

Deep reinforcement learning for multi-asset infrastructure management incorporating traffic operations adaptations and control

Prepared by:

K.G. Papakonstantinou, S. M. Stoffels, E. Miller-Hooks, M. Saifullah



U.S. DOT Region 3 University Transportation Center

r3utc.psu.edu

TABLE OF CONTENTS

TABLE OF CONTENTS.....	2
LIST OF FIGURES.....	3
LIST OF TABLES.....	4
CHAPTER 1 - INTRODUCTION.....	5
CHAPTER 2 - DEEP REINFORCEMENT LEARNING-DRIVEN MAINTENANCE PLANNING WITH TRAFFIC ADAPTATION.....	9
Overview.....	9
Markov Decision Process (MDP) and Partially observable MDP.....	9
Deep reinforcement learning solutions.....	12
DDMAC-CTDE formulation with constraints	14
Pavements	21
Bridges	25
Example network	26
Traffic assignment and user delay costs.....	27
Network constraints	33
Results.....	34
Conclusion	45
REFERENCES.....	46
APPENDIX A. MAINTENANCE ACTIVITIES AND TRANSITION PROBABILITIES.....	51
APPENDIX B. NETWORK STATE.....	55

LIST OF FIGURES

Figure 1: Constrained Deep Decentralized Multi-agent Actor Critic (DDMAC) with Centralized Training Decentralized Execution (CTDE) architecture, where $\parallel$ represents concatenation of $\hat{\mathbf{b}}$ and $\mathbf{y}$.	17
Figure 2: Modeled mean CCI for different levels of traffic.	22
Figure 3: Transition probabilities for Traffic level A, with (a) starting state = 6, (b) starting state= 5, smoothed over time with a 5-point window.	23
Figure 4: Transition probabilities in time, moving from state 9, for exogenous parameters (a) district 4, span- multi span (b) district 8, span- single span (c) district 10, span- single span.	26
Figure 5: Roadway network with various pavement and deck components.	27
Figure 6: Discretized network with 19 components consisting of 4 bridges and 15 pavement sections starting with initial conditions corresponding to (a) intact state and (b) non-intact state.	35
Figure 7: Total life cycle costs comparison of DDMAC-CTDE solutions with and w/o traffic adaptation actions for the considered network, starting from intact condition state with 5-yearly (a) budget of 600 million USD, (b) budget of 900 million USD, and (c) unrestricted budget.	36
Figure 8: Total life cycle costs comparison of DDMAC-CTDE solutions with and w/o traffic adaptation actions for the considered network, starting from non-intact condition state with 5-yearly (a) budget of 900 million USD, (b) budget of 1100 million USD, and (c) unrestricted budget.	36
Figure 9: Comparison of the total cost and its constituents with and w/o traffic adaptation cases for the network, starting from intact condition state with 5-yearly (a) budget of 600 million USD, (b) budget of 900 million USD, and (c) unrestricted budget.	37
Figure 10: Comparison of the total cost and its constituents with and w/o traffic adaptation cases for the network, starting from non-intact condition state with 5-yearly (a) budget of 900 million USD, (b) budget of 1100 million USD, and (c) unrestricted budget.	37
Figure 11: Life-cycle realization of the learned DDMAC-CTDE policy for the network, starting from intact state.	41
Figure 12: Life-cycle realization of the learned DDMAC-CTDE policy for the network, starting from nonintact state.	42
Figure 13: Network initial traffic condition (a) nominal traffic flow (q_0); (b) nominal flow to initial capacity ratio (q_0/cap_0)	43
Figure 14: A snapshot of the network's traffic condition for the intact starting state, illustrating (a) the suggested traffic actions as indicated by the adjusted capacity ratio, (b) the traffic flow, and (c) the traffic flow to capacity ratio at time $t = 14$ years with 600 million USD 5-yearly budget. The dotted line represents the component under maintenance.	43
Figure 15: A snapshot of the network's traffic condition for the non-intact starting state, illustrating (a) the suggested traffic actions as indicated by the adjusted capacity ratio, (b) the traffic flow, and (c) the traffic flow to capacity ratio at time $t = 11$ years with 900 million USD. The dotted line represents the component under maintenance.	44

LIST OF TABLES

Table 1. State discretization based on CCI values.....	22
Table 2. Maintenance action costs for asphalt pavements, reported in USD/m ²	24
Table 3. Bridge failure probability given bridge condition state.....	25
Table 4 Cost of improvement actions for bridges.....	26
Table 5: Network characteristics.	32
Table 6 Considered traffic demand for different origin and destination nodes.	33
Table 7 Comparison of objectives and imposed constraints for DDMAC-CTDE solutions with and without traffic adaptation actions for intact starting condition states and various budget scenarios.	39
Table 8 Comparison of objectives and imposed constraints for DDMAC-CTDE solutions with and without traffic adaptation actions for non-intact starting condition states and various budget scenarios.....	39
Table A-1 Maintenance activities for interstate and primary pavements under different categories (Virginia Department of Transportation 2016).....	51
Table A-2 . Minor repair transition probabilities for 6 CCI states.....	52
Table A-3. Major repair transition probabilities for 6 CCI states.....	52
Table A-4. Reconstruction transition probabilities for 6 CCI states.....	52
Table A-5. Observation probability $p(o_t st)$ given actual state st	53
Table A-6. Minor repair transition probabilities for deck states.....	53
Table A-7. Major repair transition probabilities for deck states.....	53
Table A-8. Reconstruction transition probabilities for deck states.....	54
Table A-9. Observation probability $p(o_t s_t)$ given state s_t for decks.	54
Table B-1 Network's non-intact starting condition state based on CCI and GCR for pavements and bridges.....	55

CHAPTER 1

INTRODUCTION

Preserving structural integrity through inspections and maintenance (I&M) is a critical and recurrent task that involves complex decisions strategically placed in both space and time to ensure the longevity and reliability of infrastructure assets. At the regional scale, the optimal management of entire networks amplifies the complexity of these decisions, primarily due to the vast number of interconnected assets, each with its inherent dependencies and correlations. In the context of transportation networks, the dependencies can manifest as system-level effects on traffic flow, as the individual contributions of various assets to traffic capacity collectively influence system performance, often in a nonlinear manner. Consequently, assessing the potential impacts of different combinations of I&M decisions, each linked to a specific asset, becomes an intricate process.

The optimization problem inherent in I&M presents numerous challenges, including: the exponential scaling of state/action spaces with the number of components; dependence on history related to exponentially growing decision-trees with the number of decision-steps; presence of state uncertainties induced by inherent environment stochasticity and uncertain observations; presence of constraints, due to resource scarcity and other infeasible/undesirable system responses. In addition, implementing maintenance necessitates system downtime, primarily in the form of reduced roadway capacity and occasional roadway or bridge closures. Although these maintenance activities are typically scheduled by asset class, it is essential to recognize that downtime scheduling for different asset types can impact network-wide performance. The time required for implementation of maintenance actions can extend over weeks, months, or even years, significantly affecting the transportation system performance over extended periods. Therefore, the coordination of maintenance actions across assets is paramount to minimize the overall impact on roadway users.

The optimal I&M problem has been approached by a large variety of methods in the literature, as for example in (Bocchini & Frangopol, 2011; Saydam & Frangopol, 2014; Yang & Frangopol, 2019; Chu & Chen, 2012; Mendoza, et al., 2021; Zhou & Wang, 2012), using genetic algorithms, threshold-based approaches, risk-informed techniques, renewal processes, and decision-trees, among other solution methods. Risk-based methods have in particular received much attention in the literature and practice, as in (Frangopol, et al., 1997; Faber & Stewart, 2003; Straub & Faber, 2005; Luque & Straub, 2019). Despite their important merits, many of these approaches can suffer from suboptimality and scalability issues, and despite the dynamic nature of the deterioration problem, and the uncertainty of future infrastructure condition states, several also rely on formulations that do not account for these fundamental problem properties. The computational challenges are also further exacerbated in high-dimensional settings and long decision horizons, which are often referred to as the curse of dimensionality and the curse of history in the literature (Bellman, 1957; Pineau, et al., 2003). Partially Observable Markov Decision Processes (POMDPs) offer a promising solution to address challenges in infrastructure maintenance planning. POMDPs, benefiting from dynamic programming principles, can overcome the curse of history and can

accommodate noisy real-time data. Several studies, such as in (Madanat, 1993; Madanat & Ben-Akiva, 1994; Papakonstantinou & Shinozuka, 2014; Papakonstantinou & Shinozuka, 2014b; Papakonstantinou, et al., 2016), have demonstrated their effectiveness in this context. Recent research has also explored robust solutions under model uncertainty (Arcieri, et al., 2023), as well as factored POMDPs (Kıvanç, et al., 2022), Predictive Markov Decision Processes (Guo & Liang, 2022), and combinations with dynamic Bayesian networks (Morato, et al., 2022), allowing some scalability benefits, often with global optimality guarantees. (Pineau, et al., 2003; Papakonstantinou & Shinozuka, 2014; Spaan & Vlassis, 2005; Papakonstantinou, et al., 2016; Papakonstantinou, et al., 2018).

In general, traditional MDP/POMDP approaches may still face scalability issues on larger systems with multiple components. The application of neural network parameterizations, particularly through Deep Reinforcement Learning (DRL), has thus emerged as a powerful approach to address state scalability challenges in various domains. Single-agent DRL approaches, such as the deep Q-network (DQN) algorithm in (Mnih, et al., 2015) and the proximal policy optimization (PPO) in (Schulman, et al., 2017), have led to effective solutions for complex decision-making problems, including bridge maintenance planning, rail renewal, manufacturing network optimization, and others, as for example in (Lei, et al., 2022; Mohammadi & He, 2022; Ye, et al., 2023; Arcieri, et al., 2024). As systems grow in size and complexity, single-agent DRL algorithms still face limitations, mainly due to the emerging combinatorial action spaces. To further tackle this issue, Multi-Agent DRL (MADRL) techniques, like the Deep Centralized Multi-agent Actor-Critic (DCMAC) algorithm, have been introduced (Andriotis & Papakonstantinou, 2019; Andriotis & Papakonstantinou, 2019a) and applied in transportation system problems (Zhou, et al., 2022). A DCMAC variant is presented in (Yang, 2022), with the inclusion of entropy in the actor loss, used for life-cycle management of bridge girder systems. A later variant of DCMAC, the Deep Decentralized Multi-agent Actor-Critic (DDMAC), introduced decentralized actors (Andriotis & Papakonstantinou, 2021; Morato, et al., 2023), and based on it a version with Centralized Training and Decentralized Execution (CTDE) structure, termed DDMAC-CTDE, is developed in (Saifullah, et al., 2024; Saifullah, et al., 2022), in order to be able to scale to much larger systems. DCMAC-CTDE is a deep, off-policy actor-critic algorithm that utilizes experience replay and this architecture is based on the Decentralized-POMDP (Dec-POMDP) concept (Oliehoek, et al., 2008). Alternative MADRL algorithms, including value decomposition (Sunehag, et al., 2017) and parallelized multi-agent deep Q-networks (Lee & Song, 2023), have also been suggested for deriving maintenance policies in various contexts, such as in manufacturing (Su, et al., 2021), predictive maintenance of parallel machines (Rodriguez, et al., 2022) and many more.

To address traffic-related implications related to I&M network actions, concepts like user equilibrium (UE) (Wardrop, 1952), can be employed to predict the affected traffic flows. Incorporating such concepts results in models inherently bound by nonlinearity and interdependencies among their components, reflecting the level of system correlation among the network components. Furthermore, assets within the network are also parametrically correlated, owing to factors such as material composition, structural design, local geological conditions, and varying traffic levels influenced by events and time of day. From a modeling perspective, these correlations manifest as statistical correlations in the constitutive properties and responses of different assets. Both system and statistical correlations are pivotal for optimal decision-making, implicitly introducing correlation as a fundamental consideration in the determination of optimal actions. For instance, leveraging statistical correlations allows for the utilization of information

and data obtained from specific parts of a network to update our knowledge about the condition or properties of other potentially remote network segments. Simultaneously, exploiting system correlations can lead to maintenance or intervention actions that, while temporarily shutting down certain system components, may affect traffic patterns in ways that create opportunistic intervention windows at other assets with minimal additional disruption.

In prior research, notable contributions by (Ouyang, 2007) and (Durango-Cohen & Sarutipand, 2009) incorporated traffic network effects by linking deterioration and maintenance actions with roadway conditions, such as pavement roughness, which impact driver experience and route selection. In (Ouyang, 2007) these traffic effects were integrated through a UE model embedded within their modeling framework, focusing on an optimal resurfacing plan for the highway network. However, downtime effects have not been considered in their work, and the proposed policy iteration method was limited to small-scale problems. In (Durango-Cohen & Sarutipand, 2009) the roadway segments demand was assessed as a function of their condition and capacity loss was utilized to account for the impacts of deterioration and implemented improvements on roadway conditions. Both of these studies also made the assumption that the deterioration processes were deterministic.

In (Medury & Madanat, 2013) the downtime effects of pavement improvement actions during their implementation have been considered for roadway asset management, accounting for capacity differences due to various maintenance actions. A TD-learning (SARSA) solution has been utilized in that work, based on a simplified traffic model. In the work by (Ng, et al., 2009), capacity loss during maintenance activities has also been incorporated in their maintenance planning framework, further enhancing the framework presented in (Ouyang, 2007). The authors accounted for downtime associated with related activities over increments of years and utilized a heuristic genetic algorithm for problem solution. A bi-level optimization approach with UE at the lower-level capturing deterioration and downtime effects on traffic is also presented in (Chu & Chen, 2012), with a threshold-based maintenance plan used at the upper level. The model presumes a deterministic, continuous deterioration process and a tabu search algorithm is applied to choose from various threshold options. More recently, pavement maintenance planning with integrated traffic related effects emerging due to the implementation of maintenance actions can also be seen in (Aad, et al., 2022; Prajapati, et al., 2024), with both works assuming deterministic models.

The work by (Zhou, et al., 2022) presents a new, notable approach for optimizing transportation infrastructure maintenance planning while also considering traffic impacts into the decision-making process. It expands and provides advancements over all other available works in the literature by considering uncertain deterioration mechanisms, multi-asset classes, and traffic capacity loss during improvement action executions, while also supporting a UE traffic representation and a bilevel program to incorporate the traffic effects of maintenance actions. The entire problem is formed in a MDP framework and solved by the powerful DCMAC DRL algorithm. By realistically capturing traffic, the impacts of maintenance activities are accurately accounted for in action scheduling, and results demonstrate the effectiveness of the formulation and DRL approach, showing significant reductions in both traffic congestion and total costs compared to other commonly used solution methods.

In this project, we are relying on the work in (Zhou, et al., 2022) and further significantly extend it by addressing additional complexities, such as increased system dimensionality and utilization of the DDMAC-CTDE DRL solver, partial observability in a POMDP framework, to account for the inherent uncertainty of inspection actions (Papakonstantinou, et al., 2019; Andriotis, et al., 2021), and consideration of budget, performance, and risk constraints. Most

notably, we are also now incorporating the impact of traffic on the deterioration mechanisms, and originally and uniquely in the literature, to the best of our knowledge, also formulate the problem as a joint I&M planning and related traffic adaptation framework.

While studies have considered the impact of I&M actions on traffic operations and performance, as presented until now, the potential to optimally coordinate traffic management or control strategies with I&M scheduling has little or no attention in the literature or current practice so far. We believe to be the first to address this problem in this joint context with DRL. Quantifying and controlling the repercussions of traffic-related system correlations on I&M policies, and incorporating this quantification in the joint decision-making process for I&M operations and traffic management/control interventions, will allow decision-makers to answer important questions, such as:

- (i) How should decision-makers select I&M actions on assets based on their individual impacts to traffic flows and patterns and potential traffic interventions?
- (ii) Can decision-makers efficiently and effectively adapt and intervene to control traffic patterns in order to delay deterioration of certain assets until budget availability allows their updates?
- (iii) Can traffic controls be effectively devised simultaneously to I&M activity planning to mitigate the negative effects of the activities on traffic performance, both locally and network-wide?

Optimal infrastructure management in this context also hence needs to address an important additional layer of decision complexity, arising from the joint control of traffic and I&M planning of multiple overlapping activities, together with accommodating longer-term I&M plans and shorter-term traffic operational control scales. Answering the above questions is also additionally complicated by the multiplicity of asset types existing in transportation infrastructure (e.g., pavements, bridge superstructures, substructures, etc.), each with unique deterioration and uncertainty sources.

This projects thus brings a unique control perspective to optimal I&M planning. Complementing the work in (Zhou, et al., 2022) in several directions, as explained before, additional control actions are now also considered to divert traffic from relevant locations and directly modify traffic patterns, aiming at increased expected remaining life of assets. The principal concept is thus to study how active control of traffic can directly reduce or increase the deterioration of the involved transportation assets. This control option combined with the I&M infrastructure planning can significantly improve network-level life-cycle performance, and allow for more efficient utilization of resources, by discouraging, for instance, traffic patterns that include assets that are already deteriorated and not expected to be replaced or repaired for a prolonged period of time due to budget limitations or other restrictions. In addition, the suggested joint traffic control and I&M optimization can also consider traffic adaptation plans mitigating the temporary disruptions caused during I&M actions implementation.

In the following, detailed formulation derivations are provided, and we study and demonstrate the developed framework based on a detailed network example, looking into potential benefits from the combined maintenance actions and traffic adaptation strategies, such as minimizing safety risks and maintenance costs, maximizing the overall network serviceability and availability, and minimizing the impacts of maintenance activities on traffic flows.

CHAPTER 2

DEEP REINFORCEMENT LEARNING-DRIVEN MAINTENANCE PLANNING WITH TRAFFIC ADAPTATION

OVERVIEW

This chapter presents an integrated framework for large-scale infrastructure inspections and maintenance planning with adaptive traffic control. Our approach combines Partially Observable Markov Decision Processes (POMDPs), User Equilibrium (UE) modeling, and the Deep Reinforcement Learning (DRL) algorithm, Deep Decentralized Multi-agent Actor-Critic (DDMAC) with Centralized Training and Decentralized Execution (CTDE) proposed in (Saifullah, et al., 2024) to optimize maintenance decisions while considering traffic flow adjustments. In this chapter, a complete modeling environment of both pavements and bridges is utilized based on their corresponding damage state indices, which characterize their condition states. The related stationary and nonstationary Markovian transition probabilities are obtained. Four broader categories of maintenance actions are also considered, namely *Do Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*, for both pavements and bridges. Observation actions are also defined as no-inspection and inspection for both asset types. We validate the developed framework based on a network example, which includes several bridge and pavement components with nonstationary degradation, agency-imposed constraints, and traffic delay and bridge failure risk considerations. The cost for the maintenance, traffic delays, and risk are considered as in (Saifullah, et al., 2024). The framework is shown to be effective in improving network performance and reducing maintenance costs. The results were analyzed for the cases with and without traffic adaptation actions.

