

# Capacity-Aware Planning and Scheduling in Budget-Constrained Monotonic MDPs: A Meta-RL Approach

Manav Vora<sup>1</sup>, Ilan Shomorony<sup>2</sup>, Melkior Ornik<sup>1</sup>

<sup>1</sup>Department of Aerospace Engineering and Coordinated Science Laboratory, University of Illinois Urbana Champaign, Urbana, USA

<sup>2</sup>Department of Electrical and Computer Engineering, University of Illinois Urbana Champaign, Urbana, USA  
mkvora2@illinois.edu, ilans@illinois.edu, mornik@illinois.edu

## Abstract

Many real-world sequential repair problems can be effectively modeled using monotonic Markov Decision Processes (MDPs), where the system state stochastically decreases and can only be increased by performing a restorative action. This work addresses the problem of solving multi-component monotonic MDPs with both budget and capacity constraints. The budget constraint limits the total number of restorative actions and the capacity constraint limits the number of restorative actions that can be performed simultaneously. While prior methods dealt with budget constraints, capacity constraints introduce an additional complexity in the multi-component action space that results in a combinatorial optimization problem. Including capacity constraints in prior methods leads to an exponential increase in computational complexity as the number of components in the MDP grows. We propose a two-step planning approach to address this challenge. First, we partition the components of the multi-component MDP into groups, where the number of groups is determined by the capacity constraint. We achieve this partitioning by solving a Linear Sum Assignment Problem (LSAP), which groups components to maximize the diversity in the properties of their transition probabilities. Each group is then allocated a fraction of the total budget proportional to its size. This partitioning effectively decouples the large multi-component MDP into smaller subproblems, which are computationally feasible because the capacity constraint is simplified and the budget constraint can be addressed using existing methods. Subsequently, we use a meta-trained PPO agent to obtain an approximately optimal policy for each group. To validate our approach, we apply it to the problem of scheduling repairs for a large group of industrial robots, constrained by a limited number of repair technicians and a total repair budget. Our results demonstrate that the proposed method outperforms baseline approaches in terms of maximizing the average uptime of the robot swarm, particularly for large swarm sizes. Lastly, we confirm the scalability of our approach through a computational complexity analysis across varying numbers of robots and repair technicians.

## 1 Introduction

Markov Decision Processes (MDPs) provide an efficient framework for modelling various real-world sequential decision making scenarios. One such scenario is that of sequential repair problems (Chen, Liu, and Xiahou 2021; Papakonstantinou and Shinozuka 2014), including in the context of

maintaining industrial robots (Borgi et al. 2017). With huge technical advancements in the field of robotics, industrial robots have become ubiquitous across manufacturing industries (Kibira and Qiao 2023). However, once deployed, the robots undergo wear and tear which leads to performance degradation over time (Qiao and Weiss 2018). This degradation is often stochastic due to fluctuating workloads and varying levels of wear, among other factors (Hung, Shen, and Lee 2024; Chen, Liu, and Xiahou 2021). Hence the problem of planning and scheduling maintenance of industrial robots can be modeled as a monotonic MDP (Vora, Grussing, and Ornik 2024), with the agent state modelling the monotonically decreasing performance efficiency of the robot. Restorative actions, such as repairs, can restore this state to its maximum value.

In practice, manufacturing industries often have a limited repair capacity (Perlman, Mehrez, and Kaspi 2001) while also having multiple robots performing operations simultaneously (Hassan, Liu, and Paul 2018). In this paper we consider the problem of obtaining planning and scheduling policies for budget- and capacity-constrained multi-component monotonic MDPs. Each component in this context refers to an individual system or machine, such as an industrial robot, with its own independent transition probabilities. These components are coupled due to the capacity constraints, which limit the number of restorative actions that can be performed simultaneously at any given time-step. In addition to the capacity constraint, the system may also be constrained by a shared total budget, which further complicates the optimization process. This combination of constraints results in a constrained combinatorial optimization problem, which is generally NP-hard (Papadimitriou and Steiglitz 1998). Several exact methods exist for solving constrained combinatorial optimization problems, including dynamic programming and branch-and-bound algorithms (Toth 2000). However, these methods scale poorly to large scenarios due to the exponential growth in computational complexity (Korte et al. 2011). Furthermore, significant research has previously been done on solving budget-constrained MDPs (Wu et al. 2018; Kalagarla, Jain, and Nuzzo 2021). However, the complexity of these algorithms is generally exponential in the number of states of the MDP and hence would be exponential in the number of components for a multi-component MDP, rendering them unsuitable for solving large multi-

component MDPs. In Vora, Grissing, and Ornik (2024), the authors propose an algorithm that scales linearly with the number of components of a budget-constrained multi-component POMDP. However, the algorithm involves computing an a priori budget split among the individual components. Introducing capacity constraints complicates this process by introducing an additional coupling among the component MDPs, making it difficult to perform the budget allocation effectively. Hence, this algorithm cannot be directly extended to solve budget and capacity-constrained multi-component monotonic MDPs.

To address this added complexity introduced by the capacity constraints, we propose a scalable meta-reinforcement learning approach for solving budget and capacity-constrained multi-component monotonic MDPs, comprising two key steps. The first step involves decoupling the large multi-component MDP into smaller, computationally tractable subproblems by partitioning the component MDPs into groups. The number of groups is determined by the capacity constraints: specifically, if the capacity constraint is  $k$ , we partition the components into  $k$  groups. We achieve this partitioning by formulating a Linear Sum Assignment Problem (LSAP). The LSAP optimally assigns components to groups to minimize a total cost. The cost matrix for the LSAP is designed to maximize diversity within each group in terms of the transition probabilities of the member components. By ensuring that each group contains components with varied behaviors, all groups exhibit similar aggregate characteristics. This uniformity justifies allocating the total budget proportionally among the groups based on their sizes. In the second step, we obtain the approximately optimal planning and scheduling policy for each group using a meta-PPO agent, trained on a diverse set of component groups and budget values. This two-step approach not only simplifies the capacity constraints by partitioning the components into groups but also results in a more tractable learning process by focusing on smaller MDPs with reduced dimensionality, compared to a single large multi-component MDP. Additionally, we perform a computational complexity analysis to demonstrate the scalability of this approach across varying number of components.

The key contributions of the paper are:

1. We introduce an LSAP-based partitioning method to manage the capacity and budget constraints, decoupling the large multi-component MDP into smaller MDPs.
2. We train a meta-PPO agent to obtain the planning and scheduling policies for all partitions, each representing a smaller budget-constrained multi-component MDP with simplified capacity constraints.
3. We demonstrate the scalability and effectiveness of our approach by applying it to a large-scale real-world scenario of repair scheduling for a swarm of industrial robots and performing a computational complexity analysis for varying numbers of components in the multi-component MDP.

