

Utility maximising stochastic dynamic programming: An overview

Andrew Kerr
Department of Management
University of Canterbury
New Zealand
a.kerr@mang.canterbury.ac.nz

Abstract

Many real world problems involve making repeated decisions over time in an uncertain environment. These decisions often involve a trade-off between some immediate benefit(s) and possible future benefit(s), and also take in to account the impact the decision will have on future decisions and benefits. Stochastic dynamic programming (SDP) is often used to analyse problems of this type and the objective is often to maximise the expected value of benefits, which can imply that the decision-maker is 'risk neutral'. But is this appropriate? In this paper a SDP formulation is described which accommodates risk attitudes via a utility function. The approach is discussed and illustrated for stochastic reservoir management and stochastic route choice problems.

1 Introduction

There are numerous OR problems which involve making decisions over time in an uncertain environment. These problems can quickly become complex depending on aspects such as the number and form of objectives, the way in which uncertainty is represented, and the relationships between variables. An important characteristic of many of these problems is that uncertainty is resolved during the planning process and the decision maker has the ability to make decisions which are conditional on the past history of outcomes. A suitable solution technique for these problems will therefore produce a dynamic solution rather than a static solution, where the latter does not allow for any adjustment through the planning horizon.

The general problem of interest is as follows. The planning horizon consists of $t=1\dots T$ finite stages. The finite state of the system is denoted by s^t , and is constrained in each period by $s^t \in S^t$. The decision at each stage is denoted by q^t , and is constrained by $q^t \in Q^t$. For simplicity, we assume that the state of the system is represented by a single state variable, and that a single decision is made in each stage. Uncertainty during each stage is reflected by a^t , where the probabilities of event \bar{a}^t are modelled as independent distribution or as a Markov process. The state of the system in $t+1$ is described by the function, $g^t(q^t, s^t, a^t)$ and the return in each period is $r^t(q^t, s^t, a^t)$; both functions may not be dependent on all these variables. The problem to be solved can be stated as follows.

$$\mathbf{P1} \quad \max_{q^t} \mathbb{E} \left[\sum_{t=1}^T r^t(q^t, s^t, a^t) \right] \quad (1)$$

$$\text{subject to:} \quad s^{t+1} = g^t(q^t, s^t, a^t) \quad (2)$$

$$s^t \in S^t, q^t \in Q^t \quad (3)$$

If we assume that the DM has a linear utility function, then the expected sum of the returns is equal to the sum of the expected returns [4]. A stochastic dynamic programming formulation can now be described for this problem

$$\mathbf{P2} \quad f^t(s^t, a^t) = \max_{q^t} \left(r(q^t, s^t, a^t) + \mathbb{E}[f^{t+1}(s^{t+1}, a^{t+1}) | a^t] \right) \quad (4)$$

$$\text{subject to:} \quad s^{t+1} = g^t(q^t, s^t, a^t) \quad (5)$$

$$s^t \in S^t, q^t \in Q^t \quad (6)$$

where $f^t(s^t, a^t)$ is the maximum expected total return from t to T given that the state of the system is (s^t, a^t) in period t . Note that if a^{t+1} is independent of a^t , a^t does not need to be included as a description of the state of the system. Discounting is ignored here but it could be incorporated easily.

But, what if the DM is not risk neutral (which is the implicit assumption behind taking expected values)? There are relatively few studies which consider utility maximisation in a multi-stage setting. Krautkraemer, van Kooten, and Young [12] describe one of the first applications of SDP which incorporates ‘risk’. The approach is discussed in the context of agricultural planning. A utility function is defined for the return in each period, and the total utility is the sum of the individual utilities i.e.,

$$U(r^1 + r^2 + \dots + r^T) = \sum_{t=1}^T u(r^t) | s^t \quad (7)$$

This approach accommodates preferences towards intra-period outcomes, but does not explicitly model attitudes towards outcomes over the entire planning horizon. This objective is additively separable, so a recursive relation as in **P2** can be used.

