A Markov Decision Process for Variable Selection in Branch & Bound

Paul Strang
EDF R&D, France
CNAM Paris, France
paul.strang@edf.fr

Zacharie Alès
ENSTA IP Paris, France
CNAM Paris, France
zacharie.ales@ensta.fr

Côme Bissuel
EDF R&D, France
come.bissuel@edf.fr

Olivier Juan
EDF R&D, France
olivier.juan@edf.fr

Safia Kedad-Sidhoum
CNAM Paris, France
safia.kedad_sidhoum@cnam.fr

Emmanuel Rachelson
ISAE-SUPAERO, France
emmanuel.rachelson@isae-supaero.fr

Abstract

Mixed-Integer Linear Programming (MILP) is a powerful framework used to address a wide range of NP-hard combinatorial optimization problems, often solved by Branch and bound (B&B). A key factor influencing the performance of B&B solvers is the variable selection heuristic governing branching decisions. Recent contributions have sought to adapt reinforcement learning (RL) algorithms to the B&B setting to learn optimal branching policies, through Markov Decision Processes (MDP) inspired formulations, and ad hoc convergence theorems and algorithms. In this work, we introduce BBMDP, a principled vanilla MDP formulation for variable selection in B&B, allowing to leverage a broad range of RL algorithms for the purpose of learning optimal B&B heuristics. Computational experiments validate our model empirically, as our branching agent outperforms prior state-of-the-art RL agents on four standard MILP benchmarks.

1 Introduction

Mixed-Integer Linear Programming (MILP) is a subfield of combinatorial optimization (CO), a discipline that aims at finding solutions to optimization problems with large but finite sets of feasible solutions. Specifically, mixed-integer linear programming addresses CO problems that are NP-hard, meaning that no polynomial-time resolution algorithm has yet been discovered to solve them. Mixed-integer linear programs are used to solve efficiently a vast range of high-dimensional combinatorial problems, spanning from operations research [Hillier and Lieberman, 2015] to the fields of deep learning [Tjeng et al., 2019], finance [Mansini et al., 2015], computational biology [Gusfield, 2019], and fundamental physics [Barahona, 1982]. MILPs are traditionally solved using Branch and bound (B&B) [Land and Doig, 1960], an algorithm which methodically explores the space of solutions by dividing the original problem into smaller sub-problems, while ensuring the optimality of the final returned solution. Intensively developed since the 1980s [Bixby, 2012], MILP solvers based on the B&B algorithm are high-performing tools. In particular, they rely on complex heuristics fine-tuned by experts on large heterogeneous benchmarks [Gleixner et al., 2021]. Hence, in the context of real-world applications, in which similar instances with slightly varying inputs are solved on a regular basis, there is a huge incentive to reduce B&B total solving time by

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

learning efficient tailor-made heuristics. The branching heuristic, or variable selection heuristic, which determines how to iteratively partition the space of solutions, has been found to be critical to B&B computational performance [Achterberg and Wunderling, 2013]. Over the last decade, many contributions have sought to harness the predictive power of machine learning (ML) to learn better-performing B&B heuristics [Bengio et al., 2021, Scavuzzo et al., 2024]. By using imitation learning (IL) to replicate the behaviour of a greedy branching expert at lower computational cost, Gasse et al. [2019] established a landmark result as they first managed to outperform a solver relying on human-expert heuristics. Building on the works of Gasse et al. [2019] and He et al. [2014], who proposed a Markov decision process (MDP) formulation for node selection in B&B, several contributions succeeded in learning efficient branching strategies by reinforcement [Etheve et al., 2020, Scavuzzo et al., 2022, Parsonson et al., 2022], without surpassing the performance achieved by the IL approach. Yet, if the performance of IL heuristics are caped by that of the suboptimal branching experts they learn from, the performance of RL branching strategies are, in theory, only bounded by the maximum score achievable. We note that in order to cope with dire credit assignment problems [Pignatelli et al., 2023] induced by the sparse reward model described in He et al. [2014], prior research efforts have shifted away from the traditional Markov decision process framework, finding it impractical for learning efficient branching strategies. Instead, Etheve et al. [2020], Scavuzzo et al. [2022] and Parsonson et al. [2022] have adopted unconventional MDP-inspired formulations to model variable selection in B&B.

In this work, we show that, despite improving the empirical performance of RL algorithms, these alternative formulations introduce approximations which undermine the asymptotic performance of RL branching agents in the general case. In order to address this issue, we introduce branch and bound Markov decision processes (BBMDP), a principled vanilla MDP formulation for variable selection in B&B, which preserves convergence properties brought by previous contributions without sacrificing optimality. Our new formulation allows to define a proper Bellman optimality operator, which, in turn, enables to unlock the full potential of state-of-the-art approximate dynamic programming algorithms [Hessel et al., 2017, Dabney et al., 2018, Farebrother et al., 2024] for the purpose of learning optimal B&B branching strategies. We evaluate our method on four classic MILP benchmarks, achieving state-of-the art performance and dominating previous RL agents while narrowing the gap with the IL approach of Gasse et al. [2019], as shown in Figure 1.

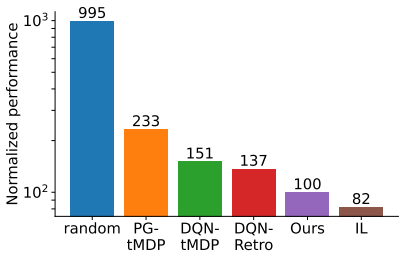


Figure 1: Normalized scores in log scale of IL, RL and random agents across the Ecole benchmark Prouvost et al. [2020].

2 Problem statement

2.1 Mixed-integer linear programming

We consider mixed-integer linear programs (MILPs), defined as:

$$P : \begin{cases} \min c^\top x \\ l \leq x \leq u \\ Ax \leq b \; ; \; x \in \mathbb{Z}^{|\mathcal{I}|} \times \mathbb{R}^{n-|\mathcal{I}|} \end{cases}$$

with n the number of variables, m the number of linear constraints, $l, u \in \mathbb{R}^n$ the lower and upper bound vectors, $A \in \mathbb{R}^{m \times n}$ the constraint matrix, $b \in \mathbb{R}^m$ the right-hand side vector, $c \in \mathbb{R}^n$ the objective function, and $\mathcal{I}$ the indices of integer variables. Throughout this document, we are interested in repeated MILPs of fixed dimension $\{P_i = (A_i, b_i, c_i, l_i, u_i)\}_{i \in N}$ sampled according to an unknown distribution $p_0 : \Omega \to \mathbb{R}^{m \times n} \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$.

In order to solve MILPs efficiently, the B&B algorithm iteratively builds a binary tree $(\mathcal{V}, \mathcal{E})$ where each node corresponds to a MILP, starting from the root node $v_0 \in \mathcal{V}$ representing the original problem P_0. The incumbent solution $\bar{x} \in \mathbb{Z}^{|\mathcal{I}|} \times \mathbb{R}^{n-|\mathcal{I}|}$ denotes the best feasible solution found at current iteration, its associated value $GUB = c^\top \bar{x}$ is called the *global upper bound* on the optimal value. The overall state of the optimization process is thus captured by the triplet $s = (\mathcal{V}, \mathcal{E}, \bar{x})$, we

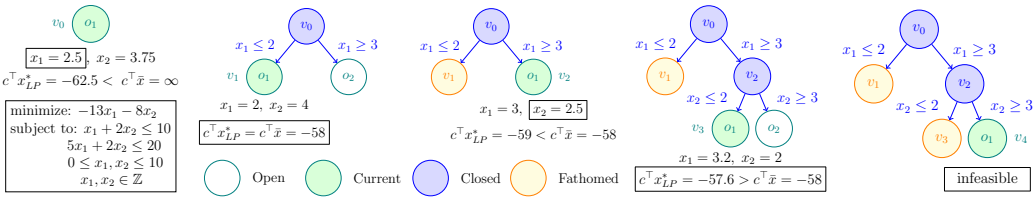


Figure 2: Solving a MILP by B&B using variable selection policy π and node selection policy ρ. Each node v_i represents a MILP derived from the original problem, each edge represents the bound adjustment applied to derive child nodes from their parent. Throughout the solving process, v_1 is pruned by integrity, v_3 is pruned by bound, and v_4 by infeasibility.

note $\mathcal{S}$ the set of all such triplets.[1] Throughout the optimization process, B&B nodes are explored sequentially. We note $\mathcal{C}$ the set of visited or closed nodes, and $\mathcal{O}$ the set of unvisited or open nodes, such that $\mathcal{V} = \mathcal{C} \cup \mathcal{O}$. Originally, $\mathcal{O} = \{v_0\}$ and $\mathcal{C} = \emptyset$. At each iteration, the node selection policy $\rho : \mathcal{S} \to \mathcal{O}$ selects the next node to explore. Since ρ necessarily defines a total order on nodes $o_i \in \mathcal{O}$, we can arrange indices such that $o_1 = \rho(s)$ denotes v_t the node currently explored at step t.

Figure 2 illustrates how B&B operates on an example. At each iteration, let $x^*_{LP} \in \mathbb{R}^n$ be the optimal solution to the linear relaxation of P_t, the problem associated with v_t. If P_t admits no solution, v_t is marked as visited and the branch is pruned by infeasibility. If $x^*_{LP} \in \mathbb{R}^n$ exists, and $GUB < c^\top x^*_{LP}$, no integer solution in the subsequent branch can improve GUB, thus v_t is marked as closed and the branch is pruned by bound. If x^*_{LP} is not dominated by $\bar{x}$ and x^*_{LP} is feasible (all integer variables in $x^*_{LP} \in \mathbb{R}^n$ have integer values), a new incumbent solution $\bar{x} = x^*_{LP}$ has been found. Hence GUB is updated and v_t is marked as visited while the branch is pruned by integrity. Else, x^*_{LP} admits fractional values for some integer variables. The branching heuristic $\pi : \mathcal{S} \to \mathcal{I}$ selects a variable x_b with fractional value $\hat{x}_b$, to partition the solution space. As a result, two child nodes (v_-, v_+), with associated MILPs $P_- = P_t \cup \{x_b \leq \lfloor \hat{x}_b \rfloor\}$ and $P_+ = P_t \cup \{x_b \geq \lceil \hat{x}_b \rceil\}$, are added to the current node.[2] Their linear relaxation is solved, before they are added to the set of open nodes $\mathcal{O}$ and v_t is marked as visited.