## 2 Preliminaries and Related Work

### 2.1 Multi-Component Markov Decision Processes

A discrete-time multi-component Markov Decision Process (MDP) consists of multiple component MDPs with individual transition probabilities as well as state and action spaces. It is defined by the 4-tuple  $(\mathcal{S}, A, T, n)$ . Here  $\mathcal{S}$  denotes the state space which is defined as  $\mathcal{S} = \prod_{i=1}^n \mathcal{S}^i$ , and  $A$  represents the action space given by  $A = \prod_{i=1}^n A^i$ , where  $\mathcal{S}^i$  and  $A^i$  denote the state and action space of component  $i$ , respectively. Similarly, the transition probability function  $T : \mathcal{S} \times A \times \mathcal{S} \rightarrow [0, 1]$  is given by  $T = \prod_{i=1}^n T^i$ , where  $T^i$  is the transition function of component  $i$  and  $T^i : \mathcal{S}^i \times A^i \times \mathcal{S}^i \rightarrow [0, 1]$ . Finally,  $n$  denotes the number of components in the multi-component MDP.

### 2.2 Budget-Constrained MDPs

Sequential decision-making problems with budget constraints are generally modeled using the Constrained MDP (CMDP) framework (Altman 2021). In a CMDP, the agent must optimize a reward function while adhering to a budget constraint that limits the cumulative cost over a fixed planning horizon. A lot of work has previously been done on solving CMDPs (Borgi et al. 2017; Xiao et al. 2019). However, CMDPs require traditional MDP descriptions and hence do not scale well, with the number of components, for large multi-component MDPs. Prior work by Boutilier and Lu (2016) presents a scalable solution for large budget-constrained multi-component MDPs, which involves effectively decoupling the multi-component MDP by allocating the shared budget among the individual component MDPs. However, introducing capacity constraints will make this budget allocation significantly more complex due to the added combinatorial complexity. Hence, this method cannot be directly extended to budget and capacity-constrained MDPs.

### 2.3 Reinforcement Learning for MDPs

Reinforcement learning (RL) has been widely used to solve MDPs (Sutton 2018; Schulman et al. 2017; Mnih et al. 2015). RL has also been employed to solve budget-constrained MDPs, where the goal is to learn policies that adhere to a predefined budget (Wu et al. 2018; Carrara et al. 2019). However, these methods tend to suffer from poor scalability due to the high dimensionality introduced by the shared resource and action constraints of a budget- and capacity-constrained multi-component MDP.

### 2.4 Capacity-Constrained MDPs

Capacity constraints, where only a limited number of actions can be performed simultaneously, introduce another layer of complexity to MDPs. Previous work by Haksar and Schwager (2018) considers the problem of solving MDPs with global capacity constraints and presents an approximate linear programming solution. While linear programs are indeed efficient and can handle large-scale problems, the challenge in our context arises from the combinatorial explosion of the state and action spaces in budget- and capacity-constrained

multi-component MDPs. As the number of components increases, the joint state and action space grow exponentially. This exponential growth results in a linear programming formulation with  $O(|S^i|^n |A^i|^n)$  variables and constraints, where  $|\cdot|$  denotes the cardinality of a finite set. Even with efficient LP solvers, solving such a large LP becomes computationally infeasible for moderate to large values of  $n$ . Therefore, the method proposed by Haksar and Schwager (2018) cannot be directly extended to the large-scale problems considered in this paper.

### 3 Problem Formulation

In this paper, we consider a multi-component monotonic MDP with budget and capacity constraints. Although the individual components of a multi-component MDP are independent in terms of the transition probabilities, the budget constraints introduce weak coupling among the component MDPs. This is because any expenditure of the budget by one component reduces the available budget for the remaining components. Furthermore, the capacity constraints introduce an additional layer of coupling among the components by restricting the number of restorative actions that can be executed at a given time step. Together, these constraints transform the problem into a constrained combinatorial optimization problem. In this paper, we consider a budget and capacity-constrained  $n$ -component monotonic MDP with the state space for each component  $i$  being  $\mathcal{S}^i = \{0, 1, \dots, \bar{s}\}$ . Here,  $\bar{s} \in \mathbb{N}_0$  denotes the maximum possible state value and  $\mathbb{N}_0$  denotes the set of non-negative integers. Furthermore,  $A^i = \{d^i, m^i\}$  is the action space for component  $i$ . The total budget is  $B \in \mathbb{N}_0$  and the capacity constraint is given by  $r \in \mathbb{N}_0$ .

At time instant  $k$ , the system state is denoted by  $s_k = (s_k^1, s_k^2, \dots, s_k^n)$ . Here,  $s_k^i \in \mathcal{S}^i$  is the state of component  $i$  at time  $k$ . The action at time  $k$  is given by  $a_k = (a_k^1, a_k^2, \dots, a_k^n)$  with the cost associated with this action being  $c_{a_k} = \sum_{i=1}^n c_{a_k^i}$ . Here,  $c_{a_k^i}$  represents the corresponding cost of action  $a_k^i$ . As mentioned in Section 2, the transition probability function for the multi-component MDP is expressed as

$$T(s_k, a_k, s_{k+1}) = \prod_{i=1}^n T^i(s_k^i, a_k^i, s_{k+1}^i).$$

For our case, performing action  $d_k^i$  leads to a decrease in the state value and costs nothing, i.e.,  $c_{d_k^i} = 0$  for all  $i, k$ . On the other hand, action  $m_k^i$  increases the state, with the maximum value bounded by  $\bar{s}$  and has a non-zero cost  $c_{m_k^i} > 0$ . Additionally, for all  $k$  and  $i$ , the state  $s_k^i = 0$  is an absorbing state. Thus, the component-wise transition function  $T^i$  follows the formulation from Vora, Grissing, and Ornik (2024):

$$T^i(s_k^i, a_k^i, s_{k+1}^i) = \begin{cases} p_1^i(s_k^i, a_k^i, s_{k+1}^i), & \text{if } a_k^i = m^i \text{ and } 0 < s_k^i \leq s_{k+1}^i, \\ p_2^i(s_k^i, a_k^i, s_{k+1}^i), & \text{if } a_k^i = d^i \text{ and } s_{k+1}^i \leq s_k^i, \\ 1, & \text{if } s_{k+1}^i = 0 = s_k^i, \\ 0, & \text{otherwise.} \end{cases}$$