Ranatunga [14] describes an SDP approach which handles preferences of the form

$$U(r^1 + r^2 + \dots + r^T) = u \left(\sum_{t=1}^T r^t \right) \quad (8)$$

This objective is non-separable because utility depends on the returns achieved in all periods, and cannot be calculated by adding the expected utility of each individual benefit. Therefore, the objective is non-separable, invalidating the use of the recursive relation as defined in **P2**, and hence dynamic programming. In order to overcome the problem of non-separability, another state variable, w^t , is defined as the accumulated returns (or ‘wealth’) up to the beginning of period t , $w^t = \sum_{t=1}^{t-1} r^t$, where $w^1 = 0$. Total

returns are therefore defined as $w^{T+1} = \sum_{t=1}^T r^t$ where $w^{t+1} = w^t + r^t$. More generally, we can state that $w^{t+1} = h^t(w^t, r^t, s^t, a^t)$. The utility function can therefore be defined over

the range of w^{T+1} which is not dependent on all the decisions made over the planning horizon, but only on the decision(s) and state of the system in T ; w^T is dependent on the decisions/state in $T-1$, and so on. Kall and Wallace [6] applied essentially the same technique to a 3-stage investment problem with a continuous wealth state variable. The non-linear utility function, defined over terminal wealth outcomes, is passed back to each stage in its function form. However, the conditions for doing so are dependent on the assumption about initial wealth and require non-negative returns. The technique would also become intractable for large problems and when utility is multidimensional. The concept of using the past history as a state variable has been mentioned in [10,11], though they do not (re-)state the problem in the way described above.

The SDP formulation to solve Problem **P1** with the objective described in Equation (8) is therefore

$$\mathbf{P2} \quad f^t(w^t, s^t) = \max_{q^t} E[f^{t+1}(w^{t+1}, s^{t+1}) | a_t] \quad (9)$$

$$\text{subject to:} \quad s^{t+1} = g^t(q^t, s^t, a^t) \quad (10)$$

$$w^{t+1} = h^t(w^t, q^t, s^t, a^t) \quad (11)$$

$$s^t \in S^t, q^t \in Q^t \quad (12)$$

$$w^1 = 0 \quad (13)$$

The terminal value function is the utility associated with the state of the system at the end of the last period: $f^{T+1}(w^{T+1}, s^{T+1}) = U(w^{T+1}, s^{T+1})$. Formulation **P2** is referred to as UMSPD: Utility Maximising Stochastic Dynamic Programming.

Ranatunga applied UMSPD to the purchase and sale of forward contracts so a risk averse DM owning thermal plant with significant intertemporal risks could hedge against price uncertainty. He only considered unidimensional utility functions which were a function of w^{T+1} . Kerr, Read, and Kaye [9] applied UMSPD to medium-term reservoir planning with inflow uncertainty. They used bidimensional utility functions which were a function of both w^{T+1} and s^{T+1} (storage). Reservoir planning is no different than any other SDP application area in that an objective of maximising an expected value is the norm, even though inflow uncertainty is frequently acknowledged as a 'risk'.

Kerr [8] recognised that UMSPD could be applied to problems which involve finding a path through a directed stochastic acyclic network. Problems with this structure are often referred to as stochastic route choice (SRC) problems. The objective of SRC problems is to maximise a function of the outcomes associated with each arc in a path. Not surprisingly, there are a relatively small number of studies which address the case where the DM has some non-linear preference or attitude towards the distribution of terminal outcomes. Even fewer produce a dynamic solution which is conditional on the outcomes of uncertain events.

The remainder of this paper includes a discussion and illustration of the application of UMSPD to a medium-term reservoir management problem (Section 2) and to a SRC problem (Section 3). A summary is presented in Section 4.

2 A reservoir planning problem

Reservoir management is concerned with the planning of reservoir releases, and the resulting hydro generation. It is an interesting and complex problem because water is storable commodity, so there is a continuous process of deciding whether to release it now, or to store it and release it at a later date, where the time frame for these decisions can range from minutes to months. In New Zealand, for example, approximately 70% of the annual national energy demand is supplied from hydro sources. Detailed planning of reservoir operation is crucial because of uncertainty about the level of natural inflows, the fact that the aggregate storage capacity is only 6 weeks (approximately) of national demand, and because thermal generation is relatively expensive.