MARKOV DECISION PROCESS (MDP) AND PARTIALLY OBSERVABLE MDPS

MDPs are controlled stochastic processes which provide solutions for optimal sequential decision-making in stochastic environments with uncertain action outcomes and exact observations. MDPs are defined by a tuple $\langle S, A, \mathbf{P}, \mathbf{C}, \gamma \rangle$ for an environment, E , where S is a set of states, A is a set of actions, $\mathbf{P}$ encapsulate the environment transition model, $\mathbf{C}$ denotes the reward model, and $\gamma \in (0, 1]$ is a scalar discount factor. At each decision step t , the agent observes the current state $s_t \in S$, takes an action $a_t \in A$, receives a reward as a result of this state and action, $c(s_t, a_t)$, and goes to the next state $s_{t+1} \in S$, according to the underlying Markovian transition probabilities

$p(s_{t+1} | s_t, a_t)$. The agent chooses actions according to its policy π , which can be deterministic, $\pi(s_t): S \rightarrow A$ or stochastic, a distribution over action space conditioned on s_t , i.e., $\pi(a_t | s_t): S \rightarrow P(A)$. For stochastic policy, π is a probability mass and density function for discrete and continuous action spaces, respectively, for deterministic policies, π , is just a scalar value function. The corresponding reward function $C_t^\pi = \sum_{i=t}^T \gamma^{i-t} c(s_i, a_i)$ is the total reward collected from time step t to the end of the planning horizon T under the policy π .

To formalize the problem of estimating the optimal policy π^* , we first define action-value function Q^π as the expected return over all potential future states and actions following policy π , conditioned on the current state-action pair, s_t, a_t as:

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{s_{t+1} \sim E, a_{t+1} \sim \pi} [C_t^\pi | s_t, a_t] \\ &= c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E, a_{t+1} \sim \pi} [Q^\pi(s_{t+1}, a_{t+1})] \end{aligned} \quad (1)$$

Similarly, a value function of a policy π is the expectation of the return from state s_t . It also can be defined as the expectation of the action-value function over all possible actions at the current step:

$$V^\pi(s_t) = \mathbb{E}_{a_t \sim \pi} [Q^\pi(s_t, a_t)] \quad (2)$$

The goal of the agent is to select the sequence of actions which maximizes the value function. An important property of MDPs is that there exists at least one deterministic optimal policy that maximizes value $V^\pi(s_t)$ under standard conditions, even though the transitions are stochastic. For a deterministic optimal policy, with a given model of transitions $p(s_{t+1} | s_t, a_t)$ and with the aid of Eqs. (1) and (2), the optimal action-value and value functions, $Q^{\pi^*}(s_t, a_t)$ and $V^{\pi^*}(s_{t+1})$, respectively, follow the Bellman equation (Bellman, 1957), and can be obtained as:

$$Q^{\pi^*}(s_t, a_t) = c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E} \left[\max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) \right] \quad (3)$$

$$\begin{aligned} V^{\pi^*}(s_t) &= \max_{a_t \in A} \{Q(s_t, a_t)\} \\ &= \max_{a_t \in A} \left\{ c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E} \left[\max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) \right] \right\} \\ &= \max_{a_t \in A} \left\{ c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E} [V(s_{t+1})] \right\} \\ &= \max_{a_t \in A} \left\{ c(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1} | s_t, a_t) V(s_{t+1}) \right\} \end{aligned} \quad (4)$$

The objective function of MDP can be described by Eq. (3) or (4), which can be achieved using value iteration, policy iteration, or linear programming formulations as presented in (Papakonstantinou & Shinozuka, 2014a).

A natural extension to MDP is partially observable MDP (POMDP), which is a generalization to the case when the environment is not perfectly observable. This is applicable in a wide variety of engineering applications where the available information is incapable of revealing the true state of the system. Further, POMDP can encode the entire history of states, actions, and observation in its belief system, which is a very efficient and sufficient feature regarding adapting policies given only the current uncertain information, without essentially also forgetting the past.

Analogous to MDPs, in the POMDP framework, the decision-maker/agent starts at a state, s_t , at a time step, t , takes an action a_t , receives a reward, c_t , transitions to the next state, s_{t+1} , and receives an observation, $o_{t+1} \in \Omega$, based on the observation probability model, $p(o_{t+1}|s_{t+1}, a_t)$ for a given state and action. More formally, a POMDP is a 7-tuple $L = \langle S, A, \mathbf{P}, \Omega, \mathbf{O}, \mathbf{C}, \gamma \rangle$ where S , A , and Ω are sets of states, actions, and possible observations, respectively, and $\mathbf{P}$, $\mathbf{O}$, and $\mathbf{C}$ are the transition, observation, and reward models, respectively.

As a result of partial observability, the agent can only form a belief $\mathbf{b}_t$ about its state, where $\mathbf{b}_t$ is a probability distribution over the set S of all possible discrete states. A Bayesian update can be used to calculate the new belief $\mathbf{b}_{t+1}$ for a given action and observation, (Papakonstantinou & Shinozuka, 2014a):

$$\mathbf{b}_{t+1} = \mathbf{b}(s_{t+1}) = p(s_{t+1} | o_{t+1}, a_t, \mathbf{b}_t) = \frac{p(o_{t+1} | s_{t+1}, a_t)}{p(o_{t+1} | \mathbf{b}_t, a_t)} \sum_{s_t \in S} p(s_{t+1} | s_t, a_t) b(s_t) \quad (5)$$

where, $p(o_{t+1} | \mathbf{b}_t, a_t)$ is the standard normalizing constant. Following Eq. (5), beliefs can be updated as new actions are performed and new observations are collected, essentially encoding the information of the entire history of actions and observations up to the current time step t . As such, a new belief $\mathbf{b}_t$ is a sufficient statistic of the history of actions and observations up to t . Thus, beliefs can be seen as alternative states of the environment, and POMDPs can be accordingly regarded as belief-MDPs. Further, the optimal value function V^{π^*} , following the optimal policy π^* takes an expression similar to Eq. (4) as:

$$V^{\pi^*}(\mathbf{b}_t) = \max_{a_t \in A} \left\{ \sum_{s_t \in S} b(s_t) c(s_t, a_t) + \gamma \sum_{o_{t+1} \in \Omega} p(o_{t+1} | \mathbf{b}_t, a_t) V^{\pi^*}(\mathbf{b}_{t+1}) \right\} \quad (6)$$

Despite the similarity with MDPs, POMDPs are more difficult to solve. The new belief-state space is continuous rather than discrete, producing a $(|S| - 1)$ – dimensional simplex. On the other hand, the optimum value function has been proved to be piece-wise linear and convex, and so can be represented by a finite number of affine hyperplanes. Thus, the decision problem is reduced to choosing a finite number of vectors, commonly known as the α -vectors. To solve this problem, point-based algorithms seek to determine a set of α -vectors iteratively which adequately supports the belief space. However, such point-based solvers can become impractical for large POMDP

problems, particularly for infrastructure systems with many components. The curse of dimensionality associated with POMDP can be alleviated using a state-of-the-art deep reinforcement learning framework, a subset of AI and machine learning methods discussed below.

DEEP REINFORCEMENT LEARNING SOLUTIONS

Reinforcement learning (RL) is a computational framework designed for goal-directed learning and decision making. It is particularly effective for solving Markov Decision Process (MDP) and Partially Observable Markov Decision Process (POMDP) problems, often structured around them (Sutton & Barto, 2018). RL can address the challenge of high-dimensional state spaces through various methods: model-free approaches, which do not rely on prior offline environment information; model-based methods that learn the underlying transition model; or hybrid combinations of these techniques.

Many of RL methods use TD error to update the value/action value function, shown as:

$$\begin{aligned} Q(s_t, a_t) &\leftarrow Q(s_t, a_t) + \beta(Q(s_t, a_t) - y_t) \\ y_t &= r(s_t, a_t) + \gamma \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) \end{aligned} \quad (7)$$

where, β is a scalar learning rate and can be chosen appropriately. The TD-error given in Equation (7) is utilized for training an RL agent taking various relevant forms. The policy gradient method (Sutton, et al., 2000) is accordingly used for learning policies in policy-based methods. Given the recent unprecedented advancement in DNN, RL algorithms, commonly known as Deep RL (DRL), have shown capabilities of discovering meaningful parameterizations in immense state spaces, outperforming human experts. Recently published works (Andriotis & Papakonstantinou, 2019; Andriotis & Papakonstantinou, 2021) based on multi-agent DRL for optimal decision making in the context of managing large infrastructure systems has shown enormous potential.

In (Andriotis & Papakonstantinou, 2019), a Deep Centralized Multi-agent Actor-Critic (DCMAC) architecture has been proposed having two centralized networks, actor (output policy) and critic (output value function based on the selected policy). It has been developed within the premises of off-policy actor-critic, DRL with replay buffer, where replay buffer is a memory comprising past experiences that the agent collected by probing its environment, stored as tuples. These experiences are used for batch training of the neural network via uniform random sampling at each step. It uses policy gradient method (Sutton, et al., 2000) to update the parameters of policy/actor network, the gradient of policy network $\mathbf{g}_{\theta^\pi}$ can be estimated using samples generated by a behavior policy $\mu \neq \pi$ as:

$$\mathbf{g}_{\theta^\pi} = \mathbb{E}_{\mathbf{s}_t \sim \mathbf{p}, \mathbf{a}_t \sim \mu} \left[w_t \left(\sum_{i=1}^n \nabla_{\theta^\pi} \log \pi_i(a_t^{(i)} | \mathbf{s}_t, \theta^\pi) \right) A^\pi(\mathbf{s}_t, \mathbf{a}_t) \right] \quad (8)$$

where, $\mathbf{s}_t = \{s_t^{(i)}\}_{i=1}^m$ state vector for m -component system, $\mathbf{a}_t = \{a_t^{(i)}\}_{i=1}^n$ is an action vector for n -agents (no. of agents and no. of components can be same or different), θ^π is a policy network weight parameter, $w_t = \min\{\delta, \pi(a_t | s_t) / \mu(a_t | s_t)\}$ is a truncated importance sampling weight and δ is a cut off value; n denotes number of agents, μ is a n -dimensional vector of agents' behavior policies, and $\mathbf{p}$ is the m -dimensional limiting state distribution under these policies, and $A^\pi(\mathbf{s}_t, \mathbf{a}_t)$

is an advantage function defined as $A^\pi(\mathbf{s}_t, \mathbf{a}_t) = Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - V^\pi(\mathbf{s}_t)$. It can be seen as a zero mean measure i.e., $\mathbb{E}_{\mathbf{a}_t \sim \pi} [A^\pi(\mathbf{s}_t, \mathbf{a}_t)] = 0$, estimating how advantageous the given action is. The advantage function can further be approximated to reduce the need for action-value function estimations, essentially following the TD-error:

$$A^\pi(\mathbf{s}_t, \mathbf{a}_t | \boldsymbol{\theta}_V) \simeq c(\mathbf{s}_t, \mathbf{a}_t) + \gamma V^\pi(\mathbf{s}_t | \boldsymbol{\theta}_V) - V^\pi(\mathbf{s}_t | \boldsymbol{\theta}_V) \quad (9)$$

where, $\boldsymbol{\theta}_V$ is a weight parameter of the value/critic network. The mean squared error is considered as a loss function for the critic network. The gradient and loss function are given as follows:

$$\begin{aligned} L_V(\boldsymbol{\theta}_V) &= \mathbb{E}_{\mathbf{s}_t \sim \rho, \mathbf{a}_t \sim \mu} \left[w_t \left(c(\mathbf{s}_t, \mathbf{a}_t) + \gamma V^\pi(\mathbf{s}_{t+1} | \boldsymbol{\theta}_V) - V^\pi(\mathbf{s}_t | \boldsymbol{\theta}_V) \right)^2 \right] \\ \mathbf{g}_{\boldsymbol{\theta}_V} &= \mathbb{E}_{\mathbf{s}_t \sim \rho, \mathbf{a}_t \sim \mu} \left[w_t \nabla_{\boldsymbol{\theta}_V} V^\pi(\mathbf{s}_t | \boldsymbol{\theta}_V) A^\pi(\mathbf{s}_t, \mathbf{a}_t | \boldsymbol{\theta}_V) \right] \end{aligned} \quad (10)$$

DCMAC network architecture operates on components beliefs b_t instead of states s_t because observations of structural and engineering systems are unlikely to disclose the system's true state; thus, we can easily replace $\mathbf{s}_t$ with $\hat{\mathbf{b}}_t = \{\mathbf{b}_t^{(i)}\}_{i=1}^m$ in Eqs. (6) - (10). DCMAC has shown important life-cycle cost improvements compared to traditional heuristically optimized policies, demonstrating its superior potential (Andriotis & Papakonstantinou, 2019). DCMAC uses policy gradient methods to update the policy parameters and works with full parameter sharing among actors, which may, in certain cases, encumber training efficiency.

To alleviate this effect, DDMAC was proposed in (Andriotis & Papakonstantinou, 2021), where a sparser parametrization of the actor-network without parameter sharing is utilized, i.e., each agent has its own unique policy network. By removing inter-agent interactions in the hidden layers, this architecture improves training efficiency for larger systems. In (Andriotis & Papakonstantinou, 2021), the DDMAC architecture is further integrated into an algorithmic scheme that supports planning under budget or other operational constraints. DDMAC is a substantial improvement over DCMAC, however, it still postulates accessibility of the entire system's state distribution to all the agents. As the input per actor is then the entire state space, this can result in training inefficiency for environments with a very large number of components/actors.

A new DDMAC version is proposed in (Saifullah, et al., 2024), that is equipped with a fully Centralized Training and Decentralized Execution (CTDE) structure, a prevalent paradigm in cooperative multi-agent reinforcement learning (RL) (Oliehoek, et al., 2008; Sunehag, et al., 2017; Foerster, et al., 2018; Rashid, et al., 2018; Gronauer & Diepold, 2022). This architecture is based on the Decentralized-POMDP (Dec-POMDP) concept (Oliehoek, et al., 2008), and fully decentralized policies are used herein, conditioned only over each of the corresponding component's belief state, making actors truly independent of each other. Actors can thus easily be parallelized for generating policies and their network parameters can be updated concurrently. The training utilizes a centralized critic, which receives the entire system's belief state to guide the actors, only in the training stage, and which is then discarded during the policy deployment. As a result, agents are given access to additional information during training, that otherwise would be hidden, and they retain this hidden knowledge during deployment. The trained policies can finally easily be deployed using locally available belief states for each agent. The CTDE framework is

powerful and efficient in cooperative settings, especially when multiple agents are present, with similar learning tasks (Gronauer & Diepold, 2022). A more detailed description of DDMAC-CTDE with constraints is provided in the following sections, largely adopted from (Saifullah, et al., 2024).

DDMAC-CTDE FORMULATION WITH CONSTRAINTS

DDMAC-CTDE adopts a fully Decentralized POMDP (Dec-POMDP) approach in fully cooperative multi-agent task (Oliehoek & Amato, 2016). A Dec-POMDP consists of an environment with multiple agents, in which the agents do not have complete observations of the environment's state and the actions taken by the other agents. More formally, Dec-POMDP is defined by a tuple $\langle \mathcal{N}_a, \mathcal{N}_c, \{S_i\}, \{A_i\}, \{\Omega_i\}, \mathbf{P}, \mathbf{O}, \mathbf{C}, \gamma \rangle$, where $\mathcal{N}_a$, is a set of agents from 1 to N_a , $\mathcal{N}_c$ is a set of components from 1 to N_c , S_k is a set of states for the k -th component and $S = S_1 \times \dots \times S_{N_c}$ is a global set of states of the system, A_i is a set of actions for agent i and $A = A_1 \times \dots \times A_{N_a}$ is the joint action space, Ω_i is a set of observations for the i -th agent and $\Omega = \Omega_1 \times \dots \times \Omega_{N_a}$ is the joint observation space, and $\mathbf{P}$, $\mathbf{O}$, and $\mathbf{C}$ are same as defined earlier in the MDP -POMDP section.

Without any loss of generality and for presentation simplicity, we henceforth assume that each agent controls exactly one corresponding system component, i.e., $N_a = N_c = N$. This assumption may easily be modified if an agent controls multiple components (Andriotis & Papakonstantinou, 2019). Then, the policy $\pi_i : \mathbf{b}^{(i)} \rightarrow P(A_i)$ is a policy followed by the i -th agent and it is a function of the local belief state vector $\mathbf{b}^{(i)}$, determined by the observation and action history of each component and calculated via Bayesian update, as described in Eq. (5). For all agents, the joint policy $\pi = \pi_1 \cdot \pi_2 \dots \pi_{N-1} \cdot \pi_N$ may be defined as $\pi : \hat{\mathbf{b}} \rightarrow P(A)$, where $\hat{\mathbf{b}} = [\mathbf{b}^{(i)}]_{i=1}^N$ is a belief matrix of the system state. The joint objective of the agents is to identify an optimal policy π^* that minimizes the discounted cumulative global cost C_0 over the planning horizon T . For concise notation, we are also using $(\cdot)'$ for $(\cdot)$ at $t+1$ time step.

Similar to DDMAC, a decentralized actor-critic architecture is selected within the premises of off-policy actor critic DRL, where each agent is assigned an actor network for policy output independent to other agents. During training, the centralized critic is using the system state distribution as input and the value function $V^\pi(\hat{\mathbf{b}})$ as the output. The expression in Eq. (6) gives the optimal value function $V^{\pi^*}(\hat{\mathbf{b}})$. Further, we can modify the Eqs. (8)-(14) as:

$$\mathbf{g}_{\theta_\pi^{(j)}} = \mathbb{E}_{b \sim p, a \sim \mu} \left[w \left(\sum_{i=1}^n \nabla_{\theta_\pi^{(j)}} \log \pi_i(a^{(i)} | \mathbf{b}^{(i)}, \theta_\pi^{(i)}) \right) A^\pi(\hat{\mathbf{b}}, \mathbf{a}) \right] \quad (11)$$

where, $\mathbf{a} = \{a^{(i)}\}_{i=1}^N$ are the actions for N -agents, $\theta_\pi^{(j)}$ is a weight parameter for the j^{th} actor network, w is truncated importance sampling factor defined earlier, $A^\pi(\hat{\mathbf{b}}, \mathbf{a})$ is an advantage

function defined also defined earlier which estimates how advantageous the given action is. We can rewrite the advantage function using belief as:

$$A^\pi(\hat{\mathbf{b}}, \mathbf{a} | \boldsymbol{\theta}_V) \simeq c(\hat{\mathbf{b}}, \mathbf{a}) + \gamma V^\pi(\hat{\mathbf{b}}' | \boldsymbol{\theta}_V) - V^\pi(\hat{\mathbf{b}} | \boldsymbol{\theta}_V) \quad (12)$$

where $\boldsymbol{\theta}_V$ are weight parameters of the value/critic network, and $\hat{\mathbf{b}}'$ is the system belief at the next time step. Similarly, the mean squared error L_V , and gradient for the critic network are given as:

$$\begin{aligned} L_V(\boldsymbol{\theta}_V) &= \mathbb{E}_{\hat{\mathbf{b}} \sim \boldsymbol{\rho}, \mathbf{a} \sim \boldsymbol{\mu}} \left[w \left(c(\hat{\mathbf{b}}, \mathbf{a}_t) + \gamma V^\pi(\hat{\mathbf{b}}' | \boldsymbol{\theta}_V) - V^\pi(\hat{\mathbf{b}} | \boldsymbol{\theta}_V) \right)^2 \right] \\ \mathbf{g}_{\boldsymbol{\theta}_V} &= \mathbb{E}_{\hat{\mathbf{b}} \sim \boldsymbol{\rho}, \mathbf{a} \sim \boldsymbol{\mu}} \left[w \nabla_{\boldsymbol{\theta}_V} V^\pi(\hat{\mathbf{b}} | \boldsymbol{\theta}_V) A^\pi(\hat{\mathbf{b}}, \mathbf{a} | \boldsymbol{\theta}_V) \right] \end{aligned} \quad (13)$$

Both actor and critic networks are trained using batch sampling from a buffer memory containing experience tuples $(\hat{\mathbf{b}}, \mathbf{a}, [\boldsymbol{\pi}_i]_{i=1}^N, c, \hat{\mathbf{b}}')$.

