Deep reinforcement learning-based life-cycle management of deteriorating transportation systems

M. Saifullah*, C.P. Andriotis†, K.G. Papakonstantinou*, and S.M. Stoffels*

*Department of Civil & Environmental Engineering, The Pennsylvania State University, University Park, PA, USA
†Faculty of Architecture & the Built Environment, Delft University of Technology, Delft, The Netherlands

ABSTRACT: Efficient life-cycle bridge asset management delineates a planning optimization problem of paramount importance for the operational reliability of transportation infrastructure. It necessitates adept inspection and maintenance policies able to reduce risks and costs while incorporating long-term stochastic deterioration models, inference under uncertain structural health data, and various probabilistic and deterministic constraints. Structural integrity management policies for individual bridges, which are mere constituents of broader complex networks, cannot be devised in isolation of the policies of other system components, such as other bridges and pavement sections, and without considering system functions and traffic considerations. Such network effects render the optimization problem even harder to solve. Currently, age- or condition-based maintenance techniques, as well as risk-based or periodic inspection plans, have been used to address this class of challenging optimization problems. However, the efficacy of these techniques is often limited by optimality-, scalability-, and uncertainty-induced complexities. In practice, infrastructure management agencies often treat interconnected systems using disjoint plans for different component types, which in general do not ensure system-level optimality. To tackle the above, the optimization problem is herein cast within constrained Partially Observable Markov Decision Processes (POMDPs), which provide a comprehensive mathematical framework for stochastic sequential decision settings under observation/monitoring data uncertainty and limited resources. For the problem solution, the DDMAC algorithm (Deep Decentralized Multi-agent Actor-Critic) is successfully used, a deep reinforcement learning algorithm well-suited for management of large multi-state multi-component systems, as illustrated in an example application of an existing transportation network in Virginia, USA. The studied network comprises several bridge and pavement components exhibiting nonstationary deterioration, and various agency-imposed constraints, and traffic delay and risk factors are considered. Comparisons against conventional management policies showcase that the DDMAC solution significantly outperforms its counterparts.

1 INTRODUCTION

Determination of Inspection and Maintenance (I&M) policies for management of multi-asset infrastructure environments requires modeling and assessment of different stochastic deterioration effects, together with adept scheduling of action sequences, able to mitigate risks and serve multi-purpose life-cycle goals. Decision-making in such complex and uncertain system settings comes with major computational challenges, due to heterogeneity of different asset classes, large number of components resulting in intractable state and action spaces, noisy observations, limited availability of resources, and performance-based constraints. Advanced I&M frameworks and their respective computational approaches must, therefore, facilitate integrated consideration of the above characteristics, a quest that needs to reach beyond the limits of existing methodologies.

There is a large variety of optimization methods that propose solutions to the I&M planning problem, ranging from threshold-based formulations with reliability analysis principles e.g., in (Saydam & Frangopol, 2014; Bocchini & Frangopol, 2011), to decision tree analysis, e.g., in (Straub & Faber, 2005), to renewal theory, e.g., in (Grall, et al., 2002; Rackwitz, et al., 2005), to stochastic optimal control, e.g., in (Madanat, 1993 ; Ellis, et al., 1995; Papakonstantinou & Shinozuka, 2014; Papakonstantinou, et al., 2018). Many of these solutions, however, suffer from optimality-, scalability-, and uncertainty-induced complexities, and are often not easily extendable to environments with constraints (deterministic or stochastic). Moreover, despite the fact that the underlying decision problem is dynamic in its nature, many optimization techniques use static formulations, with the exception of stochastic optimal control approaches which incorporate dynamic programming principles (Bellman, 1957). Due to these computational challenges, many practical techniques are prone to generating widely sub-optimal solutions, especially in settings with large dimensions and long horizons.

To address the above, in this work, the decision-making problem is cast within the joint framework of Partially Observable Markov Decision Processes (POMDP) and multi-agent Deep Reinforcement Learning (DRL). The dynamic programming principles of POMDPs mitigate the curse of history and allow adaptive reasoning in the presence of noisy real-time data. Various studies have examined and demonstrated their efficacy in I&M planning, e.g., (Papakonstantinou & Shinozuka, 2014a,b;

Papakonstantinou, et al., 2016; Memarzadeh & Pozzi, 2015; Schöbi & Chatzi, 2016), among others.

