Generalized Nested Rollout Policy Adaptation with Bias Learning

Julien Sentuc[1], Farah Ellouze[1], Jean-Yves Lucas[2], Tristan Cazenave[1]

[1] LAMSADE, Université Paris-Dauphine, PSL, CNRS, France
{julien.Sentuc,farah.Ellouze}@dauphine.eu,Tristan.Cazenave@dauphine.psl.eu
[2] OSIRIS, EDF Lab Paris-Saclay, Electricité de France, France
Jean-Yves.Lucas@edf.fr

Mots-clés : *Vehicle Routing, 3D Packing, Monte Carlo Search.*

1 Introduction

Monte Carlo Tree Search (MCTS) has been successfully applied to many games and problems. The principle underlying MCTS is learning the best move using statistics on random games.MCTS has given bith to several variations. Nested Monte Carlo Search (NMCS) [1] is a recursive algorithm which uses lower level playouts to bias its playouts, memorizing the best sequence at each level. At level 0, a Monte Carlo simulation is performed, random decisions are made until a terminal state is reached. Based on the latter, the Nested Rollout Policy Adaptation (NRPA) algorithm was introduced [3]. NRPA combines nested search, memorizing the best sequence of moves found, and the online learning of a playout policy using this sequence. Generalized Nested Rollout Policy Adaptation (GNRPA) [2] generalizes the way the probability is calculated using a temperature and a bias. It has been applied to some problems like Vehicle Routing Problem (VRP) [4]. In this work we intoduce an extension of GNRPA, namely Bias Learning GNRPA (BGNRPA). BLGNRPA automatically learns the bias weights. The goal is both to obtain better results on sets of dissimilar instances, and also to avoid some hyperparameters settings. The idea is to learn the parameters of the bias along with the policy. We applied GNRPA with bias learning to the Solomon instances of VRP and for 3D Bin Packing.

In NRPA/GNRPA each move is associated to a weight stored in an array called the policy. The goal of these two algorithms is to learn these weights thanks to the solutions found during the search, thus producing a playout policy that generates good sequences of moves.

NRPA/GNRPA use nested search. In NRPA/GNRPA, each level takes a policy as input and returns a sequence and its associated score. At any level > 0, the algorithm makes numerous recursive calls to lower levels, adapting the policy each time with the best solution to date. At level 0, NRPA/GNRPA return the sequence obtained by playout function as well as its associated score.

The playout function sequentially constructs a random solution biased by the weight of the moves until it reaches a terminal state. At each step, the function performs Gibbs sampling, choosing the actions with a probability given by the softmax function.

Let w_{ic} be the weight associated with move c in step i of the sequence. In NRPA, the probability of choosing move c at the index i is defined by :

$$p_{ic} = \frac{e^{w_{ic}}}{\sum_k e^{w_{ik}}}$$

GNRPA [2] generalizes the way the probability is calculated using a temperature tau and a bias $beta_{ic}$. The temperature makes it possible to vary the exploration/exploitation trade-off. The probability of choosing the move c at the index i then becomes :

$$p_{ic} = \frac{e^{\frac{w_{ic}}{\tau} + \beta_{ic}}}{\sum_k e^{\frac{w_{ik}}{\tau} + \beta_{ik}}}$$

For many problems, we can separate the bias into several criteria. In the case of 2 criteria β_1 and β_2 : $\beta_i c = w_1 * \beta_1 + w_2 * \beta_2$, where β_1 and β_2 describe two different characteristics of a move. The idea of learning the bias parameters w_1 and w_2 lies in adapting the importance of the different criteria along with the policy to the specific instance that we are trying to solve.

2 The Vehicle Routing Problem

The Vehicle Routing Problem is a well-known optimization problems. In this work, we used the 1987 Solomon instances [5] for the Constrained Vehicle Routing Problem with Time Windows problem (CVRPTW).

3 3D Bin Packing

The 3D bin Packing Problem is an optimization problem in which we have to store a set of boxes into one or several containers. The goal is to minimize the unused space in the containers and put the greatest possible number of items into each of them, or, alternatively, to minimize the number of container used to store all the boxes. We based our experiments on the problem modeled in the paper [6].

4 Results

Globally, BLGNRPA has better results than GNRPA and NRPA on Solomon instances of the VRP (it performs better than NRPA on all instances). BLGNRPA yields also satisfactory results on the 3D Bin Packing problem.

5 Conclusion

In this work, we introduced BLGNRPA, a new method to learn the bias weights for the GNRPA algorithm. This new method partially removes the need to choose hand-picked weights for GNRPA. Experiments show that it improves the GNRPA algorithm for two different optimization problems : the Vehicle Routing Problem and 3D Bin Packing.

Références

[1] Tristan Cazenave. Nested Monte-Carlo Search. In Craig Boutilier, editor, *IJCAI*, pages 456–461, 2009.

[2] Tristan Cazenave. Generalized nested rollout policy adaptation. In *Monte Carlo Search at IJCAI*, 2020.

[3] Christopher D. Rosin. Nested rollout policy adaptation for Monte Carlo Tree Search. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 649–654, 2011.

[4] Julien Sentuc, Tristan Cazenave, and Jean-Yves Lucas. Generalized nested rollout policy adaptation with dynamic bias for vehicle routing. In *AI for Transportation at AAAI*, 2022.

[5] Solomon. Algorithms for the vehicle routing and scheduling problems with time window constraints. In *Operations Research*, 1985.

[6] Hang Zhao, Yang Yu, and Kai Xu. Learning efficient online 3d bin packing on packing configuration trees. In *International Conference on Learning Representations*, 2022.