Deep Reinforcement Learning with Integer Optimization for Dynamic Slab Assignment Problem−Supplementary Material

Tianyang Li, Ying Meng, Lixin Tang, *Fellow, IEEE*, and Yuxuan Zhang

This is the supplementary material to the paper entitled “Deep Reinforcement Learning with Integer Optimization for Dynamic Slab Assignment Problem”, submitted to *IEEE Transactions on Control Systems Technology*.

I. SUPPLEMENTARY MATERIAL

A. Comparison Methods

*1) **Global and Greedy Method**:* As a result of DRLIP’s lack of optimality guarantees, it is crucial to design a method that considers a global policy (*GP*) to yield upper bounds at different test baseline levels. Assuming that we can observe all information of slab and order within $\mathcal{T}$ time steps at once, the planning period could be decomposed into $\lceil |T| / \mathcal{T} \rceil$ phases. With this assumption, the information on slabs and orders becomes available in each phase. This allows us to transform the dynamic decision problem into a static assignment problem, which can be solved with the integer programming (IP) solver to obtain optimal decisions. With this setting, the uncertainty is eliminated and then allow us to transform the dynamic decision problem into a static assignment problem. In each phase, the IP model of the static assignment problem can be formulated as:

$$\max\ f_{gp,\mathcal{T}} = \max\ \lambda_1 \sum_{i=1}^{|M_{\mathcal{T}}|} w_i g_i^{slab} (\sum_{j=1}^{|O_{\mathcal{T}}|} c_{ij}^{ord} x_{ij}^{ord} + c^{self} x_i^{self}) - \lambda_2 \sum_{j=1}^{|O_{\mathcal{T}}^{due}|} c^{ud} (\mathcal{C}_j - \sum_{i=1}^{|M_{\mathcal{T}}|} w_i x_{ij}^{ord}) - \lambda_3 \sum_{i=1}^{|M_{\mathcal{T}}|} c^{inv} w_i x_i^{inv} \tag{1}$$

s.t.

$$\sum_{i=1}^{|M_{\mathcal{T}}|} w_i x_{ij}^{ord} \leq \mathcal{C}_j, \qquad j = 1, \ldots, |O_{\mathcal{T}}|, \tag{2}$$

$$\sum_{j=1}^{|O_{\mathcal{T}}|} x_{ij}^{ord} + x_i^{self} + x_i^{inv} = 1, \qquad i = 1, \ldots, |M_{\mathcal{T}}|, \tag{3}$$

where $M_{\mathcal{T}}$ and $O_{\mathcal{T}}$ represent the sets of the available slabs and orders within $\mathcal{T}$ time steps, respectively. Algorithm 1 presents the full procedure of *GP* to solve DSAP. From this, one can observe that *GP* is able to provide an optimal solution for comparison when $\mathcal{T} = |T|$. In addition, it should be noted that *GP* can be regarded as the greedy method (*GM*) when $\mathcal{T} = 1$. In this case, all available slabs will be assigned to orders and self-designed orders at each time step. This not only has no inventory cost but also the unsatisfied demand cost can be minimized, and hence *GM* is a frequently used method by decision-makers for DSAP.

*2) **Stochastic Method**:* At time step t, the policy should explicitly consider future time steps in addition to the current available slabs and orders. As noted, considering more suitable orders may appear for the current slabs in the future, a practicable policy is to hold partial slabs in inventory to await future suitable orders. To this end, the stochastic method, *SM*, is proposed involved in how randomly selecting the slabs in M_t to be held in inventory at each time step. For *SM*, we prioritize the profits from the available orders as a basic guarantee for reward gains.For

