

e-print <http://www.gm.fh-koeln.de/ciopwebpub/Konen22b.d/TR-Rubiks.pdf>

# Towards Learning Rubik's Cube with N-tuple-based Reinforcement Learning

Wolfgang Konen

Technical Report,  
Computer Science Institute,  
TH Köln,  
University of Applied Sciences,  
Germany

[wolfgang.konen@th-koeln.de](mailto:wolfgang.konen@th-koeln.de)

Sep 2022,  
last update Jan 2023

## Abstract

This work describes in detail how to learn and solve the Rubik's cube game (or puzzle) in the General Board Game (GBG) learning and playing framework. We cover the cube sizes 2x2x2 and 3x3x3. We describe in detail the cube's state representation, how to transform it with twists, whole-cube rotations and color transformations and explain the use of symmetries in Rubik's cube. Next, we discuss different n-tuple representations for the cube, how we train the agents by reinforcement learning and how we improve the trained agents during evaluation by MCTS wrapping.

We present results for agents that learn Rubik's cube from scratch, with and without MCTS wrapping, with and without symmetries and show that both, MCTS wrapping and symmetries, increase computational costs, but lead at the same time to much better results. We can solve the 2x2x2 cube completely, and the 3x3x3 cube in the majority of the cases for scrambled cubes up to  $p = 15$  (QTM). We cannot yet reliably solve 3x3x3 cubes with more than 15 scrambling twists.

Although our computational costs are higher with MCTS wrapping and with symmetries than without, they are still considerably lower than in the approaches of McAleer et al. (2018, 2019) and Agostinelli et al. (2019) who provide the best Rubik's cube learning agents so far.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Motivation . . . . .	4
1.2	Overview . . . . .	5
<b>2</b>	<b>Foundations</b>	<b>6</b>
2.1	Conventions and Symbols . . . . .	6
2.1.1	Color arrangement . . . . .	6
2.1.2	Twist and Rotation Symbols . . . . .	6
2.1.3	Twist Types . . . . .	6
2.2	Facts about Cubes . . . . .	7
2.2.1	2x2x2 Cube . . . . .	7
2.2.2	3x3x3 Cube . . . . .	7
2.3	The Cube State . . . . .	8
2.4	Transformations . . . . .	9
2.4.1	Twist Transformations . . . . .	9
2.4.2	Whole-Cube Rotations (WCR) . . . . .	11
2.4.3	Color Transformations . . . . .	13
2.5	Symmetries . . . . .	15
<b>3</b>	<b>N-Tuple Systems</b>	<b>16</b>
<b>4</b>	<b>N-Tuple Representations for the Cube</b>	<b>18</b>
4.1	CUBESTATE . . . . .	18
4.2	STICKER . . . . .	18
4.3	STICKER2 . . . . .	20
4.4	Adjacency Sets . . . . .	21
<b>5</b>	<b>Learning the Cube</b>	<b>21</b>
5.1	McAleer and Agostinelli . . . . .	21
5.2	N-Tuple-based TD Learning . . . . .	24
5.2.1	Temporal Coherence Learning (TCL) . . . . .	25
5.2.2	MCTS . . . . .	25
5.2.3	Method Summary . . . . .	27
<b>6</b>	<b>Results</b>	<b>27</b>
6.1	Experimental setup . . . . .	27
6.2	Cube Solving with MCTS Wrapper, without Symmetries . . . . .	28
6.3	Number of Symmetric States . . . . .	28
6.4	The Benefit of Symmetries . . . . .	29
6.5	Computational Costs . . . . .	30
<b>7</b>	<b>Related Work</b>	<b>32</b>

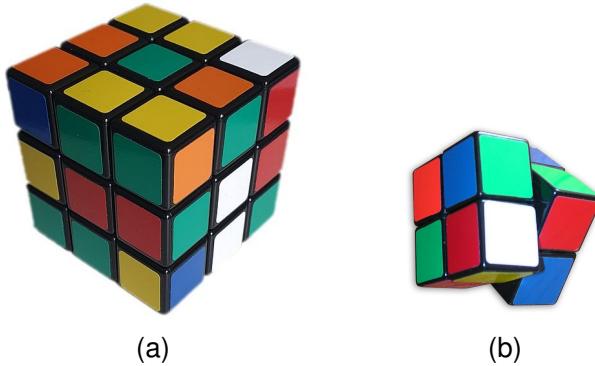
<b>8 Summary and Outlook</b>	<b>33</b>
<b>A Calculating sloc from fcol</b>	<b>37</b>
A.1 2x2x2 cube . . . . .	37
A.2 3x3x3 cube . . . . .	38
<b>B N-Tuple Representations for the 3x3x3 Cube</b>	<b>38</b>
B.1 CUBESTATE . . . . .	38
B.2 STICKER . . . . .	39
B.3 STICKER2 . . . . .	39
B.4 Adjacency Sets . . . . .	40
<b>C Hyperparameters</b>	<b>40</b>

# 1 Introduction

## 1.1 Motivation

Game learning and game playing is an interesting test bed for strategic decision making. Games usually have large state spaces, and they often require complex pattern recognition and strategic planning capabilities to decide which move is the best in a certain situation. If algorithms learn a game (or, even better, a variety of different games) just by self-play, given no other knowledge than the game rules, it is likely that they perform also well on other problems of strategic decision making.

In recent years, reinforcement learning (RL) and deep neural networks (DNN) achieved superhuman capabilities in a number of competitive games ([Mnih et al., 2015](#); [Silver et al., 2016](#)). This success has been a product of the combination of reinforcement learning, deep learning and Monte Carlo Tree Search (MCTS). However, current deep reinforcement learning (DRL) methods struggle in environments with a high number of states and a small number of reward states.



**Figure 1:** (a) Scrambled 3x3x3 Rubik's Cube. (b) 2x2x2 cube in the middle of a twist.

The Rubik's cube puzzle is an example of such an environment since the classical 3x3x3 cube has  $4.3 \cdot 10^{19}$  states and only *one* state (the solved cube) has a reward. A somewhat simpler puzzle is the 2x2x2 cube with  $3.6 \cdot 10^6$  state and again only one reward state. Both cubes are shown in Fig. 1.

The difficult task to *learn from scratch* how to solve arbitrary scrambled cubes (i.e. without being taught by expert knowledge, whether from humans or from computerized solvers) was not achievable with DRL methods for a long time. Recently, the works of [McAleer et al. \(2018, 2019\)](#) and [Agostinelli et al. \(2019\)](#) provided a breakthrough in that direction (see Sec. 5.1 and 7 for details): Their approach **DAVI** (Deep Approximate Value Iteration) learned from scratch to solve arbitrary scrambled 3x3x3 cubes.

This work investigates whether TD-n-tuple learning with much lower computational demands can solve (or partially solve) Rubik's cube as well.

## 1.2 Overview

The General Board Game (GBG) learning and playing framework ([Konen, 2019](#); [Konen and Bagheri, 2020](#); [Konen, 2022](#)) was developed for education and research in AI. GBG allows applying the new algorithm easily to a variety of games. GBG is open source and available on GitHub<sup>1</sup>. The main contribution of this paper is to take the TD-n-tuple approach from GBG ([Scheiermann and Konen, 2022](#)) that was also successful on other games (Othello, ConnectFour) and to investigate this algorithm on various cube puzzles. We will show that it can solve the 2x2x2 cube perfectly and the 3x3x3 cube partly. At the same time it has drastically reduced computational requirements compared to [McAleer et al. \(2019\)](#). We will show that wrapping the base agent with an **MCTS wrapper**, as it was done by [McAleer et al. \(2019\)](#) and [Scheiermann and Konen \(2022\)](#), is essential to reach this success.

This work is at the same time an in-depth tutorial how to represent a cube and its transformations within a computer program such that all types of cube operations can be computed efficiently. As another important contribution we will show how **symmetries** (Sec. 2.5, 6.3 and 6.4) applied to cube puzzles can greatly increase sample efficiency and performance.

The rest of this paper is organized as follows: Sec. 2 lays the foundation for Rubik’s cube, its state representation, its transformations and its symmetries. In Sec. 3 we introduce n-tuple systems and how they can be used to derive policies for game-playing agents. Sec. 4 defines and discusses several n-tuple representations for the cube. Sec. 5 presents algorithms for learning the cube: first the **DAVI** algorithm of [McAleer et al. \(2019\)](#); [Agostinelli et al. \(2019\)](#) and then our n-tuple-based TD learning (with extensions TCL and MCTS). In Sec. 6 we present the results when applying our n-tuple-based TD learning method to the 2x2x2 and the 3x3x3 cube. Sec. 7 discusses related work and Sec. 8 concludes.

---

<sup>1</sup><https://github.com/WolfgangKonen/GBG>

## 2 Foundations

### 2.1 Conventions and Symbols

We consider in this paper two well-known cube types, namely the 2x2x2 cube (pocket cube) and the 3x3x3 cube (Rubik's cube).

#### 2.1.1 Color arrangement

Each cube consists of smaller **cubies**: 8 corner cubies for the 2x2x2 cube and 8 corner, 12 edge and 6 center cubies for the 3x3x3 cube. A corner cubie has 3 **stickers** of different color on its 3 faces. An edge cubie has two, a center cubie has one sticker.

