

# Manufacturing Dispatching using Reinforcement and Transfer Learning

Shuai Zheng<sup>(✉)</sup>, Chetan Gupta, and Susumu Serita

Industrial AI Lab, Hitachi America Ltd  
Santa Clara, CA, USA

{Shuai.Zheng,Chetan.Gupta,Susumu.Serita}@hal.hitachi.com

**Abstract.** Efficient dispatching rule in manufacturing industry is key to ensure product on-time delivery and minimum past-due and inventory cost. Manufacturing, especially in the developed world, is moving towards on-demand manufacturing meaning a high mix, low volume product mix. This requires efficient dispatching that can work in dynamic and stochastic environments, meaning it allows for quick response to new orders received and can work over a disparate set of shop floor settings. In this paper we address this problem of dispatching in manufacturing. Using reinforcement learning (RL), we propose a new design to formulate the shop floor state as a 2-D matrix, incorporate job slack time into state representation, and design lateness and tardiness rewards function for dispatching purpose. However, maintaining a separate RL model for each production line on a manufacturing shop floor is costly and often infeasible. To address this, we enhance our deep RL model with an approach for dispatching policy transfer. This increases policy generalization and saves time and cost for model training and data collection. Experiments show that: (1) our approach performs the best in terms of total discounted reward and average lateness, tardiness, (2) the proposed policy transfer approach reduces training time and increases policy generalization.

**Keywords:** Reinforcement Learning · Transfer Learning · Dispatching

## 1 Introduction

In a manufacturing process, a production order moves through a sequence of job processing steps to arrive at a final product. The problem of dispatching is the assigning the next job to be processed for a given machine. Inefficient scheduling and dispatching can cause past-due cost (past-due cost is the cost when a job cannot be delivered on time) as well as inventory cost (inventory cost is the storage cost when the job is finished before due time) to go up. It is obvious that all manufacturing managers want on-time delivery and minimum past-due and inventory cost. However, achieving these goals requires efficient production scheduling that minimizes these costs. Furthermore, manufacturing is moving towards a high mix low volume product mix, which makes this even

more challenging, due to an ever-evolving product mix, causing larger variations in job types and job arrival rates.

Production scheduling problems can be categorized by shop configurations and scheduling objectives. Depending on the number of processing machines, there can be single-machine environment [1] and parallel-machine environment [2]. Depending on the number of operations (or stages, steps) of each job, there are single-operation and multi-operation environment [3]. Depending on the objective of scheduling, there are completion time based scheduling [4] (trying to increase machine efficiency) and due date based scheduling [5] (trying to be close to promised delivery dates). Multi-operation parallel-machine problems can be solved through multi-agent algorithms or can be decomposed into solving several single-operation problems [6]. In this work, we focus on dynamic dispatching for due date based objective, which has broader generalization and can be used in different shop floor settings. Dynamic dispatching is also critical for predictive maintenance tasks [7, 8], which schedule machine maintenance using data-driven and model-based approaches. Furthermore, dispatching only schedules the imminent job with the highest priority and is particularly suitable for dynamic/stochastic environments due to its low computational time in deployment/testing stage.

Traditionally to address the problem of dispatching, a lot of hyper-heuristics have been proposed and shown to be effective and reusable in different shop conditions [6]. Even though some exact solutions for deterministic scheduling problems are available, manufacturing shop floor depends on heuristics. This is because exact solutions are computationally expensive (and hence infeasible) in deployment stage and cannot solve problems for dynamic and stochastic environments. Since heuristics are very problem-specific and usually achieved by trial and error, hyper-heuristics [6] which automate the design of heuristics attract a lot of interest and have been shown effective in manufacturing as well as other industries such as: bin packing [9], vehicle routing [10], project scheduling [11]. Many hyper-heuristics are based on machine learning approaches, such as neural networks, logistic regression, decision trees, Support Vector Machines, genetic algorithms, genetic programming, reinforcement learning.

Recently, existing works have used deep reinforcement learning for scheduling problems [12, 13]. For dispatching purpose, we propose a new design to formulate the shop floor state as a 2-D matrix, incorporate job slack time into state representation. We also design lateness and tardiness rewards function for reinforcement learning. In a manufacturing shop, there are many similar production lines, where designing and maintaining a separate reinforcement learning model for each line is not feasible. To address this, we propose a transfer approach for dispatching policy using manifold alignment. Unlike discriminant subspace learning [14–16] and sparse enforced learning [17–19] where the purpose is to separate classes, manifold alignment [20] learns a subspace by matching the local geometry and preserving the neighborhood relationship within each space. In summary, the contributions of this work are:

1. The reinforcement learning module uses deep learning and policy gradient to minimize job due related cost. Compared to existing work, the novelty is that we formulate the shop floor state as a 2-D matrix, incorporate job slack time into state representation, and design lateness and tardiness rewards function for dispatching purpose.
2. The transfer learning module transfers dispatching policy between shop floors using manifold alignment. Compared to existing work, the novelty is that our approach formulates shop floor states transfer using manifold alignment, and we design a method to recover actions from sequence of states.

To test our approach, we build a simulator to simulate dynamic factory settings, where we can change factory configurations and job characteristics, such as job queue length, machine processing capacity, job arrival distribution. A simulator is needed since it is very difficult to run live experiments on an actual production line. Another option is to somehow use actual data in a shop floor, however this is a challenge: the data collected on the shop floor is mainly state and scheduling decision record data which is static in nature and cannot provide feedback on the scheduling decisions. These data can be used for hyper-heuristics rules extraction, but not for *dynamic* environment [6]. Furthermore, corporations are reluctant to share real data, because these data include trade secrets and sensitive financial information. In fact, most published works for *dynamic* production scheduling are based on simulators [6]. Using simulator to pre-train model and then deploying it into real factories can further reduce data collection and training effort.