Algorithm 1: Global Method

```
Input: available time step $\mathcal{T}$, planning period $T$, IP solver
Output: mean reward $\overline{r}$
for $t := 1$ to $|T| / \mathcal{T}$ do
    Obtain state $s_t = \{M_t, O_t\}$ based on the state transition in Eq. (8) and Eq. (9) of the main text;
    Based on $s_t$, obtain reward gain $r_{gp,t}$ by solving model (1) with IP solver;
Calculate mean reward $\overline{r} = \sum_{t=1}^{|T|/\mathcal{T}} r_{gp,t} / |T|$;
return $\overline{r}$
```

Algorithm 2: Stochastic Method

```
Input: Bernoulli distribution $B(1, \varepsilon)$, planning period $T$, IP solver
Output: mean reward $\overline{r}$
for $t := 1$ to $T$ do
    Obtain state $s_t = \{M_t, O_t\}$ based on the state transition in Eq. (8) and Eq. (9) of the main text;
    Based on $s_t$, obtain optimal solution $\overline{x}_{ij}^{ord}$ and profit $f_{sm,t}$ by solving model (4) with IP solver;
    if $M_t = \varnothing$ then
        Obtain reward gain $r_{sm,t} := f_{sm,t}$;
    else
        Generate $\psi_i^{inv}$ for slab $i \in M_t$ by random sampling from $B(1, \varepsilon)$;
        Based $\overline{x}_{ij}^{ord}$ and $\psi_i^{inv}$, calculate $l_{sm,t}$ by solving Eq. (7) with IP solver;
        Obtain reward gain $r_{sm,t} := f_{sm,t} + l_{sm,t}$;
Acquire mean reward $\overline{r} := \sum_{t=1}^{T} r_{sm,t} / |T|$;
return $\overline{r}$
```

each time step, we first assign slabs to available orders as many as possible to maximize their corresponding profits, which can be formulated as:

$$\max\ f_{sm,t} = \max\ \lambda_1 \sum_{i=1}^{|M_t|} \sum_{j=1}^{|O_t|} w_i g_i^{slab} c_{ij}^{ord} x_{ij}^{ord} - \lambda_2 \sum_{j=1}^{|O_t^{del}|} c^{ud} (\mathcal{C}_j - \sum_{i=1}^{|M_t|} w_i x_{ij}^{ord}) \tag{4}$$

s.t.

$$\sum_{i=1}^{|M_t|} w_i x_{ij}^{ord} \leq \mathcal{C}_j, \qquad j = 1, \ldots, |O_t|, \tag{5}$$

$$\sum_{j=1}^{|O_t|} x_{ij}^{ord} \leq 1, \qquad i = 1, \ldots, |M_t|. \tag{6}$$

Then, with the optimal solution $\overline{x}_{ij}^{ord}$ to model (4), we employ a random variable ψ to indicate the probability of the remaining slab being selected for inventory, in which ψ is subject to a Bernoulli distribution $B(1, \varepsilon)$ and $\varepsilon \in [0, 1]$ represents a probability parameter. Next, the other profits and cost can be computable by:

$$l_{sm,t} = \lambda_1 \sum_{i=1}^{|M_t|} c^{self} (1 - \sum_{j=1}^{|O_t|} \overline{x}_{ij}^{ord}) \left(1 - \psi_i^{inv}\right) - \lambda_3 \sum_{i=1}^{|M_t|} c^{inv} w_i (1 - \sum_{j=1}^{|O_t|} \overline{x}_{ij}^{ord}) \psi_i^{inv} \tag{7}$$

$$\text{s.t.} \quad \psi \sim B(1, \varepsilon)\,. \tag{8}$$

where ψ_i is obtained by random sampling from $B(1, \varepsilon)$. The final reward gain for *SM* is $r_{sm,t} = f_{sm,t} + l_{sm,t}$. The pseudocode of *SM* can be viewed in Algorithm 2.