Based on POMDPs, a Deep Centralized Multi-agent Actor-Critic (DCMAC) technique has been developed in (Andriotis & Papakonstantinou, 2019), which is part of the wider family of actor-critic methods (Wang, et al., 2016; Degris, et al., 2012). DCMAC makes use of the notion of belief-state MDPs, a key concept of point-based POMDP algorithms and, therefore, operates directly on the posterior probabilities of system states given previous actions and observations. Deep Decentralized Multi-agent Actor-Critic (DDMAC) (Andriotis & Papakonstantinou, 2021) proposes an architectural variant of DCMAC. In this architecture, each component is represented by a decentralized independent actor and their output is used to generate a centralized value function, which is then employed in relevant gradient calculations for updating both the actor and critic networks. As a further development, a new DDMAC version is proposed in (Saifullah, et al., in review), where a fully Centralized Training and Decentralized Execution (CTDE) concept is adopted (Lyu, et al., 2021), with decentralization at both the action and information levels, an efficient paradigm in cooperative multi-agent DRL. The architecture, termed as DDMAC-CTDE, reduces the parameter space of the policy even further by masking for every actor the other actors' input information.

In this study, a stochastically deteriorating transportation network with multiple asset classes is considered, i.e., pavement and bridge components, along with various deterministic and stochastic resource and condition constraints. The optimization is cast in a POMDP framework, utilizing a holistic modeling environment for the two classes of assets (Saifullah, et al., in review), based on their corresponding damage state indices, which characterize their condition states, and pertinent maintenance and inspection actions. The results are compared with Condition Based Maintenance (CBM) and a variant of Virginia's Department of Transportation (VDOT) I&M policy, outperforming both significantly.

2 BACKGROUND

2.1 *Partially observable Markov decision processes*

The POMDP framework is defined by 7 essential elements consisting of S, A, $\mathbf{P}$, Ω, $\mathbf{O}$, $\mathbf{C}$, and γ, where S, A and Ω are sets of states, actions, and possible observations, respectively, and $\mathbf{P}$ is the model of transitions, $\mathbf{O}$ is an observation model, $\mathbf{C}$ are the cost functions and γ is a discount factor. In POMDPs, the decision-maker (agent) starts at a state, s_t at a time step, t, takes an action a_t, receives a cost, 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, $p(o_{t+1}|s_{t+1}, a_t)$. Due to partial observability, the agent can only form a belief $\mathbf{b}_t$ about its state, where $\mathbf{b}_t$ is a probability distribution over S of all possible discrete states. A Bayesian update can be used to calculate the belief $\mathbf{b}_{t+1}$ (Papakonstantinou & Shinozuka, 2014a):

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

where probabilities $b(s_t)$, for all $s_t \in S$, form the belief vector $\mathbf{b}_t$ of length $|S|$, and the denominator of Eq. (1), $p(o_{t+1}|\mathbf{b}_t, a_t)$ is the standard normalizing constant. The goal for an agent is to choose actions at each time step that minimize its expected future discounted cumulative cost, defined by the value or action-value function (Papakonstantinou & Shinozuka, 2014a). The optimal value function for POMDPs is:

$$V^{\pi^*}(\mathbf{b}_t) = \min_{a_t \in A} \sum_{s_t \in S} b(s_t) c(s_t, a_t) + \gamma \sum_{o_{t+1} \in \Omega} p(o_{t+1} \mid \mathbf{b}_t, a_t) V^{\pi^*}(\mathbf{b}_{t+1}) \quad (2)$$

Despite existing mathematical convergence guarantees for POMDPs, traditional point-based POMDP solvers encounter scalability issues in very large state, observation, and actions spaces. Deep reinforcement learning allows us to alleviate this curse of dimensionality.

2.2 *Deep reinforcement learning and DDMAC-CTDE*

Reinforcement learning (RL) is a computational framework for evaluating and automating goal-directed learning and decision-making that is well-suited for solving MDP/POMDP problems as it is usually structured around them. RL algorithms combined with deep neural network parametrizations, give rise to DRL, which has shown capabilities of discovering powerful strategies in immense state spaces (Silver, et al., 2016; Mnih, et al., 2015).

The methods for solving RL problems can be majorly classified as value-based or policy-based learning. Value-based methods learn the state or state-action value function and act upon it by selecting the optimal action in each given state, e.g., Q-learning and DQN (Mnih, et al., 2015). In policy-based learning, policy $\pi : S \rightarrow P(A)$ is directly learned using a separate function approximator (usually a neural network). The policy gradient method is customarily used for learning policies in policy-based methods and the policy gradient, $\mathbf{g}_{\boldsymbol{\theta}^\pi}$, can be estimated in a multi-agent actor-critic setting as:

$$\mathbf{g}_{\boldsymbol{\theta}^\pi} = \mathbb{E}_{\mathbf{s}_t \sim \boldsymbol{\rho}, \mathbf{a}_t \sim \boldsymbol{\mu}} \left[w_t \left(\nabla_{\boldsymbol{\theta}^\pi} \log \pi(\mathbf{a}_t \mid \mathbf{s}_t, \boldsymbol{\theta}^\pi) \right) A^\pi(\mathbf{s}_t, \mathbf{a}_t) \right] \quad (3)$$

