Learning to Send Reinforcements: Coordinating Multi-Agent Dynamic Police Patrol Dispatching and Rescheduling via Reinforcement Learning

Waldy Joe, **Hoong Chuin Lau**[*]
School of Computing and Information Systems, Singapore Management University
waldy.joe.2018@phdcs.smu.edu.sg, hclau@smu.edu.sg

Abstract

We address the problem of coordinating multiple agents in a dynamic police patrol scheduling via a Reinforcement Learning (RL) approach. Our approach utilizes Multi-Agent Value Function Approximation (MAVFA) with a rescheduling heuristic to learn dispatching and rescheduling policies jointly. Often, police operations are divided into multiple sectors for more effective and efficient operations. In a dynamic setting, incidents occur throughout the day across different sectors, disrupting initially-planned patrol schedules. To maximize policing effectiveness, police agents from different sectors cooperate by sending *reinforcements* to support one another in their incident response and even routine patrol. This poses an interesting research challenge on how to make such complex decision of dispatching and rescheduling involving multiple agents in a coordinated fashion within an operationally reasonable time. Unlike existing Multi-Agent RL (MARL) approaches which solve similar problems by either decomposing the problem or action into multiple components, our approach learns the dispatching and rescheduling policies jointly without any decomposition step. In addition, instead of directly searching over the joint action space, we incorporate an iterative best response procedure as a decentralized optimization heuristic and an explicit coordination mechanism for a scalable and coordinated decision-making. We evaluate our approach against the commonly adopted two-stage approach and conduct a series of ablation studies to ascertain the effectiveness of our proposed learning and coordination mechanisms.

1 Introduction

This research work is motivated by a real-world problem involving coordination of patrol operations across multiple police sectors. Police patrol serves the following two key functions: to project presence (*proactive* patrol) and to respond to incidents in a timely manner (*reactive* patrol). Often, police operations are divided into multiple sectors for more effective and efficient operations. In real-world operations, incidents occur throughout the day across different sectors, disrupting initially-planned patrol schedules and necessitating re-planning decisions. To maximize policing effectiveness, police agents from different sectors cooperate by sending *reinforcements* to support one another in their incident response and even routine patrol.

We term this problem as the Multi-Agent Dynamic Police Patrol Dispatching and Rescheduling Problem (MADPRP). This is the multi-agent variant of a Dynamic Police Patrol Dispatching and Rescheduling Problem (DPRP) introduced in [Joe *et al.*, 2022]. Both problems can be modelled as a sequential decision problem [Powell, 2019] where the arrival of one or more dynamic events triggers a decision-making process which happens sequentially in response to such events.

In MADPRP, agents need to make complex decision consisting of *event-handling* (which patrol team needs to be dispatched to respond to the incident) and *re-planning* actions across space and time (rescheduling of existing patrol schedules). This problem is particularly challenging as the complex action includes both rerouting the sequence of locations to patrol (spatial) and rescheduling the time spent at each location (temporal). Such decision-making process is made even more challenging in the presence of multiple agents. Here, we define an agent as a higher-order decision-making entity that is capable of executing complex action (police sector) and usually consists of multiple sub-agents (patrol teams). This is unlike the classical examples of multi-agent settings which come in the form of multiple vehicles or machines. In this paper, we use the terms sector and agent interchangeably.

Learning-based approaches particularly Reinforcement Learning (RL) is popular in solving sequential decision problem as RL facilitates offline learning of policies and values which are computed beforehand and can be quickly executed during run-time for instantaneous decision-making. However, current Multi-Agent RL (MARL) approaches deal with problems with simple actions (either discrete or continuous). Thus, to solve problems with complex action similar to MADPRP, current MARL approaches either decompose the problem into two stages (see [Chen *et al.*, 2021]) and/or discretize the actions (see [Chen *et al.*, 2019]).

In this paper, we propose a new cooperative MARL approach that combines Multi-Agent Value Function Approxi-

[*] Corresponding Author

mation (MAVFA) with a planning heuristic to solve MADPRP directly without any decomposition step. In our approach, the learned value function is utilized by the heuristic to search for better rescheduling decision during execution. To improve scalability and to induce coordination among agents, we propose an iterative best response procedure as a decentralized optimization heuristic and an explicit coordination mechanism in this proposed approach.

This paper makes the following contributions:

- We define MADPRP as a multi-agent sequential decision problem with complex action and model it as a route-based Markov Decision Process (MDP).
- We propose a solution approach that combines MAVFA with a rescheduling heuristic to solve MADPRP that will learn policies for making dispatching and rescheduling jointly. Our proposed approach incorporates iterative best response procedure to serve as decentralized optimization heuristic and as an explicit coordination mechanism for a scalable, coordinated decision-making amongst multiple agents.
- We show experimentally that our approach outperforms the commonly used two-stage approach on a real-world problem setting and through a series of ablation studies, ascertain the effectiveness of our proposed learning and coordination mechanisms.

2 Related Works

Police Patrol Problem. Dynamic police patrol routing and scheduling problem shares many similarities with Dynamic VRP (DVRP) and existing solution approaches to DVRP could potentially be used to solve this problem [Dewinter *et al.*, 2020]. However, no prior work directly addresses dynamic police patrol problem as DVRP except for [Joe *et al.*, 2022]. Moreover, existing approaches to solve multi-agent version of DVRP are very specific to transportation or logistics problem scenario (see [Wang and Kopfer, 2013; Los *et al.*, 2020]).

Most existing works in the literature solve police patrol routing and scheduling problem or other similar problems involving emergency response in a static manner, without considering the disruption to existing plans due to occurrences of dynamic incidents [Mukhopadhyay *et al.*, 2016; Pettet *et al.*, 2022; Wang *et al.*, 2019; Wang *et al.*, 2022]. Only [Joe *et al.*, 2022; Rumi *et al.*, 2020] consider the impact of dynamic incidents to existing patrol schedules. On the other hand, existing works on multi-agent patrolling such as [Santana *et al.*, 2004; Tkach and Amador, 2021] define agent as lower-order entity such as a robot or a vehicle unlike the definition used in this paper.