SDP often applied to reservoir management problems; the scheduling horizon divides naturally into discrete time periods (the stages of the SDP). Storage is the state variable which links the stages [15]. Reservoir planning models typically minimise expected costs, or maximise expected profits, implying that the firm or DM is risk neutral. This may or may not always be a valid assumption, depending on the environment in which the reservoir is operated and any other tools which have been used to hedge against risk. Bergara and Spiller [2] describe a static economic model of New Zealand's electricity market where energy suppliers (and retailers) are all described as being risk averse, noting the significant uncertainty caused by inflow uncertainty. The now extinct Electricity Corporation of New Zealand (ECNZ) developed a planning model which incorporated this approach in a dual dynamic programming framework [3]. Aside from research motivations, there appears to be support from both theoretical and practical standpoints for considering non-'risk neutral' attitudes in the context of energy planning.

UMSDP is applied here in the context of a firm (the 'player' or DM) which operates a single reservoir in an energy market which consists of a number of other firms ('competitors') supplying energy. The player controls the quantity of reservoir release, and UMSDP is used to plan these releases for the coming year (52 weeks) given the uncertain inflows and DM's attitude to the state of the system at the end-of-horizon. Inflows are represented by independent distributions. The player is contracted for some amount of generation in each week, where the contract type is a call for differences. The contract quantity and price is known a priori, and contract revenue is received in the period that the energy is supplied. Demand must be met in each period from reservoir release and generation from the competitors. The competitors are assumed to behave as 'perfect competitors' in the sense that the player has perfect knowledge of the capacities and marginal costs of these firms when making the release decision in each period, and they will generate any feasible amount required to satisfy demand.

The objective is to maximise the DM's expected utility of end-of-horizon net wealth and end-of horizon storage. The end-of-horizon value function, f^{T+1} , is therefore described by calculating the utility of arriving at different points (accumulated wealth and storage levels) in the end-of-horizon state space. The form of utility we use here is

$$f^{T+1}(w^{T+1}, s^{T+1}) = U(w^{T+1}, s^{T+1}) = u_w(w^{T+1}) + u_s(s^{T+1}) \quad (14)$$

The UMSDP formulation is as follows

$$\text{RM1} \quad f^t(w^t, s^t) = \max_{q^t} E[f^t(w^{t+1}, s^{t+1}) | a^t] \quad (15)$$

$$\text{subject to:} \quad w^{t+1} = w^t + r^t(q^t, s^t) \quad (16)$$

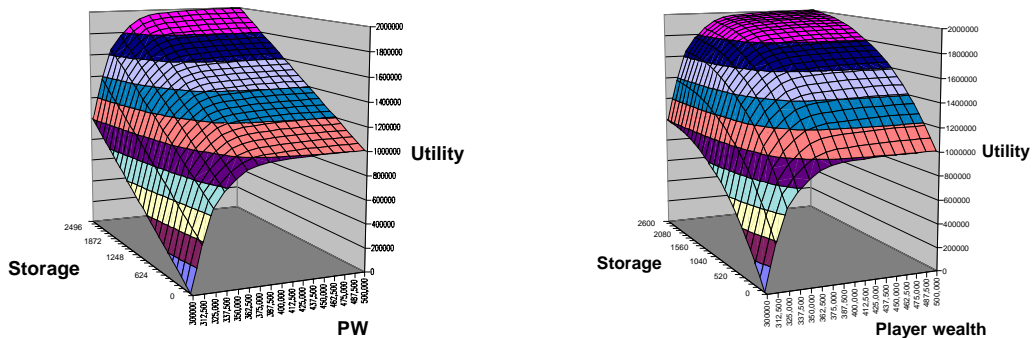
$$s^{t+1} = s^t - q^t + a^t \quad (17)$$

$$s^t \in S^t, q^t \in Q^t \quad (18)$$

$$w^1 = 0 \quad (19)$$

The utility for a particular end-of-horizon wealth and storage is calculated in (14), where u_w and u_s are concave non-decreasing utility functions which reflect risk aversion [7]. The additive form of $U(\bullet)$ described here implies that w^{T+1} and s^{T+1} are utility independent [7]. That is, the utility associated with a particular level of s^{T+1} is independent of the level of w^{T+1} , and vice versa with respect to wealth. In reality, the utility of storage and wealth would probably be utility dependent, though the form used here serves as a useful starting point. In each week, we choose the release level (q_t) that maximises the expected end of horizon utility (15) given the distribution of inflows (a_t) in the period. Equations (16) and (17) describe the state transitions for wealth and storage. Note that the calculation of $r^t(q^t, s^t)$ considers deterministic exogenous variables such as demand, competitor supply curves, and the contract quantity/price. Inflows are assumed to occur at the end of the period so $q^t > s^t$, which is a conservative representation of reality. Storage and release bounds are defined in (18) and (19), and the initial level of accumulated wealth is 0 (20). The approach described here for solving **RM1** is discrete dynamic programming. See [9] for more detail.

