & \gamma^2 & \cdots & \gamma^{m-1} & \gamma^m & \cdots & \gamma^{n-1} & 0 & \gamma^n \end{array}\right] Z = C_n' Z. \end{aligned}$$

But $Cov(C_m' Z, C_n' Z) = C_m' \Sigma_Z C_n$, where Σ_Z is the covariance matrix associated with Z. One can now address the (less messy) problem of computing, estimating, or approximating entries in Σ_Z. Generic entries are of the form (1) $Cov(r_{t+i}, r_{t+j})$, (2) $Cov(r_{t+i}, \tilde{V}(x_{t+m}))$ or $Cov(r_{t+i}, \tilde{V}(x_{t+n}))$, and (3) $Cov(\tilde{V}(x_{t+m}), \tilde{V}(x_{t+n}))$.

[1] This (Gauss-Markov) formula for w^* may be derived by, for example, using Lagrange multipliers to solve the constrained optimization problem: $\min_{\boldsymbol{w}} E\left\{\left(\sum_{k=0}^{\infty} w_k X^{[k]} - V(x_t)\right)^2\right\}$, subject to $\sum_k w_k = 1$.

Supposing for the moment that w^* could be computed, the new estimate for $V(x_t)$ is simply $\tilde{V}^{(1)}(x_t) = \sum_{k=0}^{\infty} w_k^* X^{[k]}$. Optimally weighted estimates could be constructed for the other states as well, and for each state the variance of the corresponding new $\tilde{V}^{(1)}$ decreases in a well-defined way given by Equation 3. The variance (covariance) structure of the improved $\tilde{V}^{(1)}$, in turn, enters into the calculation of the next iteration of optimal weight calculations through the generic entries of type (2) and (3) in Σ_Z listed in the previous paragraph. Conceptually, the result of following this procedure is that each state will be assigned a distinct set of weights that change value in an "open loop" fashion as value-function updates occur. The weight prescription is the "best" in the sense of expected dynamics and updates over an entire ensemble of simulations whose initial estimates are unbiased and have an initial covariance matrix specified by $\Sigma_{V^{(0)}}$. One straightforward way of obtaining initial unbiased value function estimates, $\tilde{V}^{(0)}$, would be to simply use the sample mean (computed from one or more trials) of accumulated discounted rewards along sample paths of length k starting from each state, where k is chosen to be large enough to ensure that γ^k is small. Using one long sample path to construct such estimates for more than one state necessarily implies some degree of correlation between the intial set of $\tilde{V}$'s.

3 FULL BACKUP CASE

If the expected rewards, R, and transition matrix, P, are known, then one can consider the following family of ("fully backed up") estimators for $V(x_t)$: with x_t assumed to be state I, $X^{[k]} = \sum_{i=0}^{k-1} \gamma^i \sum_{j=1}^{N} (P^i)_{Ij} R(j) + \gamma^k \sum_{j=1}^{N} (P^k)_{Ij} \tilde{V}(j), \quad k = 0, 1, 2, ...,$ —where N denotes the number of states. It can be shown, using the same methods as in the previous section, that generic entries in the covariance matrix Σ_X are given by

$$Cov(X^{[m]}, X^{[n]}) = \gamma^{m+n} \sum_{k=1}^{N} (P^m)_{Ik} \sum_{j=1}^{N} (P^n)_{Ij} Cov(\tilde{V}(k), \tilde{V}(j)), \tag{4}$$

and the covariance matrix associated with the new set of estimates has entries

$$Cov(V_I^{(1)}, V_J^{(1)}) = \sum_{m=1}^{N} \left\{ \left(\sum_{k=0}^{\infty} w_k^I \gamma^k (P^k)_{Im} \right) \left(\sum_{n=1}^{N} Cov(V(m), V(n)) \sum_{k=0}^{\infty} w_k^J \gamma^k (P^k)_{Jn} \right) \right\} \tag{5}$$

—where w_k^I denotes the k^{th} optimal weight for state I.

