Maintenance and Repair Decision Making for Infrastructure Facilities without a Deterioration Model

Pablo L. Durango-Cohen[1]

Abstract: In the existing approach to maintenance and repair decision making for infrastructure facilities, policy evaluation and policy selection are performed under the assumption that a perfect facility deterioration model is available. The writer formulates the problem of developing maintenance and repair policies as a reinforcement learning problem in order to address this limitation. The writer explains the agency-facility interaction considered in reinforcement learning and discuss the probing-optimizing dichotomy that exists in the process of performing policy evaluation and policy selection. Then, temporal-difference learning methods are described as an approach that can be used to address maintenance and repair decision making. Finally, the results of a simulation study are presented where it is shown that the proposed approach can be used for decision making in situations where complete and correct deterioration models are not (yet) available.

DOI: 10.1061/(ASCE)1076-0342(2004)10:1(1)

CE Database subject headings: Infrastructure; Stochastic models; Decision making; Rehabilitation; Maintenance.

Introduction

In the existing model-based approach for maintenance and repair decision making, policy evaluation and policy selection are performed under the assumption that a (stochastic) deterioration model is a perfect representation of a facility's physical deterioration process. This assumption raises several concerns that stem from the simplifications that are necessary to model deterioration and the uncertainties in the choice or the estimation of a model. The assumptions that deterioration is Markovian and stationary are examples of the former, while the uncertainty that exists in generating transition probabilities for the Markov decision process approach is an example of the latter. In addition, the model-based approach assumes that the data necessary to specify a deterioration model are available. This ignores the complexity, the cost, and the time required to collect reliable sets of data and, therefore, limits the effectiveness of this approach in many situations. Examples include the implementation of infrastructure management systems for developing countries or for the management of certain types of infrastructure that have not been studied extensively, such as office buildings, theme parks, or hospitals.

In this paper, the writer introduces temporal-difference (TD) learning methods, a class of reinforcement learning methods, as an approach to maintenance and repair decision making for infrastructure facilities. TD learning methods do not require a model of deterioration and, therefore, can be used to address the concerns presented in the preceding paragraph.

Maintenance and Repair Decision Making

The agency-facility interaction considered in maintenance and repair decision making for infrastructure facilities is illustrated in Fig. 1. An agency reviews facilities periodically over a planning horizon of length T. At the start of every period $t \in \{1,2,\ldots,T\}$, the agency observes the state of a facility $X_t \in \mathsf{S}$, decides to apply an action to the facility $A_t \in \mathsf{A}$, and incurs a cost $g(X_t,A_t) \in \mathcal{R}$, that depends both on the action and the facility condition. This cost structure can capture the costs of applying maintenance and repair actions as well as the facility's operating costs. In pavement management, for example, operating costs correspond to the users' vehicle operating costs. At the end of the planning horizon, the agency receives a salvage value $s(X_{T+1}) \in \mathcal{R}$ that is a function of the terminal condition of the facility.

The existing approach to maintenance and repair decision making is referred to as a model-based approach because it involves modeling the effect of actions on (changes in) condition. Policies are ‘‘evaluated’’ by using a deterioration model, a cost function $g(\cdot)$, and a salvage value function $s(\cdot)$ to predict the effect of the actions prescribed by a policy on the sum of discounted costs incurred over a planning horizon. The function of planning for maintenance and repair of infrastructure facilities is referred to as policy selection. It involves finding or constructing a policy that minimizes the sum of the predicted costs.

Existing optimization models for maintenance and repair decision making constitute applications of the ‘‘equipment replacement problem’’ introduced by Terborgh (1949). Bellman (1955) and Dreyfus (1960) formulated the problem as a dynamic control problem. Fernandez (1979) and Golabi et al. (1982) adapted and extended the formulation to address maintenance and repair decision making for infrastructure facilities and networks, respectively. Reviews of optimization models that address the management of infrastructure facilities are presented by Gendreau and Soriano (1998) and Durango (2002). The formulations can be classified as either deterministic or stochastic depending on the model used to represent deterioration. Stationary Markovian models, a class of stochastic models, are widely used and accepted

[1] Assistant Professor, Dept. of Civil and Environmental Engineering, Transportation Center, Northwestern Univ., 2145 Sheridan Rd., A335, Evanston, IL 60208. E-mail: pdc@northwestern.edu

Note. Discussion open until August 1, 2004. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on October 25, 2002; approved on July 29, 2003. This paper is part of the ***Journal of Infrastructure Systems***, Vol. 10, No. 1, March 1, 2004. ©ASCE, ISSN 1076-0342/2004/1-1–8/$18.00.

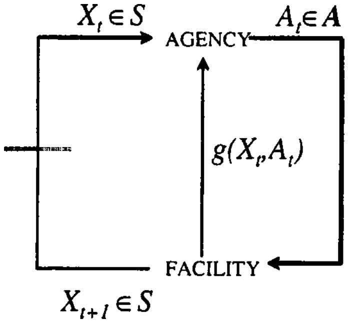

Fig. 1. Agency-facility interaction in maintenance and repair decision making

because of their strong properties and because optimal policies can be computed by solving either a linear or a dynamic program.

The research presented here constitutes an approach to maintenance and repair decision making that is radically different than the existing model-based approach. TD methods only assume that infrastructure facilities are managed under a periodic review policy. This makes them attractive because it is not necessary to make strong assumptions about deterioration. Complete coverage of reinforcement learning can be found in the works by Bertsekas (1995) and Sutton and Barto (1998).

Reinforcement Learning Framework

In this section policies, value functions, state-action value functions, and ϵ-greedy policies are defined in the context of reinforcement learning. These definitions are useful in the presentation of reinforcement learning methods that can be used to develop maintenance and repair policies for infrastructure facilities.