where, $\mathbf{s}_t = \{s_t^{(i)}\}^m$ state vector for m-component system, $\mathbf{a}_t = \{a_t^{(i)}\}^n$ is an action vector for n-agents (no. of agents and no. of components can be different), $\boldsymbol{\theta}^\pi$ is the policy network parameter vector, w_t is the importance sampling weight, $\boldsymbol{\mu}$ is a n-dimensional vector of agents' behavior policies, $\boldsymbol{\rho}$ is the m-dimensional state distribution under these policies, and $A^\pi(\mathbf{s}_t,\mathbf{a}_t)$ is the advantage function:

$$A^\pi\left(\mathbf{s}_t,\mathbf{a}_t \mid \boldsymbol{\theta}^V\right) \simeq c\left(\mathbf{s}_t,\mathbf{a}_t\right) + \gamma V^\pi\left(\mathbf{s}_t \mid \boldsymbol{\theta}^V\right) - V^\pi\left(\mathbf{s}_t \mid \boldsymbol{\theta}^V\right) \quad (4)$$

where, $\boldsymbol{\theta}^V$ are the weight parameters of the critic neural network. The mean squared error is considered as a loss function for the critic network and the relevant critic gradient can be accordingly derived.

Within this context, DDMAC, as proposed in (Andriotis & Papakonstantinou, 2021), provides an algorithm for I&M optimal planning well-suited for large multi-component systems. The framework also considers the presence of constraints through state augmentation and Lagrange multipliers. DDMAC uses a sparse parametrization of the actor-network without parameter sharing between agents (i.e., each component has its own actor-network). For even larger systems, DDMAC-CTDE formulation (Saifullah, et al., in review) is used herein, allowing for even sparser actor parametrizations. DDMAC-CTDE employs a fully decentralized logic along the lines of centralized training and decentralized execution, postulating that state accessibility for each actor network is restricted to its corresponding component. Component actions, as well as various possible sub-system actions, are assumed conditionally independent given their own state, thus the policy and its gradient are:

$$\pi\left(\mathbf{a}_t \mid \mathbf{s}_t\right) = \prod_{i=1}^{n} \pi_i\left(a_t^{(i)} \mid s_t^{(i)}\right) \quad (5)$$

$$\mathbf{g}_{\boldsymbol{\theta}^\pi} = \mathbb{E}_{\mathbf{s}_t\sim\boldsymbol{\rho},\mathbf{a}_t\sim\boldsymbol{\mu}}\left[w_t \left(\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}^\pi} \log \pi_i\left(a_t^{(i)} \mid s_t^{(i)}, \boldsymbol{\theta}^\pi\right)\right) A^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)\right] \quad (6)$$

This technically means that each control unit is seen as an autonomous agent that only utilizes component-state information to decide about its actions. For further details refer to (Saifullah, et al., in review).

3 ENVIRONMENT DESCRIPTION

3.1 *Component states*

The considered network is comprised of 85 pavement and 11 bridge components. Various indicators can describe the pavement condition, e.g., Pavement Condition Index (PCI), Critical Condition Index (CCI), International Roughness Index (IRI), and Load Related Distress Index (LDR), among many others. CCI and IRI are used in this work as they offer a joint quantification of condition, as per structural distresses and ride quality, respectively. A non-

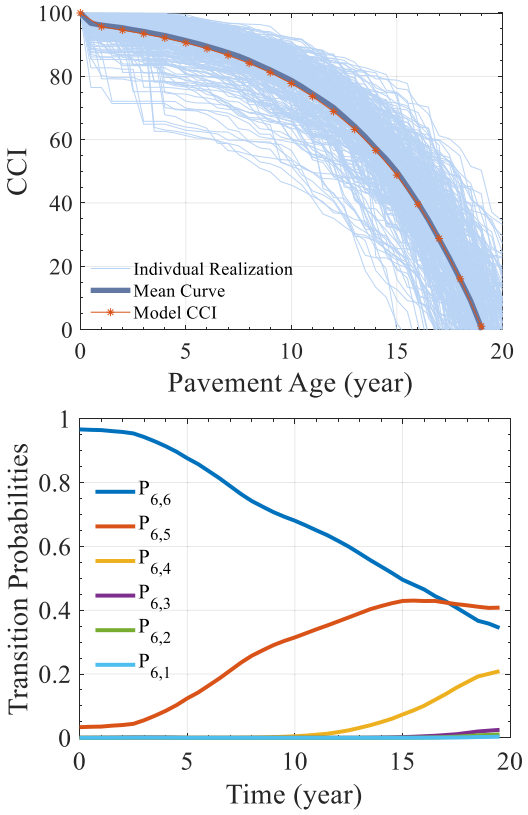

Figure 1: Fitted gamma model for CCI (top). Transition probabilities for heavy traffic, with starting state 6 (bottom).

stationary CCI model is used in this study, devised as a modified version based on a VDOT report (Katicha, et al., 2016). This model can incorporate various aspects, including different traffic levels. A gamma process is utilized, with its mean being in time equal to the modified mean CCI predictions and a relevant model variance (Katicha, et al., 2016). In Figure 1 (top), simulation results are indicatively shown for a heavy traffic level with 300 different realizations. The solid line represents the mean CCI and the red curve is the mean CCI gamma model prediction.