Cooperative MARL. The research in cooperative MARL has been to address the following two main challenges namely partial observability and scalability. The concept of Centralized Training Decentralized Execution (CTDE) was introduced to address the partial observability and since have been leveraged by many popular MARL algorithms such as COMA [Foerster *et al.*, 2018] and MADDPG [Lowe *et al.*, 2017]. To further address scalability, many works such as Value Decomposition Network (VDN) [Sunehag *et al.*, 2018], QMIX [Rashid *et al.*, 2018] and QTRAN [Son *et al.*, 2019] propose value function factorization based on Individual Global Max (IGM) assumption on top of CTDE to learn decentralized policies.

However, the current cooperative MARL approaches fall short in addressing problems with complex action directly. For instance, VDN, QMIX and COMA only solve problems with discrete actions while MADDPG addresses problems with continuous actions. Current MARL approaches solve problems with complex action by either decomposing the action into two stages and learn the policy in one of the stages, defining the actions to be either discrete or continuous, or combining both approaches.

Two-Stage Approach. Chen *et al.* [2021] propose DeepFreight, a model-free DRL-based approach to solve multi-transfer freight delivery. To solve the problem, the authors decompose the problem into two stages: truck-dispatch and request-matching and leverage on QMIX to learn the dispatch policy while implement a separate matching algorithm for the second stage. Similarly, Chen *et al.* [2022] solve a same-day delivery problem with vehicles and drones by decomposing the problem into two stages: learning assignment policy via DQN and rerouting via heuristic. Meanwhile, Chen *et al.* [2019] decomposes a dynamic courier dispatch problem into two stages namely dispatch and routing stage. The authors propose an MARL approach to learn the dispatch policy and define a discrete action space consisting of a cartesian product of the next grid to visit and the corresponding period of stay in the grid. Ma *et al.* [2021] propose a hierarchical approach to solve dynamic pickup and delivery problems by introducing two levels of agent. The first agent learns a policy to decide which orders to be released while the second agent learns a policy to choose a local search operator to reroute the vehicles. Both stages involve discrete actions.

Communication. To coordinate amongst agents, various works focus on the aspect of learning to communicate in cooperative setting (see [Jiang and Lu, 2018; Das *et al.*, 2019; Jiang *et al.*, 2020]). However, communication alone does not guarantee coordination especially when agents act simultaneously [Ruan *et al.*, 2022]. Acting simultaneously in the context of MADPRP is not ideal because it may cause uncoordinated actions resulting in poor event-handling decision. For e.g. there may be scenarios where no agent or more than one agent respond(s) to an incident and this would result in an unattended incident or delays as agents may be waiting for incidents that have already been responded by other agent.

Arising from the above-mentioned gaps in current works, we propose a cooperative MARL approach to address complex action directly with an explicit coordination mechanism in place of communication network (which is implicit in nature). Currently, there exist works that solve single-agent sequential decision problem with complex action directly [Joe and Lau, 2020; Joe *et al.*, 2022]. Inspired by these works, this paper extends the proposed idea to address multi-agent settings.

Notation	*Description*
I	A set of agents / sectors, $I \in \{1, 2, ..., \|I\|\}$
J	A set of patrol areas, $J \in \{1, 2, ..., \|J\|\}$
I_i	A set of patrol teams in sector i
J_i	A set of patrol areas in sector i where $\sum_{i \in I} \|J_i\| = \|J\|$
T	A set of time periods, $T \in \{1, 2, ..., \|T\|\}$
k	Decision epoch
t_k	Time period where k occurs
$\delta_i(k)$	A schedule of agent i at k
$\delta(k)$	A joint schedule of all agents at k where $\delta(k) = (\delta_i(k))_{i \in I}$
$\delta_{-i}(k)$	A joint schedule of all agents except for agent i at k
$(\delta_i(k), \delta_{-i}(k))$	A joint schedule where agent i follows $\delta_i(k)$ while the rest follows $\delta_{-i}(k)$
$\delta^x(k)$	A joint schedule of all agents after executing action x at k
$u^i(\delta(k))$	Payoff/utility of agent i when all agents follows $\delta(k)$
$B_i(\delta_{-i}(k))$	Best response of agent i when all other agents follow $\delta_{-i}(k)$
$x_{k,i}$	Action by agent i at k
$x_{k,-i}$	Joint action by all agents except agent i at k
x_k	Joint action by all agents at k
S_k	Joint pre-decision state at k
$S_{k,i}$	Local pre-decision state of agent i at k
S_k^x	Joint post-decision state at k
$S_{k,i}^{x,i}$	Local post-decision state at k

Table 1: Set of key notations used in this paper.

3 Problem and Model Formulation

Table 1 provides key notations and the corresponding descriptions used in this paper.

3.1 Problem Description

In MADPRP, there are $|I|$ police sectors in charge of patrolling $|J|$ patrol areas. We define each police sector as an agent; a higher-order decision-making entity which are capable of executing complex decision. Each police sector i consists of $|I_i|$ patrol teams that patrol $|J_i|$ patrol areas within its sector and each patrol shift has a duration of $|T|$ time periods. At the start of the shift, each agent is assigned to an initial patrol schedule. Throughout the shift, incidents occur dynamically and a patrol team from a certain sector is dispatched to respond to the incident which results in the need to reschedule its own and/or even the schedules of all other agents. Coordination amongst the agents is crucial as patrol teams can cross over to other sectors to respond to an incident and perform routine patrol so as to ensure that incidents are responded within target time and all patrol areas across all sectors are sufficiently patrolled.