Policy

A policy is a list that specifies a course of action for every possible contingency that an agency can encounter in managing a facility. Mathematically, a policy is a mapping from the set of states and periods $\mathsf{S}\times\{1,2,\ldots,T\}$, to the set of probability mass functions over the set of actions A. The mapping is denoted P or $P_t(x,a)$, $\forall x\in\mathsf{S}$, $a\in\mathsf{A}$, $t\in\{1,2,\ldots,T\}$. Each element of the mapping $P_t(x,a)$ is the probability that action a is taken when the state of the facility is x and the period is t. Hence, a well-defined policy P must satisfy the following basic properties:

$$P_t(x,a)\geqslant 0,\ \forall x\in\mathsf{S},\ a\in\mathsf{A},\ t\in\{1,2,\ldots,T\} \tag{1}$$

$$\sum_{a\in\mathsf{A}} P_t(x,a)=1,\ \forall x\in\mathsf{S},\ t\in\{1,2,\ldots,T\} \tag{2}$$

When a policy specifies a unique action (as opposed to a distribution over the set of actions) for each pair (x,t) in the set $\mathsf{S}\times\{1,2,\ldots,T\}$, that is, $P_t(x,a)\in\{0,1\}$, $\forall x\in\mathsf{S}$, $a\in\mathsf{A}$, $t\in\{1,2,\ldots,T\}$, it is referred to as a *deterministic* policy. Otherwise, when a policy can specify any action in the convex hull of A for each pair (x,t) in the set $\mathsf{S}\times\{1,2,\ldots,T\}$, it is referred to as a *randomized* policy. An example of a randomized policy is an ϵ-soft policy. In an ϵ-soft policy, each available action for every state has a probability of appearing that is ϵ or greater, that is, $P_t(x,a)\geqslant\epsilon$, $\forall a\in\mathsf{A}$, $x\in\mathsf{S}$, $t\in\{1,2,\ldots,T\}$.

The value P is used to denote the set of candidate policies being considered to manage the facility. It is usually assumed that every action is available regardless of the state of the facility or the period. Therefore, the number of deterministic policies in P is $|\mathsf{A}|^{|\mathsf{S}|\cdot T}$.

When a policy specifies the same probability mass function over set A in every period, it is referred to as a *stationary* policy. The subindex t can be omitted from $P_t(\cdot)$.

Return

The function R_t is used to denote the sum of discounted costs from the start of period t until the end of the planning horizon. Mathematically

$$R_t\equiv\sum_{t'=t}^{T}\delta^{t'-t}\cdot g(X_{t'},A_{t'})-\delta^{T+1-t}\cdot s(X_{T+1}),\quad \forall t\in\{1,2,\ldots,T+1\} \tag{3}$$

where $\delta\in(0,1]$=discount factor. Note that the return is a function of the random variables $X_t, X_{t+1},\ldots,X_{T+1}$ and the decision variables $A_t, A_{t+1},\ldots,A_T$.

Value Functions and State-Action Value Functions

The value function under a given policy maps each pair (x,t) in the set $\mathsf{S}\times\{1,2,\ldots,T+1\}$ to the expected return that follows an observation of the given state-period pair. For a policy P and a given state of the facility at the start of t, $X_t=x$, the value function yields the expected return that results from following P, given the current state of the facility x. The mapping is denoted $V_t^P(X_t)$. Mathematically, the value function for a policy P is defined as follows:

$$V_t^P(X_t=x)\equiv E_{\{A_t,X_{t+1},A_{t+1},X_{t+2},A_{t+2},\ldots,X_T,A_T,X_{T+1}|P\}}[R_t|X_t=x]\quad \forall x\in\mathsf{S},\ t\in\{1,2,\ldots,T+1\} \tag{4}$$

Similarly, for a policy P, a given state at the start of t, $X_t=x$, and an action for the current period $A_t=a$, a state-action value function is defined as the expected return that results from taking action a in the current period, given the state of the facility x and following policy P thereafter. The mapping is denoted $Q_t^P(X_t, A_t)$. Mathematically

$$Q_t^P(X_t=x,\ A_t=a)\equiv E_{\{X_{t+1},A_{t+1},X_{t+2},A_{t+2},\ldots,X_T,A_T,A_{T+1}|P\}}[R_t|X_t=x,A_t=a]\quad \forall x\in\mathsf{S},\ a\in\mathsf{A},\ t\in\{1,2,\ldots,T+1\} \tag{5}$$

Estimates of value functions and state-action value functions are represented with $v(\cdot)$ and $q(\cdot)$.

ϵ-Greedy Policies

A class of ϵ-soft policies known as ϵ-greedy policies is defined here. These policies are widely used in TD methods described in the next section. Let $a^*(x)$, $\forall x\in\mathsf{S}$ be $\operatorname{argmin}_{a\in\mathsf{A}}Q^P(x,a)$. The writer defines an ϵ-greedy policy with respect to policy P as a policy $\hat{P}$ such that

$$\hat{P}(x,a)=\begin{cases}1-\epsilon+\dfrac{\epsilon}{|\mathsf{A}|} & \text{if } a=a^*(x)\\[2ex] \dfrac{\epsilon}{|\mathsf{A}|} & \text{otherwise}\end{cases}\quad \forall a\in\mathsf{A},\ x\in\mathsf{S} \tag{6}$$

Thus, for a small number ϵ, an ϵ-greedy policy is a policy where the "greedy," best available actions are selected in each state with a large probability $1-\epsilon$ and where random actions are selected with a small probability ϵ.

Reinforcement Learning Methods for Infrastructure Management

In this section, the writer presents TD learning methods for policy evaluation and policy selection. This class of reinforcement learning methods can be used to address maintenance and repair decision making without a deterioration model.

For simplicity in presenting TD learning methods, it is assumed that the length of the planning horizon is infinite ($T \rightarrow \infty$), and the physical deterioration process corresponds to a stationary, Markovian process specified with transition probabilities $\Pi_{ij}(a)$, $\forall i,j \in \mathsf{S}, a \in \mathsf{A}$. A technical description of policy evaluation and policy selection in the context of this paper is presented in Appendices I and II.