Optimization

The life-cycle management problem is often accompanied by several constraints, e.g., long or short-term budgets, risk tolerance, and other imposed restrictions by various agencies. Depending on their nature, constraints are handled in this work through state augmentation or primal-dual optimization techniques based on Lagrangian multipliers. We consider the following form of the stochastic optimization problem, as in (Saifullah, et al., 2024) :

$$\begin{aligned} \pi^* &= \arg \min_{\pi \in \Pi} \mathbb{E}_{s_{0:T}, o_{0:T}, a_{0:T}} \left[\sum_{t=0}^T \gamma^t c_t \mid a_t \sim \pi(o_{0:t}, a_{0:t-1}), s_0 \sim \mathbf{b}_0 \right] \\ \text{s.t. } Z_{h,k} &= \sum_{t=0}^T \gamma^t z_{h,k}(s_t, a_t) - \alpha_{h,k} \leq 0, \quad k = 1, \dots, K \\ Z_{s,m} &= \mathbb{E}_{s_{0:T}, o_{0:T}, a_{0:T}} \left[\sum_{t=0}^T \gamma^t z_{s,m}(s_t, a_t, s_{t+1}) \right] - \alpha_{s,m} \leq 0, \quad m = 1, \dots, M \end{aligned} \quad (14)$$

where $c = c(s_t, a_t, s_{t+1})$ is the cost incurred at time t by taking action a_t , transitioning from state s_t to state s_{t+1} and taking an observation o_t ; $\gamma \in [0, 1]$ is the discount factors; $\mathbf{b}_0$ is an initial belief; T is the length of the planning horizon; $Z_{h,k}$ and $Z_{s,m}$ are the hard and soft constraints, respectively; and $z_{h,k}$ and $z_{s,m}$ are their respective auxiliary costs; $\alpha_{h,k}$ and $\alpha_{s,m}$ are real-valued scalars, with the former determining the allowable values of K hard constraint thresholds that need to be satisfied at all time steps, and the latter representing the values for M soft constraint thresholds which need to be satisfied in expectation. More details on soft and hard constraints can be seen in (Andriotis & Papakonstantinou, 2021; Saifullah, et al., 2024). The policy function belongs to a space $\pi \in \Pi_c$, which contains all possible policies that are admissible under the existing constraints of the problem. Π_c is a subset of Π , the policy space of the unconstrained problem.

The constraint structure in Eq. (14) is adaptable to a wide variety of constraint types important to infrastructure management. Hard constraints, for example, may be used to simulate a broad range of fixed resource allocation and control action availability problems, such as those with budget restrictions. In turn, soft constraints in the form of Eq. (14) may be used to represent

constraints that need to be satisfied in expectation, such as risk-based or performance-based constraints (e.g., restricting poor condition infrastructure levels, imposing goals for good condition levels, and many others).

While hard constraints can be effectively addressed through state augmentation, this approach becomes infeasible for dealing with soft constraints. The impracticality arises from the requirement to track the entire distribution of the cumulative discounted value of $z_{s,m}$ (Andriotis & Papakonstantinou, 2021) under soft constraints. Consequently, the management of probabilistic constraints necessitates the application of Lagrangian relaxation. Given the aforementioned considerations, the optimization problem can be reformulated as articulated in (Andriotis & Papakonstantinou, 2021):

$$\begin{aligned}
V &= \max_{\lambda_1, \dots, \lambda_M \geq 0} \min_{\pi \in \Pi} \mathbb{E}_{s_{0:T}, y_{0:T}, a_{0:T}, \mathbf{y}_0} \left[\sum_{t=0}^T \gamma^t \left(\bar{c}(s_t, a_t, s_{t+1}, \mathbf{y}_t) + \sum_{m=1}^M \lambda_m g_{s,m} \right) - \sum_{m=1}^M \lambda_m \alpha_{s,m} \mid a_t \sim \pi(o_{0:t}, a_{0:t-1}, \mathbf{y}_t), s_0 \sim \mathbf{b}_0, \mathbf{y}_0 \right] \\
&= \max_{\lambda_1, \dots, \lambda_M \geq 0} \min_{\pi \in \Pi} V_{\lambda}^{\pi}(\mathbf{b}_0, \mathbf{y}_0) \\
\text{s.t. } \mathbf{y}_t &= \{y_{kt}\}_{k=1}^K, y_{kt} = \sum_{\tau=0}^{t-1} \gamma^{\tau} z_{h,k}(s_{\tau}, a_{\tau}), y_{k0} = 0, \\
y_{kt} &\in [0, \alpha_{h,k}], k = 1, \dots, K
\end{aligned} \tag{15}$$

where variables y_{kt} track the discounted cumulative value of the function related to hard constraints, $z_{h,k}$, up to time step $t-1$, and $\bar{c}$ is the cost function also considering y_{kt} . Variables y_{kt} are upper bounded by $\alpha_{h,k}$. Lagrange multipliers, λ_m , constitute the dual variables of the max-min dual problem, they are positive scalars, and are associated with the soft constraints. Moreover, parameterizing the system policy and value function as:

$$\pi(\mathbf{a} \mid \hat{\mathbf{b}}, \mathbf{y}) = \prod_{i=1}^N \pi_i(a^{(i)} \mid \mathbf{b}^{(i)}, \mathbf{y}) \tag{16}$$

where $\mathbf{y} = [y_k]_{k=1}^K$ is an augmented state vector of hard constraints, $\mathbf{b}^{(i)}$ is the belief of system component i , N is the number of system components.

$$\begin{aligned}
\pi_i(a^{(i)} \mid \mathbf{b}^{(i)}, \mathbf{y}) &\simeq \pi_i(a^{(i)} \mid \mathbf{b}^{(i)}, \mathbf{y}, \boldsymbol{\theta}_{\pi}^{(i)}) \\
V_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y}) &\simeq V_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y} \mid \boldsymbol{\theta}_{\nu})
\end{aligned} \tag{17}$$

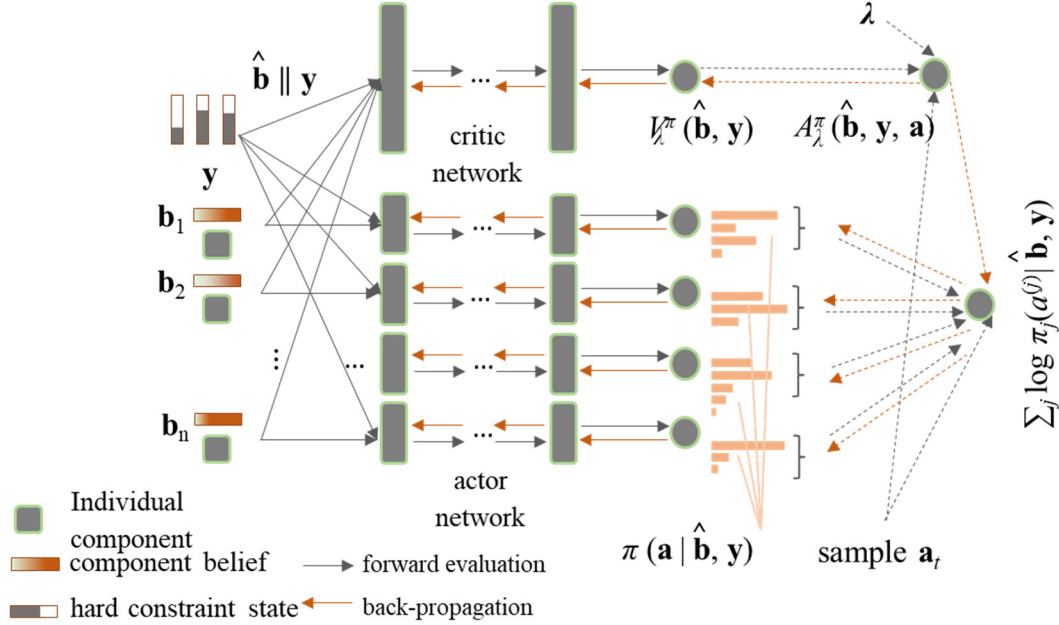


Figure 1: Constrained Deep Decentralized Multi-agent Actor Critic (DDMAC) with Centralized Training Decentralized Execution (CTDE) architecture, where $\parallel$ represents concatenation of $\hat{\mathbf{b}}$ and $\mathbf{y}$.

where, $\theta_{\pi}^{(i)}$, θ_V are weight parameters of policy and value network, respectively. Training is performed based on the collected buffer memory which contains $(\hat{\mathbf{b}}, \mathbf{y}, \mathbf{a}, [\pi_i]_{i=1}^N, \bar{c}_b, [z_{s,m}]_{m=1}^M, \hat{\mathbf{b}}', \mathbf{y}')$ tuples, where $\bar{c}_b$ is a cost collected based on the beliefs instead of states. The gradients of the policy function and the value function are computed by importance sampling as:

$$\nabla_{\theta_{\pi}^{(i)}} V_{\lambda}^{\pi} = \mathbb{E}_{\mathcal{M}} \left[w \left(\sum_{i=1}^N \nabla_{\theta_{\pi}^{(i)}} \log \pi_i(a^{(i)} | \mathbf{b}_i, \mathbf{y}, \theta_{\pi}^{(i)}) \right) A_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y}, \mathbf{a}) \right] \quad (18)$$

$$\nabla_{\theta_V} V_{\lambda}^{\pi} = \mathbb{E}_{\mathcal{M}} \left[w \nabla_{\theta_V} V_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y} | \theta_V) A_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y}, \mathbf{a}) \right] \quad (19)$$

$$\nabla_{\lambda_m} V_{\lambda}^{\pi} \simeq \sum_{t=0}^T \gamma^t z_{s,m} - \alpha_{s,m} \quad (20)$$

where w is the importance sampling weight with sample distribution a policy μ retrieved from the experience replay and target distribution the current policy A_{λ}^{π} . is the advantage function, which is herein approximated by the temporal difference:

$$A_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y}, \mathbf{a} | \theta_V) \simeq \bar{c}_b + \sum_{m=1}^M \lambda_m z_{s,m} + \gamma V_{\lambda}^{\pi}(\hat{\mathbf{b}}', \mathbf{y}' | \theta_V) - V_{\lambda}^{\pi}(\hat{\mathbf{b}}, \mathbf{y} | \theta_V) \quad (21)$$

Algorithm 1 DDMAC-CTDE with constraints

Initialize replay buffer

Initialize actor, critic, and dual parameters $\left[\boldsymbol{\theta}_{\pi}^{(j)}\right]_{j=1}^N, \boldsymbol{\theta}_V, [\lambda_m]_{m=1}^M$

for $episode = 1, \dots, N_{ep}$ **do**

for $t=1, \dots, T$ **do**

 Select action $\mathbf{a}_t$ at random according to exploration noise;

 Otherwise select action $\mathbf{a}_t \sim \boldsymbol{\mu}_t = \left[\pi_j(\cdot | \mathbf{b}_t^{(j)}, \mathbf{y}_t, \boldsymbol{\theta}_{\pi}^{(j)})\right]_j^N$

 Estimate costs $\bar{c}_{b,t}$ as:

if $t \neq T$ **then**

$$\bar{c}_{b,t} = \bar{c}_b;$$

else

$$\bar{c}_{b,t} = \bar{c}_b + \bar{c}_{S_T};$$

 calculate $z_{s,mt} = z_{s,m}$ given

 Observe $o_{t+1}^{(l)} \sim p(o_{t+1}^{(l)} | \mathbf{b}_t^{(l)}, \mathbf{y}_t, \mathbf{a}_t)$ for $l=1, 2, \dots, N$

 Compute beliefs $\mathbf{b}_{t+1}^{(l)}$ for $l=1, 2, \dots, N$

 Store tuple $(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}_t, \boldsymbol{\mu}_t, \bar{c}_{b,t}, [z_{s,mt}]_{m=1}^M, \hat{\mathbf{b}}_{t+1}, \mathbf{y}_{t+1})$ to replay buffer

end for

 Sample batch $(\hat{\mathbf{b}}_i, \mathbf{y}_i, \mathbf{a}_i, \boldsymbol{\mu}_i, \bar{c}_{b,i}, [z_{s,mi}]_{m=1}^M, \hat{\mathbf{b}}_{i+1}, \mathbf{y}_{i+1})$ from replay buffer

 If $\hat{\mathbf{b}}_i$ is terminal state $A_{\lambda,i}^{\pi} = \bar{c}_{b,i} + \sum_{m=1}^M \lambda_m z_{s,mi} - V_{\lambda}^{\pi}(\hat{\mathbf{b}}_i, \mathbf{y}_i | \boldsymbol{\theta}^V)$

 Otherwise $A_{\lambda,i}^{\pi} = \bar{c}_{b,i} + \sum_{m=1}^M \lambda_m z_{s,mi} + \gamma V_{\lambda}^{\pi}(\hat{\mathbf{b}}_{i+1}, \mathbf{y}_{i+1} | \boldsymbol{\theta}^V) - V_{\lambda}^{\pi}(\hat{\mathbf{b}}_i, \mathbf{y}_i | \boldsymbol{\theta}^V)$

 Update actor parameters $\boldsymbol{\theta}_{\pi}^{(j)}$ according to gradient:

$$\nabla_{\boldsymbol{\theta}_{\pi}^{(j)}} V_{\lambda}^{\pi} \simeq \sum_i w_i \left(\sum_{u=1}^N \nabla_{\boldsymbol{\theta}_{\pi}^{(j)}} \log \pi_u(a_i^{(u)} | \mathbf{b}_i^{(u)}, \mathbf{y}_i, \boldsymbol{\theta}_{\pi}^{(u)}) \right) A_{\lambda,i}^{\pi}$$

 Update critic parameters $\boldsymbol{\theta}^V$ according to gradient:

$$\nabla_{\boldsymbol{\theta}^V} V_{\lambda}^{\pi} \simeq \sum_i w_i \nabla_{\boldsymbol{\theta}^V} V_{\lambda}^{\pi}(\hat{\mathbf{b}}_i, \mathbf{y}_i | \boldsymbol{\theta}^V) A_{\lambda,i}^{\pi}$$

 Update dual variables $\lambda_m, m=1, \dots, M$, based on current policy return, according to gradient:

$$\nabla_{\lambda_m} V_{\lambda}^{\pi} \simeq \sum_{t=0}^T \gamma^t z_{s,mt} - \alpha_{s,m}$$

end for

The schematic is presented in [Figure 1](#), and the implementation steps are described in [Algorithm 1](#), as in ([Saifullah, et al., 2024](#)). In the subsequent sections the specific modeling of cost, and deterioration of pavements and bridges are described.

Cost modeling for transportation applications with traffic

Cost at different time steps can be decomposed into maintenance cost, c_M , inspection cost, c_I (if present), and risk cost, $c_{\mathcal{R}^\pi}$. In this study, we are not considering inspections as part of optimization hence their costs are not considered. In addition, costs relating to traffic can also be considered, such as delay cost due to maintenance, c_D and total travel time cost c_{TTC} . But other costs can also be considered, such as vehicle operating costs possibly incurred due to the system's current state, for example, extra fuel and tire costs due to the bad ride quality, environmental costs due to fuel emissions which will also depend on the system's state, and various other costs. The total cost at each decision step can be generally expressed as:

$$\begin{aligned} \bar{c}(s_t, a_{M,t}, a_{traffic,t}, s_{t+1}, V_{traffic,t}) = & \underbrace{c_M(s_t, a_{M,t})}_{\text{maintenan. cost}} + \underbrace{c_{\mathcal{R}^\pi}(s_t, a_{M,t}, s_{t+1})}_{\text{Risk cost}} + \underbrace{c_D(a_{M,t}, a_{traffic,t}, V_{traffic,t})}_{\text{Delay cost}} \\ & + \underbrace{c_{TTC}(a_{traffic,t}, V_{traffic,t})}_{\text{Total travel time cost}} \end{aligned} \quad (22)$$

where, $V_{traffic,t}$ is a traffic volume on each link/component of a transportation network at time t , $a_M \in A_M$ is the maintenance costs, $a_{traffic,t} \in A_{traffic}$ is a traffic adaptation action, For finite horizon problems with horizon T , often a terminal cost (c_{s_T}) is added to guide actions at the end of the horizon, i.e., at time step $t=T$, which is obtained from ([Saifullah, et al., 2024](#)) for this transportation application. After incorporating the budget constraint at time t and c_{s_T} at time step T , the cost can be modified as:

$$\begin{aligned} \bar{c}(s_t, a_{M,t}, a_{traffic,t}, s_{t+1}, V_{traffic,t}) = & \bar{c}_M(s_t, a_{M,t}) + \bar{c}_{\mathcal{R}^\pi}(s_t, a_{M,t}, s_{t+1}) + \bar{c}_D(a_{M,t}, a_{traffic,t}, V_{traffic,t}) \\ & + \bar{c}_{TTC}(a_{traffic,t}, V_{traffic,t}) + 1_{t=T} \bar{c}_{s_T}(s_t) \end{aligned} \quad (23)$$

The transition and observation probabilities are also affected by the budget constraint, because after utilizing the budget in a given budget cycle, the agents can only choose actions with no costs, as no maintenance and inspection actions can be performed. Further, the costs in Eqs. (22) & (23) can readily be obtained by calculating cost in expectation with respect to the belief state $\mathbf{b}$, as:

$$\begin{aligned} \bar{c}_{b,t} &= \bar{c}_{b,t}(\mathbf{b}_t, \mathbf{a}_t, \mathbf{y}_t) \\ &= E_{s_t, s_{t+1}} [\bar{c}(s_t, a_t, s_{t+1}, y_t)] \\ &= \sum_{s_t \in S} b(s_t) \sum_{s_{t+1} \in S} P(s_{t+1} | s_t, a) \bar{c}(s_t, a_t, s_{t+1}, y_t) \end{aligned} \quad (24)$$

where $\bar{c}_{b,t}$ is the estimated total cost at time t , obtained using belief state $\mathbf{b}_t$.

Budget related constraints

The funding for maintenance, dictated by the operational plan and resource allocation, must match goals tied to a budget cycle length, T_B . The total cost maintenance c_M at time τ is summarized as follows:

$$z_h(s_\tau, a_{M,\tau}) = c_M \mathbf{1}_{\tau \in \Lambda_t} \quad (25)$$

where, $\Lambda_t = \left(\lfloor t / T_B \rfloor T_B, (\lfloor t / T_B \rfloor + 1) T_B \right]$ is a budget interval of length T_B , $\lfloor x \rfloor$ is integer part of x .

For a given budget cap α_h , the maintenance cost at each time is adjusted as:

$$\bar{c}_M = \mathbf{1}_{y+z_h \leq \alpha_h} c_M \quad (26)$$

Eqs. (25) & (26) essentially affect maintenance cost, such that the accumulated cost $z_h(s_\tau, a_\tau)$ in a given budget interval T_B does not exceed the budget limit α_h for the interval. Introducing a new state variable $y = y_t$ allows the inclusion of budget constraints, which are assumed to be fully observable. This enables the agent to justify control actions based on the available budget, $\alpha_h - y_t$, at each time step.