The base case for these experiments is where the player is risk neutral in wealth and storage. Experiments were also performed using the utility functions illustrated in Table 1. The form of the utility functions used here implies a trade-off between storage and wealth, as opposed to only being risk averse towards wealth, so the impact of these functions of both components is important.



(a) 'Moderate' risk aversion – W4S0 (b) W4S2' High' risk aversion – W4S0

Table 1: Combinations of U_w and U_s

The output from the optimisation is an optimal weekly release surface which describes the optimal release should a wealth/storage pair be arrived at in that week. For concave u_w and u_s , release is non-decreasing in both storage and wealth. For the risk neutral

case, the release will change in the storage dimension but will be constant for all any value of wealth.

Simulations were performed for 20 representative years of inflow data. Table 2 shows the CDFs for end-of-horizon wealth (a) and storage (b). The impact of increasing risk aversion on the wealth CDFs is that they compress and become more upright. Minimum wealth increases, maximum wealth decreases, and mean wealth increases (\$4m) using W4S2 and decreases (\$42m) for W4S0. The standard deviations of wealth are consistent with the shapes of the CDFs, being \$45m for the RN case, \$28m for W4S2, and \$8m for W4S0. For W4S0, the effect of achieving such low variability in expected wealth (albeit with a lower mean) is to hold more water in storage, though the standard deviation of expected end-of-horizon storage is also slightly higher than the RN case (376GWh compared to 348GWh for RN). The risk averse storage CDFs are not too dissimilar from the RN CDF.

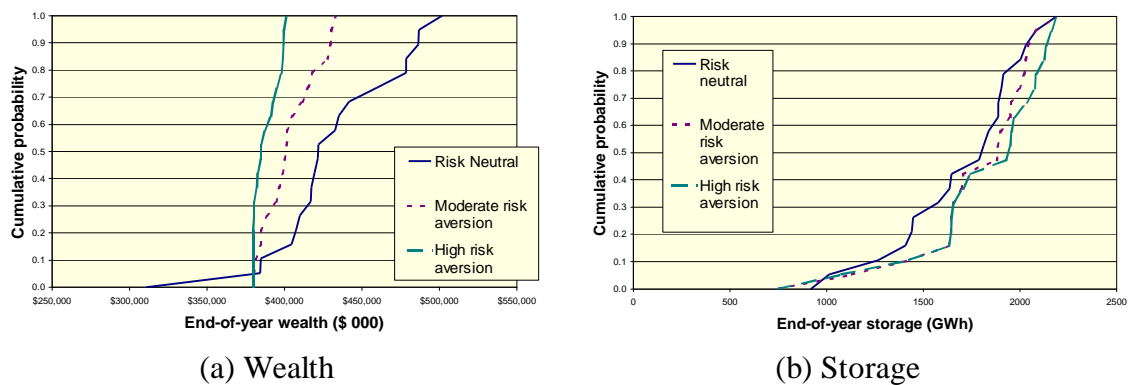


Table 2 End-of-horizon CDFs

Compared to risk neutral results, risk aversion had the expected impact on wealth, reducing the variation of end-of-horizon wealth outcomes, in some cases quite dramatically. This was at the expense of increasing the variability of end-of-horizon storage levels, though the increase was small compared to the impact on wealth. The change in mean wealth and storage was influenced by the nature of the trade-off between storage and wealth in the utility function. Overall, using a risk averse utility curve produces release schedules which substantially reduce the variability in end-of-horizon wealth outcomes while not greatly affecting end-of-horizon storage.