Schedule. Each patrol schedule includes the sequence of patrol areas to visit (routes) and when and how long to patrol each areas (schedule). It is similar to a university timetable with an additional key constraint whereby in between two different patrol areas, there must be sufficient time periods to cater for travel time. At the start of the shift, each agent is assigned to an initial patrol schedule, $\delta_i(0)$. In this problem, we assume that the initial joint schedule is available and computed independently.

Incident. A dynamic incident, ω_k occurs at decision epoch k and is described as the following tuple: $\langle \omega_k^i, \omega_k^j, \omega_k^t, \omega_k^s \rangle$ where ω_k^i refers to the sector in which the incident takes place, $\omega_k^j \in J_{\omega_k^i}$ refers to the location of the incident, $\omega_k^t \in T$ refers to the time period when the incident occurs and ω_k^s refers to the number of time periods needed to resolve the incident.

Patrol Presence. A study by Doyle *et al.* [2016] has shown that patrol presence invokes feelings of safety in people. We define patrol presence as a function of the number of effective time periods each patrol area is being patrolled in a shift. Each patrol area j needs to be patrolled for at least Q_j time periods in a given shift. We propose a presence utility function of a patrol area j, $U_p(j)$ where the utility factor of any additional patrol time periods beyond the minimum patrol time requirement decreases exponentially with a parameter β_p (see Eq. 1). This is to simulate that any additional patrol time periods beyond the minimum requirement are less effective in projecting presence. σ_j refers to the total patrol time periods in patrol area j by all teams across the patrol sectors in a given joint patrol schedule, $\delta(k)$.

$$U_p(j) = \min(\sigma_j, Q_j) + 1_A \times \sum_{i=1}^{\sigma_j - Q_j} i \times e^{-\beta_p i} \qquad (1)$$

$$\text{where } 1_A = \begin{cases} 0, \sigma_j - Q_j \leq 0 \\ 1, \sigma_j - Q_j > 0 \end{cases}$$

We define a fitness function, $f_p(\delta(k))$ to quantify the goodness of a given schedule $\delta(k)$ in terms of its ability to project presence. We represent $f_p(\delta(k))$ as a ratio of total effective patrol time of to the total time in a shift across all agents and their patrol teams (see Eq. 2). Thus, a schedule is deemed to have good patrol presence if the patrol teams spend most of the time patrolling rather than travelling between patrol areas and each patrol area is being patrolled sufficiently.

$$f_p(\delta) = \frac{\sum_{j \in J} U_p(j)}{|T| \times |I|} \qquad (2)$$

Response Time. The response time to an incident at decision epoch k, τ_k is computed as the time taken by the assigned patrol team, x_k^m to act upon the dispatch call from the point where incident occurs $(x_k^t - \omega_k^t)$ plus the travel time from its current location to the incident location. A successful incident response happens when $\tau_k \leq \tau_{target}$. We assume that any dispatch call must be acted upon within τ_{max}.

$$\tau_k = (x_k^t - \omega_k^t) + d(j_{t'}^{x_k^m}, \omega_k^j) \qquad (3)$$

$$\text{where } x_k^t - \omega_k^t \leq \tau_{max}, t' = x_k^t$$

Problem Objective. The objective is to make dispatching and rescheduling decisions at every epoch that maximize the number of successful incident responses while minimizing the reduction in patrol presence within and across all sectors.

3.2 Route-Based MDP Formulation

We use a route-based MDP modelling framework introduced by Ulmer *et al.* [2020] because, unlike in the conventional MDP, the solution to MAPDRP is an updated joint schedule and not partial schedule or planned action in the next time period. Thus, the model needs to capture the complex action and its impact in both the state and action spaces.

State. A state consists of two parts, pre-decision state S_k and post-decision state S_k^x. S_k captures the necessary information required to make dispatching and rescheduling decisions. Each state can be further categorized as local and global states. Local state refers to information unique to an agent while global state refers to shared information across the agents which may be useful to induce coordination. Each local post-decision state, $S_{k,i}$ is represented as the following tuple:

$$\langle \delta_i(k), \frac{\sigma_i(k)}{Q_j}, f_{util}(\delta_i(k)) \rangle \tag{4}$$

where $\frac{\sigma_i(k)}{Q_j}$ is the ratio of total patrol time of each patrol area covered in $\delta_i(k)$ over its minimum patrol requirement and $f_{util}(\delta_i(k))$ is a function that computes the ratio of the total patrol time over the total shift time of all patrol teams. The shared global state includes the current time, t_k and $(\frac{\sigma_j(k)}{Q_j})_{j \in J}$ which represents the ratios of total patrol time of each patrol area in all sectors over its minimum patrol requirement when all agents follow a joint schedule, $\delta(k)$. Meanwhile, the post-decision state S_k^x captures the changes to the state upon executing a decision.

Action/Decision. x_k is the action of assigning a patrol sector/agent to dispatch one of their patrol teams to an incident and updating the joint schedule of all agents at decision epoch k. x_k is represented as the following tuple: $\langle x_k^i, x_k^m, x_k^t, \delta^x(k) \rangle$ where $x_k^i \in I$ is the sector/agent assigned to respond to the incident, $x_k^m \in I_{x_k^i}$ is the dispatched patrol team belonging to the assigned agent, $x_k^t \in T$ the time period which the assigned patrol team starts to act upon the dispatch call and $\delta^x(k)$ the resulting joint schedule after executing action x_k.

Transition. There are two main transitions in the model namely, from pre-decision state, S_k to post-decision S_k^x and from S_k^x to the next pre-decision state, S_{k+1}. The transition from S_k to S_k^x takes place after executing action x_k. Meanwhile, during transition from S_k^x to S_{k+1}, a realization of a dynamic event, ω_{k+1} takes place and $S_{k+1} = (S_k^x, \omega_{k+1})$.