*3) **Deterministic Method**:* For the same considerations as *SM*, the deterministic method (*DM*) first assigns slabs to available orders to ensure reward gain, and hence the model of *DM* to maximize the profits of available orders is the same as model (4), i.e., $f_{dm,t} = f_{sm,t}$. Then, it will hold the remaining slabs with high steel grades in inventory, while the others will be allocated to self-designed orders. For convenience, let $K^h \subset \mathcal{K}$ be the set of high steel grades. Given optimal solution $\overline{x}_{ij}^{ord}$ to model (4), the other profits and costs can be acquired by:

$$\begin{aligned} l_{dm,t} = \lambda_1 \sum_{i=1}^{|M_t|} c^{self}(1 - \sum_{j=1}^{|O_t|} \overline{x}_{ij}^{ord})(1 - x_i^{inv}) - \lambda_3 \sum_{i=1}^{|M_t|} c^{inv} w_i (1 - \sum_{j=1}^{|O_t|} \overline{x}_{ij}^{ord}) x_i^{inv} \\ \text{s.t.} \quad x_i^{inv} = 1, \qquad \forall g_i^{slab} \in K^h. \end{aligned} \tag{9}$$

Likewise, the decision reward gain of *DM* can be computed as $r_{dm,t} = f_{dm,t} + l_{dm,t}$. Algorithm 3 provides the pseudocode of *DM*.

Algorithm 3: Deterministic Method

```
Input: steel grade set $K^h$, planning period $T$, IP solver
Output: mean reward $\overline{r}$
for $t := 1$ to $T$ do
    Obtain state $s_t = \{M_t, O_t\}$ based on the state transition in Eq. (8) and Eq. (9) of the main text;
    Based on $s_t$, assign slabs in $M_t$ to orders in $O_t$ for maximizing $f_{dm,t}$;
    if $M_t = \varnothing$ then
        Attain reward gain $r_{dm,t} := f_{dm,t}$;
    else
        Hold slabs in inventory according to $K^h$ and get other reward and cost $l_{dm,t}$ ;
        Calculate reward gain $r_{dm,t} := f_{dm,t} + l_{dm,t}$;
Acquire mean reward $\overline{r} := \sum_{t=1}^{T} r_{dm,t} / |T|$;
return $\overline{r}$
```

*4) **Scenario Tree Method**:* To further assess the performance of the proposed DRLIP algorithm, we also customize a scenario tree method as a baseline. The scenario tree method is widely used in stochastic programming to model uncertainty through a set of possible scenarios. It involves constructing a tree structure where each node represents a decision point and branches represent possible outcomes or scenarios. This method allows decision-makers to analyze various options based on different future states of nature.

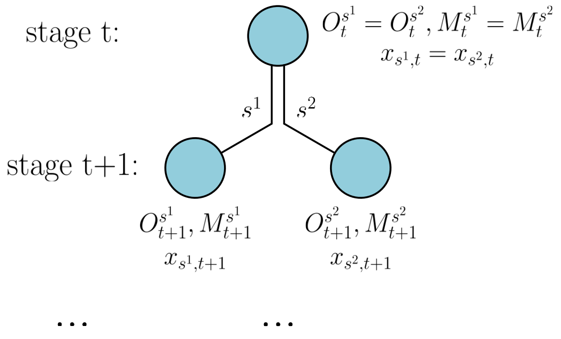
stage t:
$O_t^{s^1} = O_t^{s^2}, M_t^{s^1} = M_t^{s^2}$
$x_{s^1,t} = x_{s^2,t}$
s^1
s^2
stage t+1:
$O_{t+1}^{s^1}, M_{t+1}^{s^1}$
$x_{s^1,t+1}$
$O_{t+1}^{s^2}, M_{t+1}^{s^2}$
$x_{s^2,t+1}$
...
...

Fig. 1. An example of scenario tree with two stages, where $O_t^s, M_t^s, x_{s,t}$ are slabs, orders, and decision variables for scenario s at stage t, respectively.