To determine the transition probabilities, the CCI values are discretized into 6 condition states, with 6 being the intact state. These discretized condition states are largely adapted from the prescribed VDOT maintenance guidelines (VDOT, 2016), and the detailed description is reported in (Saifullah, et al., in review). 10^6 sequences are generated in total to obtain the transition probabilities for a given traffic level. Figure 1 (b) indicatively shows a few computed transition probabilities for heavy traffic.

The observation uncertainty for CCI is appropriately modeled by the likelihood functions $p(o_t|s_t)$, which quantify the probability of receiving an observation o_t at time t given a state s_t. A normal distribution is considered in this work as a likelihood function, with mean the actual CCI value and 3 different error variances, i.e., ∞, 72, and 18, corresponding to no-inspection, low- and high-fidelity inspections, respectively. Similarly, the IRI (in m/km)

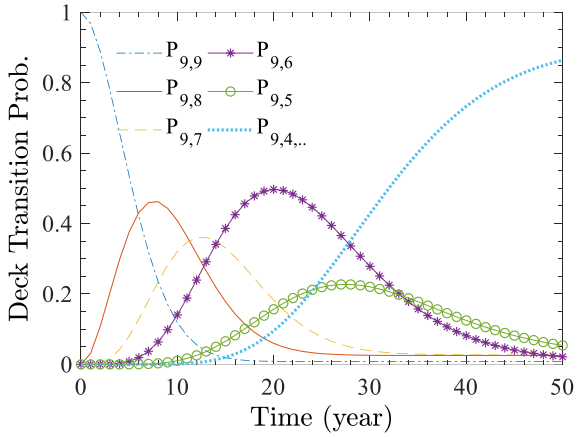

Figure 2: Transition probabilities in time, starting from state 9.

can be discretized into 5 states, with 5 being the intact state, as in (FHWA, 1999). Unlike CCI, the IRI transition model is stationary. In this case too, three different inspection activities are assumed, and the measurement errors associated with the respective inspection technologies are considered to be normally distributed with zero mean and standard deviations of ∞, 0.32, and 0.08 m/km, respectively. All resulting CCI and IRI observation probabilities are reported in (Saifullah, et al., in review).

For the objectives of this study, only the decks of bridges are considered, as they are directly influenced by traffic. To determine the serviceability of decks, 9 states are considered, with state 9 being the undamaged state, as adopted in (FHWA, 1999) and other DOTs. Condition 4 now denotes an irreversible damage state, and is thus regarded as a terminal state, as also suggested by (Manafpour, et al., 2018). The nonstationary transition probabilities are based on 30 years of in-service performance data for more than 22,000 bridges in Pennsylvania, as analyzed in (Manafpour, et al., 2018) and illustrated in Figure 2. Apart from these 6 nonstationary transitions, stationary failure probabilities are also considered, where a bridge is assumed to have a failure probability of P_f = 0.001 if it is in states 8 and 9, and P_f = 0.005 if it is in states 7, 6, 5. P_f finally reaches 0.01 if the bridge state is 4.

3.2 *Action description*

There are various guidelines for pavement maintenance from different agencies. According to (VDOT, 2016), four different maintenance actions are recommended, i.e., *Do Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*. *Minor Repair* (crack filling, moderate patching, etc.) can improve the CCI and IRI states but does not affect the rate of deterioration, *Major Repair* can improve condition states and reduce the deterioration rate by 5 years, and *Reconstruction* resets the pavement to an intact condition. A detailed description of these actions and their costs can be found in (VDOT, 2016). Maintenance actions taken at any given time will simultaneously improve both CCI and IRI indices. The maintenance action transition probabilities for CCI and IRI, their duration, and their costs are reported in (Saifullah, et al., in review).

Similar to pavements, four maintenance actions are considered for maintaining the bridge decks, i.e., Do *Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*, however, the involved performed actions are different. It is again assumed that the *Minor Repair* action does not change the rate of deterioration of the deck but it can improve the condition state of the structure. Similarly, *Major Repair* can improve both, and *Reconstruction* can reset the deck to a newly built one. The transition probabilities, action durations, and their costs are described in (Saifullah, et al., in review). Maintenance action-induced delays that can be translated to costs are considered as in (Vadakpat, et al., 2000).

There is a variety of destructive and nondestructive inspection techniques that are used for bridge decks, such as visual inspections, acoustic sensing, infrared/thermal imaging, ground penetrating radar, coring and chipping, and half-cell potential tests, among many others. Towards generality, inspection techniques are herein characterized as uninformative, low-fidelity, and high-fidelity inspection techniques, respectively. The observation probabilities for the corresponding inspections can be seen in (Saifullah, et al., in review).