3 A stochastic route choice problem

Many planning problems can be represented as discrete, directed, acyclic networks. The generic problem statement is: given an origin node, destination node, and a number of intermediary nodes all linked by arcs with uncertain lengths, determine a path from the origin to the destination that maximises some objective. These problems are often referred to as stochastic route choice (SRC) problems. Note that the length of an arc typically reflects a cost or distance or travel time and that nodes can be linked to more than one other node. (For the remainder of this discussion, arc lengths reflect travel time).

Let T_N be the time taken to reach the destination node, N . If the objective is to minimise the expected travel time, $\min E[T_N]$, the optimal path can be found using a variety of methods such as dynamic programming, complete enumeration, and integer

programming. A SDP approach involves defining $f(i)$ as is the minimum, and optimal, expected travel time (or length or cost) from node i to N . The minimum expected travel time, $f(1)$, can be found by solving:

$$\mathbf{SRC1} \quad f(i) = \min_{j \in J(i), k \in K(i,j)} \mathbb{E} [t_k + f(j)] \quad \forall i \neq I \quad (20)$$

$$f(I) = 0 \quad (21)$$

where $i=1 \dots N$ is a finite set of nodes; $k=1 \dots K$ is a finite set of directed arcs linking a pair of nodes (i,j) where $i < j$; $J(i)$ is the set of nodes succeeding i ; and $K(i,j)$ is the set of nodes linking node i to j . The objective of minimising the expected travel time implicitly assumes that the DM is risk neutral with respect to the total travel time. That is, the marginal value of a decrease in the total travel time is constant.

Consider now the situation where $U(T_N)$ reflects the DM's preferences towards the total time taken to reach the destination node, T_N . When the objective is defined this way, the optimal path must be determined by considering the expected utility of this total time. The objective is therefore no longer separable, and the problem can not be formulated using SDP formulation **SRC1**. Several authors have stated that to determine an optimal solution to this problem requires complete enumeration of all possible routes [1,13]. Algorithms have been developed to reduce the size of the network (e.g. [5,13]) or to derive a solution using simulation [1].

While the performance of some of these approaches is impressive, the major drawback is that the solutions are static in an environment where the uncertainty is dynamic. Static solutions to SRC problems are determined prior to any realisation of uncertainty, while dynamic solutions allow the DM to adjust the path once some uncertainty has been realised. Static solutions are certainly appropriate in many situations (e.g. route choice for hazardous materials), though there are many contexts in which a dynamic solution appears to be a more consistent with the decision making environment. However, there are few approaches for finding solutions to these problems because the problem is "notoriously intractable" [13].

We observe that the UMSDP technique can be applied to this problem in order to overcome the problem of the objective being a non-separable function of the arc lengths. Simply define another state variable, T_i , which is the accumulated travel time upon reaching node i (and departing from it). Because we are assuming the network is directed and acyclic, the node precedence relationships are known, so the time taken to reach node i is simply the sum of the arc lengths from the origin to i . The terminal value function is $f(N, T_N) = U(T_N)$ and the problem can be formulated as a stochastic dynamic program:

$$\mathbf{SRC2} \quad f(i, T_i) = \min_{j \in J(i), k \in K(i,j)} (\mathbb{E}[f(j, T_j) | t_k]) \quad \forall i \neq I \quad (22)$$

$$\text{subject to} \quad T_j = T_i + t_k \quad (23)$$

$$T_1 = 0 \quad (24)$$

Starting from stage $N-1$ and working backwards, $f(i, T_i)$ is evaluated for each arc and the optimal decision from each stage stored. The accumulated time variable is a

continuous variable which is discretised and evaluated at discrete values at each stage. The bounds on T_i can be calculated by working through the network, starting from the origin. Obviously the range of T_i will expand for stages further away from the origin, with the destination node having the largest range of T_i values when arc lengths are non-negative.

3.1 Example SRC problem

We now consider a small SRC problem which involves finding a utility maximising route through the network illustrated Figure 1a. The arc length distributions were normally distributed as follows: arcs 1, 4, and 6 were $N \sim (11,2)$; arcs 2 and 5 were $N \sim (21,1)$; arc 3 was $N \sim (30,10)$, and arc 7 was $N \sim (12,1)$.