We enumerate the 6 cube faces with

(ULF) = (**Up**, **Left**, **Front**) and

(DRB) = (**Down**, **Right**, **Back**).

We number the 6 colors with 0,1,2,3,4,5. My cube has these six colors

012 = wbo = (white,blue,orange) in the (ULF)-cubie<sup>2</sup> and

345 = ygr = (yellow,green,red) in the opposing (DRB)-cubie.

The solved cube in default position has colors (012345) for the faces (ULFDRB), i.e. the white color is at the **Up** face, blue at **Left**, orange as **Front** and so on. We can cut the cube such that up- and bottom-face can be folded away and have a flattened representation as shown in Figure 2.

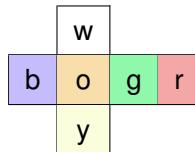


Figure 2: The face colors of the default cube in flattened representation

#### 2.1.2 Twist and Rotation Symbols

Twists of cube faces are denoted by uppercase letters U, L, F, D, R, B. Each of these twists means a 90° counterclockwise rotation.<sup>3</sup> If  $U = U^1$  is a 90° rotation, then  $U^2$  is a 180° rotation and  $U^3=U^{-1}$  is a 270° rotation.

Whole-cube rotations are denoted by lowercase letters  $u, \ell, f$ . (We do not need  $d, r, b$  here, because  $d = u^{-1}, r = \ell^{-1}$  and so on.)

Further symbols like  $f_c[i], s_\ell[i]$  that characterize a cube state will be explained in Sec. 2.3.

#### 2.1.3 Twist Types

Cube puzzles can have different twist types or twist metrics:

<sup>2</sup>We run through the faces of a cubie in counter-clockwise orientation.

<sup>3</sup>The rotation is counterclockwise when looking at the respective face

- **QTM (quarter turn metric)**: only quarter twists are allowed: e.g.  $U^1$  and  $U^{-1}$ .
- **HTM (half turn metric)**: quarter and half turns (twists) are allowed: e.g.  $U^1, U^2, U^3$ .

By *allowed* we mean what counts as one move. In QTM we can realize  $U^2$  via  $UU$  as well, but it costs us 2 moves. In HTM,  $U^2$  counts as one move.

The twist type influences **God's number** and the branching factor of the game, see Sec. 2.2.

## 2.2 Facts about Cubes

### 2.2.1 2x2x2 Cube

The **number of distinct states** for the 2x2x2 pocket cube is ([Wikipedia, 2022a](#))

$$\frac{8! \cdot 3^7}{24} = 7! \cdot 3^6 = 3,674,160 \approx 3.6 \cdot 10^6 \quad (1)$$

Why this formula? — We have 8 cubies which we can place in  $8!$  ways on the 8 cube positions. Each but the last cubie has the freedom to appear in 3 orientations, which gives the factor  $3^7$  (the last cubie is then in a fixed orientation, the other two orientations would yield illegal cube states). — Each of these raw states has the (ygr)-cubie in any of the 24 possible positions. Or, otherwise speaking, each truly different state appears in 24 whole-cube rotations. To factor out the whole-cube rotations, we count only the states with (ygr)-cubie in its default position (DRB) and divide the number of raw states by 24, q.e.d.

**God's number:** What is the minimal number of moves needed to solve any cube position? — For the 2x2x2 pocket cube, it is 11 in **HTM** (half-turn metric) and 14 in **QTM**.

**Branching factor:**  $3 \cdot 3 = 9$  in **HTM** and  $3 \cdot 2 = 6$  in **QTM**.

### 2.2.2 3x3x3 Cube

The **number of distinct states** for the 3x3x3 Cube is ([Wikipedia, 2022b](#))

$$\frac{8! \cdot 3^7 \cdot 12! \cdot 2^{11}}{2} = 43,252,003,274,489,856,000 \approx 4.3 \cdot 10^{19} \quad (2)$$

Why this formula? — We have 8 corner cubies which we can place in  $8!$  ways on the 8 cube positions. Each but the last cubie has the freedom to appear in 3 orientations, which gives the factor  $3^7$ . We have 12 edge cubies which we can place in  $12!$  ways on the edge positions. Each but the last cubie has the freedom to appear in 2 orientations, which gives the factor  $2^{11}$ . The division by 2 stems from the fact, that neither alone two corner cubies may be swapped nor alone two edge cubies may be swapped. Instead, the number of such swaps must be even (factor 2).

**God's Number:** What is the minimal number of moves needed to solve any cube position? — For the 3x3x3 Rubik's Cube, it is 20 in **HTM** (half-turn metric) and 26 in **QTM**. This is a result from [Rokicki et al. \(2014\)](#), see also <http://www.cube20.org/qtm/>.

**Branching factor:**  $6 \cdot 3 = 18$  in **HTM** and  $6 \cdot 2 = 12$  in **QTM**.

		3	2					
		0	1					
5	4	8	11	18	17	23	22	
6	7	9	10	19	16	20	21	
		14	13					
		15	12					

Figure 3: Sticker numbering for the 2x2x2 cube

### 2.3 The Cube State

A cube should be represented by objects in GBG in such a way that

- (a) cube states that are equivalent are represented by identical objects
- (b) if two cube states are equivalent, it should be easy to check this by comparing their objects
- (c) cube transformations are easy to carry out on these objects.

Condition (a) means that if two twist sequences lead to the same cube state (e.g.  $U^{-1}$  and  $UUU$ ), this should result also in identical objects. Condition (b) means, that the equality should be easy to check, given the objects. That is, a cube should *not* be represented by its twist sequence.

A cube state is in GBG represented by abstract class `CubeState` and has two describing members

$$f_c[i] = \text{fcol}[i] \quad (3)$$

$$s_\ell[i] = \text{sloc}[i] \quad (4)$$

$f_c[i] = \text{fcol}[i]$  denotes the **face color** at sticker location  $i$ . The color is one out of 0,1,2,3,4,5 for the colors w,b,o,y,g,r.

$s_\ell[i] = \text{sloc}[i]$  contains the **sticker location** of the *sticker which is in position  $i$  for the solved cube  $d$* .

Members  $f_c$  and  $s_\ell$  are vectors with 24 (2x2x2 cube) or 48 (3x3x3 cube) elements where  $i$  denotes the  $i$ th **sticker** location.

The stickers are numbered in a certain way which is detailed in Figures 3 and 4 for the flattened representations of the 2x2x2 and 3x3x3 cube, resp.

In principle, one of the two members  $f_c$  and  $s_\ell$  would be sufficient to characterize a state, since the **fcol-sloc-relation**

$$f_c[s_\ell[i]] = d.f_c[i] \quad (5)$$

holds, where  $d$  denotes the default cube. This is because  $s_\ell[i]$  transports the sticker  $i$  of the default cube  $d$  to location  $s_\ell[i]$ , i.e. it has the color  $d.f_c[i]$ . That is, we can easily

			6	5	4																			
			7			3																		
			0	1	2																			
10	9	8	16	23	22	36	35	34	46	45	44													
11		15	17		21	37		33	47		43													
12	13	14	18	19	20	38	39	32	40	41	42													
			28	27	26																			
			29		25																			
			30	31	24																			

Figure 4: Sticker numbering for the 3x3x3 cube. We do not number the center **cubies**, they stay invariant under twists.

Table 1: The three relevant twists for the 2x2x2 cube

		0	1	2	3	0	4	5	6	7	8	9	10	11	12	13	14	15	16	22	23	19	20	21	4	5
U twist	$T$	1	2	3	0	11	8	6	7	18	9	10	17	12	13	14	15	16	22	23	19	20	21	4	5	
L twist	$T$	22	1	2	21	5	6	7	4	3	0	10	11	12	13	8	9	16	17	18	19	20	14	15	23	
F twist	$T$	7	4	2	3	14	5	6	13	9	10	11	8	12	18	19	15	16	17	0	1	20	21	22	23	
$U^{-1}$	$T^{-1}$	3	0	1	2	22	23	6	7	5	9	10	4	12	13	14	15	16	11	8	19	20	21	17	18	
$L^{-1}$	$T^{-1}$	9	1	2	8	7	4	5	6	14	15	10	11	12	13	21	22	16	17	18	19	20	3	0	23	
$F^{-1}$	$T^{-1}$	18	19	2	3	1	5	6	0	11	8	9	10	12	7	4	15	16	17	13	14	20	21	22	23	

calculate  $f_c$  given  $s_\ell$ . With some more effort, it is also possible to calculate  $s_\ell$  given  $f_c$  (see Appendix A). Although one of these members  $f_c$  and  $s_\ell$  would be sufficient, we keep both because this allows to better perform assertions or cross checks during transformations.

Sometime we need the inverse function  $s_\ell^{-1}[i]$ : *Which sticker is at location  $i$ ?* It is easy to calculate  $s_\ell^{-1}$  given  $s_\ell$  with the help of the relation:

$$s_\ell^{-1}[s_\ell[i]] = i \quad (6)$$

(Note that it is *not* possible to invert  $f_c$ , because the face coloring function is not bijective.)

## 2.4 Transformations

### 2.4.1 Twist Transformations

Each basic twist is a counterclockwise<sup>4</sup> rotation of a face by  $90^\circ$ . Table 1 shows the 2x2x2 transformation functions for three basic twists. Each twist transformation can be coded in two forms:

1.  $T[i]$  (forward transformation): Which is the new location for the **sticker** being at  $i$  before the twist?