There are $|I|$ possible realizations of S_{k+1} corresponding to the number of agents that can be assigned to the handle the new dynamic event.

$$\begin{aligned} S_{k+1} =& \Big((S_{k,i}^{x_k,i}, \omega_{k+1}), (S_{k,-i}^{x_k,-i}, \emptyset) \Big) \\ =& \Big((S_{k,i}^{x_k,i}, \omega_{k+1}), (S_{k+1,-i}) \Big) \end{aligned} \tag{5}$$

Reward Function. The reward function, $R(S_k, x_k)$ is designed in such a way that high reward is given to a successful incident response while minimizing the reduction in patrol presence at the same time. We introduce $f_r(x_k)$ to quantify the response utility after executing x_k. We propose the use of exponentially decreasing function to represent the response utility similar to the one proposed by [Amador *et al.*, 2014] and [Nelke *et al.*, 2020]. In other words, the later the incident is being responded, the more severe the impact of the incident and the less effective a response would be in resolving the incident. Thus, patrol teams have more incentives to respond to the incident as early as possible.

$$R(S_k, x_k) = f_r(x_k) \times f_p(\delta^x(k)) - f_p(\delta(k)) \tag{6}$$

$$f_r(x_k) = e^{-\beta_r \times \max(0, \tau_k - \tau_{target})} \tag{7}$$

Objective Function. The objective at every decision epoch k is to select joint action x_k^* which maximizes the immediate reward and the expected future reward from yet-to-realized dynamic events which is represented by the approximated value function, $\hat{V}(S_k^x)$.

$$x_k^* = \underset{x_k \in X(S_k)}{\operatorname{argmax}} \{R(S_k, x) + \gamma \hat{V}(S_k^x)\} \tag{8}$$

4 Solution Approach

To solve the optimization problem in Eq. (8), we propose an approach that combines MAVFA based on value function factorization with rescheduling heuristic based on ejection chain heuristic, where the former learns to approximate the joint value function and the latter is used to compute the argmax. Furthermore, we incorporate an iterative best response procedure that provides a scalable, decentralized scheme for coordination among agents. We propose the use of VFA because the decision variable, x_k is complex and multi-dimensional. Off-policy Temporal Difference (TD) learning method like Q-learning and policy gradient method may not be applicable directly since the action space changes depending on the current state [Zhang and Dietterich, 2000].

Our proposed cooperative MARL approach learns the value function of joint schedule of all agents after a patrol team from a particular sector has been dispatched to attend to an incident and rescheduling actions have been performed across all sectors. In other words, the learned value function will guide the rescheduling heuristic to find dispatching and rescheduling decisions that are anticipatory and coordinated.

4.1 MAVFA based on Value Function Factorization

We assume that the joint value function approximate, $\hat{V}(S_k^x)$ can be factorized into the value function approximates of each agent as shown in Eq. 9. Similar to other value function factorization approaches like VDN and QMIX, we assume an IGM principle. This assumption is reasonable because MADPRP is not a zero-sum game. An agent which is assigned to respond to an incident may gain certain utility score but this does not directly result in other agents suffering a loss in their utility scores. IGM assumption is also valid because the total number of incidents and the total effective patrol time across all police sectors are positively correlated to the incidents responded within every sector and the effective routine patrol

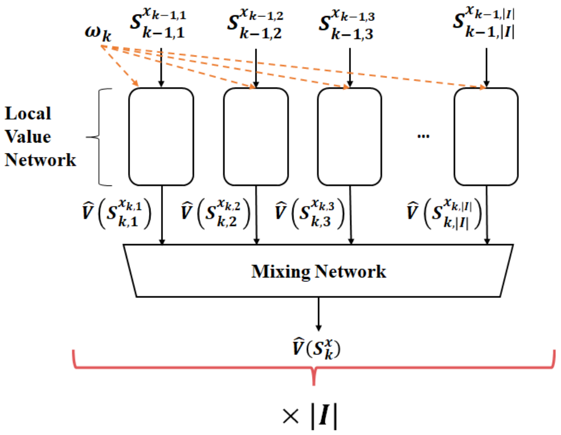
ω_k
$S_{k-1,1}^{x_{k-1,1}}$
$S_{k-1,2}^{x_{k-1,2}}$
$S_{k-1,3}^{x_{k-1,3}}$
$S_{k-1,|I|}^{x_{k-1,|I|}}$
Local Value Network
...
$\hat{V}(S_{k,1}^{x_{k,1}})$
$\hat{V}(S_{k,2}^{x_{k,2}})$
$\hat{V}(S_{k,3}^{x_{k,3}})$
$\hat{V}(S_{k,|I|}^{x_{k,|I|}})$
Mixing Network
$\hat{V}(S_k^x)$
× $|I|$

Figure 1: Given $|I|$ realizations of pre-decision state S_k, there are correspondingly $|I|$ possible variations of $\hat{V}(S_k^x)$.

time of each patrol sector respectively.

$$\begin{aligned}\hat{V}(S_k^x) =& F\Big(\hat{V}(S_{k,1}^{x_{k,1}}), \hat{V}(S_{k,2}^{x_{k,2}}), ..., \hat{V}(S_{k,|I|}^{x_{k,|1|}})\Big)\\ =& F\Big((\hat{V}(S_{k,i}^{x_{k,i}}))_{i\in I}\Big) \end{aligned} \quad (9)$$

Combining with Eq. 9, Eq. 8 can be rewritten as follows.

$$x_k^* = \underset{x_k \in X(S_k)}{\text{argmax}} \Big\{ R(S_k, x_k) + \gamma F\Big((\hat{V}(S_{k,i}^{x_{k,i}}))_{i\in I}\Big)\Big\} \quad (10)$$

As each pre-decision joint state can be represented by $|I|$ different realizations (see Eq. 5), the objective function can be further rewritten as follows:

$$\begin{aligned} x_k^* = \underset{x_k \in X(S_k)}{\text{argmax}}\ \underset{i\in I}{\text{argmax}} \Big\{ R\Big(& (S_{k-1,i}^{x_{k-1},i}, \omega_{k+1}, x_{k,i}),\\ & (S_{k,-i}, x_{k,-i})\Big)\\ & + \gamma F\Big((\hat{V}(S_{k,i}^{x_{k,i}}))_{i\in I})\Big)\Big\} \end{aligned} \quad (11)$$

We propose to represent $\hat{V}(S_k^x)$ as neural networks with parameters θ and its architecture can be found in Fig.1. Our approach learns the policy to execute complex action directly because the learned value function represents the value of a post-decision state, a state after executing both *event-handling* and *re-planning* actions.