## 2 Related work

### 2.1 Production scheduling

Hyper-heuristics are promising to handle dynamic and stochastic scheduling problems and have recently emerged as a powerful approach to automate the design of heuristics for production scheduling [6]. The basic idea of hyper-heuristics is to learn scheduling rules from a set of very good training scheduling instances. These training instances are considered optimal or perfect. Hyper-heuristics can then replicate these scheduling instances as closely as possible. Depending on the learning method, hyper-heuristics can be classified into supervised learning and unsupervised learning. Examples of supervised learning hyper-heuristics include neural networks [21], logistic regression [22], decision trees [23], Support Vector Machines [24], etc. Genetic algorithms (GA) and Genetic Programming (GP) are evolutionary computation methods and have been used for dynamic job shop scheduling problems. Reinforcement learning is an efficient algorithm to learn optimal behaviors through reward feedback information from dynamic environments [13].  $TD(\lambda)$  based reinforcement learning was used for manufacturing job shop scheduling to improve resource utilization [25]. Resource scheduling for computer clusters is also very related, such as Tetris [26] and resource scheduling in HPC [27]. RL-Mao [12] uses policy gradient reinforcement learning to reduce

computer job slowdown. The difference of our work and RL-Mao lies in that our work integrates manufacturing job slack time into state representation and the objective functions using lateness and tardiness are specifically designed for manufacturing dispatching.

## 2.2 Reinforcement learning background

In reinforcement learning, an agent interacts with an environment  $\mathcal{E}$  over many discrete time steps [28]. The state space of  $\mathcal{E}$  is defined within  $\mathcal{S}$ . At each time step  $t$ , the agent receives a state  $s_t \in \mathcal{S}$  and performs an action  $a_t \in \mathcal{A}$  following a policy  $\pi$ , where  $\mathcal{A}$  defines the action space of this agent. The agent receives a reward  $r_t$  for this action and a new state  $s_{t+1}$  is then presented to the agent. The policy  $\pi$  is a mapping function from states  $s_t$  to  $a_t$ , denoted by  $\pi(a_t|s_t)$ , which gives the probability of taking action  $a_t$ . This process continues until the agent reaches a termination state or time  $t$  exceeds a maximum threshold. The cumulative discounted reward starting from time  $t$  is defined as:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \quad (1)$$

where  $\gamma \in (0, 1]$  is a discounted factor. The goal of a reinforcement agent is to obtain a policy which maximizes the expected total discounted reward starting from time  $t = 0$ :

$$J(\pi) = \mathbb{E}[R_0|\pi]. \quad (2)$$

In policy-based model-free reinforcement learning, policy is directly parameterized as a function from states to actions,  $\pi(a|s; \theta)$ , where parameter  $\theta$  is updated using gradient ascent on  $\mathbb{E}[R_t|\pi]$ . One example of this category is the REINFORCE algorithm [29, 30]. Using the policy gradient theorem [29], the gradient with respect to  $\theta$  can be given as  $\nabla_{\theta} \log \pi(a_t|s_t; \theta) R_t$ , which is an unbiased estimate of  $\nabla_{\theta} \mathbb{E}[R_t|\pi]$ . In order to reduce the variance of this estimate, we can subtract a baseline from the return, where baseline  $b_t(s_t)$  is a learned function of state. The resulting gradient is thus given as

$$\nabla_{\theta} \log \pi(a_t|s_t; \theta) (R_t - b_t). \quad (3)$$

The term  $R_t - b_t$  is used to scale the policy gradient and can be seen as the advantage of action  $a_t$  in state  $s_t$ .

## 3 Dispatching with reinforcement learning

### 3.1 Problem description

**Objective.** Dispatching performance can be evaluated in terms of lateness  $L$  and tardiness  $TA$ . The objective of this problem is to minimize the average

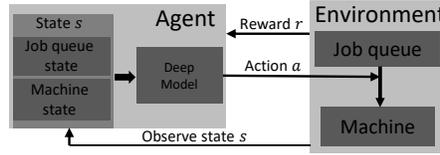


Fig. 1: Overall design of Deep Manufacturing Dispatching (DMD).

lateness and tardiness of all jobs. For one single job, let the completion time be  $c$ . Lateness is the absolute difference between job due time and job completion time:

$$L = |c - d|, \quad (4)$$

where  $L \in [0, +\infty)$ . Thus lateness considers both inventory and past-due cost. Tardiness only considers the lateness when the job is late. Tardiness  $TA$  is defined as:

$$TA = \max(c - d, 0), \quad (5)$$

where  $TA \in [0, +\infty)$ . Tardiness focuses on past-due cost, but not inventory cost.

**Constraints.** Since this is a single-operation environment, there is no precedence constraint between jobs/operations. Disjunctive constraint is enforced in this problem, so that no two jobs can be processed at the same time. If the job is being processed, it cannot be paused. All jobs are equally important.

### 3.2 Design

Figure 1 shows the overall design of proposed Deep Manufacturing Dispatching (DMD). The environment includes job queue and processing machine. At each time step, reinforcement learning agent observes a state  $s$ , which includes job queue state and machine state (schedule of next  $T$  time steps), then outputs a probability vector with respect to each action using a deep learning model as function approximator. The agent will then perform action  $a$  with the highest probability and receive a reward  $r$  from the environment.

There are  $n$  job slots and  $m$  backlog slots. Each job has a processing time (job length)  $p$  and a due time  $d$ . At each time step, the probability of arriving a new job is  $\lambda \in (0, 1)$ . When a job arrives, it will be placed on one of the  $n$  job slots randomly. If job slots are full, the job will be placed on backlog slots. For jobs placed on backlog slots, only job count can be seen and those jobs cannot be selected by dispatcher. As the backlog design for computer jobs in [12], it is reasonable that the reinforcement agent only considers jobs in job slots, not those on backlog slots, because jobs in job slots arrive earlier and get higher priority. Let  $t_{curr}$  indicate current time. Slack time *slack* of a job is defined as:

$$slack = d - t_{curr} - p. \quad (6)$$



Fig. 2: State representation: this state looks  $T = 5$  time steps ahead of current time and slack time array length is  $Z = 3$ ; there are  $n = 4$  job slots and  $m = 10$  backlog slots. The machine state (schedule of next  $T$  time steps) tells that the next 4 time steps have been scheduled. For the job in job slot 1,  $p = 3$ ,  $slack = 1$ ; for the job in job slot 3,  $p = 2$ ,  $slack = -1$ . There are 8 jobs waiting in backlog. Different colors in machine state represent different jobs.