### 3.1 Problem Statement

The main goal of this paper is to solve this budget and capacity-constrained multi-component monotonic MDP by finding a policy  $\pi$  that maximizes the minimum expected time for any component to reach the absorbing state, while satisfying the budget and capacity constraints. For each component  $i$ , the time to reach the absorbing state  $s_k^i = 0$  is denoted by  $t_{abs}^i$ . Mathematically, we are attempting to solve:

$$\begin{aligned} & \max_{\pi} \min_i \mathbb{E}[t_{abs}^i(\pi)] \\ & \text{s.t. } \sum_{k=0}^{\infty} c_{a_k}(\pi) \leq B, \\ & \sum_{i=1}^n \mathbb{1}(a_k^i(\pi)) \leq r, \quad \forall k. \end{aligned} \quad (1)$$

In (1),  $\mathbb{1}$  represents the indicator function which has a value of 1 when  $a_k^i(\pi) = m^i$  and 0 otherwise. Furthermore,  $t_{abs}^i, c_{a_k}$  and  $a_k^i$  are all functions of the policy  $\pi$ . For simplicity, we will omit this dependence on  $\pi$  in the rest of the paper. Note that we consider an infinite planning horizon in (1). This is because, given a budget constraint, the length of the horizon does not influence the optimal policy. In our experiments, however, we consider a sufficiently large finite horizon to effectively evaluate the performance of our approach.

## 4 Methodology

We will now discuss our proposed approach to obtain the approximately optimal policy for a budget and capacity-constrained multi-component monotonic MDP. Our approach follows a two-step process, as shown by the architectural overview in Figure 1. In the first step we partition the large multi-component MDP into  $r$  groups by solving a Linear Sum Assignment Problem (LSAP). The cost matrix for this LSAP is derived using statistical metrics that characterize  $T^i$  for each component  $i$ , and is designed to maximize diversity within each group by grouping components with varied transition behaviors together. This ensures that all groups have similar aggregate characteristics, which in turn validates distributing the total budget proportionally based on group sizes. After partitioning, the total budget is distributed among the groups in proportion to their sizes. The second step involves using a meta-trained reinforcement learning (RL) agent to obtain the approximately optimal policy for each group and consequently derive an optimal policy for the overall budget and capacity-constrained multi-component monotonic MDP. The following subsections provide a detailed explanation of these steps.

### 4.1 LSAP-based Partitioning of Multi-Component MDP

The problem of finding an optimal policy for the budget- and capacity-constrained multi-component monotonic MDP is that of finding the optimal solution for the constrained combinatorial optimization problem given by (1). Exact methods like integer linear programming (Schrijver 1998; Floudas

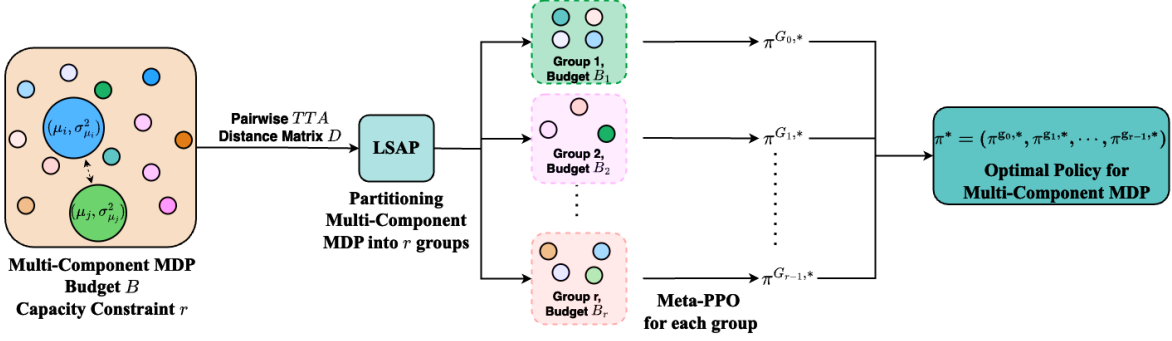


Figure 1: Architectural overview of the proposed approach.

and Lin 2005) are usually used to guarantee optimality of the solution for combinatorial optimization problems. However, these methods are not scalable for large numbers of variables (Solozabal, Ceberio, and Takáč 2020) and hence are impractical for solving large multi-component MDPs with capacity constraints. To address this challenge of scalability, we propose an LSAP-based partitioning approach. We intelligently partition the large  $n$ -component MDP with a capacity constraint  $r$  and total budget  $B$  into  $r$  groups, thereby decomposing (1) into smaller, more manageable subproblems.

The LSAP (Burkard and Cella 1999) seeks an assignment of  $n$  agents to  $n$  positions that minimizes the total assignment cost. In our context, the LSAP cost matrix is constructed using the time-to-absorption statistics (TTA) of each component, which describe the expected time for a component to reach the absorbing state in the absence of restorative actions. For component  $i$ , let  $\mu_i$  denote the expected TTA and let  $\sigma_{\mu_i}^2$  denote variance of the expected TTA. Subsequently, using the pairwise distances between the TTA statistics of the components, we construct a pairwise distance matrix  $D \in \mathbb{R}^{n \times n}$  with

$$D_{ij} = \sqrt{(\mu_i - \mu_j)^2 + (\sigma_{\mu_i}^2 - \sigma_{\mu_j}^2)^2}.$$

Thus,  $D$  captures the similarity between components  $i$  and  $j$  in terms of their respective TTA characteristics. The cost matrix for the LSAP is given by:

$$C = -D, \quad (2)$$

and the LSAP is formulated as (Crouse 2016):

$$\begin{aligned} \min_x \quad & \sum_{i=1}^n \sum_{j=1}^n C_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{j=0}^n x_{ij} = 1 \quad \forall i, \\ & \sum_{i=0}^n x_{ij} = 1 \quad \forall j, \end{aligned} \quad (3)$$

where  $x_{ij}$  is a binary decision variable indicating whether component  $i$  is assigned to position  $j$ . By minimizing the

total cost, which is the negative of the total pairwise distances, the LSAP effectively maximizes the total distances between assigned component-position pairs. This results in assigning components with the most dissimilar TTA characteristics to different positions. After solving the LSAP, we assign components to groups in a round-robin manner based on their assigned positions. So, components with assignments  $q, r+q, 2r+q$ , and so on are assigned to group  $k$ , which is denoted by  $G_q$ . This method ensures that components assigned to consecutive positions are distributed across different groups, leading to each group containing components with maximally diverse TTA characteristics. This diversity helps delay the simultaneous absorption of multiple components within a group, leading to an improved approximation of the optimal policy. Furthermore, this LSAP-based partitioning ensures uniformity in aggregate statistics across groups, thereby justifying the proportional budget distribution. Moreover, this partitioning decouples the large multi-component MDP into  $r$  smaller multi-component MDPs with a capacity constraint of 1. Thus, finding the optimal policy for each group is computationally more feasible due to the reduced combinatorial complexity resulting from the simplified capacity constraints.