Local Value Network. The architecture of the local value network, $\hat{V}(S_{k,i}^{x_{k,i}}, \theta_i)$ can be found in Fig.2. For scalability and simplicity, (homogeneous) agents share the same local network parameter θ_i. We propose an encoder network in the form of multilayer perceptrons to extract the key features of the plan in its raw form and encoding it into a lower-dimensional vector representation. Handcrafted features described in the above model formulation (Eq. 4) are concatenated to enhance the learning process.

Mixing Network. We represent F in Eq. 9 as fully-connected neural networks and we loosely refer this network as a ”Mixing Network”. We use the term ”mixing” in its general definition and it does not refer to the specific mixing network structure as proposed in QMIX. The key feature of this network is that the global state information is passed into the network to enhance coordination. Rashid *et al.* [2018] in their ablation study have shown that providing extra state information does improve performance.

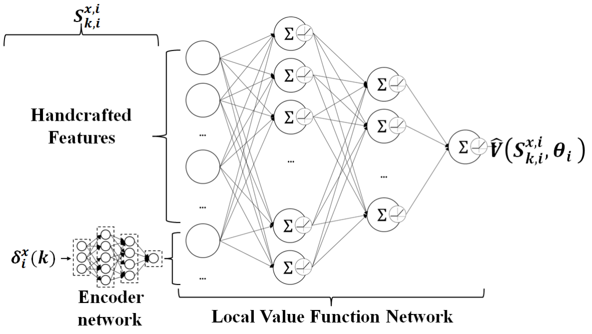

Figure 2: Local Value Network

Learning Algorithm. To learn the parameter θ, we propose to use an on-policy TD learning with experience replay (see [Joe and Lau, 2020]) with the main differences lie in how decision is being computed in line 11 and how the value function is being represented.

VFA-Guided Heuristic. To compute $\text{argmax}_{x_k \in X(S_k)}$ in Eq. 8, we propose to use rescheduling heuristic based on ejection chain [Joe *et al.*, 2022]. In this heuristic, the ejection chain consists of a sequence of defect-checking and repair operations. Insertion of an incident into a schedule potentially introduces a defect to the schedule. Repairing a defect at one part of a schedule may introduce a defect in another part. Thus, a chain of check and repair operations is formed until termination condition is met or until no defect is present. The learnt value function $\hat{V}(S_k^x)$ is used to search for a repair decision that result in a repaired schedule that maximises both immediate and future rewards. More detailed descriptions of this rescheduling heuristic can be found in [Joe *et al.*, 2022].

4.2 Iterative Best Response Procedure

To compute $\text{argmax}_{x_k \in X(S_k)}$ in a decentralized manner, we propose an iterative best response procedure. The purpose of incorporating this procedure is twofold. Firstly, it acts as a scalable, decentralized optimization heuristic. To compute this argmax directly is akin to solve the multi-agent optimization problem centrally (as a single-agent) which will be computationally expensive. Secondly, this procedure provides an explicit coordination mechanism amongst agents via asynchronous actions.

Optimization Heuristic. Lambert Iii *et al.* [2005] propose a sampled fictitious play algorithm as an optimization heuristic to solve large-scale optimization problems. Optimization problem can be formulated as a n-player game where every pure-strategy equilibrium of a game is a local optimum since no player can change its strategy to improve the objective function. The premise of this algorithm is that the problem must meet the criteria of being a potential game. Potential game possesses a pure-strategy equilibrium and has the Finite Improvement Property (FIP) [Monderer and Shapley, 1996]. Having the FIP means that every path generated by a best response procedure will converge to an equilibrium.

We formulate the optimization problem found in Eq. 11 as an I-player game Γ with agents represented as players having a finite set of strategies Δ_i and sharing the same payoff function. We define the payoff function, $u^i(\delta)$ as the utility of agent i when all agents follow a joint plan δ. We deliberately remove the index k to simplify the notation. Here, we make the assumption that the payoff of each agent is not dependent on other agents' payoff such that we can define a function, $P(\delta) = \sum_{i \in I} u^i(\delta)$. This assumption is consistent with the earlier IGM assumption. $P(\delta)$ is an *ordinal potential function* for Γ since for every $i \in I$ and for every $\delta_{-i} \in \Delta_{-i}$

$$\begin{aligned} u^i(\delta_i, \delta_{-i}) - u^i(\delta_i', \delta_{-i}) > 0 \text{ iff} \\ P(\delta_i, \delta_{-i}) - P(\delta_i', \delta_{-i}) > 0 \text{ for every } \delta_i, \delta_i' \in \Delta_i. \end{aligned} \quad (12)$$

An equilibrium of Γ is a local optimum since no player can improve its payoff by changing its individual plan. Conversely, every optimal solution, δ^* of Γ is an equilibrium since $u^i(\delta^*) \geq u^i(\delta_i, \delta^*_{-i})$ for all $i \in I$ where $\delta_i \in B_i(\delta^*_{-i})$. To search for a local optimal solution of the optimization problem in Eq. 11, we propose an iterative best response algorithm proposed by Joe and Lau [2023].