3.3 *Transportation network*

As a reference example, the Hampton Roads transportation network in Virginia, USA, is considered. The original topology and average daily traffic data of the network are used along with 11 main bridges. Each bridge is bidirectional, with the same number of lanes as in the original network, illustrated in Figure 3. The different deck types I-III are categorized based on their relevant sizes. Type I bridges have length more than 5 km, type II have lengths between 1.2-5 km, and type III are the smallest having a length less than 1.2 km.

Similarly, the network has various pavement components categorized as type I-III. Type I pavements are interstate highways, with bidirectional traffic having four lanes in each direction, thus, constituting the class of highest vehicular miles. Type II are primary highways with a bidirectional medium level of traffic, having two lanes in each direction. Lastly, type III are secondary highways with low-level bidirectional traffic and one lane in each direction. The deterioration rate of pavements is selected based on these classes, as high-volume roads have a higher rate than low-volume ones. These rates are taken from (Saifullah, et al., in review).

3.4 *Network level risks and constraints*

Risk is defined as an expected cumulative discount-

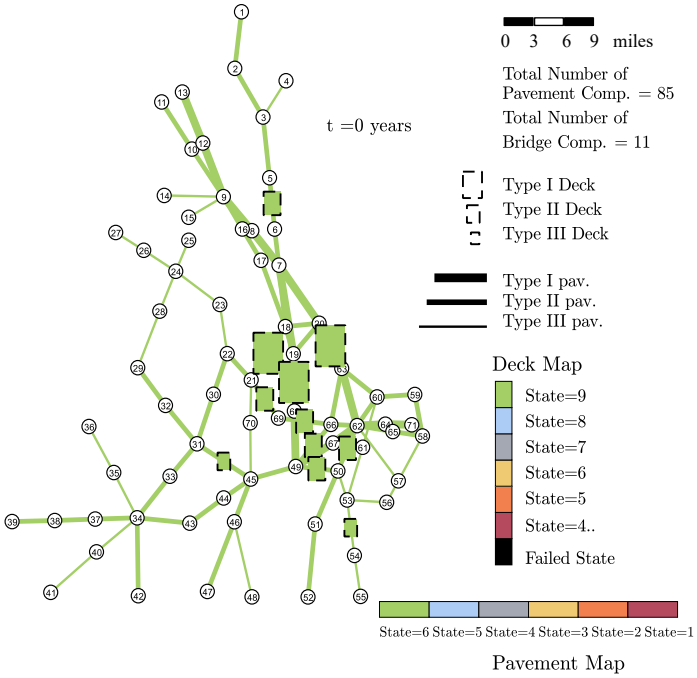

Figure 3: Hampton Roads transportation network model.

ed failure state cost over the life cycle, as in (Andriotis & Papakonstantinou, 2021). The risk cost consists of two parts: (1) accruable cost, which is taken as two times the rebuilding cost of the bridge, and (2) instantaneous cost, which is considered here as ten times the rebuilding cost of the bridge. The total risk is estimated using (i) the risk of individual bridge failures (for all network bridges), and (ii) the system-level risk, defined based on the connecting bridges over James River and York River as in (Saifullah, et al., in review). The system risk has 3 failure modes, i.e., (A) the bridge over York River fails, (B) the 3 bridges over James River fail, and (C) modes A and B occur simultaneously.

There are various constraints that are considered, based on the condition states of pavements and bridges, imposed by the FHWA and VDOT agencies. For National Highway System (NHS) bridges, no more than 10% of the total bridge deck area should be deficient (i.e., condition rating ≤ 4), and for NHS pavements, no more than 10% of lane-miles should be in poor condition (i.e., CCI<60 and IRI>2.2 m/km). Based on VDOT targets, no more than 18% of interstate and primary pavements and 35% of secondary pavements should be classified as deficient (i.e., CCI<60). Regarding serviceability, no more than 15% of interstate and primary roadways should be classified as deficient in terms of ride quality (i.e., IRI>2.2 m/km). VDOT also aims to achieve 'no' CCI lower than 35 for the interstate system (VDOT, 2019). It is essential here to mention that the above constraints are satisfied in an expectation sense (i.e., soft constraints). Therefore, the last constraint is modified here from 0 to 2%.

Finally, a budget constraint is imposed due to limited available resources. A five-year budget of $1.3 billion is allocated to Hampton Roads districts for FY2021-2026 (Nichols, 2021). This budget needs to be strictly satisfied (hard constraint) and is implemented as an augmented state of the network (Andriotis & Papakonstantinou, 2021).

4 RESULTS

This study considers a 96-component network with a total number of ~$7\text{x}10^{134}$ possible system states at any given time instant. 10 actions per component are considered which makes the total number of available actions equal to 10^{96} for the entire network at each time step. The network components start from intact states, with an episode length of 20 years, and a discount factor $\gamma = 0.97$. The DDMAC-CTDE training is performed for $1.3\text{x}10^{6}$ episodes. Training details can be found in (Saifullah, et al., in review).