## 4.2 Meta-RL for Partitioned Multi-Component Monotonic MDPs

In Section 4.1, we introduced our methodology for partitioning the components of a large budget and capacity-constrained multi-component monotonic MDP, into smaller groups. In this section, we propose a meta-RL-based approach for obtaining an approximately optimal policy for each group. Analogous to the budgeted-POMDP (bPOMDP) framework proposed in (Vora et al. 2023), we model each group as a budgeted-MDP (bMDP) to enforce adherence to budget constraints. In a bMDP, the available budget is augmented as an extra component to the state vector at each time-step. Note that due to the LSAP assignment and grouping process, the indexing of components within each group may differ from their original indices in the multi-component MDP. For notational convenience, we reindex the components within each group  $G_q$  from 1 to  $m_q$ , where  $m_q$  is the number of components in group  $G_q$ . Thus for a

group  $G_q$  with  $m_q$  components and allocated budget  $b$ , the state at time  $k$ ,  $s_k^{g_q}$ , is given by:

$$s_k^{g_q} = [s_k^{g_q,1}, s_k^{g_q,2}, \dots, s_k^{g_q,m}, b_k], \quad (4)$$

where  $s_k^{g_q,i}$  denotes the state of component  $i$  in group  $G_q$  at time  $k$  with  $i \in \{1, 2, \dots, m\}$ , and  $b_k$  denotes the budget available at time  $k$  (with  $b_0 = b$ ). In practice, however, different representations of states might affect the performance of reinforcement learning algorithms (Lesort et al. 2018). For our work, we empirically observe that the following representation helps the RL agent distinguish between the component states and the available budget more effectively:

$$s_k^{g_q} = \begin{bmatrix} s_k^{g_q,1} & s_k^{g_q,2} & \dots & s_k^{g_q,m} \\ b_k & b_k & \dots & b_k \end{bmatrix}^\top, \quad (5)$$

The action vector for group  $g_q$  at time  $k$  is denoted by  $a_k^{g_q}$  and follows the definition of  $a_k$  given in Section 3.

To obtain the approximately optimal policy for each group, we use a Proximal Policy Optimization (PPO) (Schulman et al. 2017) algorithm. The reward function for this PPO agent is defined as:

$$R(s_k^{g_q}, a_k^{g_q}) = \begin{cases} r_1 < 0, & \text{if } b_k < 0, \\ r_2 = -(H - k), & \text{if } s_k^{g_q,i} = 0 \text{ for any } i, \\ r_3 < 0, & \text{if } \sum_{i=1}^m \mathbb{1}(a_k^{g_q,i}) > 1 \\ r_4 = k - \alpha q_k, & \text{if } s_k^{g_q}, b_k > 0, \end{cases}$$

with  $q_k = \sum_{i=1}^m s_k^{g_q,i} \times \mathbb{1}(a_k^{g_q,i})$ ,  $|r_1| \geq |r_3| > |r_2| > |r_4|$  for all  $k$  and  $0 < \alpha < 1$ . Furthermore,  $H$  is the mission-specific planning horizon. Rewards  $r_1$  and  $r_3$  promote adherence to budget and capacity constraints, respectively. Furthermore,  $r_2$  penalizes the agent when one or more of the components reach the absorbing state, with the penalty being higher for smaller values of  $k$ . The reward signal  $r_4$  provides a positive reward equal to the time-step  $k$ , while imposing a penalty proportional to the state values of the components for which the agent chooses restorative actions. This signal incentivizes the agent to maintain  $s_k^{g_q} > 0$  for as long as possible, while discouraging the unnecessary usage of restorative actions on components with high state values.

To generalize this PPO agent across groups of components and budget values, we employ a meta-training procedure which involves iterative updates of the agent's policy network parameters over a randomly selected set of component groups and allocated budgets. Using this meta-PPO trained agent, we obtain an approximately optimal policy  $\pi^{g_q,*}$  for each group  $g_q$ . The overall policy for the large multi-component MDP is given by:

$$\pi^*(s_k, a_k) = (\pi^{g_0,*}(s_k^{g_0}, a_k^{g_0}), \dots, \pi^{g_{r-1},*}(s_k^{g_{r-1}}, a_k^{g_{r-1}})).$$

Since the indexing of components within each group differs from their original indices in the multi-component MDP, care must be taken when applying the final policy to ensure the correct mapping between the LSAP groups and the original components. While this policy is not guaranteed to be globally optimal for the entire multi-component MDP, our empirical results show that it performs well in practice while satisfying the budget and capacity constraints.

## 5 Implementation and Evaluation

In this section, we validate the proposed approach by determining an approximately optimal policy for a budget- and capacity-constrained multi-component monotonic MDP with a very large number of components. We compare the performance of the proposed approach against existing baselines in the context of planning and scheduling repairs for a large swarm of industrial robots. Additionally, we perform a computational complexity analysis to demonstrate the scalability of the proposed approach for varying number of components.

We consider a scenario involving the maintenance of a swarm of  $n$  industrial robots, including assembly robots, picking and packing robots and welding robots, managed by a team of  $r$  repair technicians. Each robot's health ranges from 0 to 100 and is modeled using the Condition Index (CI) (Grussing, Uzarski, and Marrano 2006). Motivated by the work of Grussing, Uzarski, and Marrano (2006) on modeling infrastructure component deterioration, we model the stochastic deterioration of each robot's CI over time using the Weibull distribution. At each time step, robots can be repaired to improve their CI. The robot swarm is considered non-operational when at least one robot has a CI of 0. The swarm is allocated a repair budget of  $B$  units for a planning horizon of 100 decision steps, with each repair costing 1 unit. Due to the limited number of repair technicians, a maximum of  $r$  robots can be repaired at any given time step. Initially, all the robots have a CI of 100. To evaluate the scalability and effectiveness of our approach, we conduct experiments for various  $(n, r)$  pairs. The objective of the repair team is to maximize the operational time of the swarm by efficiently choosing the subset of robots to repair at each time step, while adhering to the budget and capacity constraints. This objective is modeled as a budget and capacity-constrained multi-component monotonic MDP, where the state of each component MDP corresponds to the CI of an individual robot. The state and action vectors for this multi-component MDP follow the formulation given in Section 4.2. Consequently, the problem is a constrained combinatorial optimization problem, as described in (1). For a scenario with  $n = 1000$  and  $r = 300$ , this optimization problem has approximately 100,000 binary decision variables for a planning horizon of 100 steps.