If  $slack > 0$ , it means that if this job is started now, it will be completed before its due time; if  $slack < 0$ , it means that it will be completed after its due time. We now explain the design details of DMD elements: state space, action space, reward, training.

**State space.** State space includes both states of the machine and job queue. At any time step, we consider the states  $T$  time steps ahead and we use a  $Z$  length array to represent the slack time. We use a 2-D matrix to represent machine state (schedule of next  $T$  time steps) and job queue state. One example is shown in Figure 2. Value 1 in machine state means that the machine at that time step has been allocated, and value 0 means that the machine will be idle. As time proceeds one time step, the  $T$ -length array shifts up by one unit and a new 0 value is appended to the bottom. Job queue state consists of  $n$  job slots and  $m$  backlog. Different colors in machine state indicate different jobs.

Each job slot is represented using two arrays: processing time and slack time. The processing time array is represented using a  $T$  length array, with the number of 1s indicating job length  $p$ . The slack time is represented using a  $Z$  length array, where 1 means positive  $slack$  and  $-1$  means negative  $slack$ . The sum of slack time array represents job  $slack$ .

Backlog state is represented by several  $T$  length array, with a total of  $m$  slots, where each slot represents a job. In Figure 2, backlog is represented using 2 5-length array and there are 8 jobs in backlog. The 0s under machine state and backlog do not represent slack time and are padded to make the state as a complete 2-D matrix. There are 2 benefits using this 2-D representation: 1). 2-D representation can capture the relationship between job characteristics and machine status, and we can use deep learning method to discover hidden patterns from 2-D representations; 2). we can represent schedule in  $T$  time steps ahead of time.

**Action space.** At each time step, the dispatcher performs multiple actions to select a subset of jobs from  $n$  job slots until an invalid or void action is selected. The action  $a$  is a subset of  $\{\emptyset, 1, \dots, n\}$ , where  $\emptyset$  is a void action and no job slot is selected. Void action allows dispatcher to save more resources for larger jobs.

We let the dispatcher make a decision at each time step, instead of only when the machine is idle. This is because jobs are scheduled  $T$  time steps ahead. The decision of the dispatcher is to decide which job to allocate within the next  $T$  time step, not for which job to run next immediately, so we do not consider the machine is idle or not. Instead, we consider the schedule  $T$  time steps ahead.

**Reward.** We design the reward at each time step with respect to lateness and tardiness as:

$$r_L = - \sum_{j \in \mathbb{P}} \frac{L_j}{p_j}, \quad (7)$$

$$r_{TA} = - \sum_{j \in \mathbb{P}} \frac{TA_j}{p_j}, \quad (8)$$

where  $\mathbb{P}$  is set of jobs that are currently being processed by the machine. Summation of  $r_L$  and  $r_{TA}$  over all time steps for the running jobs equals to the total lateness and tardiness.

**Training.** We run the simulator multiple times to get a batch of trajectories. At each time  $t$ , we record the state  $s_t$ , action  $a_t$  and reward  $r_t$ . The discounted reward  $R_t$  at time  $t$  can then be computed using Eq.(1).  $b_t$  is a baseline function, which can be the average reward at time  $t$  of multiple trajectories.

### 3.3 Scalability and generalization

The complexity comes from training and deployment/testing stage. In training stage, the computational time mainly depends on the training of reinforcement learning agent. From Figure 3, 4, we found that the objective function converges fast. Pre-training with simulated data, transfer learning and other accelerated computing [31] can further reduce training time. In deployment/testing stage, the computational time is very fast and the same as testing time of a deep neural network. This ensures the quick response to different shop floor conditions.

## 4 Dispatching policy transfer

In reinforcement learning area, some approaches using manifold learning for cross-domain transfer learning [32–34] have been applied in games and apprentice learning, where control of games, such as Cart Pole and Simple Mass, shares similar control mechanism to balance an object. However, for dispatching problems, due to the complexity of manufacturing settings, it is not straightforward to transfer dispatching rules among factories or product lines. In our proposed Deep Manufacturing Dispatching (DMD) framework, the following features of

---

**Algorithm 1** Policy transfer learning.

---

**Input:** Source environment  $\mathcal{E}_x$ , target environment  $\mathcal{E}_y$ , source optimal policy  $\pi^{x*}$ .**Output:** Target optimal policy  $\pi^{y*}$ .

- 1: Find state projection  $\chi: \mathcal{S}^x \rightarrow \mathcal{S}^y$  using Algorithm 2.
  - 2: Following optimal policy  $\pi^{x*}$ , generate source state trajectories,  $\{s_{00}^x, s_{01}^x, \dots\}$ ,  $\{s_{10}^x, s_{11}^x, \dots\}$ , ...
  - 3: Compute target state trajectories in  $\mathcal{S}^y$ ,  $\{s_{00}^y, s_{01}^y, \dots\}$ ,  $\{s_{10}^y, s_{11}^y, \dots\}$ , ..., using Eq.(9).
  - 4: Recover policy  $\pi^y$  using Algorithm 3.
  - 5: Fine-tune policy  $\pi^y$  and get optimal policy  $\pi^{y*}$ .
- 

the training data will affect the data distribution: 1. factory setting parameters, which are used to describe job queue states and machine states, including length of processing time array  $T$ , length of slack time array  $Z$ , number of job slots  $n$ , number of backlog slots  $m$ , etc.; 2. job characteristics parameters, such as job length distribution, job arrival speed, job due time distribution. To apply a trained policy in a new factory setting or when job characteristics changes, knowledge transfer would greatly improve the performance of learning by avoiding expensive data collection process and reducing training time.

Given source environment  $\mathcal{E}_x$  with state space  $\mathcal{S}^x$  and action space  $\mathcal{A}^x$ , target environment  $\mathcal{E}_y$  with state space  $\mathcal{S}^y$  and action space  $\mathcal{A}^y$ , source optimal policy  $\pi^{x*}(a|s)$  is already learned. Transfer learning can be used to learn target optimal policy  $\pi^{y*}(a|s)$ . There are two types of policy transfer: (1) same-environment transfer, where factory setting parameters are not changed, but job characteristics parameters are changed; (2) cross-environment transfer, where factory setting parameters are changed. For cross-environment transfer, the job queue and machine states are changed, and input dimension of source policy is different from the input dimension of target policy, so source policy cannot be applied directly in new environment. To successfully and effectively transfer  $\pi^{x*}$  to  $\pi^{y*}$ , we propose the following transfer strategy as shown in Algorithms 1.