Specifically, in the customized scenario tree method, a scenario tree is built by an explicit representation of the branching process induced by the gradual observation of the newly arriving slabs and orders. The observation can be obtained by randomly sampling at each child node the number and the exact features of slabs and orders according to the given slab and order data distributions. A simple example of a scenario tree with two stages is shown in Fig. 1. The root node corresponds to the first decision stage where real slab and order data are exposed. Two connected child nodes correspond to the next decision stage, each reveals a possible sampling of slabs and orders. These two nodes are then connected to two child nodes separately in the same way. The branching process construction goes on until stage $\mathcal{T}$ is reached. The nodes on the unique path from the root to a leaf define a scenario, a prediction of the random process from stage two to stage $\mathcal{T}$. Note that scenarios sharing the same node at stage t $(t \in \{1, \dots, \mathcal{T}\})$ make the same decisions from stage 1 to t. We formalize and maximize the general profit considering all scenarios equally as below:

$$\max\ f_{\mathcal{T}}^{scenario} = \max\ \frac{1}{|S|}\sum_{s\in S}(\lambda_1 \sum_{i=1}^{|M_{\mathcal{T}}^{s}|} w_{is} g_{is}^{slab}(\sum_{j=1}^{|O_{\mathcal{T}}^{s}|} c_{ijs}^{ord} x_{ijs}^{ord} + c^{self} x_{is}^{self}) \tag{10}$$

$$- \lambda_2 \sum_{j=1}^{|O_{\mathcal{T}}^{due,s}|} c^{ud}(\mathcal{C}_{js} - \sum_{i=0}^{|M_{\mathcal{T}}^{s}|} w_{is} x_{ijs}^{ord}) - \lambda_3 \sum_{i=0}^{|M_{\mathcal{T}}^{s}|} c^{inv} w_{is} x_{is}^{inv}) \tag{11}$$

s.t.

$$\sum_{i=0}^{|M_{\mathcal{T}}^{s}|} w_{is} x_{ijs}^{ord} \leq \mathcal{C}_{js}, \qquad j = 1, \cdots, |O_{\mathcal{T}}^{due,s}|,\ s \in S, \tag{12}$$

$$\sum_{j=1}^{|O_{\mathcal{T}}^{s}|} x_{ijs}^{ord} + x_{is}^{self} + x_{is}^{inv} = 1, \qquad i = 1, \cdots, |M_{\mathcal{T}}^{s}|,\ s \in S, \tag{13}$$

$$x_{i^1 j^1 s^1}^{ord} = x_{i^2 j^2 s^2}^{ord}, \qquad i^1 = 1, \cdots, |M_{\mathcal{T}}^{s^1}|,\ j^1 = 1, \cdots, |O_{\mathcal{T}}^{due,s^1}|, \tag{14}$$

$$i^2 = 1, \cdots, |M_{\mathcal{T}}^{s^2}|,\ j^2 = 1, \cdots, |O_{\mathcal{T}}^{due,s^2}|, \tag{15}$$

$$\mathcal{N}(t, s^1) = \mathcal{N}(t, s^2), \quad s^1, s^2 \in S,\ t \in \{1, \dots, \mathcal{T}\}, \tag{16}$$

where S is the set of all scenarios and $\mathcal{N}(t, s)$ denotes the node that scenario s belongs to at stage t. Constraint set (16) represents the relationship between decisions in different scenarios.

B. *Experiment Results*

TABLE I
STATISTICAL RESULTS OF THE REWARD OBTAINED BY DRLIP AND 7 DIFFERENT GLOBAL POLICIES ON 20 SMALL-SCALE RANDOMLY GENERATED PROBLEM INSTANCES. THE BEST RESULT IS MARKED IN GRAY BACKGROUND