### 5.1 Partitioning of Multi-Component MDP

We first evaluate the performance of the LSAP-based grouping method for partitioning the robot swarm into smaller groups. This method is compared with a baseline partitioning approach, which involves randomly assigning robots to groups. The LSAP-optimized position indices of the robots are obtained by solving (3), and the assignment of robots to groups is performed using these indices, as described in Section 4.1. For each robot  $i$  in the swarm, we compute its  $TTA$  statistics by averaging the  $\mu_i$  and  $\sigma_{\mu_i}^2$  over 1000 independent Monte-Carlo simulations. The solution to (3) is obtained using the `linear_sum_assignment` function of the `scipy.optimize` module in Python, which implements a modified version of the Jonker-Volgenant algorithm (Crouse 2016).

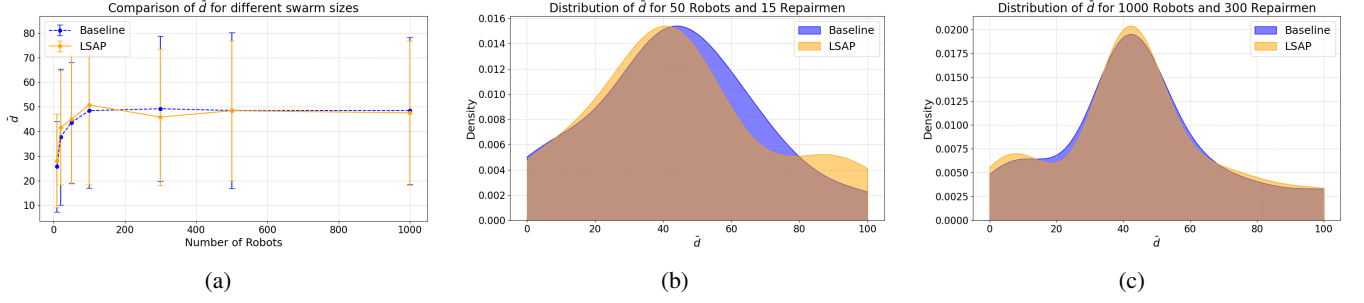


Figure 2: (a) Variation of  $\bar{d}$  values for different swarm and repair team sizes, where the error bars denote the variance of  $\bar{d}$ . (b) Distribution of  $\bar{d}$  for a swarm of 50 robots and 15 repair technicians. (c) Distribution of  $\bar{d}$  for a swarm of 1000 robots and 300 repair technicians.

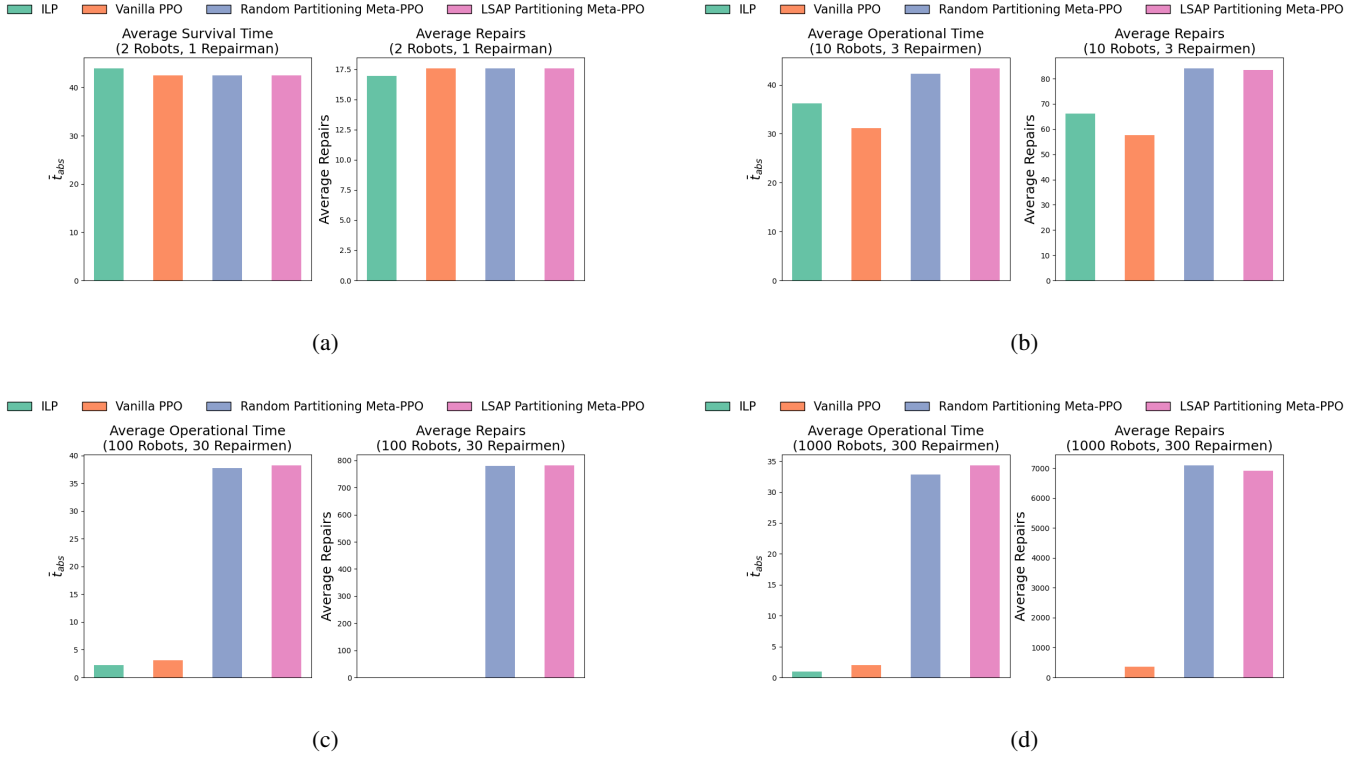


Figure 3: Average operational time and average repair counts for all four approaches across the four scenarios: (a)  $(n, r) = (2, 1)$ . (b)  $(n, r) = (10, 3)$ . (c)  $(n, r) = (100, 30)$ . (d)  $(n, r) = (1000, 30)$ .