In step 1, we want to find a state projection  $\chi$ , so that for any source state  $s^x \in \mathcal{S}^x$ , a corresponding target state is given as:

$$s^y = \chi s^x. \quad (9)$$

We will introduce how to find this projection  $\chi$  in Algorithm 2. Using computed target environment trajectories from step 3, step 4 recovers policy  $\pi^y$  using Algorithm 3.

**Find projection  $\chi$ .** Manifold alignment [20] learns a subspace by matching the local geometry within source space and target space. Given some random source states  $s_0^x, s_1^x, \dots$ , random target states  $s_0^y, s_1^y, \dots$ , manifold alignment looks for two projections  $\alpha$  and  $\beta$  to minimize cost function:

$$\begin{aligned}
C(\alpha, \beta) = & \mu \sum_{i,j} (\alpha^T s_i^x - \beta^T s_j^y)^2 W^{i,j} + 0.5 \sum_{i,j} (\alpha^T s_i^x - \alpha^T s_j^x)^2 W_x^{i,j} \\
& + 0.5 \sum_{i,j} (\beta^T s_i^y - \beta^T s_j^y)^2 W_y^{i,j}, \quad (10)
\end{aligned}$$

---

**Algorithm 2** Find projection  $\chi$  using manifold alignment.

---

**Input:** Random source states  $s_0^x, s_1^x, \dots$ , random target states  $s_0^y, s_1^y, \dots$

**Output:** Projection  $\chi$  for state transfer  $\mathcal{S}^x \rightarrow \mathcal{S}^y$ .

- 1: Compute  $W_x, W_y, W$  using Eqs.(11,13)
  - 2: Compute  $L_x, L_y$  using Eq.(21).
  - 3: Compute  $\Omega_1, \Omega_2, \Omega_3, \Omega_4$  using Eq.(23).
  - 4: Formulate matrix  $L$  and  $Z$  using Eq.(20).
  - 5: Solve Eq.(19).
  - 6: Compute  $\chi$  using Eq.(24).
- 

where  $\alpha$  and  $\beta$  project source states and target states into a space of dimension  $d_{share}$ . This is also similar to the idea in [35, 36], where different views of images use a shared regression coefficient.  $(\alpha^T s_i^x - \alpha^T s_j^y)^2 W_x^{i,j}$  minimizes the difference of  $s_i^x$  and  $s_j^y$  with weight  $W_x^{i,j}$ .  $W_x^{i,j}$  can be computed using the kernel function in Euclidean space:

$$W_x^{i,j} = \exp(-\|s_i^x - s_j^y\|). \quad (11)$$

Similarly,  $(\beta^T s_i^y - \beta^T s_j^x)^2 W_y^{i,j}$  minimizes the difference of  $s_i^y$  and  $s_j^x$  with weight  $W_y^{i,j}$ , which can be computed using Eq.(11) similarly.

$(\alpha^T s_i^x - \beta^T s_j^y)^2 W^{i,j}$  minimizes the difference of  $s_i^x$  and  $s_j^y$  in the shared space with weight  $W^{i,j}$ . To compute  $W^{i,j}$ , we can compare their knn local geometry matrix. Knn local geometry matrix  $R_{s_i^x}$  is defined as a  $(k+1) \times (k+1)$  matrix, with the  $(k_1, k_2)$ -th element as:

$$R_{s_i^x}(k_1, k_2) = \|z_{k_1} - z_{k_2}\|, \quad (12)$$

where  $z = \{s_i^x, z_1, z_2, \dots, z_k\}$ ,  $z_{k_1}$  and  $z_{k_2}$  are the  $k_1$ -th and  $k_2$ -th nearest neighbors,  $\|z_{k_1} - z_{k_2}\|$  is Euclidean distance of  $z_{k_1}$  and  $z_{k_2}$ .  $R_{s_j^y}$  can be computed similarly using Eq.(12).  $W^{i,j}$  can then be given as:

$$W^{i,j} = \exp(-dist(R_{s_i^x}, R_{s_j^y})), \quad (13)$$

$$dist(R_{s_i^x}, R_{s_j^y}) = \min_{1 \leq h \leq k!} \min(dist_1(h), dist_2(h)), \quad (14)$$

where  $dist(R_{s_i^x}, R_{s_j^y})$  is the minimum distance of  $k!$  possible permutations of the  $k$  neighbors of  $s_i^x$  and  $s_j^y$ . The distances of  $h$ -th permutation  $dist_1(h)$  and  $dist_2(h)$  is given as:

$$dist_1(h) = \|\{R_{s_j^y}\}_h - w_1 R_{s_i^x}\|_F, \quad (15)$$

$$dist_2(h) = \|R_{s_i^x} - w_2 \{R_{s_j^y}\}_h\|_F, \quad (16)$$

$$w_1 = \text{Tr} R_{s_i^x}^T \{R_{s_j^y}\}_h / \text{Tr} R_{s_i^x}^T R_{s_i^x}, \quad (17)$$

$$w_2 = \text{Tr} \{R_{s_j^y}\}_h^T R_{s_i^x} / \text{Tr} \{R_{s_j^y}\}_h^T \{R_{s_j^y}\}_h, \quad (18)$$

**Algorithm 3** Recover policy  $\pi^y$ .**Input:** State trajectories  $\{s_{00}^y, s_{01}^y, \dots\}, \{s_{10}^y, s_{11}^y, \dots\}, \dots$ **Output:** Policy  $\pi^y$ .

- 1: Compute action trajectories in  $\mathcal{A}^y$ ,  $\{a_{00}^y, a_{01}^y, \dots\}, \{a_{10}^y, a_{11}^y, \dots\}, \dots$ , using Eq.(25).
- 2: Train  $\pi^y$  with a deep model by using state trajectories as input feature, action trajectories as output class labels.

where  $\text{Tr}$  is matrix trace operator,  $w_1$  and  $w_2$  are two weights terms.