This process is repeated until $\mathcal{O} = \emptyset$ and $\bar{x}$ is returned. The dynamics of the B&B algorithm between two branching decisions can be described by the function $\kappa_\rho : \mathcal{S} \times \mathcal{I} \to \mathcal{S}$, such that $s' = \kappa_\rho(s, \pi(s))$. By design, B&B does not terminate before finding an optimal solution and proving its optimality. Consequently, optimizing the performance of B&B on a distribution of MILP instances is equivalent to minimizing the expected solving time of the algorithm. As Etheve [2021] evidenced, the variable selection strategy π is by far the most critical B&B heuristic in terms of computational performance. In practice, the total number of nodes of the B&B tree is used as an alternative metric to evaluate the performance of branching heuristics π, as it is a hardware-independent proxy for computational efficiency. Under these circumstances, given a fixed node selection strategy ρ, the optimal branching strategy π^* associated with a distribution p_0 of MILP instances can be defined as:

$$\pi^* = \arg\min_{\pi} \mathbb{E}_{P \sim p_0}(|BB_{(\pi,\rho)}(P)|), \tag{1}$$

with $|BB_{(\pi,\rho)}(P)|$ the size of the $B\&B$ tree after solving P to optimality following strategies (π, ρ).

2.2 Reinforcement learning

We consider the setting of discrete-time, deterministic MDPs [Puterman, 2014] defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, \mathcal{R})$. At each time step t, the agent observes $s_t \in \mathcal{S}$ the current state of the environment, before executing action $a_t \in \mathcal{A}$, and receiving reward $r_t = \mathcal{R}(s_t, a_t)$. The Markov transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ models the dynamics of the environment. In particular, it satisfies the Markov property: conditionally to s_t and a_t, s_{t+1} is independent of all past visited states and actions. Given a trajectory starting in state s_0 sampled according to the initial distribution p_0, the total gain is defined for all $t \geq 0$ as $G_t = \sum_{t'=t}^{\infty} \gamma^{t'-t} \mathcal{R}(s_{t'}, a_{t'})$, with $\gamma \in [0, 1]$. The

[1] To account for early resolution steps where no incumbent solution has yet been found, we define a special value for $\bar{x}$, whose $GUB = \infty$. For the sake of simplicity, we make this implicit in the remainder of the paper.

[2] $\hat{x}_b$ denotes the value of x_b in x^*_{LP}. We use the symbol $\cup$ to denote the refinement of the bound on x_b in P_t.

objective of an RL agent is to maximize the expected gain of the trajectories yielded by its action selection policy $\pi : \mathcal{S} \to \mathcal{A}$. This is equivalent to finding the policy maximizing value functions $V^\pi(s_t) = \mathbb{E}_{a_{t'} \sim \pi(s_{t'})}[G_t]$ and $Q^\pi(s_t, a_t) = \mathcal{R}(s_t, a_t) + \gamma V^\pi(s_{t+1})$. The optimal Q-value function Q^* indicates the highest achievable cumulated gain in the MDP. It satisfies the Bellman optimality equation $Q(s, a) = \mathcal{R}(s, a) + \gamma \cdot \max_{a' \in \mathcal{A}} Q(s', a')$, noting $s' = \mathcal{T}(s, a)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$. The optimal policy is retrieved by acting greedily according to the learned Q-value function: $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a)$.

2.3 Related work

Following the seminal work by Gasse et al. [2019], few contributions have proposed to build more complex neural network architectures based on transformers [Lin et al., 2022] and recurrence mechanisms [Seyfi et al., 2023] to improve the performance of IL branching agents, with moderate success. In parallel, theoretical and computational analysis [Bestuzheva et al., 2021, Sun et al., 2022] have shown that neural networks trained by imitation could not rival the tree size performance achieved by strong branching (SB), the branching expert used in Gasse et al. [2019]. In fact, low tree sizes associated with SB turn out to be primarily due to the formulation improvements resulting from the massive number of LPs solved in SB, not to the intrinsic quality of the branching decisions themselves.

Since branching decisions are made sequentially, reinforcement learning appears as a natural candidate to learn good branching policies. Etheve et al. [2020] and Scavuzzo et al. [2022] proposed the model of TreeMDP, in which state $s_i = (P_i, x^*_{LP,i}, \bar{x}_i)$ consists in the MILP associated with node v_i along with the solution of its linear relaxation and the incumbent solution at v_i. The actions available at s_i are the set of fractional variables in $x^*_{LP,i}$. Given (s_i, a_i) the tree Markov transition function produces two child node states (s_i^-, s_i^+) that can be visited in any order. Crucially, when the B&B tree is explored in depth-first-search (DFS), TreeMDP trajectories can be divided in independent subtrees, allowing to learn policies minimizing the size of each subtree independently. This helps mitigate credit assignment issues that arise owing to the length of episode trajectories. Subsequently, Parsonson et al. [2022] found the DFS node selection policy to be highly detrimental to the computational performance of RL branching strategies. Assuming that RL branching agents trained following advanced node selection strategies would perform better despite the lack of theoretical guarantee, they proposed to learn from retrospective trajectories, diving trajectories built from original TreeMDP episodes. In fact, Parsonson et al. [2022] found retrospective trajectories to alleviate the partial observability induced by the "disordered" exploration of the tree and outperform prior RL agents.

A large body of work has proposed to learn, either by imitation or reinforcement, better-performing B&B heuristics outside of variable selection [Nair et al., 2021, Paulus et al., 2022]. RL contributions in primal search [Sonnerat et al., 2022, Wu and Lisser, 2023] node selection [He et al., 2014, Etheve, 2021] and cut selection [Tang et al., 2020, Song et al., 2020, Wang et al., 2023] have all relied on the TreeMDP framework to train their agents, simply adapting the action set to the task at hand. Finally, machine learning applications in combinatorial optimization are not limited to B&B. For example, in the context of routing or scheduling problems where exact resolution rapidly becomes prohibitive, agents are trained to learn direct search heuristics [Kool et al., 2019, Grinsztajn et al., 2022, Chalumeau et al., 2023] yielding high-quality feasible solutions.

3 Branch and bound Markov decision process

By using the current B&B node as the observable state, prior attempts to learn optimal branching strategies have relied on the TreeMDP formalism to train RL agents. However, TreeMDPs are not MDPs, as they do not define a Markov process on the state random variable (for instance, a transition yields two states and is hence not a stochastic process on the state variables). As a result, this forces Etheve et al. [2020] and Scavuzzo et al. [2022] to redefine Bellman updates and derive *ad hoc* convergence theorems for TD(0), value iteration, and policy gradient algorithms. In order to leverage broader theoretical results from the reinforcement learning literature, we propose a description of variable selection in B&B as a proper Markov decision process.

3.1 Definition

The problem of finding an optimal branching strategy according to Eq. (1) can be described as a regular deterministic Markov decision process. To this end, we introduce Branch and bound Markov decision processes (BBMDP), taking $\gamma = 1$ since episodes horizons are bounded by the (finite) largest possible number of nodes:

State-action space. $\mathcal{S}$ is the set of all B&B trees $s_t = (\mathcal{V}_t, \mathcal{E}_t, \bar{x}_t)$. Note that this includes intermediate B&B trees, whose incumbent solutions $\bar{x}_t$ are yet to be proven optimal. $\mathcal{A}$ is the set of all integer variables indices $\mathcal{I}$.

Transition function. The Markov transition function is defined as $\mathcal{T} = \kappa_\rho$ with κ_ρ the branching operation described in Section 2.1. Note that if the variable associated with a_t is not fractional in x^*_{LP}, then $s_{t+1} = \mathcal{T}(s_t, a_t) = s_t$ as relaxing a variable that is not fractional has no impact on the LP relaxation. Importantly, all states for which $\mathcal{O} = \emptyset$ are terminal states.

Starting states. Initial states are single node trees, where the root node is associated to a MILP P_0 drawn according to the distribution p_0 defined in Section 2.1 (hence the use of p_0 for both the initial problem P_0 and the MDP's initial state s_0).

Reward model. We define $\mathcal{R}(s, a) = -2$ for all transitions until episode termination. Since each transition adds two B&B nodes, the overall value of a trajectory is the opposite of the number of nodes added to the B&B tree from the root node, which aligns with the definition of Eq. (1).

Unlike in TreeMDP, the current state is defined as the state of the entire B&B tree, rather than merely the current B&B node. The transition function returns a B&B tree whose open nodes are sorted according to the node selection policy ρ, thus reflecting the true dynamics of the B&B algorithm, instead of a couple of pseudo-states associated with the child nodes of the last node expansion. Note that the definition above sets BBMDPs among the specific class of MDPs called stochastic shortest path problems [Puterman, 2014].

3.2 Misconceptions when learning to explore B&B trees

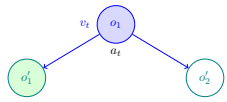
Like in TreeMDP, episode trajectories can be decomposed in independent subtree trajectories, to facilitate RL agents training. Consider a deterministic branching policy π, and let us we rewrite V^π and Q^π to exhibit the tree structure. Given an open node $o_i \in \mathcal{O}_t$, we note $T(o_i)$ the subtree rooted in o_i, and define $\bar{V}^\pi(s_t, o_i)$ the function that returns the opposite of the size of the subtree rooted in o_i, when branching according to policy π starting from state s_t until episode termination. Then V^π can be expressed as:

$$V^\pi(s_t) = \sum_{o_i \in \mathcal{O}_t} \bar{V}^\pi(s_t, o_i). \tag{2}$$

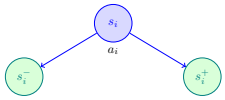
In plain words, the total number of nodes that will be added to the B&B tree past s_t is the sum of the sizes of all the subtrees $T(o_i)$ rooted in the open nodes of s_t. Because the full B&B tree is a complex object to manipulate, it is tempting to discard the tree structure in s_t and define $\bar{V}$-value functions merely as functions of $(o_i, \bar{x}_{o_i})$, rather than functions of (s_t, o_i), where $\bar{x}_{o_i} \in \mathbb{R}^n$ is the incumbent solution when o_i is processed by the B&B algorithm, at time step τ_i. The rationale for such value functions is that the size of the subtree rooted in $o_i \in \mathcal{O}_t$, for a given incumbent solution $\bar{x}_{o_i}$, should be the same, regardless of the parents of o_i, its position in the tree, or the branching decisions taken in subtrees $T(o_j)$ for $j \neq i$. **It turns out, this last statement does not always hold**, quite counter-intuitively. Let us write τ_i the time steps at which the nodes $o_i \in \mathcal{O}_t$ are selected by the node selection strategy ρ.[3] Now, consider for instance a node selection procedure ρ that performs a breadth-first search through the tree. The number of nodes in $T(o_i)$ will depend strongly on whether an improved incumbent solution $\bar{x}_{o_i}$ was found in the subtrees explored between s_t and s_{τ_i}, and, in turn, on the branching decisions taken in these subtrees. This example highlights the major importance of the node selection strategy ρ, when one wishes to define subtree sizes based solely on $(o_i, \bar{x}_{o_i})$.