Example: Figure 2 plots summed-squared error versus iteration (average results over 25 trials of initial estimates having standard normal errors) for the full back-up version of the iterative Gauss-Markov algorithm with estimator sets of size 2,3, and 5. [2] applied to a ten-state problem with $\gamma = .7$. As expected, as the number of

[2] For example, for an estimator set of size 2, $\{X^{[0]}, X^{[1]}\}$, taking $m = 0$ and $n = 1$ in Equation 4 leads to

$$\Sigma_X = \begin{bmatrix} Var(\tilde{V}(I)) & \gamma \sum_{j=1}^{N} (P)_{Ij} Cov(\tilde{V}(I), \tilde{V}(j)) \\ \gamma \sum_{j=1}^{N} (P)_{Ij} Cov(\tilde{V}(I), \tilde{V}(j)) & \gamma^2 \sum_{k=1}^{N} (P)_{Ik} \sum_{j=1}^{N} (P)_{Ij} Cov(\tilde{V}(k), \tilde{V}(j)) \end{bmatrix},$$

estimators increases, so does the rate of convergence. As $k \to \infty$, $X^{[k]}$ becomes the Neumann series for V, and the corresponding w_k dominates the other weights. Similarly, for the case of two estimators (Equation 6), the optimal weights (optimal α) place more emphasis on $X^{[1]}$ (in fact, it is not unusual for $X^{[0]}$ to have *negative* weight [3]). The following analysis further explains why an optimal weighting with an associated value of $\alpha > 1$ should not be surprising.

Equation 6 may be re-written as $V^{(k+1)} = [\alpha\gamma P + (1-\alpha)I]\, V^{(k)} + \alpha R$, and the fixed point, V^*, satisfies this equation identically. Subtracting it from both sides yields an equation for the error: $\epsilon^{(n+1)} = \underbrace{[\alpha\gamma P + (1-\alpha)I]}_{M_\alpha} \epsilon^{(k)}$. If $\mathbf{e}$ is an eigenvector for P with eigenvalue λ, then $\mathbf{e}$ is also a eigenvector for M_α with eigenvalue

$$\lambda_\alpha = \gamma\alpha\lambda + (1-\alpha). \tag{7}$$

The behavior of the error is governed by the eigenvalue of M_α having largest magnitude. By setting the magnitudes of the values of λ_{min} and λ_{max} under the mapping defined by (7) equal to one another, one arrives at the optimal choice of $\alpha^* = \frac{2}{2-\gamma(1-\lambda_{min})}$ (see Figure 3).

4 SAMPLE-BASED METHODS

Can the sample-based approach of Section 2 be made to work in way similar to the model-based approach of the last section? One could construct sample covariance matrices, $\hat{\Sigma}_X$, in the usual way, but computational experience strongly suggests that using a sample covariance matrix, which has some amount of sampling variance (error), in "optimal" weight calculations does not work well. And unfortunately, analytical approaches must contend with the problem of calculating the covariance between *mixtures* rather than simple random variables; *i.e.,* $Var(V(x_{t+1}))$ is *not* the same thing as $\sum_j P_{x_t,j} Var(V(j))$.

Suppose that R and P are unknown, but that the process has been observed as it evolves and accumulates rewards and that this information has been recorded in the following matrix and vector:

$$\hat{P} = \begin{bmatrix} \frac{n_{11}}{n_1} & \frac{n_{12}}{n_1} & \cdots & \frac{n_{1N}}{n_1} \\ \frac{n_{21}}{n_2} & \frac{n_{22}}{n_2} & \cdots & \frac{n_{2N}}{n_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{n_{N1}}{n_N} & \frac{n_{N2}}{n_N} & \cdots & \frac{n_{NN}}{n_N} \end{bmatrix} \qquad \hat{R} = \begin{bmatrix} \frac{\sum_{i=1}^{n_1} r_i(1)}{n_1} \\ \frac{\sum_{i=1}^{n_2} r_i(2)}{n_2} \\ \vdots \\ \frac{\sum_{i=1}^{n_N} r_i(N)}{n_N} \end{bmatrix} \tag{8}$$

---

and the rule for updating entries in the covariance matrix Σ_V is just the Equation 5 with "∞" replaced by "1" in the summations over k. These expressions, along with Equation 2, in essence specify how α should change in the TD(0)-type update rule:

$$V^{(k+1)}(I) = (1-\alpha)V^{(k)}(I) + \alpha \left[R(I) + \gamma \sum_{j=1}^{N} P_{Ij} V^{(k)}(j) \right] \tag{6}$$

[3] R. Sutton & S. Singh have communicated similar empirical results.

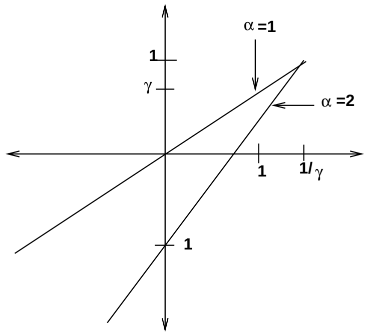

Figure 3: Geometry of optimal α.