Minimizing cost function Eq.(10) can be formulated as:

$$C(\phi) = \phi^T Z L Z^T \phi, \quad (19)$$

where

$$\phi = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, Z = \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix}, L = \begin{pmatrix} L_x + \mu\Omega_1 & -\mu\Omega_2 \\ -\mu\Omega_3 & L_y + \mu\Omega_4 \end{pmatrix}. \quad (20)$$

Columns of  $X$  are vector representations of state of  $s_0^x, s_1^x, \dots$ ; columns of  $Y$  are vector representations of state of  $s_0^y, s_1^y, \dots$ ;  $L_x$  and  $L_y$  are given as:

$$L_x = D_x - W_x, \quad L_y = D_y - W_y, \quad (21)$$

where  $D_x$  and  $D_y$  are diagonal matrix:

$$D_x^{ii} = \sum_j W_x^{ij}, \quad D_y^{ii} = \sum_j W_y^{ij}. \quad (22)$$

$\Omega_1$  and  $\Omega_4$  are diagonal matrices.  $\Omega_2$  is the same as  $W$ ,  $\Omega_3$  is the same as the transpose of  $W$ .  $\Omega_1, \Omega_2, \Omega_3$  and  $\Omega_4$  are given as:

$$\Omega_1^{ii} = \sum_j W^{ij}, \quad \Omega_2^{ij} = W^{ij}, \quad \Omega_3^{ij} = W^{ji}, \quad \Omega_4^{ii} = \sum_j W^{ji}. \quad (23)$$

$d_{share}$  is a tuning parameter indicating the size of the shared space. Finally,  $\chi$  is given as:

$$\chi = \beta^{T\dagger} \alpha^T, \quad (24)$$

where  $\dagger$  is matrix pseudo inverse.

**Recover policy  $\pi^y$ .** Algorithm 3 shows how to recover policy  $\pi^y$  given state trajectory  $\{s_{00}^y, s_{01}^y, \dots\}, \{s_{10}^y, s_{11}^y, \dots\}, \dots$ . We know that when at state  $s_{0(t-1)}^y$ , an action  $a_{0(t-1)}^y \in \mathcal{A}^y$  was taken and then the state evolved from  $s_{0(t-1)}^y$  to  $s_{0t}^y$ . However, we cannot directly get the action  $a_{0(t-1)}^y$  by comparing  $s_{0(t-1)}^y$  and  $s_{0t}^y$ . Our approach is to try all possible actions in  $\mathcal{A}^y$  and record the states  $\widetilde{s}_{0t}^y(a)$  after taking each action, then compare states  $\widetilde{s}_{0t}^y(a)$  with  $s_{0t}^y$ . To compute  $\widetilde{s}_{0t}^y(a)$ ,

when action  $a$  is taken, we remove the  $a$ -th job in job queue and set it to be 0s in matrix  $s_{0(t-1)}^y$ . We use the following equation to find the  $a_{0(t-1)}^y$ :

$$a_{0(t-1)}^y = \arg \min_{a \in \mathcal{A}^y} \|\widetilde{s}_{0t}^y(a) - s_{0t}^y\|, \quad (25)$$

where  $\|\widetilde{s}_{0t}^y(a) - s_{0t}^y\|$  is distance between states  $\widetilde{s}_{0t}^y(a)$  and  $s_{0t}^y$ .

Using Eq.(25), we can find action trajectories  $\{a_{00}^y, a_{01}^y, \dots\}$ ,  $\{a_{10}^y, a_{11}^y, \dots\}$ , ... We initialize the target policy  $\pi^y$  using a deep model, such as DNN or Deep CNN, and train policy  $\pi^y$  as a classifier from  $s^y$  to a corresponding  $a^y$ .

## 5 Experiments

### 5.1 Experiment setup and evaluation metrics

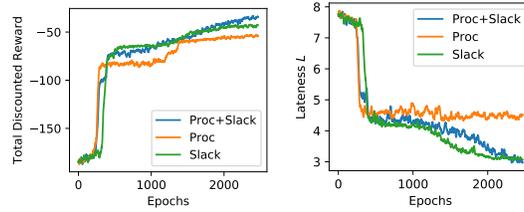
**Experiment settings.** Unless otherwise noted, the default setting is given as below: the machine looks  $T = 15$  time steps ahead, length of slack time array is  $Z = 5$ , number of job slots is  $n = 10$ , number of backlog slots is  $m = 60$ . We test a set of different values for the following job characteristics parameters:

1. **Job arrival speed**  $\lambda \in (0, 1)$  is the probability of arriving a new job at each time step, with default  $\lambda = 0.5$ .
2. **Small job probability**  $p_{small}$  indicates the probability that the new job length is short, with default  $p_{small} = 0.8$ . The short job length distribution is a random number in  $[1, 2]$ ; the long job length distribution is a random number in  $[6, 10]$ .
3. **Urgent job probability**  $p_{urgent}$  indicates probability of the job slack time is very short, with default  $p_{urgent} = 0.5$ . When the job arrives, the urgent job slack time distribution is a random number in  $[1, 5]$ ; the non-urgent job slack time distribution is a random number in  $[5, 10]$ .

At any time step, if the job queue is full, the new job will be dropped, a penalty  $-10$  will be added to reward  $r_t$ . We use Deep Neural Network (DNN) to model policy  $\pi_{a|s;\theta}$ . The input is state representation from Figure 2. The output is a probability vector of length same as action space  $\mathcal{A}$ . The activation function in hidden layer is rectified linear unit and the activation function in the last layer is Softmax. The DNN network structure we use is 2 hidden layers with 128 and 64 neurons respectively.

**Training and testing data.** For each experiment, we use the same setting to generate 10 random trajectories for training and testing respectively. Then we train each comparing method on the same training data and test it on the same testing data.