To compare the performance of the proposed partitioning approach with the baseline, we use a metric denoted by  $\bar{d}$ . For a given swarm size, we compute the average pairwise distance between robots within each group, and then further average these distances across all groups. This overall average in-group distance is denoted as  $\bar{d}$ . A higher value of  $\bar{d}$  indicates greater diversity within each group in terms of the TTA statistics of the robots. We compare the values of  $\bar{d}$  obtained using both the proposed and baseline approaches for 6 different  $(n, r)$  pairs. Table 1 shows the variation in  $\bar{d}$ , achieved using the LSAP-based and baseline partitioning approaches, for different swarm and repair team sizes, maintaining a constant  $n$  to  $r$  ratio. Figure 2a presents a graphical

representation of this variation. We observe that the LSAP-based assignment approach achieves higher values of  $\bar{d}$  for smaller swarm and repair team sizes. However, as the number of robots and repair technicians increase, both the proposed approach and baseline result in similar values of  $\bar{d}$ . This trend is further illustrated in the distribution of  $\bar{d}$  for specific scenarios. Figure 2b shows the distribution of  $\bar{d}$  for a swarm of 50 robots with 15 repair technicians, while Figure 2c shows the distribution for 1000 robots and 300 repair technicians. In the case of larger swarm sizes, the distributions for the two approaches are very similar.

The similarity in  $\bar{d}$  values for larger swarm and repair team sizes can be attributed to the inherent diversity in

Swarm Size	Repair Team Size	LSAP	Baseline
10	3	<b>28.13</b>	25.67
20	6	<b>41.51</b>	37.70
50	15	<b>44.93</b>	43.59
100	30	<b>50.73</b>	48.39
300	90	47.78	<b>49.20</b>
500	150	<b>48.42</b>	48.40
1000	300	47.51	47.51

Table 1: Average *TTA* statistics pairwise distance  $\bar{d}$  (in steps) for 6 different  $(n, r)$  pairs, achieved under the proposed and baseline partitioning approaches.

*TTA* characteristics that is naturally induced by large swarm sizes. This observation is supported by our theoretical analysis (see Section 7.1), where we prove that, under certain assumptions, any partitioning method will result in partitions with approximately equal average pairwise distances as the ratio  $\frac{n}{r}$  grows faster than  $\ln r$ .

Finally, the smooth and continuous distribution of *TTA* characteristics in our context leads to only a small difference in performance of the two approaches. However, in scenarios where data exhibits distinct clusters or broader variability, LSAP can yield significantly higher  $\bar{d}$  values by optimizing group diversity.

## 5.2 Meta-PPO for Repair Policy Synthesis

Next, we demonstrate the effectiveness of the proposed two-step approach for obtaining the approximately optimal maintenance policy for the robot swarm. As mentioned in Section 4, our approach involves partitioning the swarm into groups using an LSAP method and synthesizing a repair policy for each group using a meta-PPO agent. We compare the proposed approach against the following three baselines:

1. **Random Assignment:** Robots are randomly assigned to groups, and a meta-PPO agent is then used to determine the repair policy.
2. **Vanilla PPO:** The repair policy is derived for the entire swarm without partitioning, using a standard PPO agent.
3. **Integer Linear Programming (ILP):** The repair policy is obtained by solving the constrained combinatorial optimization problem (1) directly using the GUROBI optimizer.

In the vanilla PPO approach, a separate agent is trained for each  $(n, r)$  pair. For the ILP approach, we use GUROBI to attempt to find an optimal solution to (1). However, due to computational limitations, GUROBI can only provide optimal solutions for small problem sizes. As the problem size increases, it employs approximate methods and heuristics, which can lead to suboptimal solutions. Therefore, the ILP solution serves as a benchmark for comparison in terms of performance and optimality primarily for smaller problem instances where exact solutions can be obtained. We use the objective function of the constrained combinatorial optimization problem (1) as the metric to compare the performance of the four algorithms. This metric, denoted by  $\bar{t}_{abs}$ ,

represents the average operational time of the swarm and is averaged over 100 independent runs for each scenario. Additionally, we also compare the average number of repairs performed over the planning horizon.

We conduct experiments for four different  $(n, r)$  pairs: (2, 1), (10, 3), (100, 30) and (1000, 300). To ensure comparability, we run the ILP solver for the same number of computation steps, as the proposed approach, in each scenario. Figure 3 presents the performance of the four methods, in terms of average operational time and average number of repairs performed, for the above mentioned scenarios. We observe that for the scenario with 2 robots and 1 repairman, the ILP solution yields the highest value of  $\bar{t}_{abs}$ , as expected, since this small problem size allows for near-optimal solutions. The other three approaches—LSAP-based meta-PPO, random assignment meta-PPO, and vanilla PPO—produce slightly lower, but comparable, values of  $\bar{t}_{abs}$ . Since the capacity constraint is one in this scenario, partitioning is unnecessary, leading all three approaches to yield comparable performance. However, as the number of robots and repair technicians increases, the performance of the ILP and vanilla PPO approaches deteriorate significantly. For large scenarios, such as the one with 1000 robots and 300 repair technicians, GUROBI produces poor results within the provided time frame, as shown in Figure 3d. Even if the ILP solver were allowed up to ten times more runtime, it would still fail to find a high-quality solution, underscoring the scalability challenges of the ILP approach. The vanilla PPO and ILP-based methods scale poorly with increasing number of robots due to an exponential increase in the state and action space. In contrast, both the LSAP-based and random assignment meta-PPO approaches maintain relatively stable performance. For larger scenarios, these two approaches result in similar values of  $\bar{t}_{abs}$ . This similarity in performance is due to the fact that, as the swarm size increases, both approaches tend to generate partitions with similar diversity in terms of the robots’ *TTA* characteristics, as discussed in Section 5.1.

## 5.3 Computational Complexity Analysis

Finally, we perform a computational complexity analysis to demonstrate the scalability of the proposed approach across different  $(n, r)$  pairs. The computational complexity experiments were conducted in Python on a laptop running MacOS with an M2 chip @3.49GHz CPU and 8GB RAM. Table 2 summarizes the time taken (in seconds) for each step of the proposed approach for varying number of robots and repair technicians. We observe that the time taken for the LSAP-based partitioning step is negligible in comparison to the time required for generating policies using the meta-PPO agent, especially as the number of robots increases. Since the second step involves applying a pre-trained meta-PPO model to each group, the time complexity for this step scales linearly with the number of robots in each group. Therefore, the overall complexity of our algorithm is expected to be linear in the number of robots, i.e.,  $O(n)$ . This hypothesis is confirmed by the log-log plot shown in Figure 4, which depicts the computational complexity of the proposed approach as the number of robots increases. The trend in the



Swarm Size ( $n$ )	Repair Team Size ( $r$ )	LSAP Partitioning	Meta-PPO
2	1	0.5001	4.4594
5	2	0.3905	3.4927
10	3	0.4464	4.9642
20	6	0.3992	8.1406
50	15	0.3956	17.4025
100	30	0.8295	32.1637
500	150	0.6516	182.0258
1000	300	1.2324	290.1350

Table 2: Time taken (in seconds), averaged over 10 runs, for running each step of the proposed approach for varying number of robots and repair technicians.