Temporal-Difference Learning Methods for Policy Evaluation

Policy evaluation in TD methods does not require a facility deterioration model. That is, TD methods for policy evaluation do not use (estimates of) the transition probabilities $\Pi_{i,j}(a)$, $\forall i,j \in \mathsf{S}, a \in \mathsf{A}$ to find a solution to the system of Bellman's equations [Eq. (14)]. TD methods solve the system of equations iteratively by updating estimates of value functions and state-action value functions based on experience in managing/probing a facility and on prior estimates. The former makes these methods interaction-based methods. The latter implies that these methods can be categorized as bootstrapping methods. Policy evaluation in TD methods is performed by probing/sampling a facility for m periods. Estimates of value functions or state-action value functions are updated based on the costs incurred during the probing period as well as on prior estimates. As an example, the writer has considered a TD($m=1-$step) method for policy evaluation. In the TD($m=1-$step) method, a given estimate of the value function $v^P(x)$, $\forall x \in \mathsf{S}$ for a given policy P is updated by probing the facility during the current period. In probing the facility during the current period, an agency observes the initial state of the facility, i, the cost incurred in the period (based on i and the action a prescribed by P for i), $g(i,a)$, and the state of the facility at the end of the period, j. The value $g(i,a)+\delta v^P(j)$ is the *target* that can be constructed with the information gathered while probing the facility. A new estimate of the value function is generated by updating the prior estimate in the direction of the temporal-difference error. The temporal-difference error is given by the target minus the prior estimate. Thus, the new estimate is constructed as follows:

$$v^P(i) \leftarrow v^P(i) + \alpha[g(i,a) + \delta v^P(j) - v^P(i)] \tag{7}$$

where α denotes a step-size; and the quantity in the square brackets=temporal-difference error.

Policy evaluation can be performed by applying the procedure described previously iteratively. A complete TD($m=1-$step) algorithm for policy evaluation is presented here.

```
TD($m=1-$step) algorithm
Given a policy $p$
    Initialize estimates: $v_0^p(x)$, $\forall x \in \mathsf{S}$
    Initialize the counters for observations of each state $k(x) \leftarrow 0$, $\forall x \in \mathsf{S}$
    Let $i$ be the initial state of the facility.
Repeat for each period
    $a \leftarrow$ action prescribed by $p$ for $i$
    Take $a$, observe $g(i,a)$ and $j$
    $\text{target}_{k(i)+1}(i) \leftarrow g(i,a) + \delta v_{k(j)}^P(j)$
    $v_{k(i)+1}^P(i) \leftarrow v_{k(i)}^P(i) + \alpha_{k(i)+1}[\text{target}_{k(i)+1}(i) - v_{k(i)}^P(i)]$
    $k(i) \leftarrow k(i)+1$, $i \leftarrow j$.
```

The TD($m=1-$step) method for policy evaluation is shown to converge to the value function under P, V^P if the step sizes satisfy the following two conditions:

$$\sum_{k(x)=1}^{\infty} \alpha_{k(x)} = \infty, \quad \forall x \in \mathsf{S} \tag{8}$$

$$\sum_{k(x)=1}^{\infty} \alpha_{k(x)}^2 < \infty, \quad \forall x \in \mathsf{S} \tag{9}$$

To understand why this is so, consider the case of step sizes given by $\alpha_{k(x)} = 1/k(x)$, $\forall x \in \mathsf{S}$, which satisfy conditions [Eqs. (8) and (9)]. Note that $\forall x \in \mathsf{S}$, $k(x)=1,2,\ldots$

$$\begin{aligned}
v_{k(x)}^P(x) &= v_{k(x)-1}^P(x) + \alpha_{k(x)}[\text{target}_{k(x)}(x) - v_{k(x)-1}^P(x)] \\
&= \frac{1}{k(x)}[\text{target}_{k(x)}(x) + k(x) v_{k(x)-1}^P(x) - v_{k(x)-1}^P(x)] \\
&= \frac{1}{k(x)}[\text{target}_{k(x)}(x) + (k(x)-1) v_{k(x)-1}^P(x)] \\
&= \frac{1}{k(x)}\Bigg\{\text{target}_{k(x)}(x) + (k(x)-1)\frac{1}{k(x)-1} \\
&\qquad \times [\text{target}_{k(x)-1}(x) + (k(x)-2) v_{k(x)-2}^P(x)]\Bigg\} \\
&= \frac{1}{k(x)}[\text{target}_{k(x)}(x) + \text{target}_{k(x)-1}(x) + (k(x)-2) \\
&\qquad \times v_{k(x)-2}^P(x)] \\
&= \cdots = \frac{1}{k(x)}\left[\sum_{n=1}^{k(x)} \text{target}_n(x)\right]
\end{aligned}$$

The law of large numbers states that the last expression, the average target, converges to the expected target as $k(x) \rightarrow \infty$, $\forall x \in \mathsf{S}$. The expected targets can be used in place of the right-hand side of Eq. (16) to update value function estimates, and it is intuitively understandable that as $k(x) \rightarrow \infty$, $\forall x \in \mathsf{S}$, the TD($m=1-$step) estimates converge to the same results obtained with the fixed point iteration algorithm.