Performance constraints and risk formulation

Now we can introduce constraints to manage performance related constraint over the system's life cycle, using the performance-related random variable. The generic formulation as provided in (Andriotis & Papakonstantinou, 2021) is as follows:

$$\begin{aligned} Z_s &= \mathbb{E}_{s_{0:T}, a_{0:T}} \left[\sum_{t=0}^T \gamma^t J_t^\pi \right] - \alpha_s \\ &= J^\pi - \alpha_s \end{aligned} \quad (27)$$

where, Z_s is a constraint function, J^π is a cumulative discounted cost (or any generic parameter that needs to be bounded) accrued by following policy π . J_t^π is a cost (or the parameter value) at time step t , α_s is the cap that needs to be satisfied over the life cycle. In transportation network management, the aim is often to restrict the deterioration of infrastructure components like pavements and bridges within federally mandated limits, ensuring they remain within acceptable condition thresholds, that can be modeled through the above formulation to satisfy performance related constraints. In this study, the specific performance constraints are considered based on condition indices of the assets considered over the entire network. The details on the considered constraint are provided in the later sections.

In reliability-oriented decision-making, a significant emphasis is placed on minimizing the probability of system failure, defined as the likelihood of transitioning from a state of non-failure to failure, throughout the operational lifespan of the system. In such cases, the risk is usually defined as an expected cumulative discounted failure state cost over the life cycle, as in (Andriotis & Papakonstantinou, 2021). In the application of transportation networks, utilizing the same

definition, we will associate risk with bridge failures only, and the associated cost can be calculated as:

$$c_{\mathfrak{R}^\pi} = \mathbb{E}_{o_{0:T}} \left[\sum_{t=0}^T \gamma^t \left(c_{F_1} P_{F_{t+1}|a_{0:t}o_{0:t}} + c_{F_2} \left(P_{F_{t+1}|a_{0:t}o_{0:t}} - P_{F_t|a_{0:t}o_{0:t}} \right) \right) \right] \quad (28)$$

where $c_{\mathfrak{R}^\pi}$ is the risk cost which has two components: (1) accruable cost (c_{F_1}) and (2) instantaneous cost (c_{F_2}). The cost of instantaneous failure risk is incurred when the system enters a damage state but is not accumulated while the system remains in this state. In this study, this cost is assumed to be 10 times the cost of rebuilding the bridge. Meanwhile, the accruable cost, set at twice the bridge's rebuilding cost, is applied at each time step during the relevant sojourn time. We have only considered the risk of individual bridge failures across the network. The risk from bridge failures is calculated using their failure probability and the aforementioned costs. This risk is then incorporated directly into the cost or objective function.

PAVEMENTS

Pavement condition can be evaluated using various indicators, both internationally recognized and state-specific ones like PCI, PQI, CCI, and IRI. This study adopts a Critical Condition Index (CCI) and CCI model based on the research by (Katicha, et al., 2016). Figure 2 illustrates the mean CCI based on pavement age for different traffic levels A to E. A nonstationary gamma process is employed to estimate transition probabilities for discrete CCI levels as in (Saifullah, et al., 2024).

Fitting a nonstationary gamma process

A gamma process is utilized in this section, with its mean in time equal to the modified mean CCI predictions in Figure 2 and a relevant model variance $\sigma_s^2(t)$, primarily determined based on (Katicha, et al., 2016) and the observed data trends there. The marginal probability of CCI at every step t can be obtained by estimating the damage index DI , where $DI = 100 - \text{CCI}$, using the gamma distribution with probability density function:

$$Ga(DI | f(t), h(t)) = \frac{h(t)^{f(t)}}{\Gamma(f(t))} DI^{f(t)-1} e^{-h(t)DI} \quad (29)$$

The gamma process is parametrized here by a non-negative time-varying scale parameter function $h(t)$ and a non-negative time-varying shape parameter function $f(t)$, and $\Gamma(u) = \int_0^\infty v^{u-1} e^{-v} dv$. The relevant parameters are estimated in time based on the mean DI prediction model and the model variance $\sigma_s^2(t)$. The relationship between the gamma process parameter functions and μ_{DI}, σ_s is given as:

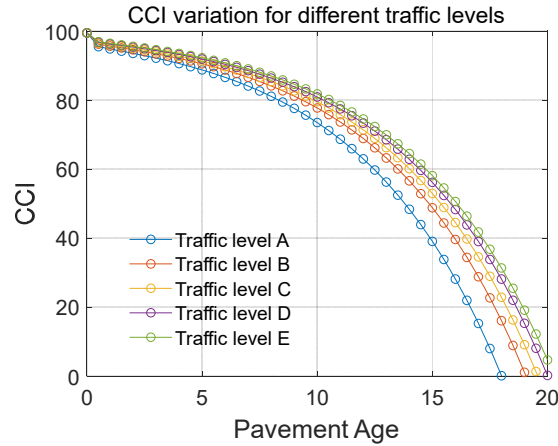


Figure 2. Modeled mean CCI for different levels of traffic.

$$\mu_{DI}(t) = \frac{f(t)}{h(t)}, \sigma_s(t) = \frac{\sqrt{f(t)}}{h(t)} \quad (30)$$

Due to their monotonicity, gamma processes are readily used as a suitable modeling choice in stochastic deterioration engineering applications and can describe continuous Markovian transitions. For time instant $t_1 < t_2$, the increment of DI follows a gamma distribution:

$$DI(t_2) - DI(t_1) \sim Ga(\cdot | f(t_2) - f(t_1), h(t_2)) \quad (31)$$

where $h(t)$ is assumed to be constant in the interval $(t_1, t_2]$. The corresponding $f(t)$ and $h(t)$ values are presented in (Saifullah, et al., 2024) for different traffic levels.

To determine the transition probabilities, the discrete condition states are first defined, as shown in Table 1. These discretized condition states in the table are herein largely adapted by the

Table 1. State discretization based on CCI values.

States of the pavement section	CCI value
$s = 6$	100-90
$s = 5$	89-80
$s = 4$	79-61
$s = 3$	60-50
$s = 2$	49-37
$s = 1$	<37

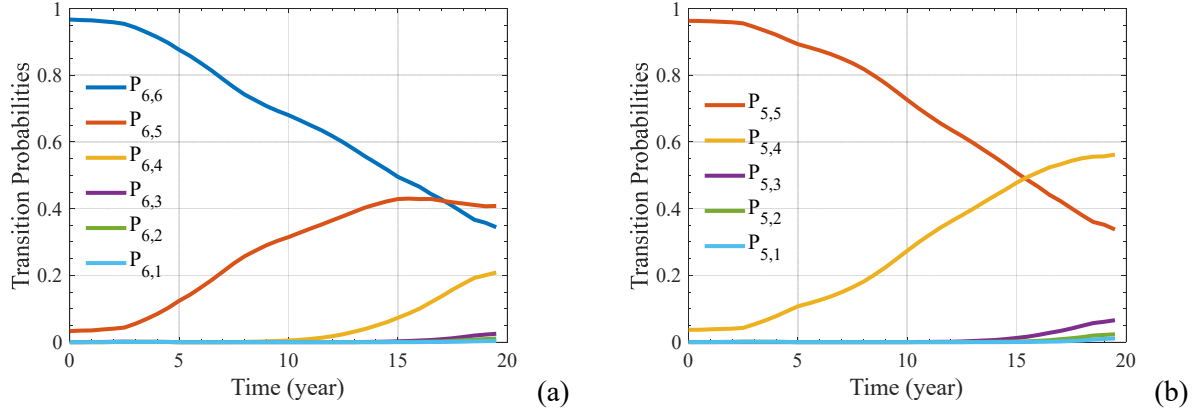


Figure 3. Transition probabilities for Traffic level A, with (a) starting state = 6, (b) starting state = 5, smoothed over time with a 5-point window.

prescribed VDOT guidelines in (Virginia Department of Transportation, 2016). Sampling is then performed by generating CCI and pavement age pairs, which are further propagated in time following the described gamma process in the previous section. In total, 10^6 sequences are generated to obtain the transition probabilities for a given traffic level. Figure 3 indicatively shows computed transition probabilities for traffic level A.

Observation probabilities for pavement inspection action

Observation uncertainty can be appropriately modeled by the likelihood functions $p(o_t|s_t)$ which quantify the probability of observing an observation o_t at time t given a state s_t at that instant. To calculate the observation probability $p(o_t|s_t)$, a normal distribution is considered with mean equal to the actual CCI value and an error variance $\sigma_{error}^2 = 18$, based on (Katicha, et al., 2016; Saifullah, et al., 2024). The error variance is linked to the variability in the measurement of CCI and the error in the reported pavement condition by an inspector. Using this inspection technique, we derive the observation probabilities as shown in Appendix A, Table A-5. For the no-inspection case, one has an equal probability of observing each state i.e., $p(o_t|s_t) = 1/6 \quad \forall o_t, s_t \in \{1, 2, \dots, 6\}$.

The cost of inspections depends on the techniques used for observing the pavement condition along with the post processing time/complexity. However, this study does not consider the inspections part of the optimization and thus the cost of inspection actions is not considered.

Maintenance, rehabilitation, and reconstruction actions in relation to CCI

The study employs four primary maintenance strategies as in (Saifullah, et al., 2024) inspired by (Virginia Department of Transportation, 2016): (*Do Nothing*, *Minor Repair*, *Major Repair*). Here, *Do Nothing* is the same as Do Nothing (DN) by VDOT, and *Minor Repair* is more aligned with VDOT's Preventive Maintenance (PM) category. *Major Repair* is, accordingly, similar to Corrective Maintenance (CM) and Restorative Maintenance (RM), and lastly *Reconstruction* is

Table 2. Maintenance action costs for asphalt pavements, reported in USD/m².

Actions	Description	Cost (USD/m ²) Interstate	Cost (USD/m ²) Primary
Do Nothing	NA	0.00	0.00
Minor Repair	Moderate patching (<10%), surface treatment, partial depth patching, thin overlay	20	16
Major Repair	Heavy patching (<20% of the pavement area), full depth patching, structural overlay	75	68
Reconstruction	Replacing the entire pavement section	350	330

Sources: (Virginia Department of Transportation, 2016; FDOT, 2020; PennDOT, 2017; ARTBA, 2020)

the same as VDOT's Reconstruction action. Further details of maintenance actions are provided in the [Appendix A, Table A-1](#) and can also be found in ([Saifullah, et al., 2024](#)).

The maintenance action transition probabilities for CCI are shown in [Appendix A, Tables A-2 to A-4](#) similar to the work presented in ([Madanat & Ben-Akiva, 1994](#)). It is assumed that the *Minor Repair* cannot change the rate of deterioration but the state of the component. The change in condition states is considered based on [Appendix A, Table A-2](#). For a *Major Repair* maintenance action, deterioration rate is reduced by 5 years or to that of a newly built pavement component, i.e., $(T_{age})_{updated} = \max(T_{age} - 5)$ years. *Major Repair* also alter the CCI conditions, with the transition probabilities shown in [Appendix A, Table A-3](#). The Reconstruction maintenance action brings the CCI condition back to the intact state (i.e., $s_t = 6$) with certainty, as shown in [Appendix A, Table A-4](#), and its deterioration rate to the initial intact value. Transitions in these tables do not include the transition because of the environment. Decoupled environmental transitions can essentially allow us to include action duration effects, among others.

Maintenance action costs

The cost of different maintenance actions is considered from (Virginia Department of Transportation, 2016; PennDOT, 2017; FDOT, 2020; ARTBA, 2020) and provided for interstate and primary roads in Table 2. It is reported in USD/m² and the total cost can be calculated using the components' length and width (which is considered as ~3.7 m or 12 feet per lane). Here, it is important to mention that the action descriptions provided in Table 2 for *Minor Repair*, *Major Repair* are indicative actions of the corresponding category and these actions are a subset of actions used in (Virginia Department of Transportation, 2016) as presented in Appendix A, Table A-1.

Table 3. Bridge failure probability given bridge condition state.

Condition State	$s_t = 9$	$s_t = 8$	$s_t = 7$	$s_t = 6$	$s_t = 5$	$s_t = 4, \dots$	$s_t = \text{failed}$
Failure prob. ($s_{t+1} = \text{failed}$)	0.001	0.001	0.005	0.005	0.005	0.01	1.0

BRIDGES

This study uses bridge decks that are categorized based on General Condition Rating (GCR), which ranges from 0-9. These GCR values are divided into 6 states (merging GCR values below 4 as one state) plus a failure state, as used in (Manafpour, et al., 2018; Saifullah, et al., 2024). The transition probabilities for decks are derived from 30-year data on over 22,000 Pennsylvania bridges, analyzed by (Manafpour, et al., 2018), considering various factors like deck length and span type, etc. Figure 4 shows indicative cases based on the various exogenous parameters for different rates of deterioration inspired by the study.

These different transition probabilities can be used when the deck deterioration is slow, medium, and fast, by moving left to right in Figure 4, respectively. Apart from these nonstationary transitions, stationary failure probabilities are also considered, as shown in Table 3.

Maintenance actions for bridge decks

Similar to pavements, four maintenance actions are considered for maintaining the bridge decks: *Do Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*. It is again assumed here that the *Minor Repair* action does not change the rate of deterioration of the deck, but it can change the condition state of the structure, as per the transition probabilities given in Appendix A, Table A-6. A *Major Repair* action can improve the deterioration rate to that of a deck younger by five years or a reset to a newly made deck, i.e., $(T_{age})_{updated} = \max(T_{age} - 5)$ years, and the change in condition state is now considered as reported in Appendix A, Table A-7. The transition probabilities for the *Reconstruction* action are shown in Appendix A, Table A-8, and reset the state and deterioration rate of the deck to a newly built one.

Maintenance action costs for bridge decks

The cost of maintenance actions has small dependence upon the current state of the bridge deck. As a result, maintenance costs are considered independent of the state of the deck here. The cost for a reconstruction action is obtained from (FHWA, 2019) and the costs for minor and major maintenance actions are considered as 15% and 45% of that, inferred from (Wells, 1995). Table 4 shows the costs of performing individual actions in USD/m². It can be observed that the unit costs of actions for bridges are higher than the unit costs of pavement actions.

Inspection actions and observation probabilities

This study focuses on two inspection actions for bridge decks: (1) no-inspection and (2) inspection, classified as uninformative and informative, respectively. The observation probabilities are

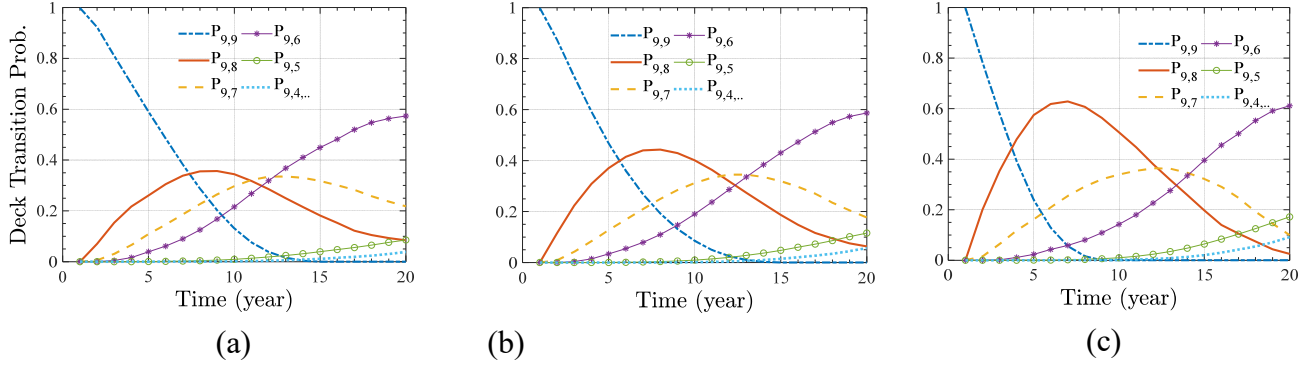


Figure 4. Transition probabilities in time, moving from state 9, for exogenous parameters
(a) district 4, span- multi span; (b) district 8, span- single span; (c) district 10, span- single span

Table 4. Cost of improvement actions for bridges

Actions	Description	Cost (USD/m ²)
Do Nothing	NA	0.00
Minor Repair	Moderate cracks filling and patching area <10% of the deck area, minor replacement of reinforcement	400.00
Major Repair	Fixing Spalls/delamination with deck area <25%, major replacement of reinforcement	1,200.00
Reconstruction	Replacing the entire deck	2,650.00

reported in (Saifullah, et al., 2024) and provided in Appendix A, Table A-9. For the no-inspection case, $p(o_i|s_i)=1/6 \ \forall o_i, s_i \in \{1,2,\dots,6\}$ is used. Again, the inspections actions are not considered as part of the action/policy space here and are kept fixed. Thus, the cost of inspection actions is not considered in this study.

EXAMPLE NETWORK

We consider in this study a 19-component transportation network, as shown in Figure 5. The network consists of 15 pavements and 4 bridges that are related to different network routes. Each bridge is bidirectional, with two different decks, types I and II, categorized based on traffic volume and route type. In this study, we only have one type I bridge that has the highest volume of traffic, with 4 lanes per direction, and it is part of high-traffic pavement components (similar to interstate highways); type II bridges usually have lower traffic volumes, with 2 lanes per direction, and are part of lower traffic pavement components (similar to primary highways), as considered in (Saifullah, et al., 2024). More details on bridge characteristics are shown in Table 5.

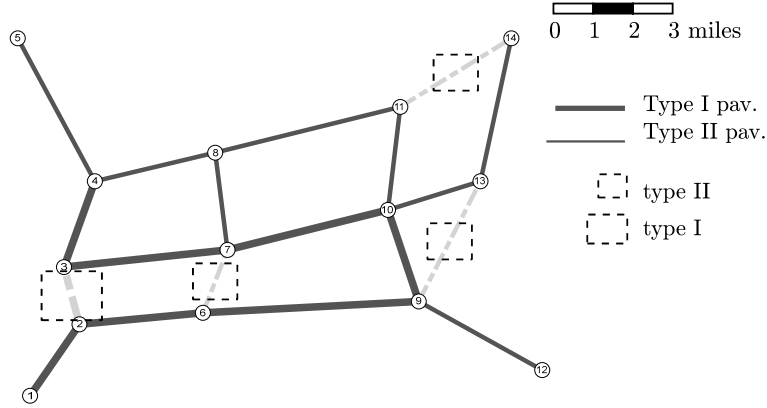


Figure 5: Roadway network with various pavement and deck components

Similarly, the network has various pavement components categorized as type I and II. Type I pavements are associated with interstate highways with bidirectional traffic having four lanes in each direction, thus, constituting the highest vehicular mileage per mile. Type II pavements correspond to the primary state highways with bidirectional medium-level traffic, having two lanes in each direction. Figure 5 represents the schematic of the network with thick links pertaining to interstate/type I highways and thinner links representing primary/type II highways. Further, the deterioration rates of pavements are being selected from Figure 2 based on the traffic levels in this work. Subsequent sections provide details on traffic assignment, user equilibrium, and traffic adaptation actions.