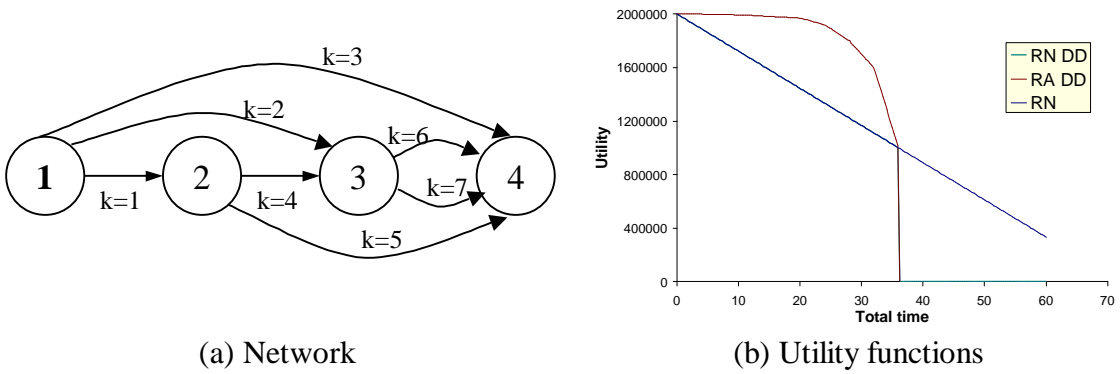


Figure 1: SRC example

Three utility functions were defined over the range of completion times (Figure 1b). The base case is when the DM is risk neutral (RN), which corresponds to the negatively sloped dashed line. The other two utility functions reflect the case where there is a limit on the completion time, and this is reflected by a ‘decreasing deadline’ utility function of the form

$$U(T_N) = \begin{cases} u(T_N) & 0 \leq T_N \leq d \\ 0 & T_N > d \end{cases} \quad (25)$$

If $T_N > d$, the entire process considered to be worthless and this is reflected by utility being zero, regardless of the level of T_N . The ‘RNDD’ and ‘RADD’ utility curves reflect risk neutral and risk averse preferences towards T_N when $T_N < d$, where $d=36$.

Discrete dynamic programming is again used to solve **SRC2** using the data and utility functions described above. The solutions are summarised in Table 3. The RN solution (arc 3) takes no account of the variability of the arc times, preferring the (combination of) arcs with the lowest mean travel time. The RNDD solution involves using arc 2 to reach node 3, then either arc 6 or 7, depending on the time of arrival at node 3 i.e., the length of arc 2. The RADD solution takes a more conservative decision at node 1, moving to node 2 via arc 1. If the realisation of t_1 is less than 15.4, arc 5 is selected to move to the destination node. Arc 1 is $N \sim (11,2)$, so the path 1-4 is most likely to occur. However, if $t_1 > 15.4$, arc 4 is selected to reach node 3, at which time another set of conditional arc selections is described. The main difference between the RADD and

RNDD policies is that risk aversion results in a more strategy which is more sensitive to the distributions of the arc lengths.

	Node 1	Node 2	Node 3
RN	Arc 3	-	-
RNDD	Arc 2	-	Arc 6: $T_3 < 19$ $21 < T_3 < 22.4$ $T_3 > 22.8$ Arc 7: $19 < T_3 < 21$ $22.4 < T_3 < 22.8$
RADD	Arc 1	Arc 5: $T_2 < 15.4$ Arc 4: $T_2 > 15.4$	Arc 6: $T_3 < 18.9$ $21.2 < T_3 < 22.4$ $T_3 > 22.8$ Arc 7: $18.9 < T_3 < 21.2$ $22.4 < T_3 < 22.8$

Table 3: SRC solution

Consider the decision process at node 3 for the RNDD and RADD cases. For low (good) T_3 where there is no chance of $T_4 > d$, the means and variances of arcs 6 and 7 are such that arc 6 has a higher expected utility than arc 7. Let \mathbf{m}_k denote the mean length of arc k . Because the means and variances of arcs 6 and 7 are similar, for T_3 near $d - \mathbf{m}_k$, there are ranges of T_3 for which each arc is preferred. However, arc 6 has a lower mean and higher variance, so for large values of T_3 there will be a higher probability that $T_4 < d$, and hence an expected utility greater than that for arc 7. Arc 6 is therefore preferred for both RADD and RNDD when T_3 is large.