Temporal-Difference Control Methods for Policy Evaluation and Policy Selection

Policy evaluation and policy selection with TD methods are performed while an agency is managing a facility. It is usually the case that there are significant costs incurred while an agency is probing the facility to perform these functions. In transportation infrastructure management, for example, these costs are of consideration because review periods are typically long (which limits opportunities to probe facilities), and future cost savings are heavily discounted. It follows that a critical component in the design and implementation of TD methods (as well as other interaction-based methods) is to devise efficient methods to perform policy evaluation and policy selection while managing a facility. Such methods must provide a systematic approach to learn as much as possible about the facility without incurring excessive costs. These often contradictory objectives correspond to the probing-optimizing dichotomy that exists in managing infrastructure facilities. One way to achieve a balance with respect to these objectives is to select ϵ-greedy actions with respect to the current estimates of the state-action value functions. By implementing this type of policy, an agency can manage facilities efficiently while ensuring adequate exploration. The writer presents a complete TD control algorithm. The algorithm is called **SARSA** due to the manner in which the update of the state-action values is performed in each period. Given initial estimates of the state-action value function under the ϵ-greedy policy, the sequence of events is as follows. First, the agency observes the initial **S**tate i and selects an **A**ction a'. Next, the agency incurs a cost (receives a **R**eward) $g(i,a')$, observes the **S**tate of the facility at the end of the period j, and selects an **A**ction a'', for the next period. Finally, the agency uses the information gathered through probing the facility to update the prior estimate of the relevant state-action value. The algorithm is presented below.

```
SARSA
Initialize $q(x,a),\ \forall x \in \mathsf{S}, a \in \mathsf{A}$
Observe $i$
Choose $\epsilon$-greedy action $a'$ based on $q(i,a),\ \forall a \in \mathsf{A}$
Repeat in each period
    Apply $a'$
    Observe $g(i,a')$, and $j$
    Choose $\epsilon$-greedy action, $a''$ based on $q(j,a),\ \forall a \in \mathsf{A}$
    $q(i,a') \leftarrow q(i,a') + \alpha[g(i,a') + \delta q(i,a'') - q(i,a')]$
    $i \leftarrow j,\ a' \leftarrow a''$
```

The convergence of SARSA to the optimal policy and the optimal state-action value function depends on the physical deterioration process and on the schemes employed to choose the parameters ϵ and α. If the assumptions presented at the start of the previous section hold, SARSA converges to the optimal policy and state-action value function, provided that the following three conditions hold:

1. Each state-action pair is visited infinitely often;
2. The policy converges to the greedy policy; and
3. The step size decreases, but not too quickly. Mathematically, the step size must satisfy Eqs. (8) and (9).

Generalizations

In this section, four generalizations to the basic TD methods presented in the previous section are described. These generalizations are important because there are many situations where they can increase the convergence rate of the basic TD algorithms. As stated earlier, this is an important consideration in the design of interaction-based methods. This is particularly important in developing maintenance and repair policies for infrastructure facilities where it is not possible to probe facilities extensively because review periods are typically long and future cost savings tend to be heavily discounted. In addition, agencies usually want to receive cost savings in the early part of the planning horizon.

TD(m−step) Methods for Policy Evaluation

The first generalization is to probe the facility for an extended period of time when evaluating a given policy. The updating rule for a general TD(m−step) method is given as follows:

$$v^p(i) \leftarrow v^p(i) + \alpha\left[\sum_{k=1}^{m} \delta^{k-1} g_k + \delta^m v^p(j) - v^p(i)\right] \quad (10)$$

where, in this case, j=state of the facility that is observed after m periods; and the sequence of costs incurred in the next m periods is given by $(g_k, k=1,\ldots,m)$. The expression $\Sigma_{k=1}^{m}\delta^{k-1}g_k + \delta^m v^p(j) = \mathrm{TD}(m-\mathrm{step})$ target.

By increasing the "probing" period m, an agency is effectively relying more on experience than on prior estimates to generate new estimates of value functions (or state-action value functions). This seems to be a good idea in situations where an agency does not have confidence in its initial estimates.

Temporal-Difference Learning Methods with Eligibility Traces

The second generalization involves making efficient use of the samples that are generated over the planning horizon to update the value function estimates. One approach is to increase the number of samples that are used to update the value function for each state that is visited. This is done by considering the samples that are generated by probing the facility for different time durations, that is, TD(m−step) samples can be generated for different values of m. As an example, the writer presents the updating rule for the case where the facility is sampled for both one and two periods ($m=1$ and $m=2$).

$$v^p(i) \leftarrow (1-\mu)\{(1-\alpha)v^p(i) + \alpha[g_1 + \delta v^p(j_1)]\} + \mu\{(1-\alpha)v^p(i) + \alpha[g_1 + \delta g_2 + \delta^2 v^p(j_2)]\} \quad (11)$$

where j_1 and j_2=states observed after one and two periods respectively; g_1 and g_2=corresponding costs; and $\mu \in [0,1]$=relative weight that is assigned to each sample. Note that if $\mu = 0.5$, the samples are weighted equally.

TD methods with eligibility traces are a generalization of this idea. The details are presented by Sutton and Barto (1998). One such algorithm is shown.

```
TD algorithm with eligibility traces
Given a policy $p$
    Initialize estimates: $v_0^p(x),\ \forall x \in \mathsf{S}$
    Initialize the memory records for each state $e(x) \leftarrow 0,\ \forall x \in \mathsf{S}$
    Let $i$ be the initial state of the facility
Repeat for each period
    $a \leftarrow$ action prescribed by $p$ for $i$
    Take $a$, observe $g(i,a)$ and $j$
    tderror$\leftarrow g(i,a) + \delta v^P(j) - v^P(i)$
    $e(i) \leftarrow e(i) + 1$
    For all $x \in \mathsf{S}$
        $v^P(x) \leftarrow v^P(x) + \alpha \cdot \mathrm{tderror} \cdot e(x)$
        $e(x) \leftarrow \delta\lambda e(x)$
        $i \leftarrow j$
```

The memory records $e(x),\ \forall x \in \mathsf{S}$ are referred to as eligibility traces. The parameter λ is used to weight the samples. Its role is similar to μ presented in the previous example. The value of α denotes the step-size. The methods are usually referred to as TD(λ) methods. In the simulation study presented next, consider the case of $\lambda = 1$, which corresponds to a facility being sampled indefinitely.

Q-Learning

The third generalization is to replace the target that is used in SARSA with $g(i,a) + \delta \min_{a \in \mathsf{A}}\{q(j,a)\}$. The new control algorithm is called Q-learning. Q-learning is an off-policy control algorithm because (with probability ϵ) the target is not specified with the action that is specified by the ϵ-greedy policy for j. The intuition behind this method is that the Q-learning target yields better estimates of optimal value functions or state-action value functions. This can decrease the number of iterations that are necessary to converge to a policy that satisfies Eq. (17) (Bellman's optimality principle).