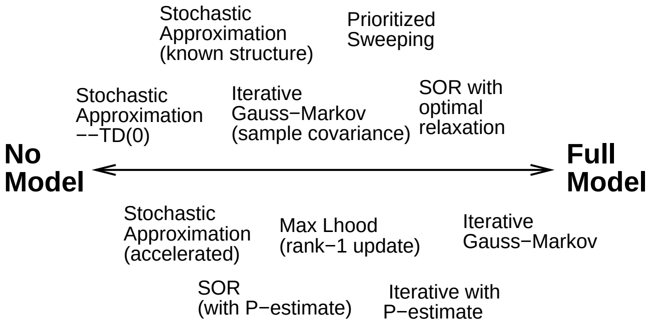
Stochastic Approximation (known structure)
Prioritized Sweeping
Stochastic Approximation --TD(0)
Iterative Gauss−Markov (sample covariance)
SOR with optimal relaxation
No Model
Full Model
Stochastic Approximation (accelerated)
Max Lhood (rank−1 update)
Iterative Gauss−Markov
SOR (with P−estimate)
Iterative with P−estimate

Figure 4: The model/no-model "continuum."

—where N is the number of states, n_{ij} is the number of $i \rightarrow j$ transitions observed, n_i is the total number of transitions from state i observed, and $r_i(k)$ denotes the i^{th} sample of reward observed in transit from state k. These are the maximum-liklihood estimates of P and R given the observed data, and one may derive confidence intervals for $\hat{P}$ and $\hat{R}$ as well (Nanthi & Wassan, 1987) (Billingsley, 1961).

Motivated by the preceding analysis of ideal values for α, consider the relaxation recursion with P and R replaced by their estimates: $V^{(k+1)} = \left[(1-\alpha)I + \alpha\gamma\hat{P}\right] V^{(k)} + \alpha\hat{R}$. V^* satisfies Equation 6 identically, and subtracting it from both sides results in an equation for the evolution of error: $V^{(k+1)} - V^* = [(1-\alpha)I + \alpha\gamma P]\,(V^{(k)} - V^*) + \alpha(\Delta R + \gamma\Delta P\hat{V}^{(k)})$, where $\Delta R = \hat{R} - R$ and $\Delta P = \hat{P} - P$ are asymptotically normal. The error equation may be approximated by $\epsilon^{(k+1)} = I\epsilon^{(k)} + (\gamma P - I)\alpha\epsilon^{(k)} + \alpha w$, where w is vector of normal, mean-zero noise. This is a disrete-time, stochastic, *bilinear* system (Mohler, 1980) that we would like to drive to zero as quickly as possible (through choice of α), and there exist advanced methods for doing so (*e.g.,* Gutman, 1981). A simple method, based on simply minimizing the expected squared error at the next step, suggests choosing $\alpha^{(k+1)} = \frac{\epsilon'(I-\gamma\hat{P})\epsilon}{\epsilon'(I-\gamma\hat{P})'(I-\gamma\hat{P})\epsilon+\sum_i \sigma_i^2}$, where σ_i is the i^{th} component of the mean zero noise, which should decay to 0 as $\hat{P}$ and $\hat{R}$ converge.

5 Discussion: TD-relaxation with Models and Approximate Models

With regard to the full backup version of the TD(0)-type recursion, Equation 6, it was shown in a Section 3 that, under certain assumptions, the α yielding fastest convergence to a fixed point has a value greater than one; *i.e.,* that the relaxation should be *over-relaxed.* On the other hand, if we have no knowlege of the transition matrix P, the usual TD(0) scheme (which is a form of stochastic approximation), having observed a transition from state i to j, in essence takes P to be a matrix of all zeroes but for a single value of one in the i, j^{th} position (*i.e.,* it uses an instantaneously sampled version of the true model), and prescribes a choice of α that is severely *under-relaxed.* This suggests that there is a continuous spectrum of optimal choices for α—that the best α depends on the degree of confidence one has in the "model," P.

Figure 4 views TD(0) as lying at the extreme no-model end of the spectrum. At the opposite end of the spectrum, methods choose their step-sizes in an optimal (minimum variance) fashion based on full knowledge of system parameters. In between these two extremes lie a mixture of methods including those based on constrained models (*e.g.,* stochastic approximation acceleration) and estimated models (*e.g.,* incremental maximum liklihood).