Explicit Coordination. Unlike communication network, the iterative best response procedure induces a more explicit form of coordination as agents take turn to respond to the other agents' actions. We note that combining communication network and iterative best response is not feasible. The presence of communication network means that an agent's state and action are dependent of other agents' states and actions. Best response procedure may not converge because other agent's payoffs will change the moment an agent performs a best response action and the inequality in Eq.12 is no longer valid.

5 Experiments

We evaluate our approach on a problem scenario involving 3 police sectors in the North-Central region of a city-state that we reside in. There are a total of $|J| = 62$ patrol areas. Each sector represents different problem complexities in terms of the ratio of patrol team per patrol area, the diversity of the patrol areas and the spatial distribution of dynamic incidents (see Table 2). Due to the classified nature of the data, synthetically-generated data based on publicly-available data source are used in this experiment[1].

5.1 Experimental Setup and Design

We divide our experiment into two phases to evaluate the impact of each of the components of our proposed approach on the solution quality and computational time through a series of ablation studies. For a fairer comparison, we use the same reward function and rescheduling heuristic.

Sector	*Parameter*	*Description*
1	$\lvert I_1 \rvert = 4$, $\lvert J_1 \rvert = 14$	High patrol team-to-area ratio, relatively homogeneous patrol densities (medium)
2	$\lvert I_2 \rvert = 4$, $\lvert J_2 \rvert = 23$	Low patrol team-to-area ratio, relatively homogeneous patrol densities (low)
3	$\lvert I_3 \rvert = 4$, $\lvert J_3 \rvert = 25$	Low patrol team-to-area ratio, more diverse patrol densities (low to high)

Table 2: Different patrol sectors representing different problem structures and complexities.

Phase 1. We evaluate the performance of our joint learning mechanism against a two-stage approach where another popular value-based RL algorithm, DQN is used to learn the dispatch policy in the first stage. We also evaluate the effect of iterative best response as a coordination mechanism by comparing it with a communication network. We run a total of 10,000 training episodes where each episode represents a given initial joint schedule and a set of dynamic events occurring throughout the planning horizon. We evaluate the solution quality based on the resulting cumulative rewards in terms of the average % improvement of incident success rate over myopic approach. Here are the models being run in this phase:

- **MAVFA-BR-H.** This is our proposed approach where BR refers to iterative best response procedure and H refers to heuristic.
- **MAVFA-C-H.** This is a model that incorporates communication network as an implicit coordination mechanism. We adapt a Attention-Based Convolutional Communication Network proposed by Jiang *et al.* [2020].
- **MAVFA-H.** This is an MAVFA model without any communication mechanism amongst agents.
- **MADQN-BR-H.** This is the two-stage approach similar to Chen *et al.* [2021]. In order to evaluate solely on the joint vs. two-stage learning mechanisms, we retain the same components, BR and H.

Phase 2. We evaluate the impact of our proposed MAVFA algorithm and iterative best response procedure in making anticipatory and coordinated decision-making during execution. We run 30 experiments to simulate one month's worth of daily operations and for each experiment, we run 20 different set of realizations of dynamic incidents to simulate different possible daily scenarios; representing a sufficiently substantial sample size for statistical evaluation. We evaluate the impact of our approach against the absence of coordination and anticipation (myopic) by running the following baseline models for comparison on top of MAVFA-BR-H and MAVFA-H:

- **VFA-H.** This model assumes each sector runs its own independent single-agent VFA without any form of communication and collaboration amongst agents.
- **BR-H.** This is a version of our approach without VFA i.e. myopic approach.

[1] Detailed descriptions of the data, the experiments and the implemented codes can be found in: https://github.com/waldyjoe/MADPRP.

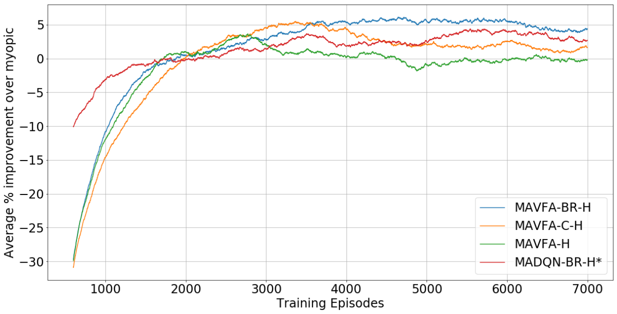

Figure 3: Average cumulative rewards over last 600 training episodes.

Model	*Success Rate*	*Time Per Decision(s)*
MAVFA-BR-H	(O) $65.0 \pm 0.7\%$ (W) $\mathbf{52.4 \pm 0.9\%}$	24.1 ± 2.1
MAVFA-H	(O) $57.3 \pm 0.6\%$ (W) $44.0 \pm 0.8\%$	9.3 ± 1.0
VFA-H	(O) $\mathbf{67.6 \pm 0.6\%}$ (W) $44.2 \pm 1.0\%$	1.4 ± 0.2
BR-H	(O) $62.2 \pm 0.8\%$ (W) $47.9 \pm 1.0\%$	23.7 ± 2.1

Table 3: Our approach statistically outperforms the other models in terms of overall success rate (O) and the success rate of the worst-performing agent (W).

We evaluate our approach on overall success rate (% of incidents responded within a target time) across all sectors and the success rate of the worst performing agent. For computational time, we evaluate based on the time taken per decision (dispatch and reschedule). We include results in terms of mean and 95% Confidence Interval (CI) of the mean since the problem and the heuristic used are stochastic in nature.

5.2 Experiment Results and Discussion

Phase 1. We observe that the cumulative rewards (represented as % improvement over myopic) stabilize after around 6000 training episodes (see Fig. 3). We assume stability where the standard deviation of the cumulative rewards are kept within 20% of the sample mean for at least 600 consecutive episodes. Fig. 3 shows that our explicit coordination mechanism results in a better overall success rate.