TD Methods with Function Approximation

The fourth generalization involves choosing a function to approximate a value function or state-action value function. This scheme is referred to as a TD method with function approximation. The samples generated by probing the facility are then used to generate/update a set of parameters that specify the function. An advantage of using this scheme is that instead of generating/updating estimates for each element of the value function ($O(|\mathsf{S}|)$ estimates) or of the state-action value function ($O(|\mathsf{S}|\cdot|\mathsf{A}|)$ estimates), it is only necessary to generate/update estimates for each of the parameters that specify the functional approximation. A disadvantage is that the convergence of this scheme is highly dependent on the quality of the approximation to the value function or state-action value function. This method is described further in the experimental design section of the case study.

This method can be classified as a model-based approach because it involves modeling the effect of actions on the sum of expected discounted costs. This is different than the existing approach that involves modeling the effect of actions on condition and assumes a correspondence between condition and costs.

Case Study: Application of Temporal-Difference Learning Methods to Development of Maintenance and Repair Policies for Transportation Infrastructure Facilities

In this section is described the implementation of TD methods for maintenance and repair decision making of infrastructure facilities. Specifically, the results of a simulation study are presented in the context of pavement management, where the writer has used the TD methods described in the previous section for the problem of fine-tuning incorrect policies. The study is meant to represent situations where there is uncertainty in specifying a deterioration model. Initially, an agency can generate a deterioration model based on available data and/or experience in managing similar facilities. An agency then chooses to either implement a maintenance and repair policy assuming that the pavement will deteriorate according to its initial beliefs or to use its initial beliefs to estimate the state-action value function and use a TD control method to fine-tune the policy while managing the pavement.

Table 1. Costs (Dollars/Lane-Yard)

Pavement condition	Maintenance and repair actions							User costs
	1	2	3	4	5	6	7	
1	0.00	6.90	19.90	21.81	25.61	29.42	25.97	∞
2	0.00	2.00	10.40	12.31	16.11	19.92	25.97	25.00
3	0.00	1.40	8.78	10.69	14.49	18.30	25.97	22.00
4	0.00	0.83	7.15	9.06	12.86	16.67	25.97	14.00
5	0.00	0.65	4.73	6.64	10.43	14.25	25.97	8.00
6	0.00	0.31	2.20	4.11	7.91	11.72	25.97	4.00
7	0.00	0.15	2.00	3.91	7.71	11.52	25.97	2.00
8	0.00	0.04	1.90	3.81	7.61	11.42	25.97	0.00

Table 2. Means and Standard Deviations of Action Effects on Change in Pavement Condition

	Deterioration model	
	Slow	Fast
Standard deviation	0.30	0.70
Action	Mean effect	
1	−0.25	−1.75
2	0.50	−0.50
3	1.75	0.25
4	3.00	1.00
5	4.25	1.75
6	5.50	2.50
7	8.00	4.00

The data for the case study are taken from empirical studies presented in the literature on pavement management. The writer considers a discount rate $(1/\delta-1)$ of 5%. As presented, TD methods assume an infinite planning horizon. In the case study we assume that an agency manages a facility over an infinite planning horizon. However, we only account for the sum of discounted costs over the first 25 years.

According to Carnahan et al. (1987), pavement condition is given by a PCI rating discretized into eight states (State 1 being failed pavement and State 8 being excellent pavement). There are seven maintenance and repair actions available in every period and for every possible condition of the pavement. The actions considered are (1) do nothing; (2) routine maintenance; (3) 1-in. overlay; (4) 2-in. overlay; (5) 4-in. overlay; (6) 6-in. overlay; and (7) reconstruction. The costs of performing actions are also taken from Carnahan et al. (1987). The operating costs considered in the study were taken from Durango and Madanat (2002) and are meant to represent the users' vehicle operating costs that are associated with the condition of the pavement. The costs are presented in Table 1 and are expressed in dollars/lane-yard.

It is assumed that the actual deterioration of the pavement is governed by one of two stationary, Markovian models: slow or fast. The transition probabilities were generated using truncated normal distributions shown by Madanat and Ben-Akiva (1994). The mean effects of applying maintenance and repair actions and the standard deviations associated with each deterioration model are presented in Table 2. The transition probabilities are presented by Durango (2002).

The optimal state-action value functions are presented in Tables 3 and 4 and are expressed in dollars/lane-yard. These poli-

Table 3. Optimal State-Action Value Functions: Slow Deterioration Model (Dollars/Lane-Yard)

	Action						
Condition	1	2	3	4	5	6	7
1	∞	∞	∞	∞	∞	∞	∞
2	∞	∞	64.06	52.40	48.00	47.62	51.80
3	61.81	55.74	47.94	40.33	40.32	41.13	48.80
4	42.73	35.42	30.06	27.59	27.71	31.50	40.80
5	25.20	19.81	17.86	15.65	19.26	23.08	34.80
6	12.93	10.31	7.79	8.94	12.74	16.55	30.80
7	7.15	4.80	4.83	6.74	10.54	14.35	28.80
8	1.59	0.87	2.73	4.64	8.44	12.25	26.80

Table 4. Optimal State-Action Value Functions: Fast Deterioration Model (Dollars/Lane-Yard)

	Action						
Condition	1	2	3	4	5	6	7
1	∞	∞	∞	∞	∞	∞	∞
2	∞	∞	∞	∞	∞	123.40	113.08
3	∞	∞	∞	∞	111.86	107.07	103.54
4	∞	∞	102.86	95.41	91.18	89.07	91.89
5	∞	87.13	82.20	76.75	75.33	75.64	84.85
6	85.71	71.08	66.33	63.68	64.47	66.86	80.78
7	71.48	60.50	58.40	57.83	60.68	64.34	78.78
8	60.13	53.43	53.35	54.69	58.42	62.23	76.78

cies and state-action value functions were computed with the generalized policy iteration algorithm presented in Appendix II.