---

<sup>4</sup>The rotation is counterclockwise when looking at this face.

		2	1				
		3	0				
23	22	5	4	8	11	18	17
6	7	9	10	19	16	20	21
		14	13				
		15	12				

**Figure 5:** The default  $2 \times 2 \times 2$  cube after twist  $U^1$

2.  $T^{-1}[i]$  (inverse transformation): Which is the (parent) location of the **sticker** that lands in  $i$  after the twist?

Example (read off from column 0 of Table 1): The L-twist transports sticker at 0 to 22:  $T[0] = 22$ . The (parent) sticker being at location 9 before the L-twist comes to location 0 after the twist:  $T^{-1}[0] = 9$ . Likewise, for the U-twist we have  $T[0] = 1$  and  $T^{-1}[0] = 3$ . We show in Fig. 5 the default cube after twist  $U^1$ .

How can we apply a twist transformation to a cube state programmatically? – We denote with  $f'_c$  and  $s'_\ell$  the new states for  $f_c$  and  $s_\ell$  after transformation. The following relations allow to calculate the transformed cube state:

$$f'_c[i] = f_c[T^{-1}[i]] \quad (7)$$

$$s'_\ell[s_\ell^{-1}[i]] = T[i] \quad (8)$$

Eq. (7) says: The new color for sticker 0 is the color of the sticker which moves into location 0 ( $f_c[9]$  in the case of an L-twist). To explain Eq. (8), we first note that  $s_\ell^{-1}[i]$  is the sticker being at  $i$  before the transformation. Then, Eq. (8) says: „The new location for the sticker being at  $i$  before the transformation is  $T[i]$ .“ For example, the L-twist transports the current sticker at location 0 to the new location  $T[0] = 22$ , i. e.  $s'_\ell[0] = 22$ .

For the 2x2x2 cube, these 3 twists U, L, F are sufficient, because  $D=U^{-1}$ ,  $R=L^{-1}$ ,  $B=F^{-1}$ . This is because the 2x2x2 cube has no center **cubies**. For the 3x3x3 cube, we need all 6 twists U, L, F, D, R, B because this cube has center cubies.

In any case, we will show in Sec. 2.4.2 that only one row in Table 1 or Table 2, say  $T$  for the U-twist, has to be known or established 'by hand'. All other twists and their inverses can be calculated programmatically with the help of Eqs. (9)-(15) that will be derived in Sec. 2.4.2.

**Table 2: The U twist for the 3x3x3 cube**

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
U twist	$T$	2	3	4	5	6	7	0	1	22	23	16	11	12	13	14	15	36	17	18	19	20	21	34	35
		24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
U twist	$T$	24	25	26	27	28	29	30	31	32	33	44	45	46	37	38	39	40	41	42	43	8	9	10	47

**Table 3: Two basic whole-cube rotations for the 2x2x2 cube**

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
<i>u</i> rotation	$T$	1	2	3	0	11	8	9	10	18	19	16	17	15	12	13	14	21	22	23	20	6	7	4	5
<i>f</i> rotation	$T$	7	4	5	6	14	15	12	13	9	10	11	8	17	18	19	16	2	3	0	1	23	20	21	22
$u^{-1}$	$T^{-1}$	3	0	1	2	22	23	20	21	5	6	7	4	13	14	15	12	10	11	8	19	19	16	17	18
$f^{-1}$	$T^{-1}$	18	19	16	17	1	2	3	0	11	8	9	10	6	7	4	5	15	12	13	14	21	22	23	20

**Normalizing the 2x2x2 Cube** As stated above, the 3 twists U, L, F are sufficient for the 2x2x2 cube. Therefore, the (DRB)-cubie will never leave its place, whatever the twist sequence formed by U, L, F is. The (DRB)-cubie has the stickers (12, 16, 20), and we can check in Table 1 that columns (12, 16, 20) are always invariant. If we have an arbitrary initial 2x2x2 cube state, we can normalize it by applying a **whole-cube rotation** such that the (ygr)-cubie moves to the (DRB)-location.

**Normalizing the 3x3x3 Cube** In the case of the 3x3x3 cube, all center cubies will be not affected by any twist sequence. Therefore, we normalize a 3x3x3 cube state by applying initially a **whole-cube rotation** such that the center cubies are in their normal position (i.e. white up, blue left and so on).

#### 2.4.2 Whole-Cube Rotations (WCR)

Each basic **whole-cube rotation** (WCR) is a counterclockwise rotation of the whole cube around the  $u, l, f$ -axis by  $90^\circ$ . Table 3 shows two of the 2x2x2 transformation functions for basic whole-cube rotations. Each rotation can be coded in two forms:

1.  $T[i]$  (forward transformation): Which is the new location for the **sticker** being at  $i$  before the twist?
2.  $T^{-1}[i]$  (inverse transformation): Which is the (parent) location of the **sticker** that lands in  $i$  after the twist?

Besides the basic rotation  $u$  there is also  $u^2$  ( $180^\circ$ ) and  $u^3 = u^{-1}$  ( $270^\circ = -90^\circ$ ).

All whole-cube rotations can be generated from these two forward rotations  $u$  and  $f$ : First, we calculate the inverse transformations via

$$T^{-1}[T[i]] = i \quad (9)$$

where  $T$  is a placeholder for  $u$  or  $f$ . Next, we calculate the missing base rotation  $\ell$  (counterclockwise around the left face) as

$$\ell = fuf^{-1} \quad (10)$$

We use here the programm-code-oriented notation „**first trafo first**“: Eq. (10) reads as „**first  $f$ , then  $u$ , then  $f^{-1}$** “.<sup>5</sup>

---

<sup>5</sup>In programm code the relation would read `cs.fTr(1).uTr().fTr(3)`. This is „**first trafo first**“, because each transformation is applied to the cube state object to the left and returns the transformed cube state object.

**Table 4: All 24 whole-cube rotations (in first-trafo-first notation)**

number	first rotation	$* u^0$	$* u^1$	$* u^2$	$* u^3$
00-03	$id$ (white up)	$id$	$u$	$u^2$	$u^3$
04-07	$f$ (green up)	$f$	$fu$	$fu^2$	$fu^3$
08-11	$f^2$ (yellow up)	$f^2$	$f^2u$	$f^2u^2$	$f^2u^3$
12-15	$f^{-1}$ (blue up)	$f^{-1}$	$f^{-1}u$	$f^{-1}u^2$	$f^{-1}u^3$
16-19	$\ell$ (orange up)	$\ell$	$\ell u$	$\ell u^2$	$\ell u^3$
20-23	$\ell^{-1}$ (red up)	$\ell^{-1}$	$\ell^{-1}u$	$\ell^{-1}u^2$	$\ell^{-1}u^3$

**Table 5: The 24 inverse whole-cube rotations (in first-trafo-first notation)**

number	first rotation	$* u^0$	$* u^1$	$* u^2$	$* u^3$
00-03	$id$ (white up)	$id$	$u^3$	$u^2$	$u^1$
04-07	$f$ (green up)	$f^{-1}$	$\ell u^3$	$fu^2$	$\ell^{-1}u$
08-11	$f^2$ (yellow up)	$f^2$	$f^2u$	$f^2u^2$	$f^2u^3$
12-15	$f^{-1}$ (blue up)	$f$	$\ell^{-1}u^3$	$f^{-1}u^2$	$\ell u$
16-19	$\ell$ (orange up)	$\ell^{-1}$	$f^{-1}u^3$	$\ell u^2$	$fu$
20-23	$\ell^{-1}$ (red up)	$\ell$	$fu^{-1}$	$\ell^{-1}u^2$	$f^{-1}u$

The other basic whole-cube rotations  $d, r, b$  are not needed, because  $d = u^{-1}$ ,  $r = \ell^{-1}$  and  $b = f^{-1}$ .

The basic whole-cube rotations are rotations of the whole cube around just one axis. But there are also composite whole-cube rotations which consists of a sequence of basic rotations.

How many different (composite) rotations are there for the cube? – A little thought reveals that there are 24 of them: To be specific, we consider the default cube where we have 4 rotations with the white face up, 4 with the blue face up, and so on. In total we have  $6 \cdot 4 = 24$  rotations since there are 6 faces. Table 4 lists all of them, together with the WCR numbering convention used in GBG.

Sometimes we need the inverse whole-cube rotations which are given in Table 5. In this table, we read for example from the element with number 5, that the WCR with key 5 (which is  $fu$  according to Table 4) has the inverse WCR  $\ell u^3$  such that

$$fu \ell u^3 = id$$

holds.

For convenience, we list in Table 6 the `<Key, InverseKey>` relation. For example, the trafo with Key=5 ( $fu$ ) has the inverse trafo with InverseKey=19 ( $\ell u^3$ ). Note that there are 10 whole-cube rotations which are their own inverse.