Consider now two open nodes o_i and o_j in $\mathcal{O}_t$. Conversely to the previous example, if one can guarantee that the subtree rooted in o_j will be solved to optimality before o_i is considered for expansion in the B&B process, then the number of nodes in $T(o_i)$ will not be affected by the

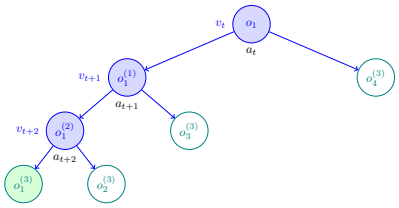
[3] Following our indexation of $o_i \in \mathcal{O}_t$, we have $t = \tau_1 < ... < \tau_i < ... < \tau_{|\mathcal{O}_t|}$.



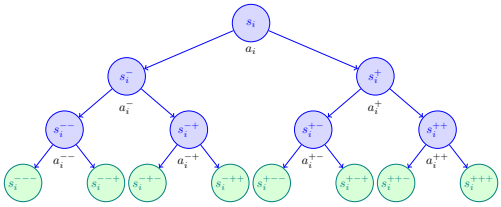
(a) 1-step BBMDP Bellman equation:
$\bar{V}^\pi(o_1, \bar{x}_{o_1}) = -2 + \bar{V}^\pi(o'_1, \bar{x}_{o'_1}) + \bar{V}^\pi(o'_2, \bar{x}_{o'_2})$



(b) 1-step TreeMDP Bellman equation:
$V^\pi(s_i) = -2 + V^\pi(s_i^-) + V^\pi(s_i^+)$



(c) k-step BBMDP Bellman equation:
$\bar{V}^\pi(o_1, \bar{x}_{o_1}) = -2k + \sum_{i=1}^{k+1} \bar{V}^\pi(o_i^{(k)}, \bar{x}_{o_i^{(k)}})$



(d) k-step TreeMDP Bellman equation:
$V^\pi(s_0) = -2^1 - 2^2 - ... - 2^k + \sum_{i=1}^{2^k} V^\pi(s_i^j)$

Figure 3: When minimizing 1-step temporal difference, TreeMDP and BBMDP yield equivalent results, see 3a, 3b. However, over k-step trajectories, the two methods diverge as shown in 3c, 3d.

branching decisions taken at any node under o_j. In fact, if o_j is solved to optimality, $\bar{x}_{o_i}$ will either not change if no feasible solution in $T(o_j)$ improves GUB, or either be the best feasible solution of the MILP associated with o_j, which does not depend on the series of actions taken in $T(o_j)$. In other words, to make sure that the size of $T(o_i)$ does only depend on the branching decisions taken in $T(o_i)$, all nodes $o_j \in \mathcal{O}_t$ must have been either fully explored or strictly unexplored at τ_i. Applying this argument recursively induces that the only node selection strategy which enables predicting a subtree size only based on $(o_i, \bar{x}_{o_i})$, is a depth-first search (DFS) exploration of the B&B tree. Therefore, we consider $\rho = DFS$ and write $\bar{V}^\pi(o_i, \bar{x}_{o_i})$ the opposite of the size of $T(o_i)$ in this context. We can now derive a refined Bellman update to train branching agents in BBMDP.

Proposition 3.1. *In DFS-BBMDP, the Bellman equation $V^\pi(s) = \mathcal{R}(s, \pi(s)) + V^\pi(s')$ yields:*

$$\bar{V}^\pi(o_1, \bar{x}_{o_1}) = -2 + \bar{V}^\pi(o'_1, \bar{x}_{o'_1}) + \bar{V}^\pi(o'_2, \bar{x}_{o'_2}), \tag{3}$$

with $o_1 = \rho(s)$, $o'_1 = \rho(\mathcal{T}(s, \pi(s))$, and o'_2 is the sibling of o'_1 in the B&B tree.

Proof. Eq. (3) follows directly from injecting Eq. (2) in the Bellman equation, and observing that most terms in the sums simplify as $\bar{V}^\pi(o_i, \bar{x}_{o_i}) = \bar{V}^\pi(o'_{i+1}, \bar{x}_{o'_{i+1}})$ for $i \geq 2$. □

Keeping the same notation convention for o_1, o'_1 and o'_2, we define $\bar{Q}^\pi(o_1, \bar{x}_{o_1}, a) = -2 + \bar{V}^\pi(o'_1, \bar{x}_{o'_1}) + \bar{V}^\pi(o'_2, \bar{x}_{o'_2})$. Analogous to $\bar{V}^\pi$, the $\bar{Q}^\pi$ function returns the opposite of the size of the subtree rooted in $o_i \in \mathcal{O}_t$ when branching on action $a \in \mathcal{A}$ at o_i and following policy π until $T(o_i)$ is fathomed. Note that if $\bar{V}^\pi$ and $\bar{Q}^\pi$ are not strictly value functions, they naturally emerge when applying Bellman equations to BBMDP value functions under $\rho = DFS$. We stress that, in depth-first search BBMDPs, it is not necessary to learn Q^* to derive π^*, since acting according to a policy minimizing the size of the subtree rooted in the current B&B node is equivalent to acting according to a global optimal policy:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a) = \arg\max_{a \in \mathcal{A}} \bar{Q}^*(o_1, \bar{x}_{o_1}, a). \tag{4}$$

3.3 Approximate dynamic programming

The previous properties enable learning π^* by training a neural network to approximate $\bar{Q}^*$ using traditional temporal difference (TD) algorithms. Notably, $\bar{Q}^*$ is easier to learn than Q^*, as it relies solely on quantities observable at time t, whereas the previous decomposition of Q^* depends on all $\bar{x}_{o_i}$ for $o_i \in \mathcal{O}_t$, which are not known at time t. Moreover, $\bar{Q}^*$ trains on much shorter trajectories than Q^*, which helps mitigate credit assignement issues.[4]

[4] On average, the length of subtree trajectories is logarithmically shorter than the length of BBMDP episodes.

Consider a transition $(s,\, a,\, r,\, s')$. Following Eq. (3), the Bellman optimality equation yields a sequence of $\bar{Q}_n$ functions via dynamic programming updates of the form:

$$\bar{Q}_{n+1}(o_1, \bar{x}_{o_1}, a) = -2 + \max_{a' \in \mathcal{A}} \bar{Q}_n(o'_1, \bar{x}_{o'_1}, a') + \max_{a'' \in \mathcal{A}} \bar{Q}_n(o'_2, \bar{x}_{o'_2}, a''). \tag{5}$$

One can also consider a k-step Bellman operator ($k \geq 1$), generalizing Eq. (5). Let π be a policy, and $s^{(k)}$ the state reached after a first application of a from s, and $k-1$ subsequent applications of π. Provided the subtree rooted in o_1 has not been fathomed within these k time steps, $s^{(k)}$ has $k+1$ new open nodes which we label $o_i^{(k)}$. Then, the Bellman update becomes:

$$\bar{Q}_{n+1}(o_1, \bar{x}_{o_1}, a) = -2k + \sum_{i=1}^{k+1} \max_{a' \in \mathcal{A}} \bar{Q}_n(o_i^{(k)}, \bar{x}_{o_i^{(k)}}, a'). \tag{6}$$

This Bellman update can be approximated using a neural network $\mathbf{q}_\theta$, whose weights are trained to minimize the k-step temporal difference loss $\mathcal{L}(\theta) = \mathbb{E}[(\mathbf{q}_\theta(o_1, \bar{x}_{o_1}, a) - \bar{Q}_{n+1}(o_1, \bar{x}_{o_1}, a))^2]$, where $\bar{Q}_{n+1}$ is bootstrapped by the previous $\mathbf{q}_\theta$. Following the work of Farebrother et al. [2024] on training value functions via classification, we also introduce a HL-Gauss cross-entropy loss adapted to the B&B setting $\mathcal{L}(\theta) = \mathbb{E}[\mathbf{q}_\theta(o_1, \bar{x}_{o_1}, a) \cdot \log p_{hist}(\bar{Q}_{n+1}(o_1, \bar{x}_{o_1}, a))]$ where p_{hist} is the function that encodes values in categorical histogram distributions, see Appendix B for complete description and theoretical motivation.

3.4 BBMDP vs TreeMDP

As it evacuates the core MDP notions of temporality and sequentiality, TreeMDP fails to describe variable selection in B&B accurately in the general case. This is illustrated in Figure 3 and further detailed in Appendix A: although the TreeMDP model is a valid approximation of BBMDP when training an RL agent to minimize temporal difference on one-step trajectories, it produces inconsistent learning schemes when evaluating temporal difference across multi-steps trajectories. In fact, repeatedly applying the tree Bellman operator from Etheve [2021] yields B&B trees that cannot, in general, be produced by DFS, as shown in Figure 3 (d).

Hence, BBMDP leverages the results first established by Etheve et al. [2020] and Scavuzzo et al. [2022] — in DFS, minimizing the whole B&B tree size is achieved when any subtree is of minimal size (Eq. (4)) — all while preserving MDP properties. Crucially, BBMDP allows to harness RL algorithms that are not compatible with the TreeMDP framework, such as k-step return temporal difference, TD(λ), or any RL algorithms using MCTS as policy improvement operator [Grill et al., 2020], as illustrated in Appendix A. In the same fashion, BBMDP can be applied to augment the pool of RL algorithms available for learning improved cut selection and primal search heuristics, simply by adapting the action set and the reward model to the task at hand. Complete exploration and comparison of such algorithms is beyond the scope of this paper and left for future work, as our primary objective here is to study the conditions under which B&B search can be cast as an MDP, and provide the community with a solid basis for principled algorithms.

4 Experiments

We now compare our branching agent against prior IL and RL approaches. For our experiments, we use the open-source solver SCIP 8.0.3 [Bestuzheva et al., 2021] as backend MILP solver, along with the Ecole library [Prouvost et al., 2020] both for instance generation and environment simulation. To foster reproducibility, our implementation and pretrained models are made publicly available at https://github.com/abfariah/bbmdp.

4.1 Experimental setup

Benchmarks. We consider the usual standard MILP benchmarks for learning branching strategies: set covering, combinatorial auctions, maximum independent set and multiple knapsack problems. We train and test on instances of same dimensions as Scavuzzo et al. [2022] and Parsonson et al. [2022], see Appendix F. As to SCIP configuration, as in previous work, we set the time limit to one hour, disable restart, and deactivate cut generation beyond root node. All the other parameters are left at their default value.