Experimental Design

The goal of the simulation study is to test the performance of different TD control algorithms in fine-tuning incorrect maintenance and repair policies. Two cases are considered:

1. Where an agency manages a pavement whose deterioration is governed by the fast model but initializes the state-action value function according to the slow model; and
2. Where the deterioration is slow and the state-action value function is initialized with the fast model.

The TD control algorithms considered in the study are: SARSA, Q-learning, TD with eligibility traces $(\lambda=1)$, and TD with function approximation.

In the TD control method with function approximation, the function that was used to approximate the state-action value function is

$$q(i,a)=\begin{cases}\infty & \text{for}\quad i=1\\ \gamma_1+\gamma_2 a+\gamma_3\dfrac{a^2}{\sqrt{i}}+\gamma_4 i+\gamma_5 ai & \text{otherwise}\end{cases}\tag{12}$$

The function was chosen to fit the optimal state-action value functions presented in Tables 3 and 4. The parameters in each case are obtained with an linear regression of the finite values of the state-action value function. A summary of the regression results is presented in Table 5. In the implementation of the control method, the parameters are updated by considering the TD$(m=1-\text{step})$ targets as additional observations of the state-action value function.

The policy that is followed for each of the methods is such that with probability $\epsilon=0.1$, an action in the set $\{a^*(i)-1,a^*(i),a^*(i)+1\}$ was chosen at random. The step size used to update the estimates of the state-action value function α was set to 0.25. Each experiment is identified by a case-algorithm pair and consisted of 100 instances of managing a pavement whose initial condition was six.

Table 5. Regression Results

	Slow model		Fast model	
R^2	0.9190		0.8916	
Adjusted R^2	0.9113		0.8776	
Coefficient	Value	*t* statistic	Value	*t* statistic
γ_1	96.08	16.30	171.84	11.65
γ_2	−15.35	−6.59	−17.53	−4.11
γ_3	1.43	5.39	1.57	4.06
γ_4	−15.24	−14.26	−17.10	−7.24
γ_5	2.34	7.76	2.24	4.03

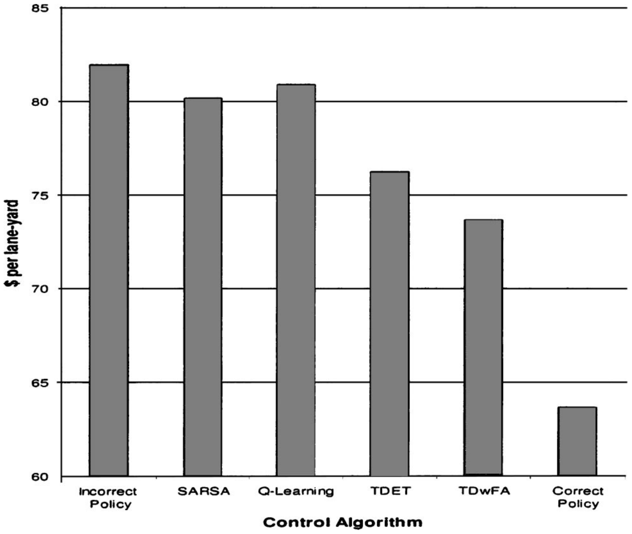

Fig. 2. Average costs for fast deterioration (Case 1)

Results

The average total discounted costs (over 25 years) for each of the experiments are shown in Figs. 2 and 3. The main observation is that for most cases the TD methods result in moderate cost savings over implementing the incorrect policy. The TD method with function approximation performed substantially better than the other control methods in both cases.

Tables 6 and 7 present the average best actions at the end of the horizon for each state to illustrate the convergence of TD methods to the optimal policy. Notice that the convergence to the optimal actions is slow. This is due to the fact that only 25 observations/samples (one per year) are used to update the state-action value function that has 37 nontrivial elements in Case 1

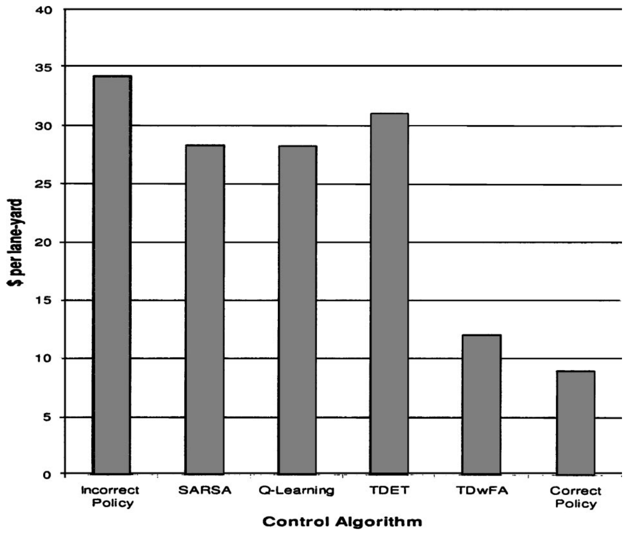

Fig. 3. Average costs for slow deterioration (Case 2)

Table 6. Average Best Action after 25 Periods (Case 1)

	Case 1: Fast deterioration					
Condition	Initial action	SARSA	Q-learning	TDET	TDwFA	Optimal action
1	7	7.0	7.0	7.0	7.00	7
2	6	6.0	6.0	6.0	5.2	7
3	5	5.0	5.0	4.9	5.3	7
4	4	4.2	4.2	4.2	6.0	6
5	4	4.0	4.0	4.0	5.0	5
6	3	3.4	3.6	3.3	4.7	4
7	2	3.2	3.1	3.1	3.6	4
8	2	2.3	2.3	3.2	2.5	3

and 40 elements in Case 2. This shortcoming is probably less important when observations come from the management of a network comprised of more than one section, because an agency generates observations from each of the sections. In addition, the use of sensors and other nondestructive evaluation techniques for condition assessment are increasing the opportunities and the cost-effectiveness of probing/sampling infrastructure facilities.