4 Summary

This paper has described an SDP approach for solving multi-stage problems where a utility function reflects preferences towards end-of-horizon outcomes. An objective of this form is non-separable because the utility associated with a particular end-of-horizon outcome is dependent on all the decisions and realisations of uncertainty over the planning horizon. Treating accumulated returns, w^t , as a state variable means that a terminal value function defined over w^{T+1} is separable. The technique was originally defined for a sequential decision process where w^{t+1} is only dependent on w^t and r^t [14]. The technique can also be applied to decision problems where w^{t+1} is a function of w^t for any $t < T+1$ i.e., a directed acyclic network [8].

The technique was illustrated for a reservoir management problem and a route choice problem. For the reservoir management problem, utility maximisation produced weekly release decisions consistent with the preferences toward the distributions of end-of-horizon storage and wealth. These results were borne out by simulation results. For the route choice problem, utility was a function of total travel time, and the optimal arc selections reflected the cumulative effect of arc variability on expected utility. In both cases, the state variable is unbounded. In addition to the appropriateness and form of

utility function(s), the accuracy of the discretisation and overall tractability of the technique are areas for future research.

References

- [1] J. F. Bard & J. E. Bennett, Arc Reduction and Path Preference in Stochastic Acyclic Networks. *Management Science* **37**(2), 198-215, 1991.
- [2] M. E. Bergara & P. T. Spiller, Competition and Direct Access in New Zealand's Electricity Market. In *Deregulation of electric utilities*, Zaccour, G. Ed., Kluwer, Boston, 1998.
- [3] M. Craddock, A. D. Shaw, & B. Graydon, Risk-Averse Reservoir Management in a Deregulated Electricity Market. In *Proceedings of the 33rd ORSNZ Conference*, Auckland, New Zealand, 157-166, 1998.
- [4] E. V. Denardo, *Dynamic Programming: Models and Applications*, Prentice-Hall, New Jersey. 1982.
- [5] A. Eiger, P. B. Mirchandani, & H. Soroush, Path Preferences and Optimal Paths in Probabilistic Networks. *Transportation Science* **19**(1), 75-84, 1985.
- [6] P. Kall & S. W. Wallace, *Stochastic Programming*, John Wiley and Sons, New York, 1994.
- [7] R. L. Keeny & H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley and Sons, New York, 1976.
- [8] A. L. Kerr, Utility Maximising Dynamic Route Selection in Acyclic Stochastic Networks. In *Proceedings of the First Western Pacific and Third Australia-Japan Workshop on Stochastic Models*, Christchurch (New Zealand), 269-277, 1999.
- [9] A. L. Kerr, E. G. Read, & R. J. Kaye, *Stochastic dynamic programming applied to medium-term reservoir management: Maximising the utility of a system supply cost minimiser*. EMRG Working Paper EMRG-WP-97-03, Department of Management, University of Canterbury, New Zealand, 1997.
- [10] J. O. S Kennedy, J. B. Hardker & J. Quiggin, Incorporating Risk Aversion into Dynamic Programming Models: Comment, *American Journal of Agricultural Economics*, **76**, 960-964, 1994.
- [11] D. M. Kreps, Decision Problems with Expected Utility Criteria, I: Upper and Lower Convergent Utility, *Mathematics of Operations Research*, **2**(1), 45-53, 1977.
- [12] J. A. Krautkraemer, G. C. van Kooten & D. L. Young, Incorporating Risk Aversion into Dynamic Programming Models, *American Journal of Agricultural Economics*, **74**, 870-878, 1992.
- [13] I. Murthy & S. Sarkar, Exact Algorithms for the Stochastic Shortest Path Problem with a Decreasing Deadline Utility Function. *European Journal of Operational Research* **103**(1), 209-229, 1997.
- [14] R. A. S. Ranatunga, *Risk averse operation of an electricity plant in an electricity market*. ME dissertation, School of Electrical Engineering, University of New South Wales, 1995.
- [15] W. W-G Yeh, Reservoir management and operations models: A state-of-the-art review, *Water Resources Research*, **21**(12), 1797-1818, 1985.