**Generating all twists from U twist** With the help of WCRs we can generate the other twists from the U twist only: We simply rotate the face that we want to twist to the up-face,

**Table 6: Whole-cube rotations: <Key, InverseKey> relation**

key	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
inv key	0	3	2	1	12	19	6	21	8	9	10	11	4	23	14	17	20	15	18	05	16	7	22	13

apply the U twist and rotate back. This reads in **first-trafo-first** notation:

$$L = f^{-1}Uf \quad (11)$$

$$F = \ell U \ell^{-1} \quad (12)$$

$$D = f^2 U f^2 \quad (13)$$

$$R = f U f^{-1} \quad (14)$$

$$B = \ell^{-1} U \ell \quad (15)$$

Thus, given the U twist from Table 1 or Table 2 and the basic WCRs given in Table 3 and Eq. (10), we can calculate all other forward transformations with the help of Eqs. (11)–(15). Then, all inverse transformations are calculable with the help of Eq. (9).

#### 2.4.3 Color Transformations

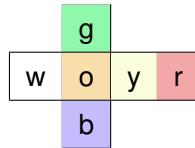
Color transformations are special transformations that allow to discover non-trivial symmetric (equivalent) states.

One way to describe a color transformation is to select a valid color permutation and to paint each sticker with the new color according to this color permutation. This is of course nothing one can do with a real cube without destroying or altering it, but it is a theoretical concept leading to an equivalent state.

Another way of looking at it is to record the twist sequence that leads from the default cube to a certain scrambled cube state. Then we go back to the default cube, make at first a whole-cube rotation (leading to a color-transformed default cube) and then apply the recorded twist sequence to the color-transformed default cube.

In any case, the transformed cube will be usually not in its normal position, so we apply finally a **normalizing operation** to it.

What are valid color permutations? – These are permutations of the cube colors reachable when applying one of the available 24 WCRs (Table 4) to the default cube. For example, if we apply WCR  $f$  (number 04) to the default cube, we get



**Figure 6: The color transformation according to WCR  $f$  (number 04)**

that is, g (green) is the new color for each up-sticker that was w (white) before and so on. The colors o and r remain untouched under this color permutation. [However, other transformations like  $fu$ ,  $fu^2$  and  $fu^3$  will change every color.]

	2	1
	3	0
23	22	5
6	7	4
14	13	
15	12	

Figure 7: The cube of Fig. 5 before color transformation.

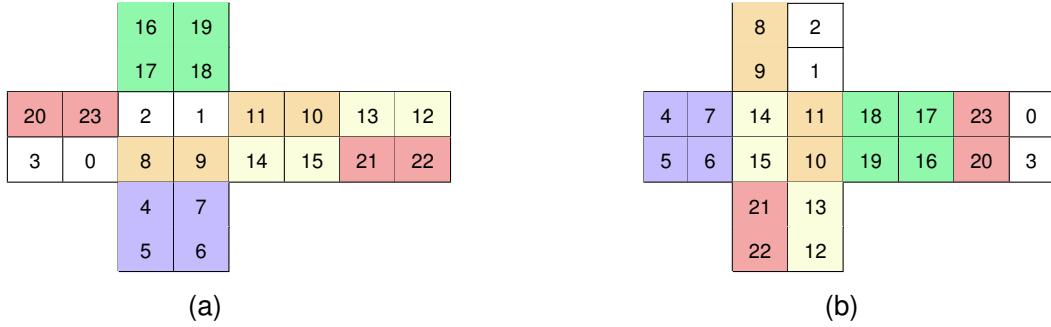


Figure 8: The cube of Fig. 7 with color transformation from Fig 6: (a) before normalization, (b) after normalization.

How can we apply a color transformation to a cube state programmatically? – We denote with  $f'$  and  $s'_\ell$  the new states for  $f$  and  $s_\ell$  after transformation. The following relations allow to calculate the transformed cube state:

$$f'_c[i] = c[f_c[i]] \quad (16)$$

$$s'_\ell[s_\ell^{-1}[i]] = T[i] \quad (17)$$

where  $c[]$  is the 6-element color trafo vector (holding the new colors for current colors 0:w, 1:b, ..., 5:r) and  $T$  is the 24- or 48-element vector of the WCR that produces this color transformation. Eq. (16) is simple: If a certain sticker has color 0 (w, white) before the color transformation, then it will get the new color  $c[0]$ , e.g. 4 (g, green), after the transformation. Eq. (17) looks complicated, but it has a similar meaning as in the twist trafo: Take  $i = 0$  as example: The new place for the sticker being at 0 before the trafo (and coming from  $s_\ell^{-1}[0]$ ) is  $T[0]$ . Therefore, we write the number  $T[0]$  into  $s'_\ell[s_\ell^{-1}[0]]$ .

A **color transformation example** is shown in Figs. 7 and 8. Fig. 7 is just a replication of Fig. 5 showing a default cube after  $U^1$  twist. The color transformation number 04 applied to the cube of Fig. 7 is shown in Fig. 8 (a)-(b) in two steps:

- (a) The stickers are re-painted and re-numbered (white becomes green, blue becomes white and so on). The structure of coloring is the same as in Fig. 7. Now the (DRB)-cubie is no longer the (ygr)-cubie, it does not carry the numbers (12,16,20).

- (b) We apply the proper **WCR** that brings the (ygr)-cubie back to the **(DRB)**-location. Compared to (a), each 4-sticker cube face is just rotated to another face, but not changed internally. We can check that the **(DRB)**-location now carries again the numbers (12,16,20), as in Fig. 7 and as it should for a normalized cube.

## 2.5 Symmetries

Symmetries are transformations of the game state (and the attached action, if applicable) that lead to equivalent states. That is, if  $s$  is a certain state with value  $V(s)$ , then all states  $s_{sym}$  being symmetric to  $s$  have the same value  $V(s_{sym}) = V(s)$  because they are equivalent. *Equivalent* means: If  $s$  can be solved by a twist sequence of length  $n$ , then  $s_{sym}$  can be solved by an equivalent twist sequence of same length  $n$ .

In the case of Rubik's cube, all whole-cube rotations (**WCRs**) are symmetries because they do not change the value of a state. But whole-cube rotations are 'trivial' symmetries because they are usually factored out by the normalization of the cube: After **2x2x2 cube normalization**, which brings the (ygr)-cubie in a certain position, or after **3x3x3 cube normalization**, which brings the center cubies in certain faces, all **WCR**-symmetric states are transformed to the same state.

Non-trivial symmetries are all color transformations (Sec. 2.4.3): In general, color transformations transform a state  $s$  to a truly different state  $s_{sym}$ , even after **cube normalization**.<sup>6</sup> Since there are 24 color transformations in Rubik's cube, there are also 24 non-trivial symmetries (including self).

Symmetries are useful to learn to solve Rubik's cube for two reasons: (a) to accelerate learning and (b) to smooth an otherwise noisy value function.

- (a) **Accelerated learning:** If a state  $s$  (or state-action pair) is observed, not only the weights activated by that state are updated, but also the weights of all symmetric states  $s_{sym}$ , because they have the same  $V(s_{sym}) = V(s)$  and thus the same reward. In this way, a single observed sample is connected with more weight updates (better sample efficiency).
- (b) **Smoothed value function:** By this we mean that the value function  $V(s)$  is replaced by

$$V^{(sym)}(s) = \frac{1}{|\mathfrak{F}_s|} \sum_{s' \in \mathfrak{F}_s} V(s') \quad (18)$$

where  $\mathfrak{F}_s$  is the set of states being symmetric to  $s$ . If  $V(s)$  were the ideal value function, both terms  $V(s)$  and  $V^{(sym)}(s)$  would be the same.<sup>7</sup> But in a real n-tuple network,  $V(s)$  is non-ideal due to n-tuple-noise (cross-talk from other states that activate the same n-tuple LUT entries). If we average over the symmetric states  $s' \in \mathfrak{F}_s$ , the noise will be dampened.

---

<sup>6</sup>In rare cases – e.g. for the solved cube – the transformed state may be identical to  $s$  or to another symmetry state, but this happens seldom for sufficiently scrambled cubes, see Sec. 6.3.

<sup>7</sup>because all  $V(s')$  in Eq. (18) are the same for an ideal  $V$

The downside of symmetries is their computational cost: In the case of Rubik’s cube, the calculation of color transformations is a costly operation. On the other hand, the number of necessary training episodes to reach a certain performance may be reduced. In the end, the use of symmetries may pay off, because the total training time may be reduced as well. In any case, we will have a better sample efficiency, since we learn more from each observed state or state-action pair. Secondly, the smoothing effect introduced with Eq. (18) can lead to better overall performance, because the smoothed value function provides a better guidance on the path towards the solved cube.

In order to balance computation time, GBG offers the option to select with  $n_{\text{Sym}}$  the number of symmetries actually used. If we specify for example  $n_{\text{Sym}}=8$  in GBG’s Rubik’s cube implementation, then the state itself and  $8-1=7$  random other (non-id) color transformations will be selected. The resulting set  $\mathfrak{F}_s$  of 8 states is then used for weight update and value function computation.

### 3 N-Tuple Systems

N-tuple systems coupled with TD were first applied to game learning by [Lucas \(2008\)](#), although n-tuples were already introduced by [Bledsoe and Browning \(1959\)](#) for character recognition purposes. The remarkable success of n-tuples in learning to play Othello ([Lucas, 2008](#)) motivated other authors to benefit from this approach for a number of other games.