To assess the DDMAC-CTDE solutions, we formulate and evaluate 2 baselines, i.e., (i) a condition-based maintenance (CBM) policy and (ii) a policy baseline following VDOT guidelines. The CBM policy is heuristically optimized to find the relevant thresholds based on the condition of each component type, i.e., bridge, interstate, primary, and secondary pavements. The policy involves the full suite of 10 actions at every even time step. However, at every odd year, action 6 is taken for every component, i.e., do-nothing and high-fidelity-inspection, as also shown in Figure 5. The detailed CBM algorithm is presented in (Saifullah, et al., in review). The VDOT

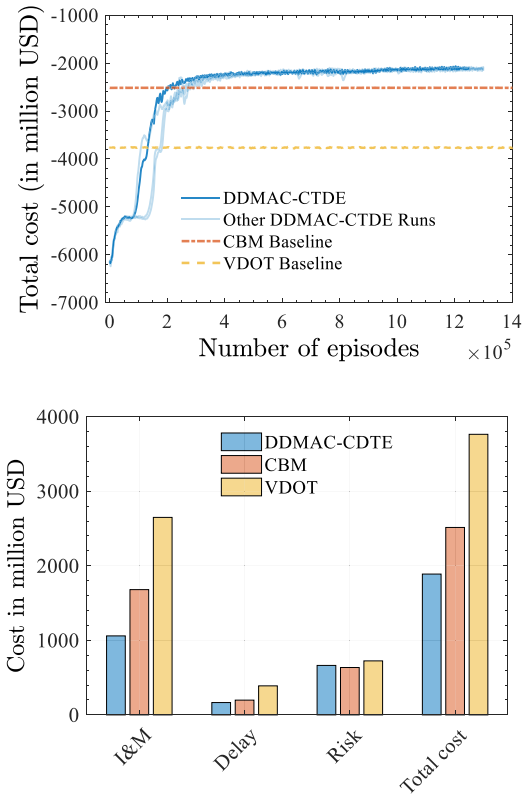

Figure 4: Total life cycle costs comparison of DDMAC-CTDE solution with CBM and VDOT policy baselines (top). Comparison of the total cost and its constituents with CBM and VDOT policy baselines (bottom).