Table 1: Performance comparison of branching agents on four standard MILP benchmarks. For each method, we report total number of B&B nodes, presolve time and total solving time outside of presolve. Lower is better, **red** indicates best agent overall, **blue** indicates best among RL agents. Presolve is common to all methods. Following prior works, we report geometrical mean over 100 test instances unseen during training and over 100 higher-dimensional transfer instances. Norm. Score denotes the aggregate average performance obtained by each agent across the four MILP benchmarks, normalized by the score of DQN-BBMDP.

Test instances

	Set Covering		**Comb. Auction**		**Max. Ind. Set**		**Mult. Knapsack**		**Norm. Score**	
Method	Node	Time	Node	Time	Node	Time	Node	Time	Node	Time
Presolve	–	4.74	–	0.90	–	1.78	–	0.20	–	–
Random	3289	5.94	1111	2.16	386.8	2.01	733.5	0.55	995	374
SB	35.8	12.93	28.2	6.21	24.9	45.87	161.7	0.69	36	2358
SCIP	62.0	2.27	20.2	1.77	19.5	2.44	289.5	**0.53**	51	253
IL	**133.8**	0.90	**83.6**	0.65	**40.1**	**0.36**	272.0	0.69	**82**	**95**
IL-DFS	136.4	**0.74**	95.5	**0.56**	69.4	0.44	472.8	1.07	114	129
PG-tMDP	649.4	2.32	168.0	0.94	153.6	0.92	436.9	1.57	233	206
DQN-tMDP	175.8	0.83	203.3	1.11	168.0	1.00	266.4	0.73	151	136
DQN-Retro	183.0	1.14	103.2	0.78	223.0	1.81	250.3	0.67	137	160
DQN-BBMDP	**152.3**	**0.77**	**97.9**	**0.62**	**103.2**	**0.69**	**236.6**	**0.66**	**100**	**100**

Transfer instances

	Set Covering		**Comb. Auction**		**Max. Ind. Set**		**Mult. Knapsack**		**Norm. Score**	
Method	Node	Time	Node	Time	Node	Time	Node	Time	Node	Time
Presolve	-	12.3	-	2.67	-	5.16	-	0.46	–	–
Random	271632	842	317235	749	215879	2102	93452	70.6	5555	2737
SB	672.1	398	389.6	255	169.9	2172	**1709**	**12.5**	9	1425
SCIP	3309	48.4	1376	14.77	3368	90.0	30620	22.1	62	90
IL	**2610**	23.1	**1309**	**9.4**	**1882.0**	**38.6**	9747	43.5	**39**	**54**
IL-DFS	3103	**22.5**	1802	10.2	3501	51.9	43224	131	75	80
PG-tMDP	44649	221	6001	30.7	**3133**	**39.5**	35614	123	298	223
DQN-tMDP	8632	71.3	20553	116	45634	477	**22631**	**65.1**	439	445
DQN-Retro	6100	59.4	2908	18.4	119478	1863	27077	79.5	494	662
DQN-BBMDP	**5651**	**46.4**	**2273**	**11.8**	7168	81.3	37098	109	**100**	**100**

Baselines. We compare our DQN-BBMDP agent against DQN-TreeMDP (DQN-tMDP) [Etheve et al., 2020], REINFORCE-TreeMDP (PG-tMDP) [Scavuzzo et al., 2022] and the current state-of-the-art DQN-Retro [Parsonson et al., 2022] agents. We also compare against the reference IL expert from Gasse et al. [2019] as well as IL-DFS, which is the same expert, only evaluated following a depth-first search node selection policy. More details on these baselines can be found in Appendix C. Finally, we report the performance of reliability pseudo cost branching (SCIP), the default branching heuristic used in SCIP, strong branching (SB) [Applegate et al., 1995], the greedy expert from which the IL agent learns from, and random branching (Random), which randomly selects a fractional variable.

Network architecture. Following prior work, we use the bipartite graph representation introduced by Gasse et al. [2019] augmented by the features proposed in Parsonson et al. [2022] to represent B&B nodes. Additionally, we use the graph convolutional architecture of Gasse et al. [2019] to parameterize our Q-value network; see Appendix E for a more detailed description.

Training & evaluation. Models are trained on instances of each benchmark separately, and evaluated on test instances and transfer instances. Validation curves can be found in Appendix H. For evaluation, we report the node and time performance over 100 test instances unseen during training, as well as on 100 transfer instances of higher dimensions (see Table 4 in Appendix F). At evaluation, performance scores are averaged over 5 seeds. Importantly, when comparing a machine learning (IL or RL) branching strategy with a standard SCIP heuristic, time performance is the only relevant criterion. In fact, when implementing one of its own branching rules, SCIP triggers a series of techniques strengthening the current MILP formulation. If these techniques effectively reduce the number of

Table 2: Ablation impact of BBMDP, HL-Gauss loss and DFS. We remove one component one at the time, and evaluate corresponding versions on 100 set covering test instances after training for 200k gradient steps as described in section 4.1.

k-step return	DQN-BBMDP	w.o. DFS	w.o. HL-Gauss	w.o. BBMDP	DQN-TreeMDP
$k = 1$	158.9	156.2 (-2%)	175.8$(+10\%)$	158.9$(+0\%)$	175.8$(+10\%)$
$k = 3$	$\mathbf{152.3}(-4\%)$	150.1 (-5%)	172.3 $(+8\%)$	162.1 $(+2\%)$	178.9 $(+13\%)$

nodes to visit, they incur computational overhead which ultimately increases SCIP overall solving time. This renders node comparisons between ML and non-ML branching strategies negligible relative to solving time evaluations, as observed by Gamrath and Schubert [2018], Scavuzzo et al. [2022].

4.2 Results

Computational results obtained on the four benchmarks are presented in Table 1. On test instances, **DQN-BBMDP consistently obtains best performance** among RL agents. When compared against prior state-of-the-art DQN-Retro, DQN-BBMDP achieves an aggregate average 27% reduction of total number of node and 38% reduction of solving time outside presolve across the four Ecole benchmarks, as reported in Figure 1. Contrary to Parsonson et al. [2022], we find DQN-Retro to yield performance comparable to DQN-tMDP. Remarkably, all RL agents outperform the SCIP solver on 3 out of 4 benchmarks in terms of solving time.

On transfer instances, DQN-BBMDP also dominates among RL agents, although it is outperformed by PG-tMDP on maximum independent set instances and by DQN-Retro on multiple knapsack instances. The aggregate performance gap between DQN-BBMDP and other RL baselines is notably wider on transfer instances, which aligns with the advantages of using a principled MDP formulation over TreeMDP. In fact, DQN-BBMDP is the first RL agent to demonstrate robust generalization capabilities on higher dimensional instances, outperforming SCIP on 3 out of 4 benchmarks.

Additional performance metrics are provided in Appendix I. A notable metric is the number of wins, *i.e.* the number of test instances that one algorithm solves faster than any other baseline. DQN-BBMDP outperforms even the IL approaches both in average solving time and total number of wins, on the set covering, combinatorial auction and multiple knapsack test instances.

4.3 Ablation study

We perform an ablation study on set covering test instances to separate the performance gain associated with BBMDP and the HL-Gauss classification loss. Since BBMDP and TreeMDP are strictly equivalent when minimizing one-step temporal difference, we evaluate the performance gap between one-step and k-step TD learning for both DQN-BBMDP and DQN-TreeMDP.

As shown in Table 2, we find that the bulk of the performance gain is brought by the use of a cross-entropy loss. Nonetheless, we observe that the use of a multi-step TD loss improves the performance of DQN-BBMDP, but degrades the performance of DQN-TreeMDP. This supports our initial claim that, in the general case, despite improving the empirical properties of RL algorithms, TreeMDP introduces approximations which hinder the asymptotic performance achievable by RL agents. Following Parsonson et al. [2022], we also evaluate the cost of opting for depth-first search instead of best estimate search, SCIP's default node selection policy, when learning branching strategies. Contrary to their work, we find DFS not to be restrictive in practice in terms of performance. We further investigate theses discrepancies in Appendix D.

5 Conclusion and perspectives

Guiding combinatorial optimization algorithms with reinforcement learning has proven challenging, including beyond mixed-integer linear programming [Berto et al., 2023]. Not only are RL agents consistently outperformed by human-expert CO heuristics or IL agents trained to mimic these experts, but their application has also been limited so far to fairly easy problem instances. This begs for a thorough study of these problems' properties, for well-posed formulations and principled

methods. In this work, we showed the theoretical and practical limits of the concept of TreeMDP for learning optimal branching strategies in MILP, highlighting in which context TreeMDPs were a valid formulation. Introducing BBMDP, we proposed a rigorous description of variable selection in B&B, unlocking the use of vanilla dynamic programming methods. In turn, the resulting DQN-BBMDP method outperformed prior RL-inspired agents on the Ecole benchmark.

Nonetheless, there remains significant room for improvement for RL approaches on both test and transfer instances, suggesting that the generalization capacity of RL agents still lags behind that of IL. We believe this to be mainly due to out of distribution effects, since transfer instances are intrinsically more complex than training instances, and Q-networks are lead to evaluate B&B nodes with subtree sizes far exceeding those encountered during training. In contrast, IL networks learn to predict the strong branching action via behavioral cloning, making them robust to such out-of-distribution scaling effects. Yet, we believe that building on a principled MDP formulation of variable selection in B&B is a stepping stone to achieve substantial acceleration of solving time for higher-dimensional MILPs in the future, also opening avenues to tackle distribution shift and domain adaptation.

References

Tobias Achterberg and Roland Wunderling. Mixed integer programming: Analyzing 12 years of progress. In *Facets of combinatorial optimization: Festschrift for martin grötschel*, pages 449–481. Springer, 2013.

David Applegate, Robert Bixby, Vašek Chvátal, and William Cook. *Finding cuts in the TSP (A preliminary report)*, volume 95. Citeseer, 1995.

Egon Balas and Andrew Ho. Set covering algorithms using cutting planes, heuristics, and subgradient optimization: a computational study. *Combinatorial Optimization*, pages 37–60, 1980.

Francisco Barahona. On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241, 1982.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, April 2021. ISSN 0377-2217. doi: 10.1016/j.ejor.2020.07.063. URL `https://www.sciencedirect.com/science/article/pii/S0377221720306895`.

David Bergman, Andre A Cire, Willem-Jan Van Hoeve, and John Hooker. *Decision diagrams for optimization*, volume 1. Springer, 2016.

Federico Berto, Chuanbo Hua, Junyoung Park, Minsu Kim, Hyeonah Kim, Jiwoo Son, Haeyeon Kim, Joungho Kim, and Jinkyoo Park. RL4CO: an Extensive Reinforcement Learning for Combinatorial Optimization Benchmark, June 2023. URL `http://arxiv.org/abs/2306.17100`. arXiv:2306.17100 [cs].