**Evaluation metrics.** To evaluate the performance and business impact, we use total discounted reward (Eq.(2)) and average lateness (Eq.(4)), tardiness (Eq.(5)). Total discounted reward includes the penalty when a job was dropped if the job queue is full. Average lateness and tardiness only consider those jobs that are successfully finished, which are straightforward and valuable to business users. The reported results are the average values over 10 random testing trajectories.



(a) Total reward. (b) Average lateness.

Fig. 3: Effectiveness of state presentation. “Proc+Slack”: using both Processing time and Slack time of Job queue in Figure 2. “Proc”: using only Processing time. “Proc”: using only Slack time.

Table 1: Total discounted reward using different objective and trajectory length.

Trajectory length	Lateness objective				Tardiness objective			
	50	100	150	200	50	100	150	200
EDF	-65.05	-116.06	-215.65	-513.53	-57.03	-95.67	-192.64	-493.14
LST	-77.11	-144.45	-260.93	-624.76	-69.09	-130.97	-247.45	-611.28
Random Forest	-65.05	-115.47	-211.37	-433.40	-57.03	-92.46	-195.20	-410.39
SVM	-65.05	-115.47	-212.48	-348.01	-57.03	-92.46	-185.50	-324.99
Neural Network	-65.05	-115.47	-164.58	-238.31	-57.03	-92.46	-137.60	-199.19
Tetris	-21.65	-89.37	-73.66	-154.15	-8.93	-16.67	-24.46	-102.53
RL-Mao	-7.00	-57.39	-250.51	-212.32	-38.86	-118.54	-223.05	-370.97
DMD	<b>-0.51</b>	<b>-34.47</b>	<b>-63.30</b>	<b>-133.14</b>	<b>-0.22</b>	<b>-4.09</b>	<b>-9.43</b>	<b>-14.55</b>

## 5.2 Effectiveness of state presentation Figure 2

Figure 3 shows the effectiveness of combining processing time and slack time information in state presentation (Figure 2). Combining both processing and slack time gives the highest total discounted lateness reward and lowest average lateness. Different to computer job resource management [12], slack time is important for manufacturing dispatching.

## 5.3 Comparison with existing algorithms

We compare the proposed Deep Manufacturing Dispatching (DMD) with 7 other dispatching policies: 2 due time related manual designed rules (EDF for Earliest-Due-First, LST for Least-Slack-Time), 3 hyper-heuristics using machine learning (random forest, SVM, and neural network with 2 hidden layers), and reinforcement learning based RL-Mao [12] and Tetris [26]. For hyper-heuristics, under each experiment setting, we choose the best heuristic rule in terms of average lateness or tardiness as ground truth rule.

Table 1 and 2 show DMD gets the highest reward and lowest average lateness and tardiness. Unless other noted, the default trajectory length is 100 time steps.

Table 2: Average lateness and tardiness using different objective and trajectory length.

Trajectory length	Average lateness				Average tardiness			
	50	100	150	200	50	100	150	200
EDF	5.29	4.44	7.02	11.68	5.77	5.38	8.71	13.89
LST	6.71	5.48	8.26	13.80	7.07	6.21	9.31	15.35
Random Forest	5.29	4.38	5.89	9.03	5.77	5.40	6.69	10.66
SVM	5.29	4.38	5.48	7.32	5.77	5.40	6.72	8.70
Neural Network	5.29	4.38	4.37	5.23	5.77	5.40	5.29	7.33
Tetris	6.13	8.08	5.73	8.20	8.25	14.17	6.20	10.07
RL-Mao	5.29	6.39	7.35	8.06	7.78	9.10	9.21	10.43
DMD	<b>2.11</b>	<b>3.16</b>	<b>3.60</b>	<b>5.01</b>	<b>1.51</b>	<b>2.24</b>	<b>3.68</b>	<b>4.14</b>

Table 3: Total discounted lateness reward with various job characteristics.

	Job arrival speed $\lambda$					Small job $p_{small}$			Urgent job $p_{urgent}$		
	0.1	0.3	0.5	0.7	0.9	0.2	0.5	0.8	0.1	0.5	0.9
EDF	-31.12	-96.99	-116.06	-317.59	-415.87	-305.49	-246.97	-116.06	-71.92	-116.06	-115.08
LST	-31.12	-126.23	-144.45	-378.46	-422.32	-316.67	-278.47	-144.45	-75.77	-144.45	-120.12
Random Forest	-31.12	-78.22	-115.47	-205.24	-246.33	-106.16	-106.33	-115.47	-71.15	-115.47	-111.95
SVM	-31.12	-78.66	-115.47	-208.10	-131.94	-154.18	-124.25	-115.47	-71.15	-115.47	-111.95
Neural Network	-31.12	-78.66	-115.47	-219.09	-260.66	-111.56	-150.48	-115.47	-71.15	-115.47	-111.95
Tetris	-28.39	-83.46	-89.37	<b>-53.08</b>	-107.22	-70.08	-48.26	-89.37	-68.74	-89.37	-69.85
RL-Mao	-13.79	-58.95	-57.39	-136.36	-178.74	-99.27	-74.24	-57.39	-49.70	-57.39	-48.11
DMD	<b>-11.12</b>	<b>-14.48</b>	<b>-34.47</b>	-62.37	<b>-106.70</b>	<b>-56.33</b>	<b>-43.26</b>	<b>-34.47</b>	<b>-38.45</b>	<b>-34.47</b>	<b>-36.86</b>

In manufacturing scheduling, where job frequency is not as high as computer jobs, this trajectory length is reasonable. We can see from Table 1 that as the length of trajectory increases, the total discounted reward decreases overall. Table 3 and 4 show total discounted reward and average lateness with different settings of  $\lambda$ ,  $p_{small}$  and  $p_{urgent}$ . Overall, for 19 settings (8 settings in Tables 1,2 and 11 settings in Tables 3,4), DMD gets best results for 18 settings on total discounted reward and 16 settings on average lateness and tardiness. Best performing methods in each column are in bold.