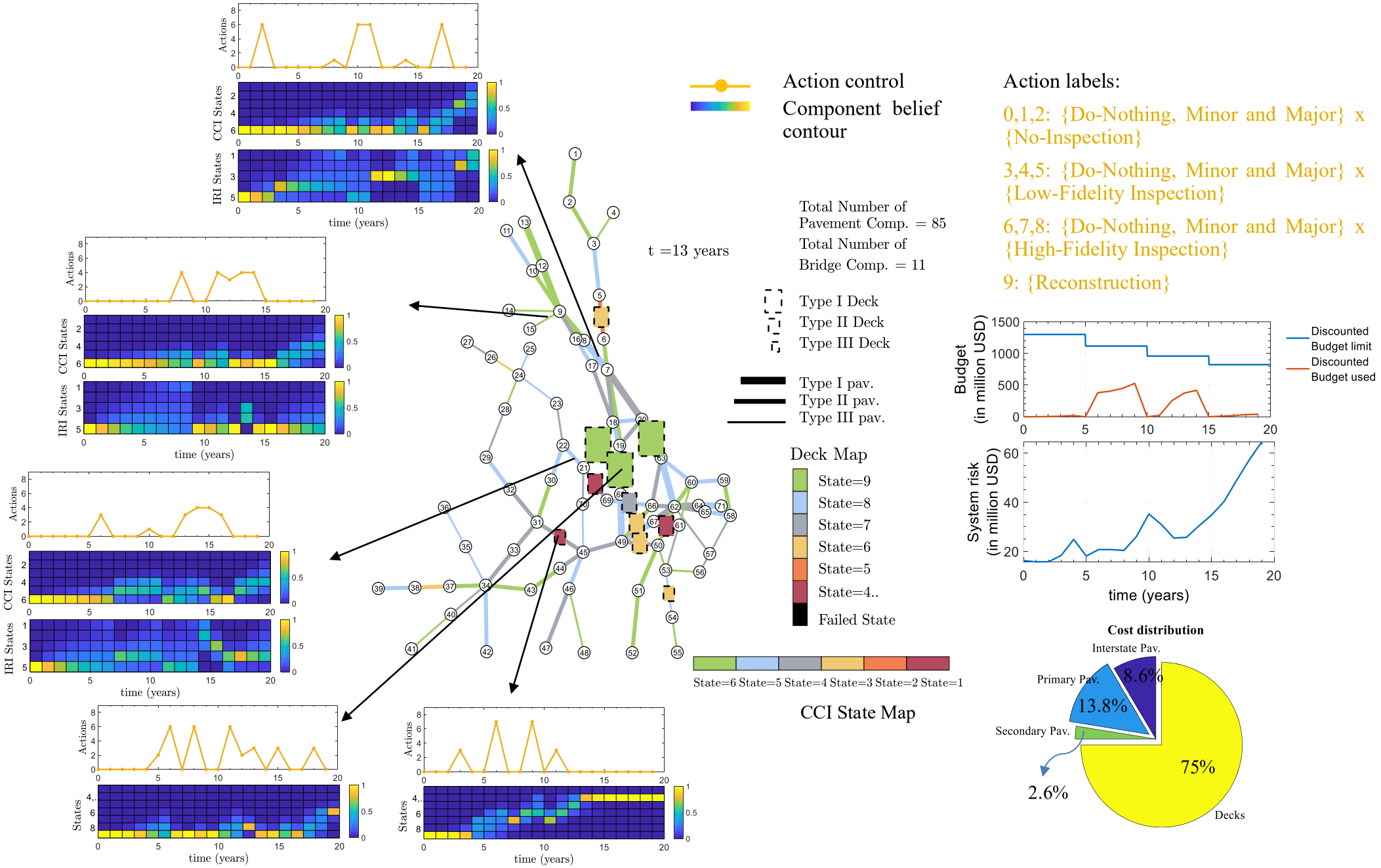

Figure 5: Life-cycle realization of the learned DDMAC-CTDE policy for the transportation network in case of intact initial conditions.

policy baseline is approximated from (VDOT, 2016) for pavement components. The original VDOT policy uses CCI and other distress metrics for action selection, but here only CCI is used. For bridge decks, the same criterion is used as for interstate components due to their similar importance.

The expected total costs during training are compared in Figure 4 (top). Figure 4 (bottom) also presents a histogram comparing the total costs with their constituents based on Monte-Carlo simulations. It can be observed that our DDMAC-CTDE solution surpasses both baselines during training and simulation by a significant margin, being 27% cheaper than the CBM policy and 48% cheaper than the VDOT policy, as given in Table 1. Table 1 also compares the average performance over 10^4 simulations in terms of poor condition states, as per the 6 different constraints discussed in Section 3.4. The performance constraints are in the rows of the table, and I, P, and S Hwy are the abbreviations of interstate, primary, and secondary highways, respectively.

To better understand how policies change over time, a detailed policy realization for some representative components is shown in Figure 5. The figure illustrates actions generated by one of the instances of the optimum policy and the evolution of component belief states is shown with contours. Additionally, Figure 5 displays the discounted budget usage over time and the 5-year budget discounted for every cycle. The budget is a hard constraint that the agents are not allowed to exceed, a requirement that is satisfied by the obtained solution. The evolution of the total risk cost associated with individual bridges and the 3 modes of system risk is also presented. Moreover, the cost distribution among different types of pavements and bridges is shown in a pie chart.

Plots with control actions represent the actions taken over time. The maintenance actions, taken at every time step, update the current belief of the system, as manifested in the next time step. The evolution of contour plots in the case of pavements shows current beliefs for both CCI and IRI states, and the current belief states at each step for two bridge decks are also shown. For example, the agent is shown to take action 7 at $t = 6$ years for a type III bridge, and then the updated belief is shown at $t = 7$ years, incorporating both I&M actions.

Table 1: Comparison of different solution schemes in terms of total cost and performance with respect to average condition states of different pavement and bridge components.

Objective & Constraints	DDMAC -CTDE	CBM policy	VDOT policy
Total budget used (billion USD)	1.86	2.54	3.62
CCI<60 and IRI>2.2m/km for I-Hwy (%)	2.0	2.9	0.0
CCI<35 of I-Hwy (%)	1.9	0.5	0.0
CCI<60 for I and P-Hwy (%)	7.3	4.7	0.1
IRI>2.2 m/km for I and P-Hwy (%)	15.0	14.0	12.0
CCI<60 for S-Hwy (%)	10.3	4.3	0.9
Bridges with condition rating ≤4 (%)	9.2	2.1	8.7

As seen in Figure 5, control actions are compatible with belief states. For example, the agents initially choose Do-Nothing actions since the belief states for both pavements and bridges initiate in the intact condition. As the conditions gradually worsen, more interventions are considered. Similarly, at the horizon end, the Do-Nothing action is optimal for pavements, as pavements do not contribute to disconnection risks, while any action without inspection can be optimal for bridges. It has also been observed that the agents maintain and inspect type I bridges more systematically. This is because type I bridges have their individual failure risk as well as mode B and mode C system failure risks associated with them.

From the pie chart, shown in Figure 5, it is observed that cost distribution is heavily skewed (as much as 75%) towards the bridge components, due to their high maintenance cost, associated risk cost, and lower traffic delay cost. Among pavements, primary highways have the largest contribution as they represent the most components in the network (47 in total). Figure 5 also shows the evolution of the system risk with time. As expected, the risk is minimal in the beginning and it increases with time, with downward jumps mainly due to the maintenance actions taken for bridges, especially of type I.