TRAFFIC ASSIGNMENT AND USER DELAY COSTS

Traffic assignment is the process of determining how demand traffic, usually defined in terms of an origin-destination matrix, is loaded onto the network, and it provides the means for computing traffic flows on the network links. Traffic assignment models based on Wardrop's principle (Wardrop, 1952) are known as user equilibrium models. User equilibrium is a concept in transportation planning that describes a state of traffic flow in which no individual driver can improve their travel time by unilaterally changing routes. In other words, user equilibrium is achieved when all drivers are using the routes that are best for them, given the current traffic conditions. (Smith, 1979; Boyles, et al., 2020) presented a variational inequality (VI) formulation:

$$(\mathbf{x} - \mathbf{x}^*)\mathbf{c}_l(\mathbf{x}^*) \geq 0 \quad (32)$$

where $\mathbf{x}$ is a vector of feasible link flows; $\mathbf{c}(\mathbf{x})$ is a vector of link costs; $\mathbf{x}^*$ indicates equilibrium link flows. When the link costs are a continuously increasing function of the link flows, another equivalent mathematical program (Beckmann, et al., 1956) can represent the UE as follows:

$$\begin{aligned}
& \min \sum_a \int_0^{x_a} tc_a(\eta) d\eta \\
& \text{s.t. } \sum_k q_k^{rs} = f^{rs}, \\
& q_k^{rs} \geq 0
\end{aligned} \tag{33}$$

where q_k^{rs} is the flow on route k between origin r and destination s ; f_{rs} the demand between r and s ; x_a the flow on link a ; $tc_a(\eta)$ is the travel cost on link a ; η represents any level of traffic flow on that link; travel cost typically increases as flow increases. This reflects the fact that the more congested a link becomes, the longer it takes to travel through it, thereby increasing travel costs. The specific form of tc_a can vary, often based on empirical data or theoretical traffic flow models.

The interpretation of the total travel cost function is that $\int_0^{x_a} tc_a(\eta) d\eta$ aggregates the travel costs across all vehicles on a link a as a function of traffic flow x_a . The solution of this program is the equilibrium route flows:

$$q_k^{rs} (tc_k^{rs} - tc^*) = 0, \quad tc_k^{rs} - tc^* \geq 0 \tag{34}$$

where tc_k^{rs} is the travel cost on route k between r and s ; tc^* is the user equilibrium route cost between r and s . Eq. (34) shows that the travel costs on each path carrying a positive flow are equal at UE.

In the context of Eq. (33) and the traffic assignment problem it represents, we are minimizing the total travel cost across the entire transportation network. This total cost is a summation of the individual travel costs on each link of the network, where the travel cost on a link is a function of the traffic flow on that link. The minimization is performed with respect to the traffic flow on each link of the network, denoted by x_a . The goal is to find the distribution of traffic flow x_a across all links a in the network that results in the lowest possible total travel cost, taking into account the congestion effects that increase travel costs as traffic flow increases.

The User Equilibrium (UE) problem can be solved through the application of the Frank-Wolfe algorithm (Frank, et al., 1956), also known as the Conditional Gradient Method, which is particularly suitable for solving large-scale network optimization problems. The Frank-Wolfe algorithm has been widely applied in transportation and traffic engineering to model and predict traffic flow patterns where no traveler can reduce their travel cost by unilaterally changing routes (LeBlanc, et al., 1975; Sheffi, 1985). The procedure involves the following steps:

Step 0: Initiate with an initial arbitrary flow, $x_a^{(0)}$, and set counter $n = 1$.

Step 1: Calculate the all-pairs shortest path in the network using the current link costs $tc_a(x_a^{(n-1)})$.

Step 2: Determine the descent direction $d^{(n)}$ by assigning all demand to the shortest paths identified in Step 1 and computing the difference in flow distribution $(y_a^{(n)} - x_a^{(n-1)})$.

Step 3: Conduct a line search to ascertain the optimal step size $\lambda^{(n)}$ that minimizes the total travel cost function.

Step 4: Update the traffic flow pattern according to $x_a^{(n)} = x_a^{(n-1)} + \lambda^{(n)}(y_a^{(n)} - x_a^{(n-1)})$.

Step 5: Evaluate the convergence criterion. If the relative change in total travel cost is below a predetermined threshold, terminate the algorithm. Otherwise, increment n and return to Step 1.

The convergence of the algorithm is assessed by the relative gap between the total travel costs of successive iterations, following a criterion that ensures the accuracy and reliability of the equilibrium state achieved (Patriksson, 2015).

The relationship between travel time, traffic volume, and user equilibrium (UE) is often described using the Bureau of Public Roads (BPR) function (Bureau of Public Roads, 1964). This function serves as the primary link-cost function in the traffic assignment algorithms employed in regional models. The BPR curve's simplicity and minimal input data requirements make it a practical choice. The curve ensures that the resulting travel time is a monotonically increasing function of the volume, which is essential for securing a singular user equilibrium solution in traffic assignment problems. Furthermore, the differentiability of the BPR curve is advantageous for the development of efficient algorithms aimed at finding the equilibrium solution. Essentially, the BPR function is an empirical equation that links travel time to traffic volume and capacity, and is defined as follows:

$$TT = TT_0 \left(1 + \alpha \left(\frac{q}{cap} \right)^\beta \right) \quad (35)$$

where: TT = Travel time on the road segment (in minutes), TT_0 = Free-flow travel time on the road segment, q = Traffic volume on the road segment (in vehicles per hour), cap = Capacity of the road segment (in vehicles per hour), and α and β are calibration parameters. In this equation, as traffic volume q approaches the capacity of the road cap , the travel time TT increases based on the parameters α and β . The values of α and β depends on specific traffic conditions and road characteristics. In this study $\alpha = 0.15$ and $\beta = 4$ are considered as predominantly used values in literature (Toledo & Bar-Gera, 2021). Capacities (in veh/hr) for each component can be obtained based on the FHWA (Federal Highway Administration, 2014) simplified guidelines as:

$$capacity = \frac{(2200 + 10(\min(70, FFS) - 50))}{1 + \%Truck / 100} \times Lanes \quad (36)$$

where, FFS is free flow speed, $\%Truck$ is the percentage of truck traffic. The free flow time (T_0) and capacity for each component are given in Table 5, and Table 6 gives the demand f_{rs} between origin node r and destination node $s \forall r, s \in 1, 2, \dots, 19$.

Traffic actions for pavement and bridges and its consequences

Capacity reduction in pavements and bridges refers to a decrease in their ability to handle traffic, often due to physical deterioration or maintenance activities. Over time, these structures experience wear and tear from traffic and environmental factors, leading to potential safety hazards and a need for repair. During maintenance, capacity is reduced as lanes need be closed, weight or height restrictions may be imposed, and through measures like speed limit reductions in the work zones. These maintenance activities, essential for long-term structural integrity and safety, temporarily disrupt the normal traffic flow and can lead to congestion and delays.

Alternatively, capacity reduction can occur without maintenance as a strategic traffic management approach. Techniques like ramp metering, dynamic speed limits, and demand management can be employed to control or redistribute traffic flow. Additionally, if maintenance cannot be performed (due to low budget), the capacity reduction serves as a strategic solution to delay the deterioration until possibly when the needed budget is available. In case the capacity reduction does not result in a decrease in traffic (hence the deterioration) of the considered component then we reverse the capacity action in our formulation.

In this work, two distinct strategies for adjusting traffic flow on pavements and bridges are examined, which involve modifying the capacity of these components. The strategy includes maintaining current capacity (no change) or reducing capacity by 50%. After selecting the capacity reduction action, the capacity of the targeted component is altered accordingly. Subsequently, a traffic assignment algorithm is employed, based on the principle of user equilibrium, to assess the impact of these capacity changes on traffic distribution across each component, in conjunction with the component maintenance activities. It is assumed that any maintenance work necessitates a 50% reduction in the capacity of the component due to lane closures. Additionally, for safety considerations, it is a customary practice to lower the speed limits within work zones, a measure that invariably affects traffic flow. This influence is accounted by integrating an additional capacity reduction in the bridge or pavement, as detailed in Equation (36), based on a speed reduction of 10 mph for components under maintenance. Thus, the recalibrated reduction in capacities for both Type I and Type II components are determined to be 52.2% of their original capacities due to maintenance. This approach allows for a comprehensive analysis of how different levels of capacity reduction influence the overall traffic patterns.

Once we get the traffic assigned to each component, we concurrently obtain the travel time for each component based on the BPR cost function. The travel time can also be converted to cost in dollars based on the user cost values provided in FHWA ([Federal Highway Administration, 2019](#)), calculated as below:

$$c_{TTC}(t) = \sum_i^N T_i(t)(c_{car}(1 - Truck_i(t)) + c_{truck}Truck_i(t)) \quad (37)$$

where, $c_{TTC}(t)$ is total travel time cost at time step t , $T_i(t)$ is travel time (in hrs) for component i , N is total number of components, c_{car} and c_{truck} are the cost obtained from FHWA and are equal to \$21.89/hr and \$29.65/hr, respectively; $Truck_i(t)$ represents the fraction of truck traffic on i^{th} component.

Effect of traffic actions on deterioration

In this study, the effect of traffic on deterioration is only considered through truck traffic as they carry the maximum load in terms of equivalent single axel load compared to the smaller vehicles, such as cars, pickup trucks, and small vans. For bridge maintenance, three sets of nonstationary Transition Probabilities (TPs) are used, corresponding to slow, moderate, and fast deterioration rates based on truck traffic. Fast deteriorating TPs, as depicted in [Figure 4 \(c\)](#), are applied when truck traffic on a bridge exceeds the nominal traffic level by 15%. Conversely, when truck traffic falls 15% below the nominal level, slow deteriorating TPs from [Figure 4 \(a\)](#) are selected. For truck traffic within a range of $\pm 15\%$ of the nominal level, moderate deterioration TPs from [Figure 4 \(b\)](#) are utilized.

In the case of pavements, there are five distinct deterioration levels, also determined by truck traffic volumes. Type I pavements with truck traffic exceeding 40% of the nominal level use the mean Critical Condition Index (CCI) curve for traffic level A ([Figure 2](#)) and its corresponding TPs ([Figure 3](#)). When truck traffic is between 110% and 140% of the nominal level, the mean CCI curve for traffic level B is then chosen. For truck traffic between 90% and 110% of the nominal, the curve for traffic level C is selected, and for truck traffic between 70% and 90% of nominal, the curve for traffic level D is used. Finally, for truck traffic below 70% of the nominal level, the curve for traffic level E is chosen.

Type II pavements follow a similar structure. If truck traffic increment is more than 20% of the nominal level, the mean CCI curve for level A is used. If truck traffic is increased 5% to 20% with respect to the nominal level, the mean CCI curve for level B is used. When the truck traffic change is accordingly between -5% and +5%, the level C curve is used, while for traffic changes between -25% and -5%, the level D curve is used, and finally when the traffic change is below -25% of the nominal, the level E curve is used. The nominal traffic level for the network is defined as the equilibrium traffic distribution when each component operates at 100% capacity without any maintenance.

In this context, while maintaining general applicability and without any loss of generality, we make the assumption that the depicted curves for various pavements serve as indicative examples for certain pavement design criteria and traffic levels. It is important to mention that we are not accounting for unanticipated traffic. It is assumed that the type I and II pavements are specifically designed to handle more traffic and are still able to follow the described deterioration curves in [Figure 2](#). It is important to note that the resulting curves may exhibit variations for alternative design criteria and traffic levels, and any additional changes can be similarly incorporated.

Table 5: Network characteristics.

Component	Length (miles)	Start node	End node	Capacity per lane (Veh/hr)	No. of lanes per direction	Total capacity (Veh/hr)	Free flow time T_0 (hrs)	Free flow speed (miles/hr)	AADTT %
1	2.16	1	2	2079.65	4	8318.58	0.033	65	0.13
2	2.92	2	6	2079.65	4	8318.58	0.045	65	0.13
3	2.30	3	4	2079.65	4	8318.58	0.035	65	0.13
4	3.88	3	7	2079.65	4	8318.58	0.060	65	0.13
5	5.10	6	9	2079.65	4	8318.58	0.078	65	0.13
6	3.92	7	10	2079.65	4	8318.58	0.060	65	0.13
7	2.44	9	10	2079.65	4	8318.58	0.038	65	0.13
8	4.07	4	5	2142.86	2	4285.71	0.074	55	0.05
9	2.93	4	8	2142.86	2	4285.71	0.053	55	0.05
10	2.49	7	8	2142.86	2	4285.71	0.045	55	0.05
11	4.52	8	11	2142.86	2	4285.71	0.082	55	0.05
12	3.39	9	12	2142.86	2	4285.71	0.062	55	0.05
13	2.64	10	11	2142.86	2	4285.71	0.048	55	0.05
14	2.30	10	13	2142.86	2	4285.71	0.042	55	0.05
15	3.71	13	14	2142.86	2	4285.71	0.067	55	0.05
16	1.50	2	3	2079.65	4	8318.58	0.023	65	0.13
17	1.70	6	7	2142.86	2	4285.71	0.031	55	0.05
18	3.38	9	13	2142.86	2	4285.71	0.062	55	0.05
19	3.15	11	14	2142.86	2	4285.71	0.057	55	0.05

Table 6 Considered traffic demand for different origin and destination nodes.

Origin nodes	Destination nodes													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	604.8	210.5	390	396.9	186.1	244.5	227.8	267.1	314.7	350.1	159.9	389.1	0
2	560.4	0	412.5	223.8	225.9	518.6	191.7	145.6	172.4	191.8	212.1	130.2	365	337.3
3	235.2	297.5	0	264.8	281.8	192.5	356.2	251.8	0	320.9	233.9	253.5	218.9	0
4	442	251.8	221.1	0	114.8	192.5	0	115.5	280.9	340	138.4	0	144.4	342.5
5	376.5	195.6	293.1	0	0	239.3	316.2	341.6	357.8	305.4	0	372	0	325.4
6	223.4	519	142.4	191.9	265.3	0	107	141	134.7	269.2	0	101.8	0	236.4
7	287.7	220.8	350.1	0	330.6	121.9	0	185.4	0	483.9	0	250.7	164.6	179.4
8	271.3	113.4	213.8	0	396.2	204.4	181.4	0	0	0	0	295.6	0	141.9
9	196.6	180.6	0	248.3	364.3	167.9	0	0	0	169.8	128.4	387	0	0
10	294.3	194.8	328.6	317.5	250.5	309.3	465.1	0	217.8	0	0	0	287.1	368
11	308.6	172.3	252.2	145.8	0	0	0	0	108.1	0	0	119.7	338.9	203.3
12	174.1	127.3	231.6	0	307.1	0	247.8	230.4	313.1	0	125.5	0	269.6	244
13	312.1	411.6	261.1	135.1	0	0	132.7	0	0	299.3	334.8	256.7	0	223.9
14	0	251.1	0	287.1	401.4	282	169.8	143.8	0	310.6	206.4	259.5	187.9	0

NETWORK CONSTRAINTS

This research incorporates three performance constraints, akin to the approach of (Saifullah, et al., 2024), delineated as follows:

- 1) A maximum of 5% of Type I pavements are allowed to be classified as in poor condition, indicated by a Condition Classification Index (CCI) below 60.
- 2) A similar cap of 5% applies to both Type I and Type II pavements being in poor condition (CCI<60).
- 3) The proportion of bridges deemed deficient, measured by deck area, must not exceed 10%.

These criteria align with the Federal Highway Administration (FHWA) and state agency standards for the maintenance of pavement and bridge infrastructure. They are incorporated as flexible, or "soft," constraints within our DRL framework for the network under study. This means that the

framework aims to achieve these performance targets on average over the network system's lifespan. Furthermore, the study evaluates various budget scenarios, renewed every five years, for different initial states of the infrastructure. Unlike the performance constraints, these budgetary considerations are treated as "hard" constraints, factoring into the network's state in a manner that is fully observable and directly influences the decision-making process.

RESULTS

This study examines a complex, dynamic 19-component transportation network, shown in [Figure 5](#). The problem is characterized by stochastic, non-stationary dynamics, and partially observable state conditions through uncertain inspections. The network is evaluated over a decision period of 20 steps (years), with a discount rate of $\gamma = 0.970$. For civil engineering systems, discount factors typically range from 0.93 to 0.98. Higher discount factors make the decision problem more challenging because they increase the effective horizon of important decisions. The specific characteristics of the network and relevant traffic data are detailed in [Table 5](#) and [Table 6](#).

This network includes 15 pavement sections, with 8 designated as Type I, comparable to interstate highways with multiple lanes and high traffic volumes, and 7 as Type II, akin to primary highways. Additionally, the infrastructure comprises 4 bridges: one Type I bridge situated on a Type I route and three Type II bridges on Type II routes. The details on pavement and bridge characteristics can be seen in the [Table 5](#). The deterioration rates of pavements and bridges are selected from [Figure 2](#) and [Figure 4](#), based on the different truck traffic levels in this work, without loss of generality, as mentioned in previous sections. Traffic volumes are also related to user delay costs associated with maintenance actions, as explained in more detail earlier.

There are 4 maintenance actions per component considered, with action labels 0-3, where {0, 1, 2, 3} represent Do Nothing, Minor Repair, Major Repair, and Reconstruction actions, respectively. Further, the traffic adaptation actions through capacity changes are also incorporated into the action space, where we have 100% and 50% available capacities to select from for a component only when the Do-Nothing action is selected. Otherwise, we select a 52.2% capacity reduction (2.2% accounts for a 10 mph speed reduction in the work zones). Additionally, inspection frequencies are set as annual for Type I components and biennial for Type II components, integrating these operational strategies into the decision-making framework.

The network's evaluation includes a variety of scenarios based on different initial conditions and budget allocations, with three 5-year budget cycles proposed for both intact and non-intact starting conditions (given in [Table B-1](#) of [Appendix B](#)). For networks that are initially intact, budgets of 600, 900, and unlimited million USD are allocated and renewed every five years. For the non-intact starting state, budgets of 900, 1,100, and unlimited million USD are used, spanning scenarios from low to high funding levels. These budgetary considerations aim to elucidate the influence of financial constraints on total costs and maintenance policies. [Figure 6 \(a\)](#) and [\(b\)](#) visually represent the network in both its intact and non-intact starting condition states. This comprehensive approach, combining detailed analysis of infrastructure characteristics, maintenance strategies, and budgetary frameworks, offers insights into the optimal management and operation of transportation networks under varying conditions and constraints.

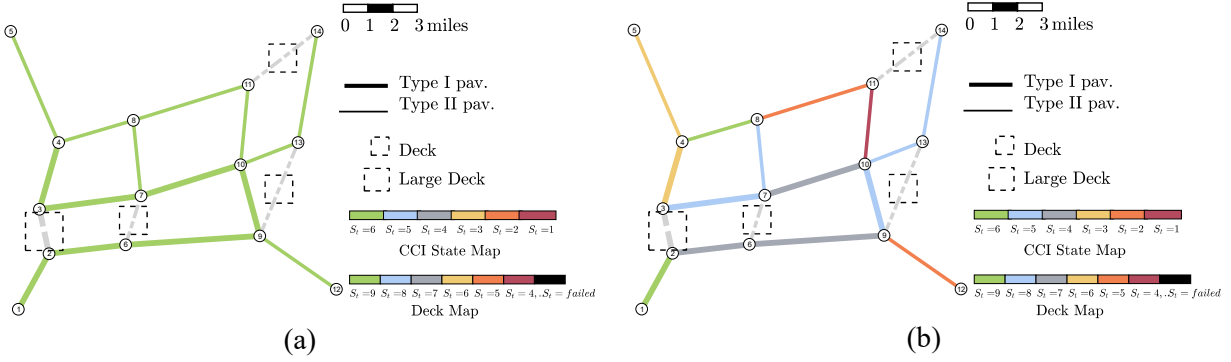


Figure 6: Discretized network with 19 components consisting of 4 bridges and 15 pavement sections starting with initial conditions corresponding to (a) intact state and (b) non-intact state.