The main goal of n-tuple systems is to map a highly non-linear function in a low dimensional space to a high dimensional space where it is easier to separate ‘good’ and ‘bad’ regions. This can be compared to the kernel trick of support-vector machines. An n-tuple is defined as a sequence of  $n$  cells of the board. Each cell can have  $m$  positional values representing the possible states of that cell.<sup>8</sup> Therefore, every n-tuple will have a (possibly large) look-up table indexed in form of an  $n$ -digit number in base  $m$ . Each entry corresponds to a feature and carries a trainable weight. An n-tuple system is a system consisting of  $k$  n-tuples. As an example we show in Fig. 9 an n-tuple system consisting of four 8-tuples.

Let  $\Theta$  be the vector of all weights  $\theta_i$  of the n-tuple system.<sup>9</sup> The length of this vector may be large number, e.g.  $m^n k$ , if all  $k$  n-tuples have the same length  $n$  and each cell has  $m$  positional values. Let  $\Phi(s)$  be a binary vector of the same length representing the feature occurrences in state  $s$  (that is,  $\Phi_i(s) = 1$  if in state  $s$  the cell of a specific n-tuple as indexed by  $i$  has the positional value as indexed by  $i$ ,  $\Phi_i(s) = 0$  else). The value function of the n-tuple network given state  $s$  is

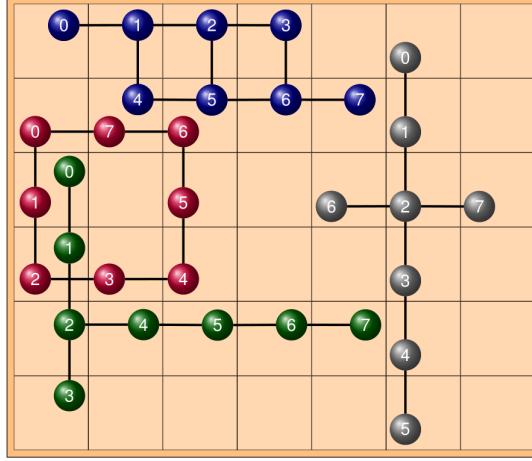
$$V(s) = \sigma(\Phi(s) \cdot \Theta) \quad (19)$$

with transfer function  $\sigma$  which may be a sigmoidal function or simply the identity function.

---

<sup>8</sup>A typical example is a 2-player board game, where we usually have 3 positional values {0: empty, 1: player1, 2: player2 }. But other, user-defined values are possible as well.

<sup>9</sup>The index  $i$  indexes three qualities: an n-tuple, a cell in this n-tuple and a positional value for this cell.



**Figure 9: Example n-tuples:** We show 4 random-walk 8-tuples on a 6x7 board. The tuples are selected manually to show that not only snake-like shapes are possible, but also bifurcations or cross shapes. Tuples may or may not be symmetric.

An agent using this n-tuple system derives a policy from the value function in Eq. (19) as follows: Given state  $s$  and the set  $A(s)$  of available actions in state  $s$ , it applies with a forward model  $f$  every action  $a \in A(s)$  to state  $s$ , yielding the next state  $s' = f(s, a)$ . Then it selects the action that maximizes  $V(s')$ .

Each time a new agent is constructed, all n-tuples are either created in fixed, user-defined positions and shapes, or they are formed by *random walk*. In a *random walk*, all cells are placed randomly with the constraint that each cell must be adjacent<sup>10</sup> to at least one other cell in the n-tuple.

Agent training proceeds in the TD-n-tuple algorithm as follows: Let  $s'$  be the actual state generated by the agent and let  $s$  be the previous state generated by this agent. TD(0) learning adapts the value function with model parameters  $\Theta$  through (Sutton and Barto, 1998)

$$\Theta \leftarrow \Theta + \alpha \delta \nabla_{\Theta} V(s) \quad (20)$$

Here,  $\alpha$  is the learning rate and  $V$  is in our case the n-tuple value function of Eq. (19).  $\delta$  is the usual TD error (Sutton and Barto, 1998) after the agent has acted and generated  $s'$ :

$$\delta = r + \gamma V(s') - V(s) \quad (21)$$

where the sum of the first two terms, reward  $r$  plus the discounted value  $\gamma V(s')$ , is the desirable target for  $V(s)$ .

---

<sup>10</sup>The form of adjacency, e. g. 4- or 8-point neighborhood or any other (might be cell-dependent) form of adjacency, is user-defined.

## 4 N-Tuple Representations for the Cube

In order to apply n-tuples to cubes, we have to define a board in one way or the other on which we can place the n-tuples. This is not as straightforward as in other board games, but we are free to invent abstract boards. Once we have defined a board, we can number the board cells  $k = 0, \dots, K-1$  and translate a cube state into a BoardVector: A **BoardVector**  $\mathbf{b}$  is a vector of  $K$  non-negative integer numbers  $b_k \in \{0, \dots, N_k - 1\}$ . Each  $k$  represents a board cell and every board cell  $k$  has a predefined number  $N_k$  of position values.<sup>11</sup>

A **BoardVector** is useful to calculate the feature occurrence vector  $\Phi(s)$  in Eq. (19) for a given n-tuple set: If an n-tuple contains board cell  $k$ , then look into  $b_k$  to get the position value for this cell  $k$ . Set  $\Phi_i(s) = 1$  for that index  $i$  that indexes this n-tuple cell and this position value.

In the following we present different options for boards and **BoardVectors**. We do this mainly for the 2x2x2 cube, because it is somewhat simpler to explain. But the same ideas apply to the 3x3x3 cube as well, they are just a little bit longer. Therefore, we defer the lengthy details of the 3x3x3 cube to Appendix B.

### 4.1 CUBESTATE

A natural way to translate the cube state into a board is to use the flattened representation of Fig. 11 as the board and extract from it the 24-element vector  $\mathbf{b}$ , according to the given numbering. The  $k$ th element  $b_k$  represents a certain cubie face location and gets a number from  $\{0, \dots, 5\}$  according to its current face color  $f_c$ . The solved cube is for example represented by  $\mathbf{b} = [0000 1111 2222 \dots 5555]$ .

This representation CUBESTATE is what the BoardVecType CUBESTATE in our GBG-implementation means: Each board vector is a copy of `fcol`, the face colors of all cubie faces. `fcol` is also the vector that uniquely defines each cube state. An upper bound of possible combinations for  $\mathbf{b}$  is  $6^{24} = 4.7 \cdot 10^{18}$ . If we factor out the (DRB)-cubie, which always stays at its home position, we can reduce this to 21 board cells with 6 positional values, leading to  $6^{21} = 2.1 \cdot 10^{16}$  weights. Both numbers are of course way larger than the true number of distinct states (Sec. 2.2.1) which is  $3.6 \cdot 10^6$ . This is because most of the combinations are dead weights in the n-tuple LUTs, they will never be activated during game play.

The dead weights occur because many combinations are not realizable, e.g. three white faces in one cubie or any of the  $6^3 - 8 \cdot 3 = 192$  cubie-face-color combinations that are not present in the real cube. The problem is that the dead weights are scattered in a complicated way among the active weights and it is thus not easy to factor them out.

### 4.2 STICKER

McAleer et al. (2019) had the interesting idea for the 3x3x3 cube that 20 **stickers** (cubie faces) are enough. To characterize the full 3x3x3 cube, we need only one (not 2 or 3) sticker

<sup>11</sup>In GBG package `ntuple2` (base for agent TDNTuple3Agt), all  $N_k$  have to be the same. In package `ntuple4` (base for agent TDNTuple4Agt), numbers  $N_k$  may be different for different  $k$ .

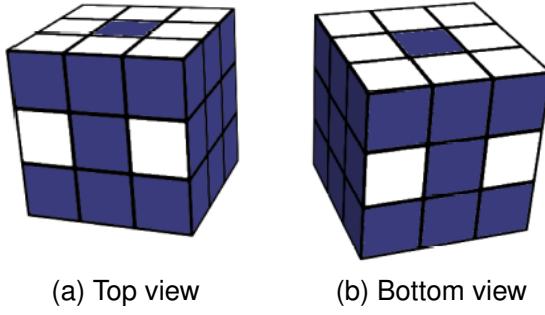


Figure 10: The sticker representation used to reduce dimensionality: Stickers that are used are shown in white, whereas ignored stickers are dark blue (from McAleer et al. (2019)).

	3	2						
	0	1						
5	4	8	11	18	17	23	22	
6	7	9	10	19	16	20	21	
	14	13						
	15	12						

**Figure 11:** Tracked stickers for the 2x2x2 cube (white), while ignored stickers are blue.

for every of the 20 cubies, as shown in Fig. 10. This is because the location of one sticker uniquely defines the location and orientation of that cubie. We name this representation STICKER in GBG.

Translated to the  $2 \times 2 \times 2$  cube, this means that 8 stickers are enough because we have only 8 cubies. We may for example track the 4 top stickers 0,1,2,3 plus the 4 bottom stickers 12,13,14,15 as shown in Fig. 11 and ignore the 16 other stickers. Since we always normalize the cube such that the (DRB)-cubie with sticker 12 stays in place, we can reduce this even more to **7 stickers** (all but sticker 12).