We also observe that our approach seems to only outperform the two-stage approach by a slight margin. However, as indicate by the symbol * beside MADQN-BR-H, this model has been pre-trained with one additional round of training. This is because we observe that 10,000 episodes were not sufficient for this model to learn effectively as it is only able to achieve an -2% improvement over myopic. Thus, given the same number of training episodes, our proposed approach will outperform it by a bigger margin.

Phase 2. Our proposed approach is statistically able to produce decisions that result in higher overall success rate and success rate of the worst-performing agent as compared to the other baselines except for VFA-H (see Table 3). Our proposed coordination mechanism account to about 13% increase in overall success rate (vs. MAVFA-H) while collaboration amongst agent significantly increases the success rate of the worst performing agent by $> 18\%$ (vs. VFA-H).

The overall success rate of VFA-H is higher because Sector 1 has a higher patrol team-to-area ratio which increases its ability to respond to local incident quickly, skewing the average result. Cooperation in light of limited resources inevitably means that some sort of compromise is needed which in this, our approach is able to ensure that every sector's success rate is at least of a certain reasonable threshold ($> 50\%$).

Although our proposed approach is slower than the other models, it is able to compute the decision within an operationally realistic time of $< 30s$ (see Table 3). In fact, our value function approximation steps and iterative best response procedure account for $< 1s$ and $< 15s$ of additional computation time per decision respectively.

Discussion. The magnitude of our improvements may not seem substantial (about 5%). In reality however, a 5% improvement translates into 3 more incidents responded within target time, which is quite significant in the law enforcement context. Given the additional complexity of our problem (routing and scheduling) and multi-agent setting, an improvement of 5% is comparable with the performances of various offline methods in the literature that solve DVRP with stochastic customers (see [Ritzinger *et al.*, 2016]).

Although the comparison against existing cooperative MARL approaches would strengthen our evaluation attempt, such comparisons are not so straightforward. Given that these approaches require the problem or action to be decomposed into two stages, the rescheduling heuristic chosen would have to be different. In addition, these approaches assume simultaneous actions by all agents which mean that additional step is needed to ensure proper handling of dynamic event to prevent incident being ignored or more than one agent responding to one incident. Thus, modifications to these approaches such as action-masking are needed and these may result in deviations from these approaches' original design. The eventual solution quality of these modified approaches need to be assessed more carefully as any improvement may come from the rescheduling heuristic used.

6 Conclusion and Future Works

We presented a pioneering effort on a cooperative MARL approach to solve MADPRP, an instance of multi-agent sequential decision problem with complex action, directly. Moving forward, there are many opportunities to further evaluate and build upon the ideas proposed. For example, our proposed approach can be evaluated on other multi-agent sequential decision problem settings. This will require exploration of more context-specific network architecture. It would also be interesting to compare our approach with existing cooperative MARL approaches. However, as mentioned earlier, additional care is required in designing the experiment to ensure fair comparison, which we hope to address in the future.

Ethics Statement

AI and data-driven approaches in law enforcement context can be a double-edged sword; they may result in significant benefits in maintaining law and order but also possible ethical and adverse societal impacts. Any initiative that is driven by data is inevitably vulnerable to data manipulation and abuse by human operators [Richardson *et al.*, 2019]. The existing public reputation of the law enforcement agency itself would also determine how the proposed initiative is perceived and the impact to the society prior to implementation. These factors exist with or without AI and their impacts may vary from country to country, city to city.

Although our approach aims to achieve better operational efficiency and service quality for public good, we acknowledge that, for this approach to be adopted into an operational system, there could be unintended harms caused by inherent biases in the data. Police may patrol certain areas more often due to historical bias (for e.g. based on past police operations) and they can be heavily influenced by how the authority perceives those areas. In addition, the training data may also rely on incident reporting through police public hotline (such as 911 calls) which is triggered by the public rather than by the law enforcement agency itself. Such data is also very much influenced by the inherent biasness that exists in the society for e.g. such calls, more often than not come from segments of society who are more privileged or educated. This may result in overpolicing areas that are poor, marginalized or of a certain demographic. Thus, there is a need for decision makers to work with various community stakeholders to address these concerns prior to implementation of such data-driven approaches. For example, should there be a risk of data bias, such (historical crime) data could be removed or preprocessed from the training dataset with the aim to have a learnt policy that will discount or downplay certain biasness that exists in that particular dataset.

That being said, we acknowledge that AI does not replace human operators totally but it should be used to inform human decision-makers. Sound policies and tight processes still need to be in place and human operators ultimately should be accountable in applying AI and data-driven approaches for the good of society.

References

[Amador *et al.*, 2014] Sofia Amador, Steven Okamoto, and Roie Zivan. Dynamic multi-agent task allocation with spatial and temporal constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[Chen *et al.*, 2019] Yujie Chen, Yu Qian, Yichen Yao, Zili Wu, Rongqi Li, Yinzhi Zhou, Haoyuan Hu, and Yinghui Xu. Can sophisticated dispatching strategy acquired by reinforcement learning? In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1395–1403, 2019.

[Chen *et al.*, 2021] Jiayu Chen, Abhishek K Umrawal, Tian Lan, and Vaneet Aggarwal. Deepfreight: A model-free deep-reinforcement-learning-based algorithm for multi-transfer freight delivery. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 31, pages 510–518, 2021.

[Chen *et al.*, 2022] Xinwei Chen, Marlin W Ulmer, and Barrett W Thomas. Deep q-learning for same-day delivery with vehicles and drones. *European Journal of Operational Research*, 298(3):939–952, 2022.

[Das *et al.*, 2019] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*, pages 1538–1546. PMLR, 2019.

[Dewinter *et al.*, 2020] Maite Dewinter, Christophe Vandeviver, Tom Vander Beken, and Frank Witlox. Analysing the police patrol routing problem: A review. *ISPRS International Journal of Geo-Information*, 9(3):157, 2020.