Experimental setup

For the purposes of this numerical investigation, the actor and critic networks consist of fully connected layers. Each actor network (one per component) has two hidden layers with 200 neurons each, utilizing *ReLU* activation, and an output layer with 16 neurons using a *SoftMax* activation function. The centralized critic network features two hidden layers with 500 neurons each, also employing *ReLU* activation, and an output layer with 1 neuron using a linear activation function. The training involves 2×10^6 episodes, each lasting 20 time steps, with a buffer memory size of 15,000 episodes. The Adam optimizer (Kingma & Ba, 2014) is utilized for stochastic gradient descent on the network parameter space, with learning rates being gradually adjusted from 0.001 and 0.0001 to 0.0001 and 0.00001 for the critic and actor, respectively. The learning rate of Lagrange multipliers is set to 0.001. The exploration noise is linearly annealed from 100% to 5% at the first 100,000 episodes of the training process and then reduced to 2% afterwards.

DDMAC-CTDE solutions

The training of the DDMAC-CTDE network over 2×10^6 episodes spanned a duration of 10 to 14 days. Figure 7 and Figure 8 show the training curves for the considered transportation network, with both intact and non-intact initial states. Specifically, Figure 7(a) to Figure 7(c) represent the total training cost under the scenario of an intact initial state, comparing scenarios with and without traffic adaptation across three distinct 5-year budget allocations: (a) 600 million USD, (b) 900 million USD, and (c) an unlimited budget. These illustrations reveal that incorporating traffic adaptation strategies consistently results in lower total costs compared to scenarios omitting traffic considerations. Notably, the difference in total costs increases from 1% to 3% as the budget reduces from unrestricted to 600 million USD.

Similarly, Figures 8 (a) to 8 (c) show the total training costs in scenarios starting with non-intact condition states, again evaluating the impact of traffic adaptation under the three budgetary conditions: (a) 900 million USD, (b) 1,100 million USD, and (c) an unlimited budget. The trend here is similar to that of the intact case, indicating that traffic adaptation leads to reduced overall

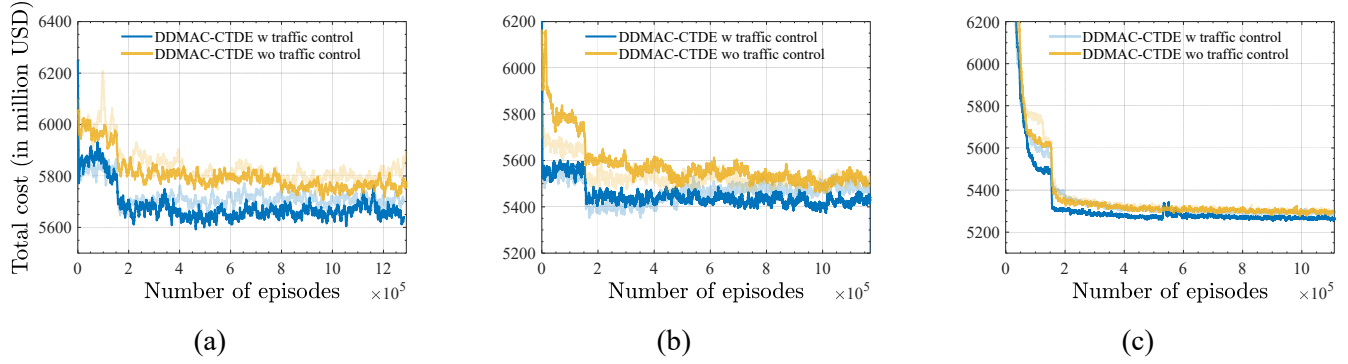


Figure 7: Total life cycle costs comparison of DDMAC-CTDE solutions with and w/o traffic adaptation actions for the considered network, starting from intact condition state with 5-yearly (a) budget of 600 million USD, (b) budget of 900 million USD, and (c) unrestricted budget.

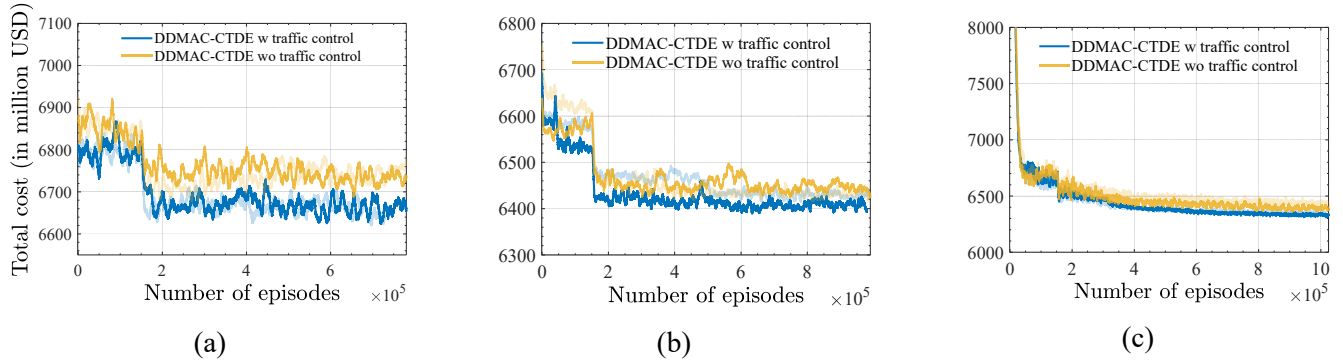


Figure 8: Total life cycle costs comparison of DDMAC-CTDE solutions with and w/o traffic adaptation actions for the considered network, starting from non-intact condition state with 5-yearly (a) budget of 900 million USD, (b) budget of 1,100 million USD, and (c) unrestricted budget.

total costs, with the cost difference increasing as the budget is reduced. In scenarios with lower budgets, the cost difference is slightly more pronounced for non-intact states, which is consistent with the increased financial requirements for maintaining the network and the potential need to alter traffic, thereby reducing deterioration.

These costs can also be observed in [Table 7](#) and [Table 8](#). For example, in the non-intact scenario with a 900 million USD budget, the inclusion of traffic actions yields a 3.5% cost improvement vs. the 3% improvement observed in the intact starting condition scenario with a 600 million USD budget case. For the next budget categories, the respective improvements stand at 2% for both starting conditions. These tables also show that without the terminal cost, the difference between the costs with and without traffic actions is even greater for both intact and non-intact cases. These observations are consistent from a theoretical standpoint. When additional actions are included in the agent's action space, the agent gains more options to choose from. If these additional actions are less relevant, the agent may simply opt not to select them. Therefore, adding extra actions to the agent's action space will result in either the same or an improved policy in terms of total cost.

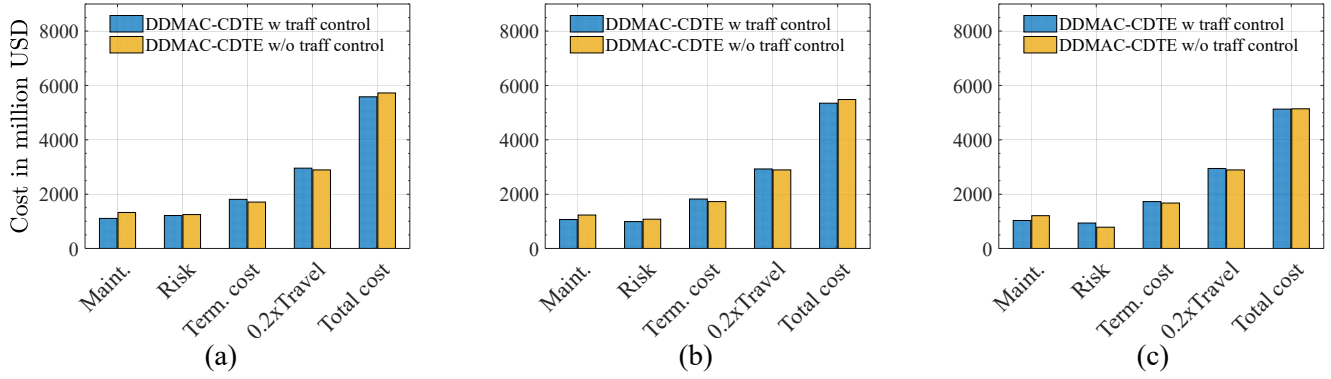


Figure 9: Comparison of the total cost and its constituents with and w/o traffic adaptation cases for the network, starting from intact condition state with 5-yearly (a) budget of 600 million USD, (b) budget of 900 million USD, and (c) unrestricted budget.

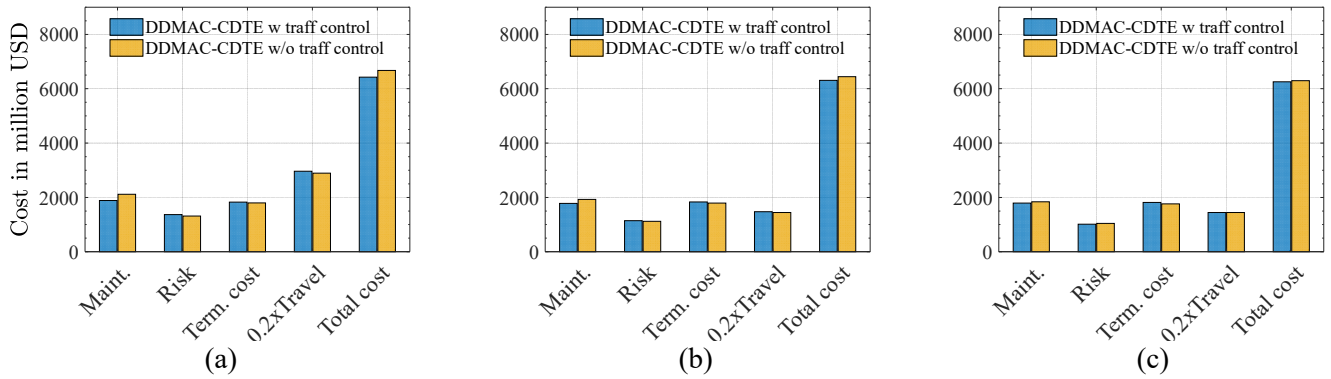


Figure 10: Comparison of the total cost and its constituents with and w/o traffic adaptation cases for the network, starting from non-intact condition state with 5-yearly (a) budget of 900 million USD, (b) budget of 1100 million USD, and (c) unrestricted budget.

Figures 9 and 10 show the total cost constituents using 10^4 simulations containing maintenance, delay, risk, terminal, and travel time costs for both intact and non-intact starting states, respectively. Figures 9(a)-(c) represent bar plot comparisons of these costs for the 5-yearly budget levels 600 million, 900 million, and unrestricted amount, respectively, for the intact case, and Figures 10 (a)-(c) similarly represent the bar plots for 900 million, 1100 million USD, unrestricted budget, respectively. Maintenance costs are notably higher (20%, 15%, and 17%) when traffic control actions are absent, compared to when they are present in the intact case with 600, 900 million USD, and unrestricted budgets, respectively. Similarly, these costs are higher by 12%, 8%, and 3% for the non-intact initial state. We observe these maintenance savings due to the fact that we can redirect traffic from the most expensive to repair components to other less expensive to repair components, slowing down the deterioration process where needed, for example, the large (type I) bridge connecting nodes 1 and 2. These results demonstrate high to moderate cost savings for the maintenance agency, particularly in tighter budget scenarios, except for intact starting condition state with unrestricted budget scenario, where agent saved more in maintenance cost by choosing less conservative policy and incurred a high risk cost. Travel time cost, displayed with a factor of 0.2 for visualization purposes, carries an actual weight of 0.1 in the

total cost calculation. We choose 0.1 to ensure comparability with other costs and avoid dominance in the solution.

Table 7 and Table 8 present a comparative analysis of the average outcomes from 10^4 simulations, focusing on three distinct performance constraints. These constraints are organized in the table rows, demonstrating that the DDMAC-CTDE algorithm effectively meets all specified constraints under both starting conditions and budget levels, with and without traffic adaptation measures. Some exceptions are noted in the scenario with an intact starting condition under a 600 million USD budget where traffic adaptation is applied; here, the performance slightly exceeds the threshold for total deficient type-I pavements ($CCI < 60$), surpassing the 5% cap by 0.43%. In scenarios without traffic adaptation, the agent exceeds the thresholds for total deficient type-I pavements ($CCI < 60$) by 0.49% and 0.46% for the 600 million USD and unrestricted budget conditions, respectively. Additionally, while starting from an intact condition, it also exceeds the threshold set for total deficient pavements (cap 5%) by 0.41%.

Additionally, under non-intact starting conditions, the DDMAC-CTDE agent with traffic control allowed the network to exceed the threshold for total deficient ($CCI < 60$) pavements (capped at 5%) by 0.05% for the 1,100 million USD budget case. In contrast, the DDMAC-CTDE agent without traffic control exceeded the same threshold by 0.35% and 0.37% for budgets of 900 million and 1,100 million USD, respectively. However, all these exceedances are less than 1% of the cap values, indicating that the agent adheres closely to these constraint thresholds. The tables also demonstrate the tight 95% confidence bounds on all the estimated values.

Figure 11 and Figure 12 show the details of the realization instances for the intact starting condition with 600 million USD as a 5-yearly budget and the non-intact starting condition with a 900 million 5-yearly budget, respectively. The figures illustrate actions generated by one of the instances of the optimum policy and the evolution of belief states as a contour. These figures also depict the discounted budget utilization over time, as well as the discounted 5-year budget for each cycle. The budget is a hard constraint that the agents cannot exceed, and this requirement is satisfied by the resulting solutions. However, out of 4 budget cycles, the budget got exhausted at least once, indicating a tight budget scenario. In the case starting from an intact condition, the budget was most heavily utilized during the second and third cycles, attributed to minimal deterioration in the initial cycle. Conversely, for scenarios beginning from non-intact conditions, significant budget allocation was necessary in the first three cycles to address the initial deficiencies and deterioration rates of the components. The figures also elucidate the temporal progression of three performance constraints within the context of pavement and bridge condition states throughout the network's lifecycle. It is seen that the three performance thresholds are respected for the most part of the life cycle. However, for these indicative realizations, the performance constraints are never violated on average over the horizon period. Additionally, these figures provide insights into the time-dependent risk costs associated with bridges. In the intact start case, the risk is minimal initially and increases with time, with downward jumps mainly due to the maintenance actions taken for bridges. However, for non-intact starting case, it starts with a high value and goes down significantly in the middle of the considered horizon (associated with major repairs in bridges) and then increases until the end of horizon. Plots with control actions (in yellow) represent the maintenance performed over time. These maintenance actions, taken at each step, update the current belief of the system, which is manifested in the subsequent step.

Table 7 Comparison of objectives and imposed constraints for DDMAC-CTDE solutions with and without traffic adaptation actions for intact starting condition states and various budget scenarios.

	5-yearly Budget					
	600 mill USD		900 mill USD		unrestricted	
	w traffic adaptation	w/o traffic adaptation	w traffic adaptation	w/o traffic adaptation	w traffic adaptation	w/o traffic adaptation
Total cost incurred (in billion USD)	5.55 ($\pm 2.8 \times 10^{-2}$)	5.73 (+3%) ($\pm 3.5 \times 10^{-2}$)	5.35 ($\pm 1.5 \times 10^{-2}$)	5.48 (+2%) ($\pm 2.7 \times 10^{-2}$)	5.14 ($\pm 2.1 \times 10^{-3}$)	5.19 (+1%) ($\pm 2.1 \times 10^{-3}$)
Total cost incurred w/o terminal cost (in billion USD)	3.77 ($\pm 2.8 \times 10^{-2}$)	3.95 (+5%) ($\pm 3.5 \times 10^{-2}$)	3.53 ($\pm 1.5 \times 10^{-2}$)	3.68 (+4%) ($\pm 2.8 \times 10^{-2}$)	3.42 ($\pm 2.2 \times 10^{-3}$)	3.47 (+1.5%) ($\pm 2.1 \times 10^{-3}$)
CCI<60 and for type I pav. (% cap 5%)	5.43 (± 0.07)	5.49 (± 0.07)	5.0 (± 0.06)	2.95 (± 0.05)	4.97 (± 0.05)	5.46 (± 0.05)
CCI<60 for type I and II pav. (% cap 5%)	4.88 (± 0.05)	5.41 (± 0.05)	4.86 (± 0.04)	2.75 (± 0.03)	4.19 (± 0.03)	3.69 (± 0.03)
Decks with condition rating ≤ 4 (% cap 10%)	8.5 (± 0.13)	5.72 (± 0.11)	4.25 (± 0.08)	5.45 (± 0.10)	9.75 (± 0.04)	6.55 (± 0.05)

Table 8 Comparison of objectives and imposed constraints for DDMAC-CTDE solutions with and without traffic adaptation actions for non-intact starting condition states and various budget scenarios.

	5-yearly Budget					
	900 mill USD		1100 mill USD		unrestricted	
	w traffic adaptation	w/o traffic adaptation	w traffic adaptation	w/o traffic adaptation	w traffic adaptation	w/o traffic adaptation
Total cost incurred (in billion USD)	6.49 ($\pm 2.2 \times 10^{-2}$)	6.72 (+3.5%) ($\pm 2.2 \times 10^{-2}$)	6.32 ($\pm 9.9 \times 10^{-3}$)	6.45 (+2%) ($\pm 1.0 \times 10^{-2}$)	6.25 ($\pm 2.9 \times 10^{-3}$)	6.30(+0.8%) ($\pm 3.0 \times 10^{-3}$)
Total cost incurred w/o terminal cost (in billion USD)	4.60 ($\pm 2.2 \times 10^{-2}$)	4.88 (+6%) ($\pm 2.2 \times 10^{-2}$)	4.47 ($\pm 1.0 \times 10^{-2}$)	4.65(+4%) ($\pm 1.0 \times 10^{-2}$)	4.44 ($\pm 1.9 \times 10^{-3}$)	4.53 (+2%) ($\pm 2.1 \times 10^{-3}$)
CCI<60 for type-I (% cap 5%)	3.08 (± 0.05)	4.92 (± 0.06)	4.21 (± 0.05)	4.41 (± 0.05)	2.70 (± 0.04)	2.84 (± 0.04)
CCI<60 for type-I and type-II pav. (% cap 5%)	5.0 (± 0.04)	5.35 (± 0.05)	5.05 (± 0.04)	5.37 (± 0.04)	4.45 (± 0.04)	4.80 (± 0.04)
Decks with condition rating ≤ 4 (% cap 10%)	9.22 (± 0.12)	8.01 (± 0.12)	10.0 (± 0.11)	9.38 (± 0.10)	9.23 (± 0.12)	9.98 (± 0.13)

Note: The estimated 95% confidence bounds are provided in parentheses, calculated based on 10^4 simulations. The bold values indicate the difference in costs with and without traffic adaptation actions for DDMAC-CTDE solutions.

Additionally, the traffic action in terms of capacity change has only two levels: no-change in capacity and a 50% reduction. Both capacity change actions are available with the *Do-Nothing* action. Each maintenance action results in a permanent capacity reduction of 52.2% for that time step. As mentioned earlier, these capacities affect the traffic equilibrium and the respective components' deterioration rates.