Instance index	*GP-1* ($\mathcal{T}$=2)	*GP-2* ($\mathcal{T}$=3)	*GP-3* ($\mathcal{T}$=5)	*GP-4* ($\mathcal{T}$=6)	*GP-5* ($\mathcal{T}$=10)	*GP-6* ($\mathcal{T}$=15)	*GP-7* ($\mathcal{T}$=30)	DRLIP	Gap (%)
1	3.46*e*+02	4.37*e*+02	5.53*e*+02	5.33*e*+02	6.29*e*+02	6.64*e*+02	7.84*e*+02	6.40*e*+02	1.83*e*+01
2	3.45*e*+02	4.49*e*+02	5.14*e*+02	5.63*e*+02	6.26*e*+02	6.32*e*+02	7.79*e*+02	5.68*e*+02	2.71*e*+01
3	3.32*e*+02	4.50*e*+02	5.89*e*+02	5.84*e*+02	6.66*e*+02	6.89*e*+02	8.00*e*+02	5.94*e*+02	2.57*e*+01
4	4.72*e*+02	5.73*e*+02	6.85*e*+02	7.16*e*+02	7.69*e*+02	7.75*e*+02	9.17*e*+02	7.12*e*+02	2.24*e*+01
5	2.88*e*+02	4.23*e*+02	4.68*e*+02	5.33*e*+02	5.43*e*+02	5.95*e*+02	7.41*e*+02	6.04*e*+02	1.85*e*+01
6	3.15*e*+02	3.62*e*+02	4.82*e*+02	4.70*e*+02	5.41*e*+02	5.73*e*+02	6.95*e*+02	5.89*e*+02	1.52*e*+01
7	3.51*e*+02	4.32*e*+02	5.21*e*+02	5.59*e*+02	6.28*e*+02	6.70*e*+02	8.17*e*+02	6.85*e*+02	1.61*e*+01
8	2.98*e*+02	3.80*e*+02	4.74*e*+02	4.99*e*+02	5.74*e*+02	5.99*e*+02	7.46*e*+02	5.87*e*+02	2.14*e*+01
9	3.97*e*+02	4.58*e*+02	5.70*e*+02	5.94*e*+02	6.73*e*+02	7.07*e*+02	8.37*e*+02	6.70*e*+02	2.00*e*+01
10	3.35*e*+02	4.70*e*+02	5.72*e*+02	6.04*e*+02	6.31*e*+02	6.60*e*+02	7.82*e*+02	6.27*e*+02	1.98*e*+01
11	3.11*e*+02	4.16*e*+02	5.17*e*+02	5.57*e*+02	6.08*e*+02	6.32*e*+02	7.60*e*+02	6.12*e*+02	1.95*e*+01
12	3.49*e*+02	4.28*e*+02	5.36*e*+02	5.86*e*+02	6.21*e*+02	7.21*e*+02	8.35*e*+02	6.72*e*+02	1.96*e*+01
13	2.88*e*+02	3.30*e*+02	4.53*e*+02	4.32*e*+02	5.02*e*+02	5.51*e*+02	6.69*e*+02	5.98*e*+02	1.06*e*+01
14	3.29*e*+02	4.80*e*+02	5.70*e*+02	5.96*e*+02	6.38*e*+02	7.17*e*+02	8.35*e*+02	6.94*e*+02	1.68*e*+01
15	3.07*e*+02	3.66*e*+02	5.18*e*+02	5.17*e*+02	5.86*e*+02	6.25*e*+02	7.55*e*+02	5.18*e*+02	3.14*e*+01
16	3.19*e*+02	3.88*e*+02	5.18*e*+02	5.46*e*+02	5.94*e*+02	6.41*e*+02	7.73*e*+02	6.50*e*+02	1.60*e*+01
17	4.14*e*+02	4.51*e*+02	5.57*e*+02	5.80*e*+02	6.49*e*+02	6.90*e*+02	8.44*e*+02	6.08*e*+02	2.79*e*+01
18	3.18*e*+02	4.19*e*+02	5.33*e*+02	5.42*e*+02	6.33*e*+02	6.64*e*+02	7.99*e*+02	5.68*e*+02	2.89*e*+01
19	3.05*e*+02	3.98*e*+02	5.01*e*+02	5.37*e*+02	6.01*e*+02	6.06*e*+02	7.57*e*+02	5.03*e*+02	3.36*e*+01
20	4.02*e*+02	4.82*e*+02	5.56*e*+02	5.81*e*+02	6.32*e*+02	6.69*e*+02	7.93*e*+02	6.29*e*+02	2.08*e*+01
MEAN	3.41*e*+02	4.30*e*+02	5.34*e*+02	5.57*e*+02	6.17*e*+02	6.54*e*+02	7.86*e*+02	6.16*e*+02	2.15*e*+01
SD	4.60*e*+01	5.19*e*+01	5.05*e*+01	5.61*e*+01	5.45*e*+01	5.33*e*+01	5.36*e*+01	5.38*e*+01	5.76*e*+00