How to lay out this representation as a board? – McAleer et al. (2019) create a rectangular one-hot-encoding board with  $7 \times 21 = 147$  cells (7 rows for the stickers and 21 columns for the locations) carrying only 0's and 1's. This is fine for the approach of McAleer et al. (2019), where they use this board as input for a DNN, but not so nice for n-tuples. Without constraints, such a board amounts to  $2^{147} = 1.7 \cdot 10^{44}$  combinations, which is unpleasantly large (much larger than in CUBESTATE).<sup>12</sup>

STICKER has more dead weights than CUBESTATE, so it seems like a step back. But the point is, that the dead weights are better structured: If for example sticker 0 appears at column 1 then this column and the two other columns for the same cubie are automatically

<sup>12</sup>A possible STICKER **BoardVector** for the default cube would read  $\mathbf{b} = [1000000\ 0100000\ 0010000\ \dots]$ , meaning that location 0 has the first sticker, location 1 has the second sticker, and so on. In any STICKER **BoardVector** there are only 7 columns carrying exactly one 1, the other carry only 0's. Every row carries exactly one 1.

**Table 7: The correspondence *corner location*  $\leftrightarrow$  STICKER2 for the solved cube. The yellow colored cells show the location of the 7 (2x2x2) and 8 (3x3x3) corner stickers that we track.**

2x2x2	location	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
3x3x3	location	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46
STICKER2	corner	a	b	c	d	a	d	h	g	a	g	f	b	e	f	g	h	e	c	b	f	e	h	d	c
	face ID	1	1	1	1	2	3	2	3	3	2	3	2	1	1	1	1	2	2	3	2	3	3	2	3

forbidden for all other stickers. Likewise, if sticker 1 is placed in another column, another set of 3 columns is forbidden, and so on. We can use this fact to form a much more compact representation STICKER2.

### 4.3 STICKER2

As the analysis in the preceding section has shown, the 21 location columns of STICKER cannot carry the tracked stickers in arbitrary combinations. Each cubie (represented by 3 columns in STICKER) carries only exactly *one* sticker. We can make this fact explicit by choosing another representation for the 21 locations:

$$\text{corner location} = (\text{corner cubie}, \text{face ID}).$$

That is, each location is represented by a pair: corner cubie a,b,c,d,f,g,h (we number the top cubies with letters a,b,c,d and the bottom cubies with letters e,f,g,h and omit e because it corresponds to the (DRB)-cubie) and a face ID. To number the faces with a **face ID**, we follow the convention that we start at the top (bottom) face with face ID 1 and then move counter-clockwise around the corner cubie to visit the other faces (2,3). Table 7 shows the explicit numbering in this new representation.

To represent a state as board vector we use now a much smaller board shown in Table 8: Each cell in the first row has 7 position values (the letters) and each cell in the second row has 3 position values (the **face IDs**). We show in Table 8 the board vector for the default cube,  $b = [abcdfgh\ 1111111]$ . Representation STICKER2 allows for  $7^7 \cdot 3^7 = 1.8 \cdot 10^9$  combinations in total, which is much smaller than STICKER and CUBESTATE.

**Table 8: STICKER2 board representation for the default 2x2x2 cube. For the **BoardVector**, cells are numbered row-by-row from 0 to 16.**

corner	a	b	c	d	f	g	h	7 positions
face ID	1	1	1	1	1	1	1	3 positions

STICKER2 has some dead weights remaining, because the combinations can carry the same letter multiple times, which is not allowed for a real cube state. But this rate of dead weights is tolerable.

It turns out that STICKER2 is in all aspects better than CUBESTATE or STICKER. Therefore, we will only report the results for STICKER2 in the following.

## 4.4 Adjacency Sets

To create n-tuples by random walk, we need adjacency sets (sets of neighbors) to be defined for every board cell  $k$ .

For CUBESTATE, the board is the flattened representation of the  $2 \times 2 \times 2$  cube (Fig. 3). The adjacency set is defined as the 4-point neighborhood, where two stickers are neighbors if they share a common edge on the cube, i.e. are neighbors on the cube.

For STICKER2, the board consists of 16 cells shown in Table 8. Here, the adjacency set for cell  $k$  contains all other cells different from  $k$ .

Again, the details of ideas similar to Sec. 4.1–4.4, but now for the  $3 \times 3 \times 3$  cube, are shown in Appendix B.1–B.4.

## 5 Learning the Cube

### 5.1 McAleer and Agostinelli

The works of [McAleer et al. \(2018, 2019\)](#) and [Agostinelli et al. \(2019\)](#) contain up to now the most advanced methods for learning to solve the cube from scratch. [Agostinelli et al. \(2019\)](#) introduces the cost-to-go function for a general Marko decision process

$$J(s) = \min_{a \in A(s)} \sum_{s'} P^a(s, s') (g^a(s, s') + \gamma J(s')) \quad (22)$$

where  $P^a(s, s')$  is the probability of transitioning from state  $s$  to  $s'$  by taking action  $a$  and  $g^a(s, s')$  is the cost for this transition. In the Rubik's cube case, we have deterministic transitions, that is  $s' = f(s, a)$  is deterministically prescribed by a forward model  $f$ . Therefore, the sum reduces to one term and we specialize to  $\gamma = 1$ . Furthermore, we set  $g^a(s, s') = 1$ , because only the length of the solution path counts, so that we get the simpler equation

$$J(s) = \min_{a \in A(s)} (1 + J(s')) \quad \text{with} \quad s' = f(s, a). \quad (23)$$

Here,  $A(s)$  is the set of available actions in state  $s$ . We additionally set  $J(s^*) = 0$  if  $s^*$  is the solved cube. To better understand Eq. (23) we look at a few examples: If  $s_1$  is a state one twist away from  $s^*$ , Eq. (23) will find this twist and set  $J(s_1) = 1$ . If  $s_2$  is a state two twists away from  $s^*$  and all one-twist states have already their correct labels  $J(s_1) = 1$ , then Eq. (23) will find the twist leading to a  $s_1$  state and set  $J(s_2) = 1 + 1 = 2$ . While iterations proceed, more and more states (being further away from  $s^*$ ) will be correctly labeled, once their preceding states are correctly labeled. In the end we should ideally have

$$J(s_n) = n.$$

However, the number of states for Rubik's cube is too large to store them all in tabular form. Therefore, [McAleer et al. \(2019\)](#) and [Agostinelli et al. \(2019\)](#) approximate  $J(s)$  with a deep neural network (DNN). To train such a network in the Rubik's cube case, they

---

**Algorithm 1** DAVI algorithm (from Agostinelli et al. (2019)). Input:  $B$ : batch size,  $K$ : maximum number of twists,  $M$ : training iterations,  $C$ : how often to check for convergence,  $\epsilon$ : error threshold. Output:  $\Theta$ , the trained neural network parameters.

---

```

1: function DAVI( $B, K, M, C, \epsilon$ )
2:    $\Theta \leftarrow \text{INITIALIZENETWORKPARAMETERS}$ 
3:    $\Theta_C \leftarrow \Theta$ 
4:   for  $m = 1, \dots, M$  do
5:      $X \leftarrow \text{GENERATESCRAMBLEDSTATES}(B, K)$  ▷  $B$  scrambled cubes
6:     for  $x_i \in X$  do
7:        $y_i \leftarrow \min_{a \in A(s)} [1 + j_{\Theta_C}(f(x_i, a))]$  ▷ cost-to-go function, Eq. (23)
8:        $(\Theta, \text{loss}) \leftarrow \text{TRAIN}(j_{\Theta}, X, y)$  ▷ loss = MSE( $j_{\Theta}(x_i), y_i$ )
9:       if ( $m \bmod C = 0$  & loss <  $\epsilon$ ) then
10:         $\Theta_C \leftarrow \Theta$ 
11:   return  $\Theta$ 

```

---

introduce **Deep Approximate Value Iteration (DAVI)**<sup>13</sup> shown in Algorithm 1. The network output  $j_{\Theta}(s)$  is trained in line 8 to approximate the (unknown) cost-to-go  $J(s)$  for every state  $s = x_i$ . The main trick of DAVI is, as Agostinelli et al. (2019) write: „For learning to occur, we must train on a state distribution that allows information to propagate from the goal state to all the other states seen during training. Our approach for achieving this is simple: each training state  $x_i$  is obtained by randomly scrambling the goal state  $k_i$  times, where  $k_i$  is uniformly distributed between 1 and  $K$ . During training, the cost-to-go function first improves for states that are only one move away from the goal state. The cost-to-go function then improves for states further away as the reward signal is propagated from the goal state to other states through the cost-to-go function.“

Agostinelli et al. (2019) use in Algorithm 1 two sets of parameters to train the DNN: the parameters  $\Theta$  being trained and the parameters  $\Theta_C$  used to obtain improved estimates of the cost-to-go function. If they did not use this two separate sets, performance often „saturated after a certain point and sometimes became unstable. Updating  $\Theta_C$  only after the error falls below a threshold  $\epsilon$  yields better, more stable, performance.“ (Agostinelli et al., 2019) To train the DNN, they used  $M = 1\,000\,000$  iterations, each with batch size  $B = 10\,000$ . Thus, the trained DNN has seen ten billion cubes ( $10^{10}$ ) during training, which is still only a small subset of the  $4.3 \cdot 10^{19}$  possible cube states.