Ksenia Bestuzheva, Mathieu Besançon, Wei-Kun Chen, Antonia Chmiela, Tim Donkiewicz, Jasper van Doornmalen, Leon Eifler, Oliver Gaul, Gerald Gamrath, Ambros Gleixner, Leona Gottwald, Christoph Graczyk, Katrin Halbig, Alexander Hoen, Christopher Hojny, Rolf van der Hulst, Thorsten Koch, Marco Lübbecke, Stephen J. Maher, Frederic Matter, Erik Mühmer, Benjamin Müller, Marc E. Pfetsch, Daniel Rehfeldt, Steffan Schlein, Franziska Schlösser, Felipe Serrano, Yuji Shinano, Boro Sofranac, Mark Turner, Stefan Vigerske, Fabian Wegscheider, Philipp Wellner, Dieter Weninger, and Jakob Witzig. The SCIP Optimization Suite 8.0. Technical Report, Optimization Online, December 2021. URL `http://www.optimization-online.org/DB_HTML/2021/12/8728.html`.

Robert E. Bixby. A brief history of linear and mixed-integer programming computation. *Documenta Mathematica*, 2012:107–121, 2012.

Felix Chalumeau, Shikha Surana, Clément Bonnet, Nathan Grinsztajn, Arnu Pretorius, Alexandre Laterre, and Tom Barrett. Combinatorial optimization with policy adaptation using latent space search. *Advances in Neural Information Processing Systems*, 36:7947–7959, 2023.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

Marc Etheve. *Solving repeated optimization problems by Machine Learning*. phdthesis, HESAM Université, December 2021. URL https://theses.hal.science/tel-03675471.

Marc Etheve, Zacharie Alès, Côme Bissuel, Olivier Juan, and Safia Kedad-Sidhoum. Reinforcement Learning for Variable Selection in a Branch and Bound Algorithm. In Emmanuel Hebrard and Nysret Musliu, editors, *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, Lecture Notes in Computer Science, pages 176–185, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58942-4. doi: 10.1007/978-3-030-58942-4_12.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop Regressing: Training Value Functions via Classification for Scalable Deep RL, March 2024. URL http://arxiv.org/abs/2403.03950. arXiv:2403.03950 [cs, stat].

Alex S Fukunaga. A branch-and-bound algorithm for hard multiple knapsack problems. *Annals of Operations Research*, 184(1):97–119, 2011.

Gerald Gamrath and Christoph Schubert. Measuring the impact of branching rules for mixed-integer programming. In *Operations Research Proceedings 2017: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR), Freie Universiät Berlin, Germany, September 6-8, 2017*, pages 165–170. Springer, 2018.

Maxime Gasse, Didier Chetelat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact Combinatorial Optimization with Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d14c2267d848abeb81fd590f371d39bd-Abstract.html.

Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christophel, Kati Jarck, Thorsten Koch, and Jeff Linderoth. MIPLIB 2017: data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021. Publisher: Springer.

Jean-Bastien Grill, Florent Altché, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, and Rémi Munos. Monte-Carlo Tree Search as Regularized Policy Optimization, July 2020. URL http://arxiv.org/abs/2007.12509. arXiv:2007.12509 [cs, stat].

Nathan Grinsztajn, Daniel Furelos-Blanco, and Thomas D. Barrett. Population-Based Reinforcement Learning for Combinatorial Optimization, October 2022. URL http://arxiv.org/abs/2210.03475. arXiv:2210.03475 [cs].

Dan Gusfield. *Integer linear programming in computational and systems biology: an entry-level text and course*. Cambridge University Press, 2019.

He He, Hal Daume III, and Jason M Eisner. Learning to Search in Branch and Bound Algorithms. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://papers.nips.cc/paper/2014/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining Improvements in Deep Reinforcement Learning, October 2017. URL http://arxiv.org/abs/1710.02298. arXiv:1710.02298 [cs].

Frederick S Hillier and Gerald J Lieberman. *Introduction to operations research*. McGraw-Hill, 2015.

Ehsan Imani and Martha White. Improving Regression Performance with Distributional Losses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2157–2166. PMLR, July 2018. URL https://proceedings.mlr.press/v80/imani18a.html. ISSN: 2640-3498.

Wouter Kool, Herke van Hoof, and Max Welling. Attention, Learn to Solve Routing Problems!, February 2019. URL http://arxiv.org/abs/1803.08475. arXiv:1803.08475 [cs, stat].

AH Land and AG Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960.

Kevin Leyton-Brown, Mark Pearson, and Yoav Shoham. Towards a universal test suite for combinatorial auction algorithms. In *Proceedings of the 2nd ACM conference on Electronic commerce*, pages 66–76, 2000.

Jiacheng Lin, Jialin Zhu, Huangang Wang, and Tao Zhang. Learning to branch with Tree-aware Branching Transformers. *Knowledge-Based Systems*, 252:109455, September 2022. ISSN 0950-7051. doi: 10.1016/j.knosys.2022.109455. URL https://www.sciencedirect.com/science/article/pii/S0950705122007298.

Renata Mansini, Wodzimierz Ogryczak, M Grazia Speranza, and EURO: The Association of European Operational Research Societies. *Linear and mixed integer programming for portfolio optimization*, volume 21. Springer, 2015.

Vinod Nair, Sergey Bartunov, Felix Gimeno, Ingrid von Glehn, Pawel Lichocki, Ivan Lobov, Brendan O'Donoghue, Nicolas Sonnerat, Christian Tjandraatmadja, Pengming Wang, Ravichandra Addanki, Tharindi Hapuarachchi, Thomas Keck, James Keeling, Pushmeet Kohli, Ira Ktena, Yujia Li, Oriol Vinyals, and Yori Zwols. Solving Mixed Integer Programs Using Neural Networks, July 2021. URL http://arxiv.org/abs/2012.13349. arXiv:2012.13349 [cs, math].

Christopher W. F. Parsonson, Alexandre Laterre, and Thomas D. Barrett. Reinforcement Learning for Branch-and-Bound Optimisation using Retrospective Trajectories, December 2022. URL http://arxiv.org/abs/2205.14345. arXiv:2205.14345 [cs].

Max B. Paulus, Giulia Zarpellon, Andreas Krause, Laurent Charlin, and Chris Maddison. Learning to Cut by Looking Ahead: Cutting Plane Selection via Imitation Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17584–17600. PMLR, June 2022. URL https://proceedings.mlr.press/v162/paulus22a.html. ISSN: 2640-3498.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *arXiv preprint arXiv:2312.01072*, 2023.

Antoine Prouvost, Justin Dumouchelle, Lara Scavuzzo, Maxime Gasse, Didier Chételat, and Andrea Lodi. Ecole: A Gym-like Library for Machine Learning in Combinatorial Optimization Solvers, November 2020. URL http://arxiv.org/abs/2011.06069. arXiv:2011.06069 [cs, math].

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Lara Scavuzzo, Feng Yang Chen, Didier Chételat, Maxime Gasse, Andrea Lodi, Neil Yorke-Smith, and Karen Aardal. Learning to branch with Tree MDPs, October 2022. URL http://arxiv.org/abs/2205.11107. arXiv:2205.11107 [cs, math].

Lara Scavuzzo, Karen Aardal, Andrea Lodi, and Neil Yorke-Smith. Machine Learning Augmented Branch and Bound for Mixed Integer Linear Programming, February 2024. URL http://arxiv.org/abs/2402.05501. arXiv:2402.05501 [cs, math].

Tom Schaul. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Mehdi Seyfi, Amin Banitalebi-Dehkordi, Zirui Zhou, and Yong Zhang. Exact Combinatorial Optimization with Temporo-Attentional Graph Neural Networks, November 2023. URL http://arxiv.org/abs/2311.13843. arXiv:2311.13843 [cs].

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, December 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/10.1126/science.aar6404.

Jialin Song, Ravi Lanka, Yisong Yue, and Bistra Dilkina. A General Large Neighborhood Search Framework for Solving Integer Linear Programs, December 2020. URL http://arxiv.org/abs/2004.00422. arXiv:2004.00422 [cs, math, stat].

Nicolas Sonnerat, Pengming Wang, Ira Ktena, Sergey Bartunov, and Vinod Nair. Learning a Large Neighborhood Search Algorithm for Mixed Integer Programs, May 2022. URL http://arxiv.org/abs/2107.10201. arXiv:2107.10201 [cs, math].

Haoran Sun, Wenbo Chen, Hui Li, and Le Song. Improving Learning to Branch via Reinforcement Learning. July 2022. URL https://openreview.net/forum?id=z4D7-PTxTb.

Yunhao Tang, Shipra Agrawal, and Yuri Faenza. Reinforcement Learning for Integer Programming: Learning to Cut. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9367–9376. PMLR, November 2020. URL https://proceedings.mlr.press/v119/tang20a.html. ISSN: 2640-3498.

Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Zhihai Wang, Xijun Li, Jie Wang, Yufei Kuang, Mingxuan Yuan, Jia Zeng, Yongdong Zhang, and Feng Wu. Learning Cut Selection for Mixed-Integer Linear Programming via Hierarchical Sequence Model. February 2023. URL https://openreview.net/forum?id=Zob4P9bRNcK.

Dawen Wu and Abdel Lisser. A deep learning approach for solving linear programming problems. *Neurocomputing*, 520:15–24, February 2023. ISSN 09252312. doi: 10.1016/j.neucom.2022.11.053. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231222014412.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by the theoretical and experimental results presented in sections 3 and 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main limitations are adressed in section 1 and 5. Moreover, section 4.3 as well as appendix D also discuss the limitations associated with the use of DFS in BBMDP.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proposition 3.1 clearly states the conditions for its validity, and its proof is provided.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Comprehensive experimental details and references are provided in section 4 and appendices F, G C, H, I to allow for straightforward experiment reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: In order to enhance reproducibility, we share our implementations of the different RL baselines with the community, as well as pretrained network weights.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental details are provided in section 4 while a comprehensive list of hyperparameters is provided in appendix G.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Standard deviation for our experimental results are provided in appendix I.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All hardware ressource information are provided in appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have conducted our research in confirmity with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All authors of assets are duly cited in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A BBMDP vs TreeMDP

Side-by-side comparison

TreeMDP offers an intuitive and computationaly efficient framework for training RL agents to learn enhanced variable selection strategies in branch and bound. However, by discarding core MDP concepts such as temporality and sequentiality, it fails to accommodate the diversity of RL algorithms. For example, defining the k-step return temporal difference target in TreeMDP is challenging. Since branching on action a_i at state s_i produces two child next states s_i^- and s_i^+, applying recursively the tree Bellman operator from Etheve [2021] to V^π yields $V^\pi(s_i) = -\sum_{j=1}^{k} 2^j + \sum_{j=1}^{2^k} V^\pi(s_i^j)$. The corresponding TreeMDP trajectory is depicted in Figure 3 (d) for $k = 3$. More generally, it is important to note that such trajectories cannot be obtained when solving a MILP using B&B with a depth-first search node selection policy, since the subtree rooted in s_i^- must be fully fathomed before the node s_i^+ can be considered for expansion. Therefore, the TreeMDP k-step return TD target defined above stems from approximations in the B&B dynamics, which undermines the best performance achievable by k-step tMDP-DQN, as demonstrated in the ablation study in Section 4.3. Table 3 further highlights the key differences between BBMDP and TreeMDP.