Summary and Conclusions

In this paper, the writer introduces temporal-difference learning methods, a class of reinforcement learning methods, as an approach to address maintenance and repair decision making for infrastructure facilities without a deterioration model. In temporal-difference learning, policies are evaluated directly by predicting the effects of actions on costs. These methods use the sequence of costs that follows the application of an action to update estimates of value functions or state-action value functions. This differs from the existing approach to decision making where policy evaluation involves modeling the effect of actions on condition and predicting future costs by assuming that there is a correspondence between condition and costs.

A case study in pavement management is presented where the implementation of temporal-difference learning methods is effective in fine-tuning incorrect policies, which result in savings over a 25-year horizon. As a whole, the results appear interesting when we consider that the implementation is based on samples that come from one facility (as opposed to a network of facilities) and that no substantial effort was spent on the choice of the parameters α, ϵ, and λ. The temporal-difference method with function approximation performed better than the other methods and probably warrants further study.

Table 7. Average Best Action after 25 Periods (Case 2)

	Case 2: Slow deterioration					
Condition	Initial action	SARSA	Q-learning	TDET	TDwFA	Optimal action
1	7	7.0	7.0	7.0	7.0	7
2	7	7.0	7.0	7.0	5.0	6
3	7	7.0	7.0	7.0	6.0	5
4	6	6.0	6.0	6.0	6.0	4
5	5	5.0	5.0	5.0	5.0	4
6	4	4.0	4.0	4.0	4.0	3
7	4	4.0	4.0	4.0	3.0	2
8	3	2.6	2.6	2.8	1.7	2

This research presents an approach to maintenance and repair decision making that is radically different. It provides an alternative approach that could be used to assess the costs associated with generating reliable data for the choice and specification of a deterioration model. The methods presented only assume that the infrastructure facility is managed under a periodic review policy. This makes the methodology attractive because strong assumptions about deterioration are not necessary. For example, the existing approach to maintenance and repair decision making usually assumes that deterioration is stationary and Markovian. This is in spite of empirical evidence to the contrary.

Acknowledgments

The writer acknowledges the many comments and suggestions provided by Samer Madanat and Stuart Dreyfus at the University of California, Berkeley.

Appendix I. Policy Evaluation over an Infinite Planning Horizon

The return at t under a given policy cannot be determined with certainty until the end of the planning horizon because it depends on the realization of $\{A_t, X_{t+1}, A_{t+1}, X_{t+2}, \ldots, X_{T+1}\}$. Policy evaluation involves predicting the return under a given policy. In the context of maintenance and repair decision making, the expected discounted sum of costs under a policy, i.e., the value function for a policy, is used as the cost predictor. Therefore, policy evaluation corresponds to finding the value function for a policy. From the definitions presented and the assumptions that deterioration is Markovian and stationary, the writer shows that

$$\begin{aligned}
V_t^P(X_t) &= E_{\{A_t, X_{t+1}, A_{t+1}, X_{t+2}, A_{t+2}, \ldots, X_T, A_T, X_{T+1}|P\}}[R_t(X_t, A_t)|X_t] \\
&= E_{\{A_t, X_{t+1}, A_{t+1}, X_{t+2}, A_{t+2}, \ldots, X_T, A_T, X_{T+1}|P\}}[g(X_t, A_t) \\
&\quad + \delta R_{t+1}(X_{t+1}, A_{t+1})|X_t] = E_{\{A_t, X_{t+1}|P\}}[g(X_t, A_t) \\
&\quad + \delta V_{t+1}^P(X_{t+1})|X_t] = \sum_{a \in \mathsf{A}} P_t(X_t, a) g(X_t, a) \\
&\quad + \delta \sum_{a \in \mathsf{A}} P_t(X_t, a) \sum_{j \in \mathsf{S}} \Pi_{X_t, j}(a) V_{t+1}^P(j)
\end{aligned} \tag{13}$$

The equations that are obtained from evaluating Eq. (13) for each pair (X_t, t) in the set $\mathsf{S} \times \{1,2, \ldots, T\}$ are referred to as Bellman's equations for policy P. Policy evaluation consists of finding a solution to this system of equations.

By considering the special case of evaluating a stationary policy over an infinite planning horizon, the set of Bellman's equations can be rewritten as follows:

$$\begin{aligned}
V_t^P(X_t) &= \sum_{a \in \mathsf{A}} P(X_t, a) g(X_t, a) + \delta \sum_{a \in \mathsf{A}} P(X_t, a) \sum_{j \in \mathsf{S}} \Pi_{X_t, j}(a) \\
&\quad + V_{t+1}^P(j), \\
&\qquad \forall X_t \in \mathsf{S}, \quad t \in \{1,2, \ldots, T\}
\end{aligned} \tag{14}$$

It is assumed that the costs are bounded, i.e., $\exists M$: $|g(x,a)| \leqslant M$, $\forall x \in \mathsf{S}$, $a \in \mathsf{A}$, and $|s(x)| \leqslant M$, $\forall x \in \mathsf{S}$. It can be shown that each of the limits, $\lim_{T \to \infty} V_t^P(x)$, $\forall x \in \mathsf{S}$, $t \in \{1,2, \ldots, T\}$, exists and is finite. Furthermore, it can be shown that

$\lim_{T\to\infty} V_t^P(x) = \lim_{T\to\infty} V_t^P(x),\ \forall x \in \mathsf{S},\ t,\ t' \in \{1,2,\ldots,T\}$. The intuition for this is that every time that a state of the facility is observed, an agency finds itself in the same situation because the condition of the facility is the same and there are still infinite periods left until the end of the planning horizon. The proofs of these results appear in the work by Ross (1992). The writer lets $V^P(x) \equiv \lim_{T\to\infty} V_t^P(x),\ \forall x \in \mathsf{S},\ t \in \{1,2,\ldots,T\}$. In this special case, policy evaluation involves finding the solution of the following system of $|\mathsf{S}|$ equations and unknowns

$$V^P(x) = \sum_{a\in\mathsf{A}} P(x,a)g(x,a) + \delta \sum_{a\in\mathsf{A}} P(x,a) \sum_{j\in\mathsf{S}} \Pi_{x,j}(a) V^P(j),\quad \forall x \in \mathsf{S} \qquad (15)$$