The combined effect of both actions on condition states is illustrated through the evolution of contours. For pavements, these contours show the current beliefs for the CCI states, while for decks, they show the current beliefs based on the considered discretized GCR states. For example, in [Figure 11](#), the DDMAC-CTDE agent implements a Major repair on the component connected by nodes (2, 3) at $t = 5, 14, 16$ years. The belief gets updated according to the Major Repair action and the environment effect transitions, selected based on the final truck traffic (estimated after traffic assignment) in the next time step. Recognizing it as a critical Type I bridge component (have larger traffic volume and deteriorate faster and are the most expensive to maintain), the agent initially opts to maintain the component's capacity. However, by $t = 5$ years, the agent strategically decreases its capacity by 50% to slow down its deterioration throughout its lifecycle. Similarly, for component connected by nodes (4,5), agent only takes one Major repair at $t = 5$ years and for the rest of the life cycle only managed its traffic by changing its capacity.

In contrast, for the component linked by nodes (1, 2), the agent enacts Minor and Major repairs at $t = 6$ and $t = 18$ years, respectively. Since this is a corner component, its traffic cannot be altered because the demand must be met, and no other component can fulfill the demand. Therefore, for corner components, the agent is not allowed to select capacity changes unless it is a repair action. The corner components are (1, 2), (4, 5), and (9, 12).

In [Figure 12](#), the agent consistently prefers Major repairs for both types of bridges. However, for Type I and Type II pavements, the agent prefers a mix of Minor and Major repairs depending on the specific needs of different components. Only a selection of these components is shown in the figure. Unlike the intact case, the agent begins by spending aggressively to reduce risk, which can be seen in the reduction in risk from a high value of 100 million to 45 million USD from $t = 0$ to $t = 1$ years. Additionally, there is a noticeable jump in risk at $t = 5$ years, which corresponds to an abrupt violation of the bridge threshold, as indicated in the constraint plot. The agent manages this situation by taking major actions on bridges, as seen in the action control plots (yellow), where both bridges receive a Major repair recommendation.

Similar to the previous case, for the Type I bridge connecting nodes (2, 3), the agent takes more traffic actions due to its high traffic volume. It recommends capacity reduction at 11 time steps, which constitutes about 55% of the total considered horizon (including during the 3 repairs), indicating traffic reduction and thus slower deterioration. For the Type I pavement shown in the figure, the agent opts for minor but frequent actions. This component started from state 5, which is close to the intact state, with an age of 7 years (as indicated in the [Table B-1](#) of [Appendix B](#)). Since this is a Very Good condition defined in [Table 1](#), and Major repairs cost significantly more than Minor repairs, the agent's actions are justifiable. On the other hand, the Type II pavement connecting nodes (10, 11) shown in the figure starts in condition state 1 ($CCI < 37$) and has a deterioration rate of 18 years (as indicated in the [Table B-1](#) of [Appendix B](#)), thus being among the worst possible states. Therefore, the agent decides to take multiple Major repairs over the lifecycle for this pavement component.



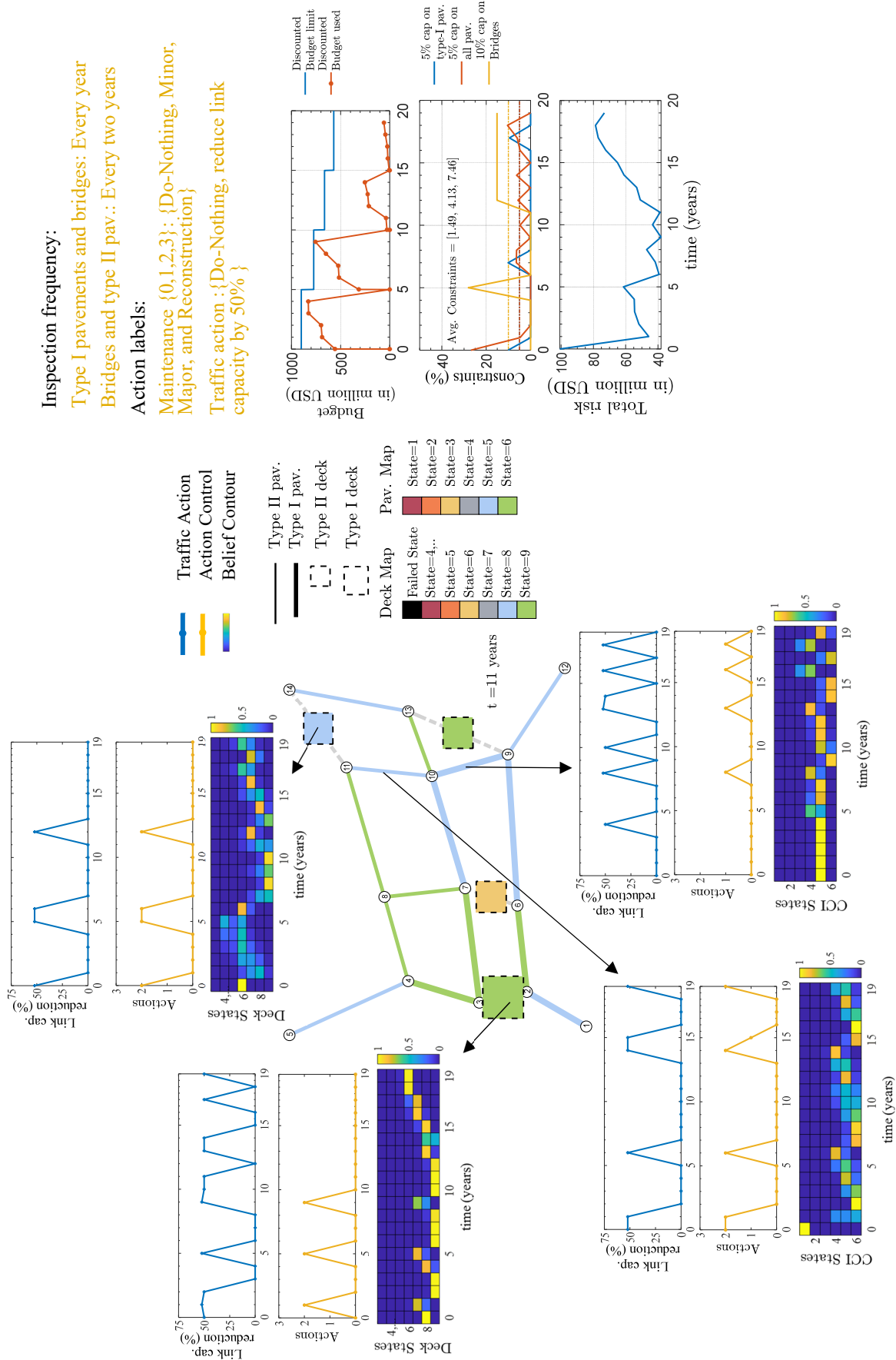


Figure 12: Life-cycle realization of the learned DDMAC-CTDE policy for the network, starting from nonintact state.

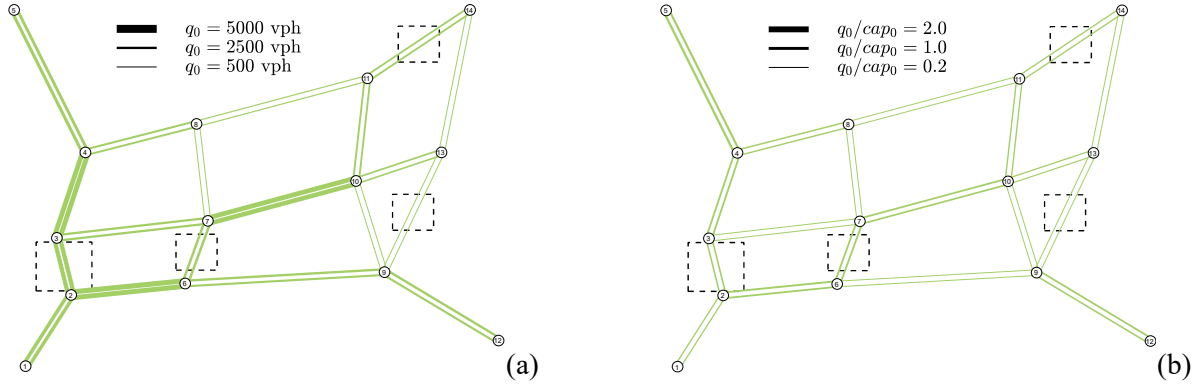


Figure 13: Network initial traffic condition (a) nominal traffic flow (q_0); (b) nominal flow to initial capacity ratio (q_0/cap_0)

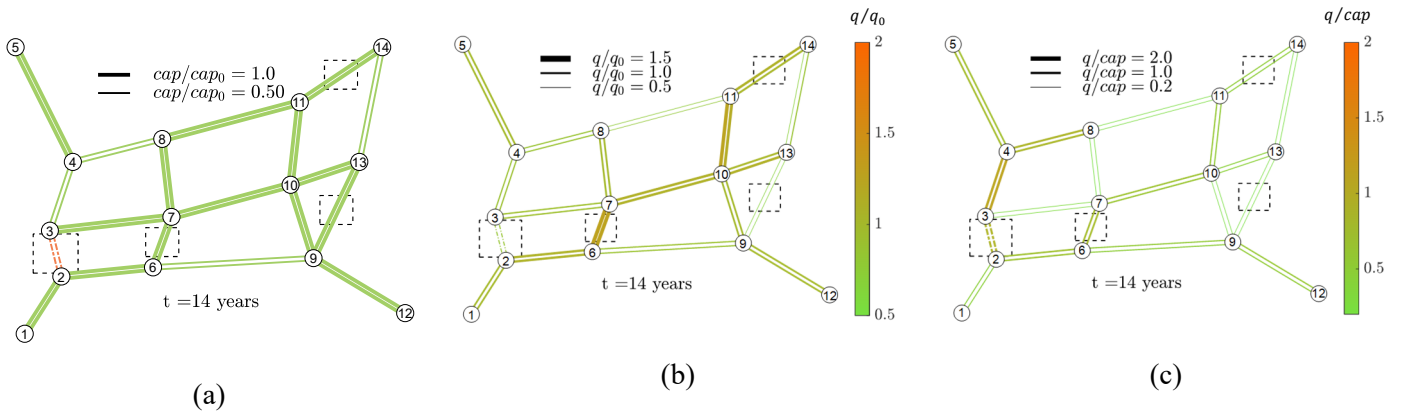


Figure 14: A snapshot of the network's traffic condition for the intact starting state, illustrating (a) the suggested traffic actions as indicated by the adjusted capacity ratio, (b) the traffic flow, and (c) the traffic flow to capacity ratio at time $t = 14$ years with 600 million USD 5-yearly budget. The dotted line represents the component under maintenance.

In general, it is observed that in the non-intact start case, the agent takes many maintenance actions over the components' lifecycle to satisfy the imposed performance constraints, compared to the intact start case, as expected. Figure 13 displays the nominal two-way traffic conditions of the network when no actions are taken by the agent. The traffic flow distribution is based on the user equilibrium, utilizing the origin-destination matrix and corresponding traffic demands from Table 6. This serves as a base traffic in selecting different deterioration curves (shown in Figure 2) based on the change in traffic levels, as discussed in previous sections. Figure 13(a) depicts the nominal traffic flow (q_0) for each component in vehicles per hour (vph). Figure 13(b) presents the traffic flow normalized by the initial capacities (cap_0) as outlined in Table 5.

The effects of traffic actions under intact and non-intact starting conditions are illustrated in Figure 14 and Figure 15, respectively. Figure 14 and Figure 15 (a) show the capacity changes recommended by the DDMAC-CTDE agent at time steps $t = 14$ and $t = 11$ years for the respective starting conditions. The maintenance activities are indicated in orange with dashed lines.

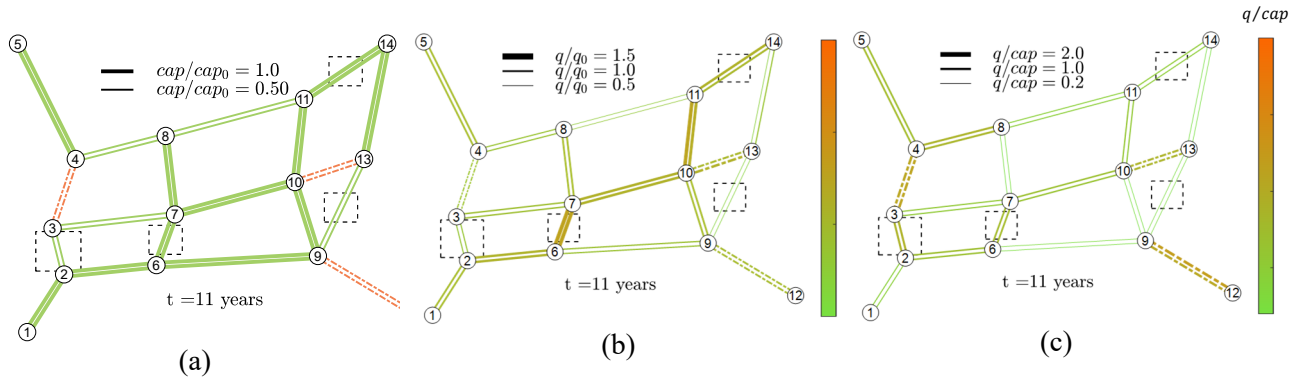


Figure 15: A snapshot of the network's traffic condition for the non-intact starting state, illustrating (a) the suggested traffic actions as indicated by the adjusted capacity ratio, (b) the traffic flow, and (c) the traffic flow to capacity ratio at time $t = 11$ years with 900 million USD. The dotted line represents the component under maintenance.

For the intact starting condition, [Figure 14 \(a\)](#) demonstrates the capacity modifications among various components: the component connecting nodes (2, 3) is under maintenance, with a mandated 52.2% reduction in capacity. The components connected by nodes (3, 4), (4, 8), (6, 9), and (13, 14) have their capacities reduced to 50%, while the rest of the components remain at 100% of their initial capacities. Similarly, for the non-intact starting condition, [Figure 15 \(a\)](#) demonstrates the capacity modifications among various components: the components connecting nodes (3, 4), (10, 13), and (9, 12) are under maintenance, with a mandated 52.2% reduction in capacity. The components connected by nodes (2, 3), (3, 7), (4, 8), (8, 11), and (9, 13) have their capacities reduced to 50%, while the rest of the components remain at 100% of their initial capacities.

[Figure 14](#) and [Figure 15 \(b\)](#) present the normalized traffic flow (q) with respect to the nominal traffic flow (q_0), as shown in [Figure 14 \(a\)](#). This illustration employs varying line thicknesses and a color-map to depict continuous values interpolated between and extending beyond three discrete levels, ranging from 0.5 to 2.0. It is important to note that during training, a reduction in the capacity of a component does not necessarily decrease the traffic flow through that component. In such cases, we reversed the capacity reduction action and re-ran the User Equilibrium. For the converged policy, the agent learned this behavior and subsequently largely avoided such actions.

Initially, it was assumed that the agent would always try to reduce the capacities of Type I components since they carry higher traffic volumes and deteriorate faster. For example, for the Type I bridge linked by nodes (2, 3), the agent employs the 50% capacity reduction option for most of its lifespan (see [Figure 11](#)). However, this relationship is more complex due to the dynamics of the entire network traffic, as the components are interconnected. A capacity change in one component affects the entire traffic distribution across the network. For instance, examining both [Figure 14 \(a\)](#) and [\(b\)](#), we observe that many components experience a traffic reduction even though they are at 100% capacity. Examples include the components connecting nodes (3, 7), (8, 11), and (9, 13), with normalized traffic values of $\{0.89, 0.88\}$, $\{0.76, 0.79\}$, and $\{0.66, 0.66\}$ in each direction, respectively. Similarly, examining [Figure 15 \(a\)](#) and [\(b\)](#), we find components (6, 9) and (13, 14) with reduced normalized traffic values of $\{0.87, 0.88\}$ and $\{0.69, 0.85\}$ in each

direction, respectively. This likely occurs because nearby components had their capacities reduced, which can be seen from the figures.

The agent's capacity reduction actions are also influenced by the shortest routes and their alternatives from different nodes. If a route is significantly shorter compared to alternatives and carries substantial traffic, it is identified as a critical route. The agent minimizes traffic disruptions on such routes by opting for fewer capacity reduction actions, thereby keeping traffic delays to a minimum. One such critical route is from node 1 to node 14, via nodes 1-2-6-7-10-11-14. This means the agent avoids reducing the capacities of links along this path, unless necessary. This strategy is evident in both realizations presented in [Figure 11](#) and [Figure 12](#), where the components (10, 11) and (11, 14), part of this route, only experience capacity reductions during repairs.

[Figure 14](#) and [Figure 15 \(c\)](#) illustrate the ratio of traffic flow to current capacity, providing an indication of the congestion map. A ratio greater than 1 signifies that traffic demand exceeds road capacity, indicating potential congestion, while a ratio less than 1 implies uncongested traffic. The corner components, as defined earlier, will always carry the same traffic. Therefore, during maintenance, they will have twice the q/cap (see the component connecting nodes (9, 12) in [Figure 15 \(c\)](#)) compared to the nominal q_0/cap_0 case, which is shown in [Figure 13 \(b\)](#). In some cases, even though traffic is reduced, we still observe an increased q/cap ratio due to the 50% capacity reduction. A common example is the component connecting nodes (3, 4). As a Type I pavement that has high demand, traffic does not decrease proportionately to the capacity reduction, resulting in an increased q/cap ratio in both intact and non-intact cases. However, most of the components are at or below the capacity in both cases.

CONCLUSION

This study undertakes the intricate challenge of multi-asset infrastructure management, considering the impact of maintenance actions on traffic operations and performance, as well as the unique incorporation of traffic operations adaptations and control within the inspection and maintenance planning framework. The employment of Deep Reinforcement Learning (DRL) techniques has enabled the development of a sophisticated methodology able to offer optimized maintenance policies and traffic adaptation strategies for infrastructure networks. Realistic, partial observability of asset conditions through inspections is also supported, since the framework is equipped to fully support Partially Observable Markov Decision Processes (POMDPs). Many operational, budget, and performance-based targets and constraints can also be imposed within the devised DRL formulation. As successfully demonstrated in an example application, integrating traffic adaptation capabilities with maintenance planning can successfully reduce total costs and, more importantly, maintenance costs, particularly in tight budget scenarios, while effectively meeting all performance targets and constraints. Overall, this study thus supports decision-makers in answering several important complex questions, offering a new promising perspective toward advancing transportation infrastructure management, enabling several possible future directions.