#### 5.4 Dispatching policy transfer

In step 1 of Algorithm 1, we generate 2000 random states from source environment and target environment respectively for each transfer setting. The following four transfer learning settings are considered: (1). Same-environment: Source  $\lambda = 0.5$ ; Target  $\lambda = 0.7$ . (2). Same-environment: Source  $\lambda = 0.5$ ,  $p_{small} = 0.8$ ; Target  $\lambda = 0.7$ ,  $p_{small} = 0.5$ . (3). Cross-environment: Source  $n = 10$ ; Target  $n = 15$ . (4). Cross-environment: Source  $n = 10$ ,  $m = 60$ ; Target  $n = 15$ ,  $m = 30$ . In all figures, *noTransfer* curve is trained from scratch under target setting, *Transfer* curve is initialized using the output of Algorithm 3 under source set-

Table 4: Average lateness with various job characteristics.

	Job arrival speed $\lambda$					Small job $p_{small}$			Urgent job $p_{urgent}$		
	0.1	0.3	0.5	0.7	0.9	0.2	0.5	0.8	0.1	0.5	0.9
EDF	5.33	5.00	4.44	16.29	19.41	21.05	12.83	4.44	3.26	4.44	4.67
LST	5.33	6.28	5.48	20.89	22.54	21.18	15.00	5.48	4.00	5.48	4.94
Random Forest	5.33	4.00	4.38	8.21	12.73	8.11	5.42	4.38	<b>2.88</b>	4.38	4.32
SVM	5.33	4.16	4.38	8.21	<b>4.02</b>	9.47	5.68	4.38	<b>2.88</b>	4.38	4.32
Neural Network	5.33	4.16	4.38	10.30	15.43	<b>7.61</b>	6.36	4.38	<b>2.88</b>	4.38	4.32
Tetris	5.33	7.67	8.08	6.58	12.78	7.75	6.30	8.08	7.68	8.08	6.26
RL-Mao	4.50	6.65	6.39	9.35	10.87	11.62	8.19	6.39	5.15	6.39	5.83
DMD	<b>3.45</b>	<b>3.21</b>	<b>3.16</b>	<b>4.51</b>	9.74	7.70	<b>4.32</b>	<b>3.16</b>	4.55	<b>3.16</b>	<b>4.03</b>

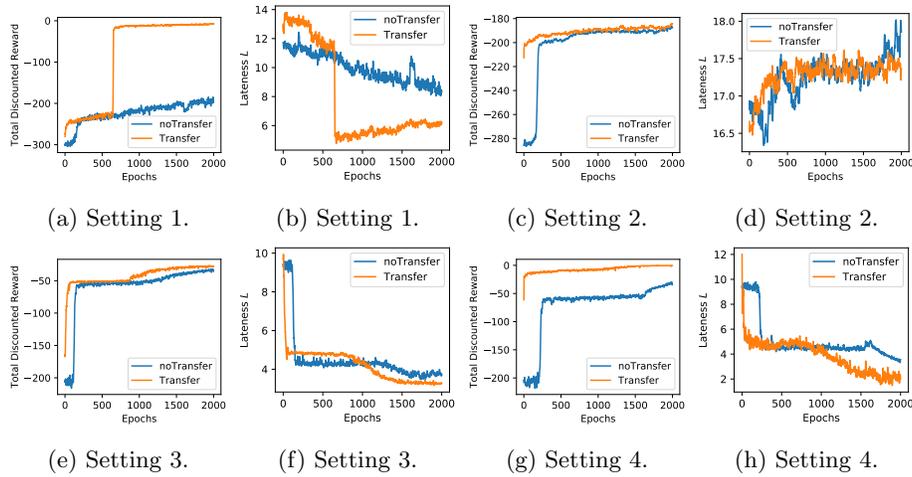


Fig. 4: Policy transfer evaluation using Deep Manufacturing Dispatching (DMD).

ting, then trained under target setting. Figure 4 shows total discounted reward and average lateness of the 4 transfer settings using DMD. We find that policy transferring reduces training time and gives larger total discounted reward and smaller average lateness compared to training from scratch. Table 5 shows the results using 3 hyper-heuristics. Best results are highlighted for each transfer setting. Note that, as mentioned in Section 5.1, dropped job penalty is considered in total discounted reward, but not in average lateness. We find that policy transfer gives competitive or better results than training from scratch using hyper-heuristics. This shows the effectiveness of Algorithm 1.

## 6 Conclusion

Dispatching is a difficult yet important problem. It is not straightforward due to the complexity of manufacturing settings. We showed the promising applica-

Table 5: Policy transfer evaluation for hyper-heuristics.

Transfer setting		noTransfer				Transfer			
		1	2	3	4	1	2	3	4
Total discounted lateness	Random Forest	-205.24	-216.44	-120.22	-120.22	<b>-140.02</b>	<b>-205.97</b>	<b>-115.47</b>	<b>-115.47</b>
	SVM	-208.10	<b>-189.27</b>	<b>-115.47</b>	<b>-115.47</b>	<b>-128.26</b>	-213.40	<b>-115.47</b>	<b>-115.47</b>
	Neural Network	-219.09	-229.46	<b>-114.15</b>	-121.65	<b>-174.52</b>	<b>-204.55</b>	-115.47	<b>-115.47</b>
Average lateness	Random Forest	8.21	14.05	4.63	4.63	<b>5.08</b>	<b>10.63</b>	<b>4.38</b>	<b>4.38</b>
	SVM	8.21	13.05	<b>4.38</b>	<b>4.38</b>	<b>5.13</b>	<b>10.96</b>	<b>4.38</b>	<b>4.38</b>
	Neural Network	10.30	16.74	<b>4.38</b>	4.72	<b>7.00</b>	<b>13.40</b>	<b>4.38</b>	<b>4.38</b>

tion of combining reinforcement learning and transfer learning in manufacturing industry. The transfer learning module increases the generalization of learned dispatching rules and saves cost and time for data collection and model training.