Under the assumptions presented, the system is known to have a unique solution. The proof appears in the work by Bertsekas (1995). In the context of policy evaluation, the system of Bellman's equations is usually solved iteratively. The methods are based on the property that the value functions evaluated for each element in S are, by definition, fixed points of Bellman's equations. These equations constitute a contraction mapping (Bertsekas 1995; Ross 1992). Therefore, any sequence of value function estimates that is generated by iteratively evaluating Bellman's equations (for arbitrary initial estimates) converges to the value functions. That is, for arbitrary $v_0^P(x),\ \forall x \in \mathsf{S},\ \lim_{k\to\infty} v_k^P(x) = V^P(x),\ \forall x \in \mathsf{S}$, where

$$v_{k+1}^P(x) \leftarrow \sum_{a\in\mathsf{A}} P(x,a)g(x,a) + \delta \sum_{a\in\mathsf{A}} P(x,a) \times \sum_{j\in\mathsf{S}} \Pi_{x,j}(a) v_K^P(j),\quad \forall x \in \mathsf{S},\quad k = 0,1,2,\ldots \qquad (16)$$

A particularly interesting choice of initial estimates for the value functions would be to set $v_0^P(x) = -s(x),\ \forall x \in \mathsf{S}$. In this case, $v_k^P(x) = V_k^P(x),\ \forall x \in \mathsf{S},\ k \in \{1,2,\ldots,T\}$.

The procedure described by the last set of equations can be implemented iteratively to obtain estimates of the value functions. This algorithm is known as the fixed-point iteration algorithm. In dynamic programming, the algorithm is called the policy evaluation algorithm. This algorithm can be adapted to obtain estimates of state-action value functions. The process of generating a sequence of estimates where each estimate is a function of prior estimates is known as bootstrapping.

Appendix II. Policy Selection over an Infinite Planning Horizon

In the case of an infinite planning horizon, it can be shown that there exists an optimal policy that is stationary and deterministic. A proof appears in the work by Ross (1992). The necessary and sufficient conditions for a stationary and deterministic policy P^* to be optimal are given by the following version of Bellman's optimality principle:

$$V^{P^*}(x) = \min_{a\in\mathsf{A}} \{Q^{P^*}(x,a)\},\quad \forall x \in \mathsf{S} \qquad (17)$$

The process of constructing an optimal policy can be performed iteratively by improving an arbitrary initial policy until the set of Eq. (17) is satisfied. An example of an algorithm that can be used to perform policy selection is presented.

```
Generalized policy iteration algorithm
Let $p$ be an arbitrary initial policy
Policy evaluation
Find $Q^p(x,a),\ \forall x \in \mathsf{S},\ a \in \mathsf{A}$
policy_stable$\leftarrow 1$
Policy iteration
For each $x \in \mathsf{S}$
    $\{b \leftarrow a$, for $p(x,a) = 1\}$
    If $b \neq a^*(x)$, then $\{p(x,a) \leftarrow 1$, for $a = a^*(x)$, $p(x,a) \leftarrow 0$,
    otherwise$\}$, policy_stable$\leftarrow 0$
If policy_stable$= 1$ then stop or else go to policy evaluation
```

where for a given policy P, $a^*(x) \equiv \operatorname{argmin}_{a\in\mathsf{A}} \{Q^P(x,a)\},\ \forall x \in \mathsf{S}$.

Under the assumptions presented earlier, it can be shown that the generalized policy iteration algorithm converges to an optimal policy that is stationary and deterministic in, at most, $|\mathsf{S}| \cdot |\mathsf{A}|$ iterations. Furthermore, it can be shown that the value functions under successive policies decrease. These statements follow from a result that is known in dynamic programming as the policy improvement theorem. A proof is presented by Bertsekas (1995).

References

Bellman, R. E. (1955). "Equipment replacement policy." *J. Soc. Ind. Appl. Math.*, 8(3), 133–146.

Bertsekas, D. (1995). *Dynamic programming and optimal control*, Athena Scientific, Belimont, Mass.

Carnahan, J., Davis, W., Shahin, M., Keane, P., and Wu, M. (1987). "Optimal maintenance decisions for pavement management." *J. Transp. Eng.*, 113(5), 554–572.

Dreyfus, S. (1960). "A generalized equipment replacement study." *J. Soc. Ind. Appl. Math.*, 8(3), 425–435.

Durango, P. (2002). "Adaptive optimization models for infrastructure management." PhD thesis, Univ. of California, Berkeley, Berkeley, Calif.

Durango, P., and Madanat, S. (2002). "Optimal maintenance and repair policies in infrastructure management under uncertain facility deterioration rates: An adaptive control approach." *Transp. Res., Part A: Policy Pract.*, 36, 763–778.

Fernandez, J. (1979). "Optimal dynamic investment policies for public facilities: The transportation case." PhD thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Gendreau, M., and Soriano, P. (1998). "Airport pavement management systems: An appraisal of existing methodologies." *Transp. Res., Part A: Policy Pract.*, 32(3), 197–214.

Golabi, K., Kulkarni, R., and Way, G. (1982). "A statewide pavement management system." *Interfaces*, 12(6), 5–21.

Madanat, S., and Ben-Akiva, M. (1994). "Optimal inspection and repair policies for infrastructure facilities." *Transp. Sci.*, 28(1), 55–61.

Ross, S. (1992). *Applied probability models with optimization applications*, Dover, New York.

Sutton, R., and Barto, A. (1998). *Reinforcement learning: An introduction*, MIT Press, Cambridge, Mass.

Terborgh, G. (1949). *Dynamic equipment replacement policy*, McGraw-Hill, New York.