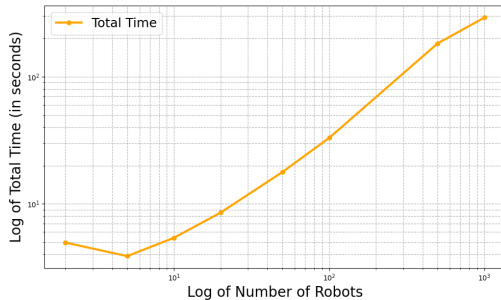


Figure 4: Log-log plot showing the computational complexity of the proposed approach for varying number of robots.

log-log plot demonstrates the linear scalability of our algorithm with respect to the swarm size. Additionally, Figure 5 presents a heatmap that captures the variation in computational time for different combinations of  $(n, r)$  pairs. We observe that the time taken to run the proposed approach is more sensitive to changes in the number of robots than to changes in the number of repair technicians. This finding indicates that the overall computational complexity is dominated by the number of robots, with the impact of the number of repair technicians being relatively small.

## 6 Conclusions

In this paper we present a computationally efficient and scalable algorithm for solving very large budget and capacity-constrained multi-component monotonic MDPs. For such an MDP, the individual component MDPs are coupled due to the shared budget as well as due to the capacity constraints which limit the number of restorative actions that can be performed at a given time step. Existing approaches that decouple the multi-component MDP by pre-allocating the budget, cannot be directly applied due to the additional combinatorial complexity introduced by the capacity constraints. To address this challenge, we first partition the large multi-component MDP into groups, with the number of groups being equal to the capacity limit. This partitioning is achieved using a Linear Sum Assignment approach, which ensures that the components within each group exhibit as diverse transition probability behaviors as possible. The total budget is then distributed among the groups in proportion to their

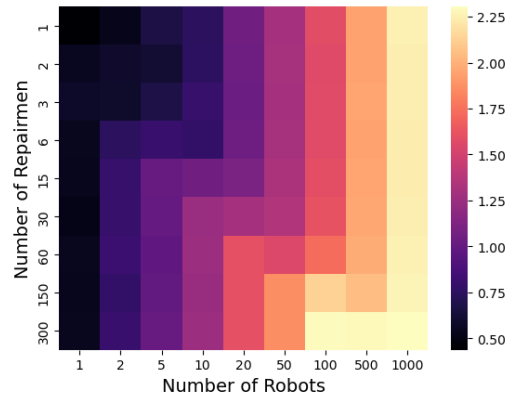


Figure 5: Heatmap showing the time taken for the proposed approach, across different numbers of robots and repair technicians. Values are plotted on a logarithmic scale to better capture the variation.

size. Finally, we use a meta-PPO agent to obtain an approximately optimal policy for each group. Experimental results from a real-world scenario, involving the maintenance of a swarm of industrial robots by a team of repair technicians, demonstrate that our approach outperforms the baselines, especially in larger problem instances. The performance gap is significantly higher for non-partitioning approaches, underscoring the importance of partitioning the multi-component MDP. Additionally, the computational complexity analysis shows that the proposed method scales linearly with the number of components, making it highly scalable for large-scale applications. Future work will focus on extending the algorithm’s capabilities to scenarios involving hierarchical budget constraints along with the capacity constraints.

## References

- Altman, E. 2021. *Constrained Markov Decision Processes*. Routledge.
- Borgi, T.; Hidri, A.; Neef, B.; and Naceur, M. S. 2017. Data analytics for predictive maintenance of industrial robots. In *2017 International Conference on Advanced Systems and Electric Technologies*, 412–417.
- Boutilier, C.; and Lu, T. 2016. Budget Allocation Using



Weakly Coupled, Constrained Markov Decision Processes. In *UAI*.

Burkard, R. E.; and Cela, E. 1999. Linear assignment problems and extensions. In *Handbook of combinatorial optimization: Supplement volume A*, 75–149. Springer.

Carrara, N.; Leurent, E.; Laroche, R.; Urvoy, T.; Maillard, O.-A.; and Pietquin, O. 2019. Budgeted reinforcement learning in continuous state space. In *32nd Advances in Neural Information Processing Systems*.

Chen, Y.; Liu, Y.; and Xiahou, T. 2021. A deep reinforcement learning approach to dynamic loading strategy of repairable multistate systems. *IEEE Transactions on Reliability*, 71(1): 484–499.

Crouse, D. F. 2016. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4): 1679–1696.

Floudas, C. A.; and Lin, X. 2005. Mixed integer linear programming in process scheduling: Modeling, algorithms, and applications. *Annals of Operations Research*, 139: 131–162.

Grussing, M. N.; Uzarski, D. R.; and Marrano, L. R. 2006. Condition and reliability prediction models using the Weibull probability distribution. In *Applications of Advanced Technology in Transportation*, 19–24.

Haksar, R. N.; and Schwager, M. 2018. Controlling large, graph-based MDPs with global control capacity constraints: An approximate LP solution. In *57th Conference on Decision and Control*, 35–42.

Hassan, M.; Liu, D.; and Paul, G. 2018. Collaboration of multiple autonomous industrial robots through optimal base placements. *Journal of Intelligent & Robotic Systems*, 90: 113–132.

Hung, Y.-H.; Shen, H.-Y.; and Lee, C.-Y. 2024. Deep reinforcement learning-based preventive maintenance for repairable machines with deterioration in a flow line system. *Annals of Operations Research*, 1–21.

Kalagarla, K. C.; Jain, R.; and Nuzzo, P. 2021. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *35th AAAI Conference on Artificial Intelligence*, 9, 8030–8037.

Kibira, D.; and Qiao, G. 2023. Degradation Modeling of a Robot Arm to Support Prognostics and Health Management. In *International Manufacturing Science and Engineering Conference*.

Korte, B. H.; Vygen, J.; Korte, B.; and Vygen, J. 2011. *Combinatorial Optimization*. Springer.

Lesort, T.; Díaz-Rodríguez, N.; Goudou, J.-F.; and Filliat, D. 2018. State representation learning for control: An overview. *Neural Networks*, 108: 379–392.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Papadimitriou, C. H.; and Steiglitz, K. 1998. *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation.

Papakonstantinou, K. G.; and Shinozuka, M. 2014. Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory. *Reliability Engineering & System Safety*, 130: 202–213.

Perlman, Y.; Mehrez, A.; and Kaspi, M. 2001. Setting expediting repair policy in a multi-echelon repairable-item inventory system with limited repair capacity. *Journal of the Operational Research Society*, 52(2): 198–209.

Qiao, G.; and Weiss, B. A. 2018. Quick health assessment for industrial robot health degradation and the supporting advanced sensing development. *Journal of Manufacturing Systems*, 48: 51–59.