The "incremental maximum liklihood algorithm" works as follows: Suppose that, after having observed a number of transitions, $\hat{P}$ (and $\hat{R}$) estimates have been constructed as in Equation 8, $(I-\gamma\hat{P})^{-1}$ has been computed, and that we now observe a transition from state i to state j. The $i \rightarrow j$ transition changes only the i^{th} row of $\hat{P}$, and the corresponding change in $(I-\gamma\hat{P}_{new})^{-1}$ can be computed without starting from scratch by making use of the following well-known "rank-one update"

Theorem: *Let A be a matrix with inverse $N = A^{-1}$. Let h be a column vector and β a row vector. Then $\bar{N} = (A - h\beta)^{-1}$ exists iff $\beta N h \neq 1$ and, if so,*

$$\bar{N} = N + \frac{(Nh)(\beta N)}{1-\beta N h}.$$

Let h be a column vector of all zeros but for a "1" in the i^{th} position, and $\beta = -\frac{\gamma}{n_i(n_i+1)}[\; n_{i1} \quad n_{i2} \quad \cdots \quad n_{ij}-n_i \quad \cdots \quad n_{iN} \;]$. Then

$$(I-\gamma\hat{P}_{new})^{-1} = (I-\gamma\hat{P}_{old})^{-1} + \frac{[i^{th} \;\; column \;\; of \;\; (I-\gamma\hat{P}_{old})^{-1}][\beta(I-\gamma\hat{P}_{old})^{-1}]}{1-[i^{th} \;\; entry \;\; of \;\; \beta(I-\gamma\hat{P}_{old})^{-1}]}.$$

This formula gives rise to an online update rule for $(I-\gamma P)^{-1}$ that takes approximately N^2 operations per observation. The update gives the *exact* inverse. Gauss-Seidel takes number-of-contractions $\times N^2$ per observation, where number-of-contractions depends on the accuracy desired and is usually small but may occasionally spike to $> N$ operations. Prioritized Sweeping (Moore, 1993) makes use of maximum liklihood estimates (8) as well and empirically outperforms Gauss-Seidel. It could be argued that the rank-one update algorithm is relatively straightforward to understand/implement; there are no parameters like Moore's ϵ or β, and convergence analysis is straightforward since it is directly coupled to the convergence of $\hat{P}$ and $\hat{R}$ to the true system parameters.

Some final comments regarding Figure 4: Since the early 1950's when Stochastic Approximation was first introduced, there have been a number of schemes proposed for accelerating convergence. Among them, (Kesten, 1958) suggests maintaining step size at nomimal values until a change in sign of successive estimates occurs; (Jacobs, 1988) is recent related work. (Venter, 1967)'s method accelerates the rate of convergence by, in effect, estimating the slope of the underlying regression function at the desired root. More recently, (Dupuis & Simha, 1991) have suggested taking multiple samples at given operating points to reduce variance. Apart from these acceleration schemes for improving the standard algorithm, it should be noted that stochastic approximation is a rather general method for computing roots of noisy (nonlinear) regression functions. In the case of the value-estimation problem, however, the sought-after value occurs at the intersection of a collection of *linear* manifolds— the standard algorithm could be improved upon by exploiting this fact (in Venter's method, for example, each data point contributes information about the slope of the regression function at its root).

Acknowledgements

Thanks to Andy Barto and to the members of the Adaptive Networks Laboratory. This work was supported in part by the National Science Foundation under Grant ECS-9214866 to Professor Barto.

References

A. Barto & M. O. Duff (1994). Monte Carlo Matrix Inversion and Reinforcement Learning. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kaufmann.

P. Billingsley (1961). *Statistical Inference for Markov Processes* Univerity of Chicago Press.

P. Dupuis & R. Simha (1991). On sampling controlled stochastic approximation. *IEEE-TAC* 36:915-924.

P. Gutman (1981). Stabilizing Controllers for Bilinear Systems. *IEEE-TAC* 26(4), pp. 917-922.

R. Jacobs (1988). Increased Rates of Convergence Through Learning Rate Adaptation. *Neural Networks* 1:295-307.

H. Kesten (1958). Accelerated Stochastic Approximation. *Ann. Math. Statist.* 29:41-59.

S. Lavenberg & P. Welch (1981). A perspective on the use of Control Variables to increase the efficiency of Monte-Carlo simulations. *Management Science* 27:322-335.

R. Mohler & W. Kolodzki (1980). An Overview of Stochastic Bilinear Control Systems. *IEEE-TAC* pp. 917-922.

A. Moore & C. Atkeson (1993). Prioritized Sweeping: Reinforcement Learning with Less Data. *Machine Learning* 13:103-130.

K. Nanthi & M. Wassan (1987). *Statistical Estimation for Stochastic Processes*, Queen's Papers in Pure and Applied Mathematics, no. 78.

R. Sutton (1988). Learning to Predict by the Method of Temporal Differences. *Machine Learning* 3:9-44.

J. Venter (1967). An Extension of the Robbins-Monro procedure. *Ann. Math. Statist.* 38:181-190.

C. Watkins (1989). Learning from Delayed Rewards. PhD Thesis Cambridge University.