REFERENCES

- Aad, M., Abbas, M. & Wang, L., 2022. *A simulation-optimization framework for integrated infrastructure condition and traffic management strategies*. IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 2869-2874.
- Andriotis, C.P. & Papakonstantinou, K.G., 2019a. *Life-cycle policies for large engineering systems under complete and partial observability*. 13th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP13).
- Andriotis, C.P., Papakonstantinou, K.G. & Chatzi, E.N., 2021. Value of structural health information in partially observable stochastic environments. *Structural Safety*, Volume 93, p. 102072.
- Andriotis, C. P. & Papakonstantinou, K. G., 2019. Managing engineering systems with large state and action spaces through deep reinforcement learning. *Reliability Engineering & System Safety*, Volume 191, p. 106483.
- Andriotis, C. P. & Papakonstantinou, K. G., 2021. Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints. *Reliability Engineering & System Safety*, Volume 212, p. 107551.
- Arcieri, G. et al., 2023. Bridging POMDPs and Bayesian decision making for robust maintenance planning under model uncertainty: An application to railway systems. *Reliability Engineering & System Safety*, Volume 239, p. 109496.
- Arcieri, G. et al., 2024. POMDP inference and robust solution via deep reinforcement learning: An application to railway optimal maintenance. *Machine Learning*, pp. 1-29.
- ARTBA, 2020. *FAQ: How much does it cost to build a mile of road?* <https://www.artba.org/about/faq/>, The American Road & Transportation Builders Association.
- Beckmann, M., McGuire, C. B. & Winsten, C. B., 1956. *Studies in the Economics of Transportation*. New Haven, CT: Yale University Press.
- Bellman, R. E., 1957. *Dynamic programming*. Princeton University Press.
- Bocchini, P. & Frangopol, D. M., 2011. A probabilistic computational framework for bridge network optimal maintenance scheduling. *Reliability Engineering & System Safety*, 96(2), pp. 332-49.
- Boyles, S. D., Lownes, N. E. & Unnikrishnan, A., 2020. *Transportation network analysis*. Austin, TX: Version 0.91 (Public beta).
- Bureau of Public Roads, 1964. *Traffic Assignment Manual*., Washington, DC: Urban Planning Division, U.S. Department of Commerce Transportation Research Board, 2000.
- Chu, J. & Chen, Y., 2012. Optimal threshold-based network-level transportation infrastructure life cycle management with heterogeneous maintenance actions. *Transportation Research Part B: Methodological*, 46(9), pp. 1123-1143.
- Durango-Cohen, P. & Sarutipand, P., 2009. Maintenance optimization for transportation systems with demand responsiveness. *Transportation Research Part C: Emerging Technologies*, 17(4), p. 337-348.

- Faber, M. H. & Stewart, M. G., 2003. Risk assessment for civil engineering facilities: critical overview and discussion. *Reliability Engineering & System Safety*, 80(2), pp. 173-84.
- FDOT, 2020. *Cost Per Mile Models for Long Range Estimating*, Florida Department of Transportation.
- Federal Highway Administration, 2014. *Simplified Highway Capacity Calculation Method for the Highway Performance Monitoring System*. Available at: <https://www.fhwa.dot.gov/policyinformation/pubs/pl18003/chap02.cfm>
- Federal Highway Administration, 2019. *Work Zone Road User Costs - Concepts and Applications: Chapter 2. Work Zone Road User Costs*. Available at: <https://ops.fhwa.dot.gov/wz/resources/publications/fhwahop12005/sec2.htm>
- FHWA, 2019. *Bridge Replacement Unit Costs*, Federal Highway Administration.
- Foerster, J. et al., 2018. Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32.
- Frangopol, D. M., Lin, K. Y. & Estes, A. C., 1997. Life-cycle cost design of deteriorating structures. *Journal of Structural Engineering*, 123(10), pp. 1390-401.
- Frank, M., Wolfe, P. & others, 1956. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2), pp. 95-110.
- Gronauer, S. & Diepold, K., 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pp. 1-49.
- Guo, C. & Liang, Z., 2022. A predictive Markov decision process for optimizing inspection and maintenance strategies of partially observable multi-state systems. *Reliability Engineering & System Safety*, Volume 226, p. 108683.
- Katicha, S. W. et al., 2016. *VDOT: Development of Enhanced Pavement Deterioration Curves*, Virginia Department of Transportation.
- Kingma, D. P. & Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kıvanç, İ., Özgür-Ünlüakın, D. & Bilgiç, T., 2022. Maintenance policy analysis of the regenerative air heater system using factored POMDPs. *Reliability Engineering & System Safety*, Volume 219, p. 108195.
- LeBlanc, L. J., Morlok, E. K. & Pierskalla, W. P., 1975. An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research*, 9(5), pp. 309-318.
- Lee, D. & Song, J., 2023. Risk-informed operation and maintenance of complex lifeline systems using parallelized multi-agent deep Q-network. *Reliability Engineering & System Safety*, p. 109512.
- Lei, X., Xia, Y., Deng, L. & Sun, L., 2022. A deep reinforcement learning framework for life-cycle maintenance planning of regional deteriorating bridges using inspection data. *Structural and Multidisciplinary Optimization*, 65(5), p. 149.
- Luque, J. & Straub, D., 2019. Risk-based optimal inspection strategies for structural systems using dynamic Bayesian networks. *Structural Safety*, Volume 76, pp. 60-80.

- Madanat, S., 1993. Optimal infrastructure management decisions under uncertainty. *Transportation Research Part C: Emerging Technologies*, 1(1), pp. 77-88.
- Madanat, S. & Ben-Akiva, M., 1994. Optimal inspection and repair policies for infrastructure. *Transportation Science*, 28(1), pp. 55-62.
- Manafpour, A. et al., 2018. Stochastic analysis and time-based modeling of concrete bridge deck deterioration. *Journal of Bridge Engineering*, 23(9), p. 04018066.
- Medury, A. & Madanat, S., 2013. Incorporating network considerations into pavement management systems: A case for approximate dynamic programming. *Transportation Research Part C: Emerging Technologies*, pp. 134-150.
- Mendoza, J., Bismut, E., Straub, D. & Köhler, J., 2021. Risk-Based fatigue design considering inspections and maintenance. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 7(1), p. 04020055.
- Mnih, V. et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540), pp. 529--533.
- Mohammadi, R. & He, Q., 2022. A deep reinforcement learning approach for rail renewal and maintenance planning. *Reliability Engineering & System Safety*, Volume 225, p. 108615.
- Morato, P., Andriotis, C.P., Papakonstantinou, K.G. & Rigo, P., 2023. Inference and dynamic decision-making for deteriorating systems with probabilistic dependencies through Bayesian networks and deep reinforcement learning. *Reliability Engineering & System Safety*, 235(0951-8320), p. 109144.
- Morato, et al., 2022. Optimal inspection and maintenance planning for deteriorating structural components through dynamic Bayesian networks and Markov decision processes. *Structural Safety*, Volume 94, p. 102140.
- Ng, M., Lin, D. & Waller, S., 2009. Optimal long-term infrastructure maintenance planning accounting for traffic dynamics. *Journal of Computer-aided Civil and Infrastructure Engineering*, 24(7), p. 459-469.
- Oliehoek, F. A. & Amato, C., 2016. *A concise introduction to decentralized POMDPs*. Springer.
- Oliehoek, F. A., Spaan, M. T. & Vlassis, N., 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, Volume 32, pp. 289--353.
- Ouyang, Y., 2007. Pavement resurfacing planning for highway networks: parametric policy iteration approach. *Journal of Infrastructure Systems*, Volume 13, pp. 65-71.
- Papakonstantinou, K.G., Andriotis, C.P., Gao, H. & Chatzi, E.N., 2019. *Quantifying the value of structural monitoring for decision making*. 13th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP13).
- Papakonstantinou, K. G., Andriotis, C. P. & Shinozuka, M., 2016. *Point-based POMDP solvers for life-cycle cost minimization of deteriorating structures*. The 5th International Symposium on Life-Cycle Civil Engineering (IALCCE 2016).
- Papakonstantinou, K. G., Andriotis, C. P. & Shinozuka, M., 2018. POMDP and MOMDP solutions for structural life-cycle cost minimization under partial and mixed observability. *Structure and Infrastructure Engineering*, 14(7), pp. 869-882.

- Papakonstantinou, K. G. & Shinozuka, M., 2014a. Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory. *Reliability Engineering & System Safety*, Volume 130, pp. 202-213.
- Papakonstantinou, K. G. & Shinozuka, M., 2014b. Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part II: POMDP implementation. *Reliability Engineering & System Safety*, Volume 130, pp. 214-224.
- Papakonstantinou, K. & Shinozuka, M., 2014. Optimum inspection and maintenance policies for corroded structures using partially observable Markov decision processes and stochastic, physically based models. *Probabilistic Engineering Mechanics*, Volume 37, pp. 93-108.
- Patriksson, M., 2015. *The traffic assignment problem: models and methods*. Dover Publications.
- PennDOT, 2017. *Road Maintenance and Preservation (MaP)*, Pennsylvania Department of Transportation.
- Pineau, J., Gordon, G. & Thrun, S., 2003. *Point-based value iteration: An anytime algorithm for POMDPs*. IJCAI, pp. 1025-1032.
- Prajapati, A., Sengupta, A. & Guler, S. I., 2024. Accounting for traffic dynamics in pavement maintenance scheduling. *Transportation Research Record*, 2678(7), pp. 361-375.
- Rashid, T. et al., 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *International Conference on Machine Learning*, pp. 4295--4304.
- Rodriguez, M. L. R. et al., 2022. Multi-agent deep reinforcement learning based predictive maintenance on parallel machines. *Robotics and Computer-Integrated Manufacturing*, Volume 2022, p. 102406.
- Saifullah, M., Andriotis, C.P., Papakonstantinou, K.G. & Stoffels, S., 2022. *Deep reinforcement learning-based life-cycle management of deteriorating transportation systems*. IABMAS, pp. 293-301.
- Saifullah, M., Papakonstantinou, K. G., Andriotis, C. P. & Stoffels, S. M., 2024. Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management. *arXiv preprint arXiv:2401.12455*.
- Saydam, D. & Frangopol, D., 2014. Risk-based maintenance optimization of deteriorating bridges. *Journal of Structural Engineering*, 141(4), p. 04014120.
- Schulman, J. et al., 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Sheffi, Y., 1985. *Urban transportation networks*. New Jersey: Prentice-Hall.
- Smith, M., 1979. The existence, uniqueness and stability of traffic equilibria. *Transportation Research 13B*, pp. 295-304.
- Spaan, M. T. & Vlassis, N., 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, Volume 24, pp. 195-220.
- Straub, D. & Faber, M. H., 2005. Risk based inspection planning for structural systems. *Structural Safety*, 27(4), pp. 335-355.
- Su, J., Adams, S. & Beling, P., 2021. Value-decomposition multi-agent actor-critics. *Proceedings of the AAAI conference on Artificial Intelligence*, 35(13), pp. 11352--11360.

- Sunehag, P. et al., 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.
- Sutton, R. & Barto, A., 2018. *Reinforcement Learning: An Introduction*. 2nd ed. The MIT Press.
- Sutton, R. S., McAllester, D. A., Singh, S. P. & Mansour, Y., 2000. *Policy gradient methods for reinforcement learning with function approximation*. Denver, CO, s.n., pp. 1057--1063.
- Sutton, R. S., McAllester, D. A., Singh, S. P. & Mansour, Y., 2000. *Policy gradient methods for reinforcement learning with function approximation*. Advances in neural information processing systems 12, pp. 1057--1063.
- Toledo, O. & Bar-Gera, H., 2021. *BPR cost and objective functions*. Available at: <https://rdrr.io/rforge/travelr/man/bpr.function.html>
- Virginia Department of Transportation, 2016. *VDOT: Supporting Document for the Development and Enhancement of the Pavement Maintenance Decision Matrices Used in the Needs-Based Analysis*, Virginia Department of Transportation.
- Wardrop, J. G., 1952. Road paper. Some theoretical aspects of road traffic research.. *Proceedings of the Institution of Civil Engineers*, 1(3), pp. 325--362.
- Wells, D. T., 1995. *Technical Assistance Report: Maintenance, Repair, and Rehabilitation Unit Costs for Pontis*, Virginia Transportation Research Council.
- Yang, D. Y., 2022. Adaptive risk-based life-cycle management for large-scale structures using deep reinforcement learning and surrogate modeling. *Journal of Engineering Mechanics*, 148(1), p. 04021126.
- Yang, D. Y. & Frangopol, D. M., 2019. Life-cycle management of deteriorating civil infrastructure considering resilience to lifetime hazards: A general approach based on renewal-reward processes. *Reliability Engineering & System Safety*, Volume 183, pp. 197-212.
- Ye, Z., Cai, Z., Yang, H., Si, S., & Zhou, F., 2023. Joint optimization of maintenance and quality inspection for manufacturing networks based on deep reinforcement learning. *Reliability Engineering & System Safety*, Volume 236, p. 109290.
- Zhou, G. & Wang, L., 2012. Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation. *Transportation Research Part C: Emerging Technologies*, 21(1), pp. 287--305.
- Zhou, W., Miller-Hooks, E., Papakonstantinou, K.G., Stoffels, S., & McNeil, S., 2022. A reinforcement learning method for multiasset roadway improvement scheduling considering traffic impacts. *Journal of Infrastructure Systems*, 28(4), p. 04022033.

APPENDIX A. MAINTENANCE ACTIVITIES AND TRANSITION PROBABILITIES

Table A-1 Maintenance activities for interstate and primary pavements under different categories (Virginia Department of Transportation, 2016).

Activity Category	Activities
Do Nothing (DN)	N/A
Routine Maintenance (PM)	1. Minor Patching (<5 of Pavement Area; Surface Patching; Depth 2")
	2. Crack Sealing
	3. Thin Treatments (Chip Seal, Slurry Seal, Latex, Thin Hotmix Asphalt Concrete (THMACO), 'Novachip' etc.)
Corrective Maintenance (CM) Major action	1. Moderate Patching (<10 of pavement area; Partial Depth Patching; Depth 6")
	2. Partial Depth Patching (<10 of Pavement Area; Depth 4"-6") and Surface Treatment
	3. Partial Depth Patching (<10 of Pavement Area; Depth 4"-6") and Thin (≤ 2 ") AC Overlay
	4. ≤ 2 " Milling and ≤ 2 " AC Overlay
Restorative Maintenance (RM)	1. Heavy Patching (<20 of Pavement Area; Full Depth Patching; Depth 12")
	2. ≤ 4 " Milling and Replace with ≤ 4 " AC Overlay
	3. Full Depth Patching (<20 of Pavement Area; Full Depth Patching; Depth 9"-12") and 4" AC Overlay
	4. Cold In-Place Recycling
Rehabilitation /Reconstruction (RC)	1. Mill, Break and Seat and 9"-12" AC Overlay
	2. Reconstruction
	3. Full Depth Reclamation

Table A-2 . Minor Repair transition probabilities for 6 CCI states.

Condition State	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 6$	0.97	0.03	0.00			
$s_t = 5$	0.87	0.10	0.03			
$s_t = 4$	0.40	0.47	0.10	0.03		
$s_t = 3$		0.40	0.47	0.10	0.03	
$s_t = 2$			0.40	0.47	0.10	0.03
$s_t = 1$				0.40	0.47	0.13
Deterioration rate	Does not change					

Table A-3. Major Repair transition probabilities for 6 CCI states.

Condition State	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 6$	1.00	0.00	0.00			
$s_t = 5$	0.96	0.04	0.00			
$s_t = 4$	0.80	0.20	0.00			
$s_t = 3$	0.65	0.25	0.10			
$s_t = 2$	0.50	0.30	0.20			
$s_t = 1$	0.40	0.30	0.30			
Deterioration rate	Reset by 5 years					

Table A-4. Reconstruction transition probabilities for 6 CCI states.

Condition State	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 6$	1.00	0.00	0.00			
$s_t = 5$	1.00	0.00	0.00			
$s_t = 4$	1.00	0.00	0.00			
$s_t = 3$	1.00	0.00	0.00			
$s_t = 2$	1.00	0.00	0.00			
$s_t = 1$	1.00	0.00	0.00			
Deterioration rate	Reset to new pavement section					

Table A-5. Observation probability $p(o_t|s_t)$ given actual state s_t

Actual state	$p(o_t = 6 s_t)$	$p(o_t = 5 s_t)$	$p(o_t = 4 s_t)$	$p(o_t = 3 s_t)$	$p(o_t = 2 s_t)$	$p(o_t = 1 s_t)$
$s_t = 6$	0.801	0.197	0.002	0.000	0.000	0.000
$s_t = 5$	0.153	0.664	0.183	0.000	0.000	0.000
$s_t = 4$	0.001	0.078	0.822	0.099	0.000	0.000
$s_t = 3$	0.000	0.000	0.149	0.693	0.158	0.000
$s_t = 2$	0.000	0.000	0.001	0.137	0.718	0.144
$s_t = 1$	0.000	0.000	0.000	0.000	0.042	0.958

Table A-6. Minor Repair transition probabilities for deck states.

Condition State	$s_t = 9$	$s_t = 8$	$s_t = 7$	$s_t = 6$	$s_t = 5$	$s_t = 4, \dots$	$s_t = failed$
$s_t = 9$	0.97	0.03	0.00				
$s_t = 8$	0.85	0.12	0.03				
$s_t = 7$	0.40	0.45	0.12	0.03			
$s_t = 6$		0.40	0.45	0.12	0.03		
$s_t = 5$			0.40	0.45	0.12	0.03	
$s_t = 4, \dots$				0.40	0.45	0.15	
$s_t = failed$							1.0
Deterioration rate	Does not change						

Table A-7. Major Repair transition probabilities for deck states.

Condition State	$s_t = 9$	$s_t = 8$	$s_t = 7$	$s_t = 6$	$s_t = 5$	$s_t = 4, \dots$	$s_t = failed$
$s_t = 9$	1.00	0.00	0.00				
$s_t = 8$	0.95	0.05	0.00				
$s_t = 7$	0.80	0.20	0.00				
$s_t = 6$	0.60	0.30	0.10				
$s_t = 5$	0.40	0.40	0.20				
$s_t = 4, \dots$	0.30	0.40	0.30				
$s_t = failed$							1.0
Deterioration rate	Reset by 5 years						

Table A-8. Reconstruction transition probabilities for deck states.

Condition State	$s_t = 9$	$s_t = 8$	$s_t = 7$	$s_t = 6$	$s_t = 5$	$s_t = 4, \dots$	$s_t = \text{failed}$
$s_t = 9$	1.00	0.00	0.00				
$s_t = 8$	1.00	0.00	0.00				
$s_t = 7$	1.00	0.00	0.00				
$s_t = 6$	1.00	0.00	0.00				
$s_t = 5$	1.00	0.00	0.00				
$s_t = 4, \dots$	1.00	0.00	0.00				
$s_t = \text{failed}$	1.00	0.00	0.00				
Deterioration rate	Reset to newly built deck						

Table A-9. Observation probability $p(o_t|s_t)$ given state s_t for decks.

Actual state (s_t)	$p(o_t=9 s_t)$	$p(o_t=8 s_t)$	$p(o_t=7 s_t)$	$p(o_t=6 s_t)$	$p(o_t=5 s_t)$	$p(o_t=4, \dots s_t)$	$p(o_t=\text{failed} s_t)$
$s_t = 9$	0.90	0.10					
$s_t = 8$	0.10	0.80	0.10				
$s_t = 7$		0.10	0.80	0.10			
$s_t = 6$			0.10	0.80	0.10		
$s_t = 5$				0.10	0.80	0.10	
$s_t = 4, \dots$					0.10	0.90	
$s_t = \text{failed}$							1.0

APPENDIX B. NETWORK STATE

This appendix shows the non intact condition state of the network.

Table B-1. Network's non-intact starting condition state based on CCI and GCR for pavements and bridges.

Component	Start Node	End Node	Condition State	Deterioration rate (in years)
1	1	2	6	5
2	2	6	4	3
3	3	4	3	12
4	3	7	5	3
5	6	9	4	17
6	7	10	4	4
7	9	10	5	7
8	4	5	3	12
9	4	8	6	4
10	7	8	5	3
11	8	11	2	16
12	9	12	2	11
13	10	11	1	18
14	10	13	5	14
15	13	14	5	13
16	2	3	8	3
17	6	7	6	5
18	9	13	5	7
19	11	14	6	10