Table 3: Side by side comparison of BBMDP and TreeMDP frameworks, complementing Figure 3.

	BBMDP	TreeMDP
MDP	Yes	No
State s	$s_t = (\mathcal{V}_t, \mathcal{E}_t, \bar{x}_t)$	$s_i = (P_i, x^*_{LP,i}, \bar{x}_i)$
Action a	$a_t \in \mathcal{I}$	$a_i \in \mathcal{I}$
Reward r	-2	-1
Next state s'	$s_{t+1} = (\mathcal{V}_{t+1}, \mathcal{E}_{t+1}, \bar{x}_{t+1})$	s_i^-, s_i^+
k-step next state $s^{(k)}$	$s_{t+k} = (\mathcal{V}_{t+k}, \mathcal{E}_{t+k}, \bar{x}_{t+k})$	s_i^1, s_i^2, ..., $s_i^{2^k}$
$\mathcal{B}(V^\pi) = V^\pi$	$\bar{V}^\pi(o_1, \bar{x}_{o_1}) = -2 + \bar{V}^\pi(o_1', \bar{x}_{o_1'}) + \bar{V}^\pi(o_2', \bar{x}_{o_2'})$	$V^\pi(s_i) = -2 + V^\pi(s_i^-) + V^\pi(s_i^+)$
$\mathcal{B}^k(V^\pi) = V^\pi$	$\bar{V}^\pi(o_1, \bar{x}_{o_1}) = -2k + \sum_{i=1}^{k+1} \bar{V}^\pi(o_i^{(k)}, \bar{x}_{o_i^{(k)}})$	$V^\pi(s_i) = -\sum_{j=1}^{k} 2^j + \sum_{j=1}^{2^k} V^\pi(s_i^j)$

Learning to branch with MCTS

MCTS-based RL algorithms have achieved remarkable performance in combinatorial settings, particularly in board games Silver et al. [2018]. Interstingly, by providing a rigorous MDP formulation for variable selection in B&B, BBMDP enables the adapation of MCTS-based learning algorithms to the B&B framework, overcoming the incompatibilities that persisted within TreeMDP. In fact, when searching for the next node to expand, MCTS traverses the tree following a UCT exploration criterion:

$$a^k = \arg\max_{a \in \mathcal{A}} \left[Q(s,a) + \pi(s,a) \cdot \frac{\sqrt{\sum_{b \in \mathcal{A}} N(s,b)}}{1 + N(s,a)} \left(c_1 + \log(\frac{\sum_{b \in \mathcal{A}} N(s,b)}{c_2}) \right) \right] \tag{7}$$

where $N(s,a)$ corresponds to the number of visits of the node (s,a), and c_1, c_2 are search hyperparameters. This formulation balances exploration and exploitation, guiding the search toward promising states while ensuring sufficient exploration of the tree. However in TreeMDP, such an exploration criteria leads to inconsistencies, as the exploration between node s_i^- and s_i^+ is balanced based on their estimated value. Since both s_i^- and s_i^+ originate from the same branching decision a_i, differentiating between them for expansion is irrelevant. In B&B, branching on a_i implies that s_i^- and s_i^+ are both necessarily added to the B&B tree. Consequently, in TreeMDP, MCTS could guide the search towards a suboptimal action a_i based on the assumption that it produces a state s_i^- of small subtree size, while failing to account for the potentially enormous subtree size of s_i^+.

This fundamental inconsistency hinders the effective application of MCTS within the TreeMDP framework.

B HL-Gauss Loss

As they investigated the uneven success met by complex neural network architectures such as Transformers in supervised versus reinforcement learning, Farebrother et al. [2024] found that training agents using a cross-entropy classification objective significantly improved the performance and scalability of value-based RL methods. However, replacing mean squared error regression with cross-entropy classification requires methods to transform scalars into distributions and distributions into scalars. Farebrother et al. [2024] found the Histogram Gaussian loss (HL-Gauss) [Imani and White, 2018], which exploits the ordinal structure of the regression task by distributing probability mass on multiple neighboring histogram bins, to be a reliable solution across multiple RL benchmarks. Concretely, in HL-Gauss, the support of the value function $\mathcal{Z}_v \subset \mathbb{R}$ is divided in m_b bins of equal width forming a partition of $\mathcal{Z}_v$. Bins are centered at $\zeta_i \in \mathcal{Z}_v$ for $1 \leq i \leq m_b$, we use $\eta = (\zeta_{max} - \zeta_{min})/m_b$ to denote their width. Given a scalar $z \in \mathcal{Z}_v$, we define the random variable $Y_z \sim \mathcal{N}(\mu = z, \sigma^2)$ and note respectively ϕ_{Y_z} and Φ_{Y_z} its associate probability density and cumulative distribution function. z can then be encoded into a histogram distribution on $\mathcal{Z}_v$ using the function $p_{hist} : \mathbb{R} \to [0,1]^{m_b}$. Explicitly, p_{hist} computes the aggregated mass of ϕ_{Y_z} on each bin:

$$p_{hist}(z) = (p_i(z))_{1 \leq i \leq m_b} \text{ with } p_i(z) = \int_{\zeta_i - \frac{\eta}{2}}^{\zeta_i + \frac{\eta}{2}} \phi_{Y_z}(y)dy = \Phi_{Y_z}(\zeta_i + \frac{\eta}{2}) - \Phi_{Y_z}(\zeta_i - \frac{\eta}{2})$$

Conversely, histogram distributions $(p_i)_{1 \leq i \leq m_b}$ such as the ones outputted by agents' value networks can be converted to scalar simply by computing the expectation: $z = \sum_{i=1}^{m_b} p_i \cdot \zeta_i$.

BBMDP is a challenging setting to adapt HL-Gauss, as the support for value functions spans over several order of magnitude. In practice, we observe that for train instances of the Ecole benchmark, $\mathcal{Z}_v = [-10^6, -2]$. Since value functions predict the number of node of binary trees built with B&B, it seems natural to choose bins centered at $\zeta_i = -2^i$ to partition $\mathcal{Z}_v$. In order to preserve bins of equal size, we consider distributions on the support $\psi(\mathcal{Z}_v)$ with $\psi(z) = \log_2(-z)$ for $z \in \mathcal{Z}_v$, such that $\psi(\mathcal{Z}_v)$ is efficiently partitioned by bins centered at $\zeta_i = i$ for $1 \leq i \leq m_b$. Thus, in BBMDP histograms distributions are given by $p_{hist}(z) = (p_i \circ \psi(z))_{1 \leq i \leq m_b}$ for $z \in \mathcal{Z}_v$, and can be converted back to $\mathcal{Z}_v$ through $z = \sum_{i=1}^{m_b} p_i \cdot \psi^{-1}(\zeta_i)$ with $\psi^{-1}(z) = -2^z$.

C Baselines

Imitation learning We trained and tested IL agents using the official Ecole re-implementation of Gasse et al. [2019] shared at `https://github.com/ds4dm/learn2branch-ecole/tree/main`.

DQN-TreeMDP Since there is no publicly available implementation of Etheve et al. [2020], we re-implemented DQN-TreeMDP and trained it on the four Ecole benchmarks, using when applicable the same network architectures and training parameters as in DQN-BBMDP and DQN-Retro. We share implementation and trained network weights to the community.

PG-tMDP We used the official implementation of Scavuzzo et al. [2022] to evaluate PG-TreeMDP. For each benchmark, we used the tMDP+DFS network weights shared at `https://github.com/lascavana/rl2branch`.

DQN-Retro As Parsonson et al. [2022] only trained on set covering instances, we took inspiration from the official implementation shared at `https://github.com/cwfparsonson/retro_branching` to train and evaluate DQN-Retro agents on the four Ecole benchmarks. Importantly, we trained and tested DQN-Retro following SCIP's default node selection strategy, see Appendix D for more details. We share our re-implementation and trained network weights with the community.

Importantly, on the multiple knapsack benchmark, given the high computational cost associated with opting for a depth-first search node selection policy, all baselines excepted IL-DFS are evaluated following SCIP default node selection policy.

D BBMDP vs Retro branching

In their work, Parsonson et al. [2022] proposed to train RL agents on retrospective trajectories built from TreeMDP episodes, in order to leverage the state-of-the art node selection policies implemented in MILP solvers. When reproducing their work, we found several discrepancies with the results they stated. First, the performance gap between DQN-Retro and DQN-TreeMDP [Etheve et al., 2020] turned out to be much narrower than expected. On test set covering instances, the only benchmark on which the two agents are compared in Parsonson et al. [2022], we even found DQN-TreeMDP to perform better. Second, Parsonson et al. [2022] found that adopting a best-first-search (BeFS) node selection strategy at evaluation time greatly improved the performance of DQN-TreeMDP on test set covering instances, indicating that abandoning DFS during training could be beneficial. However, in our experiments, we observed a 20% performance drop when replacing DFS by BeFS at evaluation time. After thorough examination of both Parsonson et al. [2022]'s article and implementation, we found that the baseline labeled as DQN-TreeMDP (FMSTS-DFS in their article) was quite distant from the branching agent originally described in [Etheve et al., 2020]. In fact, in Parsonson et al. [2022], the Etheve et al. [2020] branching agent is not trained on TreeMDP trajectories, but on retrospective trajectories built from TreeMDP episodes, using a DFS construction heuristic. Therefore, Parsonson et al. [2022] could not conclude on the superiority of retro branching over TreeMDP, nor could they assess the limitations of DFS-based RL agents. In contrast, our contribution provides compelling evidence that, while DFS is generally expected to hinder the training performance of RL agents due to its reputation as a suboptimal node selection policy, the theoretical guarantees brought by DFS in BBMDP enable to surpass prior state-of-the-art non-DFS agents. We believe this is because optimizing the node selection policy has less influence on tree size performance compared to optimizing the variable selection policy, as evidenced by Etheve [2021].