Schrijver, A. 1998. *Theory of Linear and Integer Programming*. John Wiley & Sons.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Solozabal, R.; Ceberio, J.; and Takáč, M. 2020. Constrained combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:2006.11984*.

Sutton, R. S. 2018. Reinforcement Learning: An Introduction. *A Bradford Book*.

Toth, P. 2000. Optimization engineering techniques for the exact solution of NP-hard combinatorial optimization problems. *European Journal of Operational Research*, 125(2): 222–238.

Vora, M.; Grussing, M. N.; and Ornik, M. 2024. Solving Truly Massive Budgeted Monotonic POMDPs with Oracle-Guided Meta-Reinforcement Learning. *arXiv preprint arXiv:2408.07192*.

Vora, M.; Thangeda, P.; Grussing, M. N.; and Ornik, M. 2023. Welfare Maximization Algorithm for Solving Budget-Constrained Multi-Component POMDPs. *IEEE Control Systems Letters*, 7: 1736–1741.

Wu, D.; Chen, X.; Yang, X.; Wang, H.; Tan, Q.; Zhang, X.; Xu, J.; and Gai, K. 2018. Budget constrained bidding by model-free reinforcement learning in display advertising. In *27th ACM International Conference on Information and Knowledge Management*, 1443–1451.

Xiao, S.; Guo, L.; Jiang, Z.; Lv, L.; Chen, Y.; Zhu, J.; and Yang, S. 2019. Model-based constrained MDP for budget allocation in sequential incentive marketing. In *28th ACM International Conference on Information and Knowledge Management*, 971–980.

## 7 Appendix

### 7.1 Almost Sure Asymptotic Contraction of Partition Averages

**Theorem 1.** Let  $X \in \mathbb{R}^{nk \times nk}$  be a symmetric random matrix where the entries  $\{X_{ij} \mid i \leq j\}$  are independently and identically distributed (i.i.d.) standard normal random variables, i.e.,  $X_{ij} \sim \mathcal{N}(0, 1)$  for  $i \leq j$ , and  $X_{ji} = X_{ij}$  for  $i > j$ . Let  $\mathcal{G} = (V, E)$  be a complete graph with  $nk$  nodes, where  $V$  denotes the set of vertices and  $E$  denotes the set of edges, and the adjacency matrix is given by  $X$ .

Partition the node set  $V$  into  $n$  disjoint subsets  $V_1, V_2, \dots, V_n$ , each of size  $k$ . For each  $t \in \{1, 2, \dots, n\}$ , let  $E_t$  be the edges in the subgraph induced by  $V_t$ . Let  $S_t$  denote the average of the edge weights within subgraph  $V_t$ .

Then, for any fixed  $\epsilon > 0$ , if  $k = \omega(\ln n)$  (i.e.,  $k$  grows faster than  $\ln n$ ), the probability that there exist  $t \neq l$  such that  $|S_t - S_l| > \epsilon$  tends to zero as  $n \rightarrow \infty$ . That is,

$$\lim_{n \rightarrow \infty} \max_{V_1, V_2, \dots, V_n} P(\exists t \neq l \text{ such that } |S_t - S_l| > \epsilon) = 0. \quad (6)$$

*Proof.* The number of edges in  $E_t$  is

$$|E_t| = \binom{k}{2} = \frac{k(k-1)}{2}. \quad (7)$$

Also, since  $S_t$  is the average of the edge weights within the subgraph induced by  $V_t$ , we have:

$$S_t = \frac{1}{|E_t|} \sum_{(i,j) \in E_t} X_{ij} = \frac{1}{\binom{k}{2}} \sum_{(i,j) \in E_t} X_{ij}. \quad (8)$$

For each  $t$ ,  $S_t$  is the average of  $|E_t| = \binom{k}{2}$  i.i.d. standard normal random variables  $X_{ij} \sim \mathcal{N}(0, 1)$ . Therefore,  $S_t$  is normally distributed with mean zero and variance

$$\sigma^2 = \frac{1}{|E_t|} = \frac{2}{k(k-1)}. \quad (9)$$

Similarly, the difference between any two such averages  $S_t$  and  $S_l$  (for  $t \neq l$ ) is also normally distributed with mean zero and variance  $2\sigma^2$ :

$$S_t - S_l \sim \mathcal{N}(0, 2\sigma^2). \quad (10)$$

Using the Chernoff bound we have:

$$\begin{aligned} P(|S_t - S_l| > \epsilon) &\leq P(|S_t| > \frac{\epsilon}{2}) + P(|S_l| > \frac{\epsilon}{2}) \\ &\leq 4 \exp\left(-\frac{\epsilon^2 k(k-1)}{32}\right). \end{aligned} \quad (11)$$

There are at most  $\binom{nk}{k}$  possible choices for  $S_t$ . Using the upper bound on the binomial coefficient, we have:

$$\binom{nk}{k} \leq \left(\frac{enk}{k}\right)^k = \exp(k \ln(en)).$$

Therefore, the total number of  $(S_t, S_l)$  pairs is at most

$$\binom{nk}{k}^2 \leq \exp(2k \ln(en)). \quad (12)$$

Applying the union bound over all partitions and all pairs  $(S_t, S_l)$  within each partition, the probability that there exists a partition with at least one pair  $(S_t, S_l)$  such that  $|S_t - S_l| > \epsilon$  is bounded by

$$\begin{aligned} &\max_{V_1, V_2, \dots, V_n} P(\exists t \neq l \text{ such that } |S_t - S_l| > \epsilon) \\ &\leq \binom{nk}{k} \cdot \binom{nk}{k} \cdot 4 \exp\left(-\frac{\epsilon^2 k(k-1)}{32}\right). \end{aligned} \quad (13)$$

Simplifying (13), we get:

$$\begin{aligned} &\max_{V_1, V_2, \dots, V_n} P(\exists t \neq l \mid |S_t - S_l| > \epsilon) \\ &\leq 4 \exp\left(2k(1 + \ln n) - \frac{\epsilon^2 k(k-1)}{32}\right). \end{aligned} \quad (14)$$

The exponent in (14) tends to  $-\infty$  if  $k$  grows faster than  $\ln n$ , i.e.,  $k = \omega(\ln n)$ . This means that the right-hand side (RHS) of (14) goes to 0.

Thus, if  $k$  grows faster than  $\ln n$ , the probability that any two group averages  $S_t$  and  $S_l$  differ by more than  $\epsilon$  tends to zero. This implies that, asymptotically, all  $S_t$  are approximately equal with high probability regardless of how the partitions  $V_t$  are formed. Therefore, under these conditions, any partitioning method will result in groups with similar average pairwise distances, making the partitioning algorithm-agnostic.  $\square$