## References

1. Park, J., Nguyen, S., Zhang, M., Johnston, M.: Genetic programming for order acceptance and scheduling. In: Evolutionary Computation (CEC), 2013 IEEE Congress on. pp. 1005–1012. IEEE (2013)
2. Jakobović, D., Jelenković, L., Budin, L.: Genetic programming heuristics for multiple machine scheduling. In: European Conference on Genetic Programming. pp. 321–330. Springer (2007)
3. Garey, M.R., Johnson, D.S., Sethi, R.: The complexity of flowshop and jobshop scheduling. *Mathematics of operations research* **1**(2), 117–129 (1976)
4. Vazquez-Rodriguez, J.A., Ochoa, G.: On the automatic discovery of variants of the neh procedure for flow shop scheduling using genetic programming. *Journal of the Operational Research Society* **62**(2), 381–396 (2011)
5. Mascia, F., López-Ibáñez, M., Dubois-Lacoste, J., Stützle, T.: From grammars to parameters: Automatic iterated greedy design for the permutation flow-shop problem with weighted tardiness. In: International Conference on Learning and Intelligent Optimization. pp. 321–334. Springer (2013)
6. Branke, J., Nguyen, S., Pickardt, C.W., Zhang, M.: Automated design of production scheduling heuristics: A review. *IEEE Transactions on Evolutionary Computation* **20**(1), 110–124 (2016)
7. Mobley, R.K.: An introduction to predictive maintenance. Elsevier (2002)
8. Zheng, S., Ristovski, K., Farahat, A., Gupta, C.: Long short-term memory network for remaining useful life estimation. In: Prognostics and Health Management (ICPHM), 2017 IEEE International Conference on. pp. 88–95. IEEE (2017)
9. Özcan, E., Parkes, A.J.: Policy matrix evolution for generation of heuristics. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation. ACM (2011)
10. Vonolfen, S., Beham, A., Kommenda, M., Affenzeller, M.: Structural synthesis of dispatching rules for dynamic dial-a-ride problems. In: International Conference on Computer Aided Systems Theory. pp. 276–283. Springer (2013)
11. Frankola, T., Golub, M., Jakobovic, D.: Evolutionary algorithms for the resource constrained scheduling problem. In: Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on. pp. 715–722. IEEE (2008)

12. Mao, H., Alizadeh, M., Menache, I., Kandula, S.: Resource management with deep reinforcement learning. In: Proceedings of the 15th ACM Workshop on Hot Topics in Networks. pp. 50–56. ACM (2016)
13. Chen, X., Hao, X., Lin, H.W., Murata, T.: Rule driven multi objective dynamic scheduling by data envelopment analysis and reinforcement learning. In: Automation and Logistics (ICAL), 2010 IEEE International Conference on. pp. 396–401. IEEE (2010)
14. Zheng, S., Ding, C.: Kernel alignment inspired linear discriminant analysis. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 401–416. Springer Berlin Heidelberg (2014)
15. Zheng, S., Nie, F., Ding, C., Huang, H.: A harmonic mean linear discriminant analysis for robust image classification. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 402–409. IEEE (2016)
16. Zheng, S., Ding, C., Nie, F., Huang, H.: Harmonic mean linear discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering* (2018). <https://doi.org/10.1109/TKDE.2018.2861858>
17. Zheng, S., Ding, C.: Sparse classification using group matching pursuit. *Neurocomputing* **338**, 83–91 (2019). <https://doi.org/10.1016/j.neucom.2019.02.001>
18. Zheng, S., Ding, C.: Minimal support vector machine. *arXiv preprint arXiv:1804.02370* (2018)
19. Zheng, S., Ding, C., Nie, F.: Regularized singular value decomposition and application to recommender system. *arXiv preprint arXiv:1804.05090* (2018)
20. Wang, C., Mahadevan, S.: Manifold alignment without correspondence. In: *IJCAI*. vol. 2, p. 3 (2009)
21. Weckman, G.R., Ganduri, C.V., Koonce, D.A.: A neural network job-shop scheduler. *Journal of Intelligent Manufacturing* **19**(2), 191–201 (2008)
22. Ingimundardottir, H., Runarsson, T.P.: Supervised learning linear priority dispatch rules for job-shop scheduling. In: *International conference on learning and intelligent optimization*. pp. 263–277. Springer (2011)
23. Li, X., Olafsson, S.: Discovering dispatching rules using data mining. *Journal of Scheduling* **8**(6), 515–527 (2005)
24. Shiue, Y.R.: Data-mining-based dynamic dispatching rule selection mechanism for shop floor control systems using a support vector machine approach. *International Journal of Production Research* **47**(13), 3669–3690 (2009)
25. Zhang, W., Dietterich, T.G.: A reinforcement learning approach to job-shop scheduling. In: *IJCAI*. vol. 95, pp. 1114–1120. Citeseer (1995)
26. Grandl, R., Ananthanarayanan, G., Kandula, S., Rao, S., Akella, A.: Multi-resource packing for cluster schedulers. *ACM SIGCOMM Computer Communication Review* **44**(4), 455–466 (2015)
27. Zheng, S., Shae, Z.Y., Zhang, X., Jamjoom, H., Fong, L.: Analysis and modeling of social influence in high performance computing workloads. In: *European Conference on Parallel Processing*. pp. 193–204. Springer Berlin Heidelberg (2011)
28. Sutton, R.S., Barto, A.G., et al.: *Reinforcement learning: An introduction*. MIT press (1998)
29. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**(3-4), 229–256 (1992)
30. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: *Advances in neural information processing systems*. pp. 1057–1063 (2000)

31. Zheng, S., Vishnu, A., Ding, C.: Accelerating deep learning with shrinkage and recall. In: 2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS). pp. 963–970. IEEE (2016)
32. Ammar, H.B., Eaton, E., Ruvolo, P., Taylor, M.E.: Unsupervised cross-domain transfer in policy gradient reinforcement learning via manifold alignment. In: Proc. of AAAI (2015)
33. Joshi, G., Chowdhary, G.: Cross-domain transfer in reinforcement learning using target apprentice. arXiv preprint arXiv:1801.06920 (2018)
34. Pan, S.J., Yang, Q., et al.: A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**(10), 1345–1359 (2010)
35. Zheng, S., Cai, X., Ding, C., Nie, F., Huang, H.: A closed form solution to multi-view low-rank regression. In: AAAI. pp. 1973–1979 (2015)
36. Zheng, S.: Machine Learning: Several Advances in Linear Discriminant Analysis, Multi-View Regression and Support Vector Machine. Ph.D. thesis, The University of Texas at Arlington (2017)