[Doyle *et al.*, 2016] Maria Doyle, Louise Frogner, Henrik Andershed, and Anna-Karin Andershed. Feelings of safety in the presence of the police, security guards, and police volunteers. *European journal on criminal policy and research*, 22:19–40, 2016.

[Foerster *et al.*, 2018] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Jiang and Lu, 2018] Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems*, 31, 2018.

[Jiang *et al.*, 2020] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. In *International Conference on Learning Representations*, 2020.

[Joe and Lau, 2020] Waldy Joe and Hoong Chuin Lau. Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, pages 394–402, 2020.

[Joe and Lau, 2023] Waldy Joe and Hoong Chuin Lau. Coordinating multi-party vehicle routing with location congestion via iterative best response. In *Multi-Agent Systems: 18th European Conference, EUMAS 2021. Special issue in SN Computer Science, 4:157*. Springer Nature, 2023.

[Joe *et al.*, 2022] Waldy Joe, Hoong Chuin Lau, and Jonathan Pan. Reinforcement learning approach to solve dynamic bi-objective police patrol dispatching and rescheduling problem. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 32, pages 453–461, 2022.

[Lambert Iii *et al.*, 2005] Theodore J Lambert Iii, Marina A Epelman, and Robert L Smith. A fictitious play approach to large-scale optimization. *Operations Research*, 53(3):477–489, 2005.

[Los *et al.*, 2020] Johan Los, Frederik Schulte, Margaretha Gansterer, Richard F Hartl, Matthijs TJ Spaan, and Rudy R

Negenborn. Decentralized combinatorial auctions for dynamic and large-scale collaborative vehicle routing. In *International Conference on Computational Logistics*, pages 215–230. Springer, 2020.

[Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[Ma *et al.*, 2021] Yi Ma, Xiaotian Hao, Jianye Hao, Jiawen Lu, Xing Liu, Tong Xialiang, Mingxuan Yuan, Zhigang Li, Jie Tang, and Zhaopeng Meng. A hierarchical reinforcement learning based optimization framework for large-scale dynamic pickup and delivery problems. *Advances in Neural Information Processing Systems*, 34:23609–23620, 2021.

[Monderer and Shapley, 1996] Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

[Mukhopadhyay *et al.*, 2016] Ayan Mukhopadhyay, Chao Zhang, Yevgeniy Vorobeychik, Milind Tambe, Kenneth Pence, and Paul Speer. Optimal allocation of police patrol resources using a continuous-time crime model. In *International conference on decision and game theory for security*, pages 139–158. Springer, 2016.

[Nelke *et al.*, 2020] Sofia Amador Nelke, Steven Okamoto, and Roie Zivan. Market clearing–based dynamic multi-agent task allocation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1):1–25, 2020.

[Pettet *et al.*, 2022] Geoffrey Pettet, Ayan Mukhopadhyay, Mykel J Kochenderfer, and Abhishek Dubey. Hierarchical planning for dynamic resource allocation in smart and connected communities. *ACM Transactions on Cyber-Physical Systems*, 6(4):1–26, 2022.

[Powell, 2019] Warren B Powell. A unified framework for stochastic optimization. *European Journal of Operational Research*, 275(3):795–821, 2019.

[Rashid *et al.*, 2018] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.

[Richardson *et al.*, 2019] Rashia Richardson, Jason Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 94(192), 2019.

[Ritzinger *et al.*, 2016] Ulrike Ritzinger, Jakob Puchinger, and Richard F Hartl. A survey on dynamic and stochastic vehicle routing problems. *International Journal of Production Research*, 54(1):215–231, 2016.

[Ruan *et al.*, 2022] Jingqing Ruan, Linghui Meng, Xuantang Xiong, Dengpeng Xing, and Bo Xu. Learning multi-agent action coordination via electing first-move agent. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 32, pages 624–628, 2022.

[Rumi *et al.*, 2020] Shakila Khan Rumi, Wei Shao, and Flora D Salim. Realtime predictive patrolling and routing with mobility and emergency calls data. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 964–968, 2020.

[Santana *et al.*, 2004] Hugo Santana, Geber Ramalho, Vincent Corruble, and Bohdana Ratitch. Multi-agent patrolling with reinforcement learning. In *Autonomous Agents and Multiagent Systems, International Joint Conference on*, volume 4, pages 1122–1129. IEEE Computer Society, 2004.

[Son *et al.*, 2019] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5887–5896. PMLR, 2019.

[Sunehag *et al.*, 2018] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087, 2018.

[Tkach and Amador, 2021] Itshak Tkach and Sofia Amador. Towards addressing dynamic multi-agent task allocation in law enforcement. *Autonomous Agents and Multi-Agent Systems*, 35(1):1–18, 2021.

[Ulmer *et al.*, 2020] Marlin W Ulmer, Justin C Goodson, Dirk C Mattfeld, and Barrett W Thomas. On modeling stochastic dynamic vehicle routing problems. *EURO Journal on Transportation and Logistics*, 9(2):100008, 2020.

[Wang and Kopfer, 2013] Xin Wang and Herbert Kopfer. Dynamic collaborative transportation planning: A rolling horizon planning approach. In *International Conference on Computational Logistics*, pages 128–142. Springer, 2013.

[Wang *et al.*, 2019] Wanyuan Wang, Zichen Dong, Bo An, and Yichuan Jiang. Toward efficient city-scale patrol planning using decomposition and grafting. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[Wang *et al.*, 2022] Wanyuan Wang, Hansi Tao, and Yichuan Jiang. Efficient online city-scale patrolling by exploiting offline model-based coordination policy. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13805–13818, 2022.

[Zhang and Dietterich, 2000] Wei Zhang and Thomas G Dietterich. Solving combinatorial optimization tasks by reinforcement learning: A general methodology applied to resource-constrained scheduling. *Journal of Artificial Intelligence Reseach*, 1:1–38, 2000.