The heuristic function of the trained DNN alone cannot solve 100% of the cube states. Especially for higher twist numbers  $k_i$ , an additional solver or search algorithm is needed. This is in the case of McAleer et al. (2019) a Monte Carlo Tree Search (MCTS), similar to AlphaZero (Silver et al., 2017), which uses the DNN as the source for prior probabilities. Agostinelli et al. (2019) use instead a variant of A\*-search, which is found to produce solutions with a shorter path in a shorter runtime than MCTS.

---

<sup>13</sup>More precisely, McAleer et al. (2019) use Autodidactic Iteration (ADI), a precursor to DAVI, very similar to DAVI, just a bit more complicated to explain. Therefore, we describe here only DAVI.

---

**Algorithm 2** TD-n-tuple algorithm for Rubik's cube. Input:  $p_{max}$ : maximum number of twists,  $M$ : training iterations,  $E_{train}$ : maximum episode length during training,  $c$ : negative cost-to-go,  $R_{pos}$ : positive reward for reaching the solved cube  $s^*$ ,  $\alpha$ : learning rate.  $j_\Theta(s)$ : n-tuple network value prediction for state  $s$ . Output:  $\Theta$ , the trained n-tuple network parameters.

---

```

1: function TDNTUPLE( $p_{max}, M, E_{train}, c, R_{pos}$ )
2:    $\Theta \leftarrow \text{INITIALIZENETWORKPARAMETERS}$ 
3:   for  $m = 1, \dots, M$  do
4:      $p \sim U(1, \dots, p_{max})$             $\triangleright$  Draw  $p$  uniformly random from  $\{1, 2, \dots, p_{max}\}$ 
5:      $s \leftarrow \text{SCRAMBLE}(\text{SOLVEDCUBE}(p))$             $\triangleright$  start state
6:     for  $k = 1, \dots, E_{train}$  do
7:        $s_{new} \leftarrow \arg \max_{a \in A(s)} V(s')$  with  $s' = f(s, a)$  and
8:        $V(s') = c + \begin{cases} R_{pos} & \text{if } s' = s^* \\ j_\Theta(s') & \text{if } s' \neq s^* \end{cases}$ 
9:       Train network  $j_\Theta$  with Eq. (20) to bring  $V(s)$  closer to target  $T = V(s_{new})$ :
```

$$V(s) \leftarrow V(s) + \alpha(T - V(s))$$

```

10:     $s \leftarrow s_{new}$ 
11:    if ( $s = s^*$ ) then
12:      break            $\triangleright$  break out of  $k$ -loop
13:    return  $\Theta$ 
```

---

## 5.2 N-Tuple-based TD Learning

To solve the Rubik’s cube in GBG we use an algorithm that is on the one hand inspired by **DAVI**, but on the other hand more similar to traditional reinforcement learning schemes like temporal difference (TD) learning. In fact, we want to use in the end the same TD-FARL algorithm ([Konen and Bagheri, 2021](#)) that we use for all other GBG games.

We show in Algorithm 2 our method, that we will explain in the following, highlighting also the similarities and dissimilarities to **DAVI**.

First of all, instead of minimizing the positive cost-to-go as in **DAVI**, we maximize in lines 7-8 a value function  $V(s')$  with a negative cost-to-go. This maximization is functionally equivalent, but more similar to the usual TD-learning scheme. The negative cost-to-go, e.g.  $c = -0.1$ , plays the role of the positive 1 in Eq. (23).

Secondly, we replace the DNN of **DAVI** by the simpler-to-train n-tuple network  $j_\Theta$  with STICKER2 representation as described in Sec. 3 and 4. That is, each time  $j_\Theta(s')$  is requested, we first calculate for state  $s'$  the **BoardVector** in STICKER2 representation, then the occurrence vector  $\Phi(s')$  and the value function  $V(s')$  according to Eq. (19).

The central equations for  $V(s')$  in Algorithm 2, lines 7-8, work similar to Eq. (23) in **DAVI**: If  $s = s_1$  is a state one twist away from  $s^*$ , the local search in  $\arg \max V(s')$  will find this twist and the training step in line 9 moves  $V(s)$  closer to  $c + R_{pos}$ .<sup>14</sup> Likewise, neighbors  $s_2$  of  $s_1$  will find  $s_1$  and thus move  $V(s_2)$  closer to  $2c + R_{pos}$ . Similar for  $s_3, s_4, \dots$  under the assumption that a ‘known’ state is in the neighborhood. We have a clear gradient on the path towards the solved cube  $s^*$ . If there are no ‘known’ states in the neighborhood of  $s_n$ , we get for  $V(s_n)$  what the net maximally estimates for all those neighbors. We pick the neighbor with the highest estimate, wander around randomly until we hit a state with a ‘known’ neighbor or until we reach the limit  $E_{train}$  of too many steps.

Note that Algorithm 2 is different from **DAVI** insofar that it follows the path  $s \rightarrow s' \rightarrow \dots$  as prescribed by the current  $V$ , which may lead to a state sequence ‘wandering in the unknown’ until  $E_{train}$  is reached. In contrast to that, **DAVI** generates many start states  $s_0$  drawn from the distribution of training set states and trains the network just on pairs  $(s_0, T)$ , i.e. they do just **one step** on the path. We instead follow the full path, because we want the training method for Rubik’s cube to be as similar as possible to the training method for other GBG games.<sup>15</sup>

Algorithm 2 is basically the same algorithm as GBG uses for other games. The only differences are (i) the cube-specific start state selection borrowed from **DAVI** (a 1-twist start state has the same probability as a 10-twist start state) and (ii) the cube-specific reward in line 8 of Algorithm 2 with its negative cost-go-go  $c$  which is however a common element of many RL rewards.

Algorithm 2 currently learns with only one parameter vector  $\Theta$ . However, it could be extended as in **DAVI** to two parameter vectors  $\Theta$  and  $\Theta_C$ . The weight training step in line

---

<sup>14</sup>It is relevant, that  $R_{pos}$  is a **positive** number, e.g. 1.0 (and not 0, as it was for **DAVI**). This is because we start with an initial n-tuple network with all weights set to 0, so the initial response of the network to any state is 0.0. Thus, if  $R_{pos}$  were 0, a one-twist state would see all its neighbors (including  $s^*$ ) initially as responding 0.0 and would not learn the right transition to  $s^*$ . With  $R_{pos} = 1.0$  it will quickly find  $s^*$ .

<sup>15</sup>We note in passing that we tested the **DAVI** variant with  $E_{train} = 1$  for our TD-n-tuple method as well. However, we found that this method gave much worse results, so we stick with our GBG method here.

9 is done with the help of Eq. (20) for  $\Theta$  using the error signal  $\delta$  of Eq. (21).

There are two extra elements, TCL and MCTS, that complete our n-tuple-based TD learning. They are described in the next two subsections.

### 5.2.1 Temporal Coherence Learning (TCL)

The TCL algorithm developed by Beal and Smith [Beal and Smith \(1999\)](#) is an extension of TD learning. It replaces the global learning rate  $\alpha$  with the weight-individual product  $\alpha\alpha_i$  for every weight  $\theta_i$ . Here, the adjustable learning rate  $\alpha_i$  is a free parameter set by a pretty simple procedure: For each weight  $\theta_i$ , two counters  $N_i$  and  $A_i$  accumulate the sum of weight changes and the sum of absolute weight changes. If all weight changes have the same sign, then  $\alpha_i = |N_i|/A_i = 1$ , and the learning rate stays at its upper bound. If weight changes have alternating signs, then the global learning rate is probably too large. In this case,  $\alpha_i = |N_i|/A_i \rightarrow 0$  for  $t \rightarrow \infty$ , and the effective learning rate will be largely reduced for this weight.

In our previous work ([Bagheri et al., 2015](#)) we extended TCL to  $\alpha_i = g(|N_i|/A_i)$  where  $g$  is a transfer function being either the identity function (standard TCL) or an exponential function  $g(x) = e^{\beta(x-1)}$ . It was shown in [Bagheri et al. \(2015\)](#) that TCL with this exponential transfer function leads to faster learning and higher win rates for the game ConnectFour.

### 5.2.2 MCTS

We use Monte Carlo Tree Search (MCTS) ([Browne et al., 2012](#)) to augment our trained network during testing and evaluation. This is the method also used by [McAleer et al. \(2019\)](#) and by AlphaGo Zero ([Silver et al., 2017](#)), but they use it also during training.

MCTS builds iteratively a search tree starting with a tree containing only the start state  $s_0$  as the root node. Until the iteration budget is exhausted, MCTS does the following: In every iteration we start from the root node and select actions following the tree policy until we reach a yet unexpanded leaf node  $s_\ell$ . The tree policy is implemented in our MCTS wrapper according to the UCB formula ([Silver et al., 2017](#)):

$$a_{new} = \arg \max_{a \in A(s)} \left( \frac{W(s, a)}{N(s, a)} + U(s, a) \right) \quad (24)$$

$$U(s, a) = c_{puct} P(s, a) \frac{\sqrt{\varepsilon + \sum_{b \in A(s)} N(s, b)}}{1 + N(s, a)} \quad (25)$$

Here,  $W(s, a)$  is the accumulator for all backpropagated values that arrive along branch  $a$  of the node that carries state  $s$ . Likewise,  $N(s, a)$  is the visit counter and  $P(s, a)$  the prior probability.  $A(s)$  is the set of actions available in state  $s$ .  $\varepsilon$  is a small positive constant for the special case  $\sum_b N(s, b) = 0$ : It guarantees that in this special case the maximum of  $U(s, a)$  is given by the maximum of  $P(s, a)$ . The prior probabilities  $P(s, a)$  are obtained

---

**Algorithm 3** TD-n-tuple training algorithm. Input: see Algorithm 2. Output:  $\Theta$ : trained n-tuple network parameters.