5 CONCLUSIONS

In this work, the I&M problem of a large deteriorating bridge-pavement network with 96-components is formulated within a POMDP-DRL framework, including risks and other condition and budget related constraints. Pavement states are defined by CCI and IRI metrics and bridge states are defined by deck condition ratings. Due to immensely large state and action spaces, the problem is solved with a newly and originally developed DRL algorithmic approach named Deep Decentralized Multi-agent Actor Critic with Centralized Training and Decentralized Execution (DDMAC-CTDE) which uses sparse parametrizations and local component state information for actor networks to obtain near optimal solutions. The optimal life-cycle policies are compared against a Condition-Based Maintenance (CBM) policy and an adapted VDOT policy. The DDMAC-CTDE solution is shown to surpass the two baselines by 27% and 48%, respectively, satisfying all the considered constraints.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the U.S. National Science Foundation under CAREER Grant No. 1751941 and LEAP-HI Grant No. 2053620, and the Center for Integrated Asset Management for Multimodal Transportation Infrastructure Systems, 2018 U.S. DOT Region 3 University Center. Dr. Andriotis would further like to acknowledge the support of the TU Delft AI Labs program.

6 REFERENCES

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.

Bellman, R. E., 1957. *Dynamic programming.* Mineola, NY: 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.

Degris, T., White, M. & Sutton, R. S., 2012. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839.*

Ellis, H., Jiang, M. & Corotis, R. B., 1995. Inspection, maintenance, and repair with partial observability. *Journal of Infrastructure Systems,* 1(2), pp. 92-99.

FHWA, 1999. *Status of the Nation's Highways, Bridges and Transit: Conditions and Performance, Report to Congress,* Washington D.C.: Federal Highway Administration.

Grall, A., Bérenguer, C. & Dieulle, L., 2002. A condition-based maintenance policy for stochastically deteriorating systems. *Reliability Engineering & System Safety,* 76(2), pp. 167-180.

Katicha, S. W. et al., 2016. VDOT: Development of enhanced pavement deterioration curves.

Lyu, X., Xiao, Y., Daley, B. & Amato, C., 2021. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.04402.*

Madanat, S., 1993 . Optimal infrastructure management decisions under uncertainty. *Transportation Research Part C: Emerging Technologies,* 1(1), pp. 77-88.

Manafpour, A. et al., 2018. Stochastic analysis and time-based modeling of concrete bridge deck deterioration. *Journal of Bridge Engineering,* 23(9), p. 04018066.

Memarzadeh, M. & Pozzi, M., 2015. Integrated inspection scheduling and maintenance planning for infrastructure systems. *Computer-Aided Civil and Infrastructure Engineering,* 31(6), pp. 403-415.

Mnih, V. et al., 2015. Human-level control through deep reinforcement learning. *Nature,* 518(7540), pp. 529--533.

Nichols, K. M., 2021. *The state of transportation in Hampton Roads 2020,* Chesapeake: Hampton Roads, TPO.

Papakonstantinou, K. G., Andriotis, C. P. & Shinozuka, M., 2016. *POMDP solutions for monitored structures.* Pittsburgh, PA, IFIP WG-7.5 Conference on Reliability and Optimization of Structural Systems.

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. G. & 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.

Rackwitz, R., Lentz, A. & Faber, M. H., 2005. Socio-economically sustainable civil engineering infrastructures by optimization. *Structural Safety,* 27(3), pp. 187-229.

Saifullah, M., Papakonstantinou, K. G., Andriotis, C. P. & Stoffels, S. M.. Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management. *Manuscript under review.*

Saydam, D. & Frangopol, D., 2014. Risk-based maintenance optimization of deteriorating bridges. *Journal of Structural Engineering,* 141(4), p. 04014120.

Schöbi, R. & Chatzi, E. N., 2016. Maintenance planning using continuous-state partially observable Markov decision processes and non-linear action models. *Structure and Infrastructure Engineering,* 12(8), pp. 977-994.

Silver, D. et al., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature,* 529(7587), pp. 484--489.

Straub, D. & Faber, M. H., 2005. Risk based inspection planning for structural systems. *Structural Safety,* 27(4), pp. 335-355.

Vadakpat, G., Stoffels, S. M. & Dixon, K., 2000. Road User Cost Models for Network-Level Pavement Management. *Transportation Research Record ,* 1699(1), pp. 49-57.

VDOT, 2016. *Supporting document for the development and enhancement of the pavement maintenance decision matrices used in the needs-based analysis,* Richmond: Virginia Department of Transportation.

VDOT, 2019. *Maintenance and operations comprehensive review,* Richmond: Virginia Department of Transportation.

Wang, Z. et al., 2016. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224.*