E Neural network

State representation Following the works of Gasse et al. [2019], MILPs are best represented by bipartite graphs $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{C}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ where $\mathcal{V}_\mathcal{G}$ denotes the set of variable nodes, $\mathcal{C}_\mathcal{G}$ denotes the set of constraint nodes, and $\mathcal{E}_\mathcal{G}$ denotes the set of edges linking variable and constraints nodes. Nodes $v_\mathcal{G} \in \mathcal{V}_\mathcal{G}$ and $c_\mathcal{G} \in \mathcal{C}_\mathcal{G}$ are connected if the variable associated with $v_\mathcal{G}$ appears in the constraint associated with $c_\mathcal{G}$. Given a MILP P, defined as in Section 2.1, its associate bipartite representation $\mathcal{G}$ has $|\mathcal{G}| = |\mathcal{V}_\mathcal{G}| + |\mathcal{C}_\mathcal{G}| = n + m$ nodes. We use bipartite graphs to encode the information contained in $(o_i, \bar{x}_{o_i})$, as defined in section 3.2. In our experiments, IL and PG-tMDP agents use the list of features of Gasse et al. [2019] to represent variable nodes, constraint nodes and edges, while DQN-BBMDP, DQN-TreeMDP and DQN-Retro agents also make use of the additional features introduced by Parsonson et al. [2022].

Network architecture All RL agents utilize the graph convolutional network architecture described in Scavuzzo et al. [2022] and Parsonson et al. [2022]. In DQN-BBMDP, the architecture differs slightly, with the final layer outputting distribution vectors in $\mathbb{R}^{m_b}$ instead of scalar values in $\mathbb{R}$.

F Instance dataset

Instance datasets used for training and evaluation are decribed in Table 4. We trained and tested on instances of same dimensions as Scavuzzo et al. [2022] and Parsonson et al. [2022]. As a reminder, the size of action set $\mathcal{A}$ is equal to the number of integer variables in P. Consequently, action set sizes in the Ecole benchmark range from 30 to 480 for train / test instances and from 50 to 980 for transfer instances.

Table 4: Instance size for each benchmark. Performance is evaluated on test instances that match the size of the training instances, as well as on larger instances, to further assess the generalization capacity of our agents. Last two columns indicate the approximate number of integer variables after presolve, both for train / test and transfer instances.

Benchmark	Generation method	Parameters	Parameter value		# Int. variables	
			Train / Test	Transfer	Train / Test	Transfer
Combinatorial auction	Leyton-Brown et al. [2000]	Items Bids	100 500	200 1000	100	200
Set covering	Balas and Ho [1980]	Items Sets	500 1000	1000 1000	100	130
Maximum independent set	Bergman et al. [2016]	Nodes	500	1000	480	980
Multiple knapsack	Fukunaga [2011]	Items Knapsacks	100 6	100 12	30	50

G Training pipeline

DQN Implementation In Algorithm 1, we provide a description of DQN-BBMDP training pipeline. Our DQN implementation includes several Rainbow-DQN features [Hessel et al., 2017]: double DQN [Van Hasselt et al., 2016], n-step learning and prioritized experience replay (PER) Schaul [2015]. Moreover, as DQN-BBMDP learns distributions representing Q-values, it integrates elements of [Bellemare et al., 2017].

Algorithm 1 DQN-BBMDP

for $t = 0...N-1$ **do**
  Draw randomly an instance $P \sim p_0$.
  Solve P by acting following a combined ϵ-greedy and Boltzman exploration according to $\mathbf{q}_{\theta_t}$.
  Collect transitions along the generated tree $(s_i, a_i, \sum_{j=1}^{k} r_{i+j}, s_{i+k})$ and store them into a replay buffer $\mathcal{B}_{replay}$.
  Update θ_t using the loss described in Sec 3.3 on transition batches drawn from $\mathcal{B}_{replay}$.
end for

Exploration We train our agents following Boltzmann and ϵ-greedy exploration combined. Concretely, agents select actions uniformly from $\mathcal{A}$ with probability ϵ, while following a Boltzmann exploration strategy with temperature τ for the remaining probability $1-\epsilon$. The decay rates for ϵ and τ are listed in Table 5.

Reward model In Section 3.1, we defined $\mathcal{R}(s,a) = -2$ for all transition, so that the overall value of a trajectory matched the size of the B&B tree. In practice, all negative constant reward model yield equivalent optimal policies in BBMDP, therefore, we chose to implement $\mathcal{R}(s,a) = -1$ for all RL baselines in order to allow clearer comparison between BBMDP and TreeMDP agents.

Training parameters Table 5 provides the list of hyperparameters used to train DQN agents on the Ecole benchmarks. To allow fair comparisons, when applicable, we keep SCIP parameters, training parameters and network architectures fixed for all DQN-agents.

H Validation curves

All experiments were conducted on an NVIDIA DGX A100 system equipped with 8× A100 40GB GPUs, 2× AMD EPYC 7742 64-core CPUs (128 threads total), and 1 TB of DDR4 RAM. We present validation curves for DQN-BBMDP, DQN-TreeMDP and DQN-Retro in Figure 4. For each benchmark we trained for 200k gradient steps, which took approximately 2 days for combinatorial auction instances, 3 days for set covering instances, 5 days for multiple knapsack instances and 7

Table 5: Training parameters for all DQN branching agents. For DQN-Retro, we take $\gamma = 0.99$ as in Parsonson et al. [2022].

Module	Training parameter	Value
Q-learning	Batch size	128
	Optimizer	Adam
	k-step return	3
	Learning rate l_r	5×10^{-5}
	Discount factor γ	1.0
	Agent steps per network update	10
	Soft target network update τ_{net}	10^{-4}
Replay buffer	Buffer minimum size $\lvert\mathcal{B}_{replay}\rvert_{init}$	20×10^3
	Buffer maximum capacity $\lvert\mathcal{B}_{replay}\rvert_{max}$	100×10^3
Prioritized experience replay	PER α	0.6
	PER β_{init}	0.4
	PER β_{final}	1.0
	$\beta_{init} \rightarrow \beta_{final}$ learner steps	100×10^3
	Minimum experience priority	10^{-3}
Exploration	Start exploration probability ϵ_{init}	1.0
	Minimum exploration probability ϵ_{min}	2.5×10^{-2}
	ϵ-decay	10^{-4}
	Start temperature τ_{init}	1.0
	Minimum temperature τ_{min}	10^{-3}
	τ-decay	10^{-5}
HL-Gauss (only for DQN-BBMDP)	z_{min}	-1
	z_{max}	16
	m_b	18
	σ	0.75

days for maximum independent set instances. As shown in Figure 4, DQN-BBMDP training was interrupted before final convergence on 3 out of 4 benchmarks, hinting that performance could likely be improved by training for more steps.

I Additional computational results

In this section, we include further computational results on instances of the Ecole benchmark.

Table 6 provides additional performance metrics to compare the different baselines across test and transfer benchmarks. For each benchmark, we report the number of wins and the average rank of each baseline across 100 evaluation instances. The number of wins is defined as the number of instances where a baseline solves a MILP problem faster than any other baseline. When multiple baselines fail to solve an instance to optimality within the time limit, their performance is ranked based on final dual gap.

Finally, Table 7 recapitulates the computational results presented in Table 1, and provides for each baseline the per-benchmark standard deviation over five seeds, as well as the fraction of test instances solved to optimality within the time limit.

Table 6: Additional performance metrics for each baseline on train / test and transfer instance benchmarks, see Appendix F for instance details. For each benchmark, we report the number of wins, and the average rank of each baseline across the 100 evaluation instances. We also report for each baseline the fraction of test instances solved to optimality within time limit. The number of wins is defined as the number of instances where a baseline solves a MILP problem faster than all other baselines. When multiple baselines fail to solve an instance to optimality within time limit, their performance is ranked based on final dual gap.

	Train / Test			Transfer		
Method	Solved	Wins	Rank	Solved	Wins	Rank
SCIP	100/100	8/100	5.7	100/100	10/100	3.3
IL	100/100	1/00	3.8	100/100	29/100	1.8
IL-DFS	100/100	35/100	2.1	**100/100**	**58/100**	**1.7**
PG-tMDP	100/100	0/100	6.6	78/100	0/100	6.8
DQN-tMDP	100/100	11/100	2.9	96/100	0/100	5.0
DQN-Retro	100/100	1/100	4.9	98/100	0/100	5.1
DQN-BBMDP	100/100	**44/100**	**1.9**	**100/100**	**3/100**	**4.2**

Set covering

	Train / Test			Transfer		
Method	Solved	Wins	Rank	Solved	Wins	Rank
SCIP	100/100	14/100	4.1	100/100	14/100	3.51
IL	100/100	16/100	3.5	**100/100**	**47/100**	**1.7**
IL-DFS	100/100	31/100	2.6	100/100	20/100	2.5
PG-tMDP	100/100	0/100	5.7	100/100	0/100	5.8
DQN-tMDP	100/100	0/100	6.2	100/100	0/100	6.5
DQN-Retro	100/100	10/100	3.6	100/100	1/100	4.6
DQN-BBMDP	100/100	**34/100**	**2.3**	**100/100**	**18/100**	**2.9**

Combinatorial Auction

	Train / Test			Transfer		
Method	Solved	Wins	Rank	Solved	Wins	Rank
SCIP	100/100	9/100	5.7	100/100	7/100	4.5
IL	100/100	**72/100**	**1.6**	100/100	**57/100**	**1.7**
IL-DFS	100/100	10/100	2.4	100/100	0/100	3.2
PG-tMDP	100/100	0/100	4.9	**100/100**	**36/100**	**1.8**
DQN-tMDP	100/100	1/100	4.8	85/100	0/100	6.1
DQN-Retro	100/100	**6/100**	5.4	22/100	0/100	6.7
DQN-BBMDP	100/100	2/100	**3.3**	95/100	0/100	4.2

Maximum Independent Set

	Train / Test			Transfer		
Method	Solved	Wins	Rank	Solved	Wins	Rank
SCIP	100/100	**88/100**	**1.4**	**100/100**	**60/100**	**1.9**
IL	100/100	1/100	4.5	100/100	6/100	3.4
IL-DFS	100/100	1/100	5.8	98/100	0/100	6.0
PG-tMDP	100/100	0/100	6.0	98/100	5/100	5.0
DQN-tMDP	100/100	1/100	3.5	99/100	**14/100**	**3.5**
DQN-Retro	100/100	3/100	3.5	98/100	9/100	3.8
DQN-BBMDP	100/100	**6/100**	**3.3**	**100/100**	6/100	4.3

Multiple Knapsack

Table 7: Computational performance comparison on four MILP benchmarks. Following prior works, we report geometrical mean over 100 instances, averaged over 5 seeds, as well as per-benchmark standard deviations.