TABLE II
STATISTICAL RESULTS OF THE NET PROFITS OBTAINED BY DRLIP AND THE OTHER METHODS ON 20 PRACTICAL PROBLEM INSTANCES. THE BEST RESULT IS MARKED IN GRAY BACKGROUND

Instance index	Greedy	*SM-4*	*SM-5*	*DM-7*	DRLIP
1	3.04*e*+03	3.28*e*+03	3.75*e*+03	2.40*e*+03	5.12*e*+03
2	2.81*e*+03	3.27*e*+03	3.25*e*+03	2.95*e*+03	4.54*e*+03
3	3.10*e*+03	3.43*e*+03	3.63*e*+03	2.77*e*+03	4.75*e*+03
4	3.41*e*+03	2.91*e*+03	3.48*e*+03	1.90*e*+03	4.83*e*+03
5	2.77*e*+03	4.33*e*+03	5.04*e*+03	3.30*e*+03	4.71*e*+03
6	2.77*e*+03	2.42*e*+03	3.53*e*+03	2.09*e*+03	5.69*e*+03
7	3.43*e*+03	3.46*e*+03	4.46*e*+03	2.67*e*+03	4.15*e*+03
8	3.14*e*+03	3.28*e*+03	4.57*e*+03	3.09*e*+03	4.69*e*+03
9	2.69*e*+03	3.19*e*+03	3.94*e*+03	2.84*e*+03	5.36*e*+03
10	3.06*e*+03	3.05*e*+03	3.26*e*+03	1.97*e*+03	5.02*e*+03
11	2.69*e*+03	3.54*e*+03	4.25*e*+03	2.42*e*+03	4.90*e*+03
12	2.69*e*+03	3.40*e*+03	3.63*e*+03	2.57*e*+03	5.37*e*+03
13	2.95*e*+03	2.29*e*+03	2.16*e*+03	1.92*e*+03	4.79*e*+03
14	2.16*e*+03	2.27*e*+03	2.60*e*+03	2.34*e*+03	5.56*e*+03
15	2.23*e*+03	2.06*e*+03	2.17*e*+03	2.38*e*+03	5.20*e*+03
16	2.66*e*+03	2.95*e*+03	3.15*e*+03	2.33*e*+03	5.48*e*+03
17	2.49*e*+03	2.23*e*+03	3.43*e*+03	1.28*e*+03	4.87*e*+03
18	2.97*e*+03	2.75*e*+03	3.72*e*+03	2.80*e*+03	4.54*e*+03
19	2.53*e*+03	2.74*e*+03	3.14*e*+03	1.76*e*+03	4.02*e*+03
20	2.35*e*+03	2.91*e*+03	3.09*e*+03	1.53*e*+03	5.03*e*+03
MEAN	2.80*e*+03	2.99*e*+03	3.51*e*+03	2.37*e*+03	4.93*e*+03
SD	3.41*e*+02	5.43*e*+02	7.16*e*+02	5.17*e*+02	4.30*e*+02
Time (s)	3.36*e*+02	3.65*e*+02	6.65*e*+02	5.04*e*+02	4.09*e*+02