	Train / Test			Transfer		
Method	Nodes	Time	Solved	Nodes	Time	Solved
Random	$3289 \pm 4.2\%$	$5.9 \pm 4.3\%$	100/100	$270365 \pm 9.5\%$	$811 \pm 7.9\%$	60/100
SB	$35.8 \pm 0.0\%$	$12.93 \pm 0.0\%$	100/100	$672.1 \pm 0.0\%$	$398 \pm 0.2\%$	82/100
SCIP	$62.0 \pm 0.0\%$	$2.27 \pm 0.0\%$	100/100	$3309 \pm 0.0\%$	$48.4 \pm 0.1\%$	100/100
IL	$\mathbf{133.8 \pm 1.0\%}$	$0.90 \pm 4.8\%$	100/100	$\mathbf{2610 \pm 0.7\%}$	$23.1 \pm 1.5\%$	**100/100**
IL-DFS	$136.4 \pm 1.8\%$	$\mathbf{0.74 \pm 5.3\%}$	100/100	$3103 \pm 2.0\%$	$\mathbf{22.5 \pm 3.1\%}$	**100/100**
PG-tMDP	$649.4 \pm 0.7\%$	$2.32 \pm 2.4\%$	100/100	$44649 \pm 3.7\%$	$221 \pm 4.1\%$	78/100
DQN-tMDP	$175.8 \pm 1.1\%$	$0.83 \pm 4.5\%$	100/100	$8632 \pm 4.9\%$	$71.3 \pm 5.8\%$	96/100
DQN-Retro	$183.0 \pm 1.2\%$	$1.14 \pm 4.1\%$	100/100	$6100 \pm 4.2\%$	$59.4 \pm 4.2\%$	98/100
DQN-BBMDP	$\mathbf{152.3 \pm 0.6\%}$	$\mathbf{0.77 \pm 5.6\%}$	100/100	$\mathbf{5651 \pm 2.2\%}$	$\mathbf{46.4 \pm 3.3\%}$	**100/100**

Set covering

	Train / Test			Transfer		
Method	Nodes	Time	Solved	Nodes	Time	Solved
Random	$1111 \pm 4.3\%$	$2.16 \pm 6.6\%$	100/100	$354650 \pm 6.7\%$	$814 \pm 7.1\%$	64/100
SB	$28.2 \pm 0.0\%$	$6.21 \pm 0.1\%$	100/100	$389.6 \pm 0.0\%$	$255 \pm 0.2\%$	88/100
SCIP	$20.2 \pm 0.0\%$	$1.77 \pm 0.1\%$	100/100	$1376 \pm 0.0\%$	$14.77 \pm 0.1\%$	100/100
IL	$\mathbf{83.6 \pm 0.8\%}$	$0.65 \pm 7.3\%$	100/100	$\mathbf{1309 \pm 1.6\%}$	$\mathbf{9.4 \pm 2.2\%}$	**100/100**
IL-DFS	$95.5 \pm 0.9\%$	$\mathbf{0.56 \pm 7.1\%}$	100/100	$1802 \pm 2.0\%$	$10.2 \pm 1.8\%$	100/100
PG-tMDP	$168.0 \pm 2.8\%$	$0.94 \pm 6.0\%$	100/100	$6001 \pm 2.7\%$	$30.7 \pm 2.4\%$	100/100
DQN-tMDP	$203.3 \pm 4.2\%$	$1.11 \pm 4.0\%$	100/100	$20553 \pm 3.8\%$	$116 \pm 3.9\%$	100/100
DQN-Retro	$103.2 \pm 1.2\%$	$0.78 \pm 7.5\%$	100/100	$2908 \pm 1.7\%$	$18.4 \pm 2.7\%$	100/100
DQN-BBMDP	$\mathbf{97.9 \pm 1.2\%}$	$\mathbf{0.62 \pm 8.5\%}$	100/100	$\mathbf{2273 \pm 1.9\%}$	$\mathbf{11.8 \pm 2.0\%}$	**100/100**

Combinatorial auction

	Train / Test			Transfer		
Method	Nodes	Time	Solved	Nodes	Time	Solved
Random	$386.8 \pm 5.4\%$	$2.01 \pm 4.8\%$	100/100	$215879 \pm 6.7\%$	$2102 \pm 6.2\%$	25/100
SB	$24.9 \pm 0.0\%$	$45.87 \pm 0.4\%$	100/100	$169.9 \pm 0.2\%$	$2172 \pm 0.9\%$	15/100
SCIP	$19.5 \pm 0.0\%$	$2.44 \pm 0.4\%$	100/100	$3368 \pm 0.0\%$	$90.0 \pm 0.2\%$	100/100
IL	$\mathbf{40.1 \pm 3.45\%}$	$\mathbf{0.36 \pm 3.1\%}$	100/100	$\mathbf{1882 \pm 4.0\%}$	$\mathbf{38.6 \pm 3.2\%}$	**100/100**
IL-DFS	$69.4 \pm 6.5\%$	$0.44 \pm 4.8\%$	100/100	$3501 \pm 2.7\%$	$51.9 \pm 2.6\%$	100/100
PG-tMDP	$153.6 \pm 5.0\%$	$0.92 \pm 2.6\%$	100/100	$\mathbf{3133 \pm 4.6\%}$	$\mathbf{39.5 \pm 3.8\%}$	**100/100**
DQN-tMDP	$168.0 \pm 5.6\%$	$1.00 \pm 3.4\%$	100/100	$45634 \pm 7.4\%$	$477 \pm 5.1\%$	85/100
DQN-Retro	$223.0 \pm 4.1\%$	$1.81 \pm 3.6\%$	100/100	$119478 \pm 6.1\%$	$1863 \pm 4.8\%$	22/100
DQN-BBMDP	$\mathbf{103.2 \pm 9.3\%}$	$\mathbf{0.62 \pm 6.8\%}$	100/100	$7168 \pm 5.3\%$	$81.3 \pm 4.2\%$	95/100

Maximum independent set

	Train / Test			Transfer		
Method	Nodes	Time	Solved	Nodes	Time	Solved
Random	$733.5 \pm 13.0\%$	$0.55 \pm 6.9\%$	100/100	$93452 \pm 14.3\%$	$70.6 \pm 9.2\%$	99/100
SB	$161.7 \pm 0.0\%$	$0.69 \pm 0.1\%$	100/100	$\mathbf{1709 \pm 0.5\%}$	$\mathbf{12.5 \pm 0.9\%}$	**100/100**
SCIP	$289.5 \pm 0.0\%$	$0.53 \pm 0.2\%$	100/100	$30260 \pm 0.0\%$	$22.14 \pm 0.2\%$	100/100
IL	$272.0 \pm 12.9\%$	$0.69 \pm 8.5\%$	100/100	$9747 \pm 7.5\%$	$46.5 \pm 6.6\%$	100/100
IL-DFS	$472.8 \pm 13.0\%$	$1.07 \pm 9.0\%$	100/100	$43224 \pm 9.0\%$	$131 \pm 8.6\%$	98/100
PG-tMDP	$436.9 \pm 21.2\%$	$1.57 \pm 16.9\%$	100/100	$35614 \pm 14.3\%$	$123 \pm 15.4\%$	98/100
DQN-tMDP	$266.4 \pm 7.2\%$	$0.73 \pm 4.6\%$	100/100	$\mathbf{22631 \pm 8.6\%}$	$\mathbf{65.1 \pm 5.5\%}$	99/100
DQN-Retro	$250.3 \pm 9.5\%$	$0.67 \pm 5.0\%$	100/100	$27077 \pm 8.8\%$	$79.5 \pm 6.2\%$	98/100
DQN-BBMDP	$\mathbf{236.6 \pm 6.4\%}$	$\mathbf{0.66 \pm 2.7\%}$	100/100	$37098 \pm 7.0\%$	$109 \pm 4.9\%$	**100/100**

Multiple knapsack

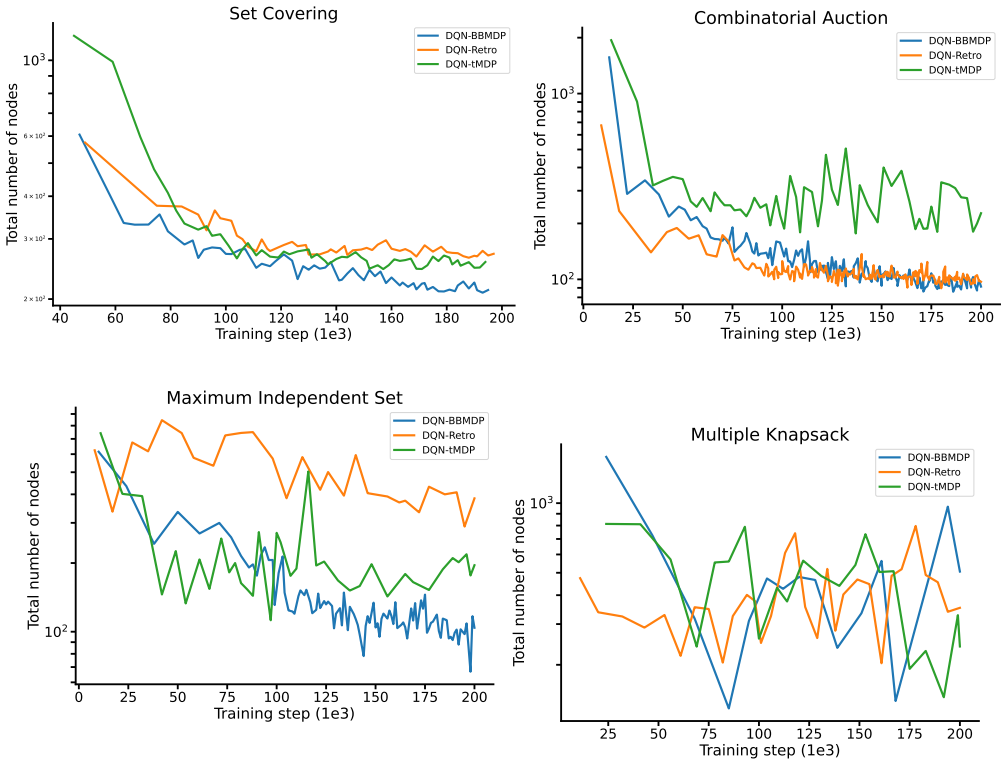


Figure 4: Validation curves for DQN-BBMDP, DQN-Retro and DQN-tMDP agents, in log scale. Throughout training, agents are evaluated on 20 validation instances after each batch of 100 training instances solved. Note that on the multiple knapsack benchmark, none of the agents reach convergence.