---

```

1: function TDNTUPLETRAIN( $p_{max}$ ,  $M$ ,  $E_{train}$ ,  $c$ ,  $R_{pos}$ )
2:    $\Theta \leftarrow \text{INITIALIZENETWORKPARAMETERS}$ 
3:   INITIALIZETCLPARAMETERS       $\triangleright$  Set TCL-accumulators  $N_i = A_i = 0, \alpha_i = 1 \forall i$ 
4:   for  $m = 1, \dots, M$  do
5:     Perform one  $m$ -iteration of Algorithm 2 with learning rates  $\alpha\alpha_i$  instead of  $\alpha$ 
6:      $N_i \leftarrow N_i + \Delta\theta_i$  and  $A_i \leftarrow A_i + |\Delta\theta_i|$            $\triangleright$  Update TCL-accumulators
7:      $\triangleright$  where  $\Delta\theta_i$  is the last term in Eq. (20)
8:      $\alpha_i \leftarrow |N_i|/A_i \quad \forall i$  with  $A_i \neq 0$ 
9:   return  $\Theta$ 

```

---

**Algorithm 4** Evaluation algorithm with MCTS solver. Input: trained n-tuple network  $j_\Theta$ ,  $p$ : number of scrambling twists,  $B$ : batch size,  $E_{eval}$ : maximum episode length during evaluation,  $I$ : number of MCTS-iterations,  $c_{PUCT}$ : relative weight for  $U(s, a)$  in Eq. (24),  $d_{max}$ : maximum MCTS tree depth. Output: solved rate.

---

```

1: function TDNTUPLEVAL( $j_\Theta$ ,  $p$ ,  $B$ ,  $E_{eval}$ ,  $I$ ,  $c_{PUCT}$ ,  $d_{max}$ )
2:    $X \leftarrow \text{GENERATESCRAMBLEDCUBES}(B, p)$                        $\triangleright B$  scrambled cubes
3:    $C_{solved} \leftarrow 0$ 
4:   for  $x_i \in X$  do
5:      $s \leftarrow x_i$ 
6:     for  $k = 1, \dots, E_{eval}$  do
7:        $T \leftarrow \text{PERFORMMCTSSEARCH}(s, I, c_{PUCT}, d_{max}, j_\Theta)$ 
8:        $a \leftarrow \text{SELECTMOSTVISITEDACTION}$ 
9:        $s \leftarrow f(s, a)$ 
10:      if  $(s = s^*)$  then
11:         $C_{solved} \leftarrow C_{solved} + 1$ 
12:        break                                 $\triangleright$  break out of  $k$ -loop
13:   return  $C_{solved}/B$                    $\triangleright$  percentage solved

```

---

by sending the trained network's values of all follow-up states  $s' = f(s, a)$  with  $a \in A(s)$  through a softmax function (see Sec. 3).<sup>16</sup>

Once an unexpanded leaf node  $s_\ell$  is reached, the node is expanded by initializing its accumulators:  $W(s, a) = N(s, a) = 0$  and  $P(s, a) = p_{s'}$  where  $p_{s'}$  is the softmax-squashed output  $j_\Theta(s')$  of our n-tuple network for each state  $s' = f(s, a)$ . The value of the node is the network output of the best state  $j_\Theta(s_{best}) = \max_{s'} j_\Theta(s')$  and this value is backpropagated up the tree.

More details on our MCTS wrapper can be found in [Scheiermann and Konen \(2022\)](#).

---

<sup>16</sup>Note that the prior probabilities and the MCTS iteration are only needed at test time, so that we – different to AlphaZero – do not need MCTS during self-play training.

### 5.2.3 Method Summary

We summarize the different ingredients of our n-tuple-based TD learning method in Algorithm 3 (training) and Algorithm 4 (evaluation).

In line 5 of Algorithm 3 we perform one  $m$ -iteration of Algorithm 2 which does an update step for weight vector  $\Theta$ , see Eq. (20). All weights of activated n-tuple entries get a weight change  $\Delta\theta_i$  equal to the last term in Eq. (20) where the global  $\alpha$  is replaced by  $\alpha\alpha_i$ .

Line 2 in Algorithm 4 generates a set  $X$  of  $B$  scrambled cube states. Line 7 builds for each  $x_i \in X$  an MCTS tree (see Sec. 5.2.2) starting from root node  $x_i$  and line 8 selects the most visited action of the root node. If the goal state  $s^*$  is not found during  $E_{eval}$   $k$ -loop trials, this  $x_i$  is considered as not being solved.

## 6 Results

### 6.1 Experimental setup

We use for all our GBG experiments the same RL method based on n-tuple systems and TCL. Only its hyperparameters are tuned to the specific game, as shown below. We refer to this method/agent as **TCL-base** whenever it alone is used for game playing. If we wrap such an agent by an MCTS wrapper with a given number of iterations, then we refer to this as **TCL-wrap**.

We investigate two variants of Rubik’s Cube: 2x2x2 and 3x3x3. We trained all TCL agents by presenting them  $M = 3\,000\,000$  cubes scrambled with  $p$  random twists, where  $p$  is chosen uniformly at random from  $\{1, \dots, p_{max}\}$ . Here,  $p_{max} = 13$  [16] for 2x2x2 and  $p_{max} = 9$  [13] for 3x3x3, where the first number is for **HTM**, while the second number in square brackets is for **QTM**. With these  $p_{max}$  cube twists we cover the complete cube space for 2x2x2, where God’s number (Sec. 2.2) is known to be 11 [14]. But we cover only a small subset in the 3x3x3 case, where God’s number is known to be 20 [26] (Rokicki et al., 2014).<sup>17</sup> We train 3 agents for each cube variant  $\{2x2x2, 3x3x3\} \times \{\text{HTM, QTM}\}$  to assess the variability of training.

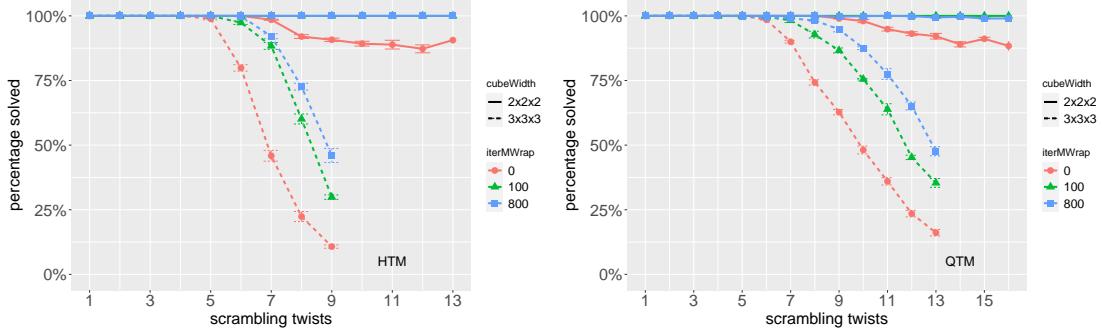
The hyperparameters of the agent for each cube variant were found by manual fine-tuning. For brevity, we defer the exact explanation and setting of all parameters to Appendix C.

We evaluate the trained agents for each  $p$  on 200 scrambled cubes that are created by applying the given number  $p$  of random scrambling twists to a solved cube. The agent now tries to solve each scrambled cube. A cube is said to be *unsolved* during evaluation if the agent cannot reach the solved cube in  $E_{eval} = 50$  steps.<sup>18</sup>

---

<sup>17</sup>We limit ourselves to  $p_{max} = 9$  [13] in the 3x3x3 **HTM** [**QTM**] case, because our network has not enough capacity to learn all states of the 3x3x3 Rubik’s cube. Experiments with higher twist numbers during training did not improve the solved-rates.

<sup>18</sup>During training, we use lower maximum episode lengths  $E_{train}$  (see Appendix C) than  $E_{eval} = 50$  in order to reduce computation time (in the beginning, many episodes cannot be solved, and 50 would waste a lot of computation time). But  $E_{train}$  is always at least  $p_{max} + 3$  in order to ensure that the agent has a fair chance to solve the cube and collect the reward.



**Figure 12: Percentage of solved cubes as a function of scrambling twists  $p$  for the trained TD-N-tuple agent wrapped by MCTS wrapper with different numbers of iterations. The red curves are *TCL-base* without wrapper, the other colors show different forms of *TCL-wrap*. Twist type is **HTM** (left) and **QTM** (right). Each point is the average of 3 independently trained agents.**

## 6.2 Cube Solving with MCTS Wrapper, without Symmetries

The trained TD-N-tuple agents learn to solve the cubes to some extent, as the red curves *TCL-base* in Fig. 12 show, but they are in many cases (i.e.  $p > p_{max}/2$ ) far from being perfect. These are the results from training each agent for 3 million episodes, but the results would not change considerably, if 10 million training episodes were used.

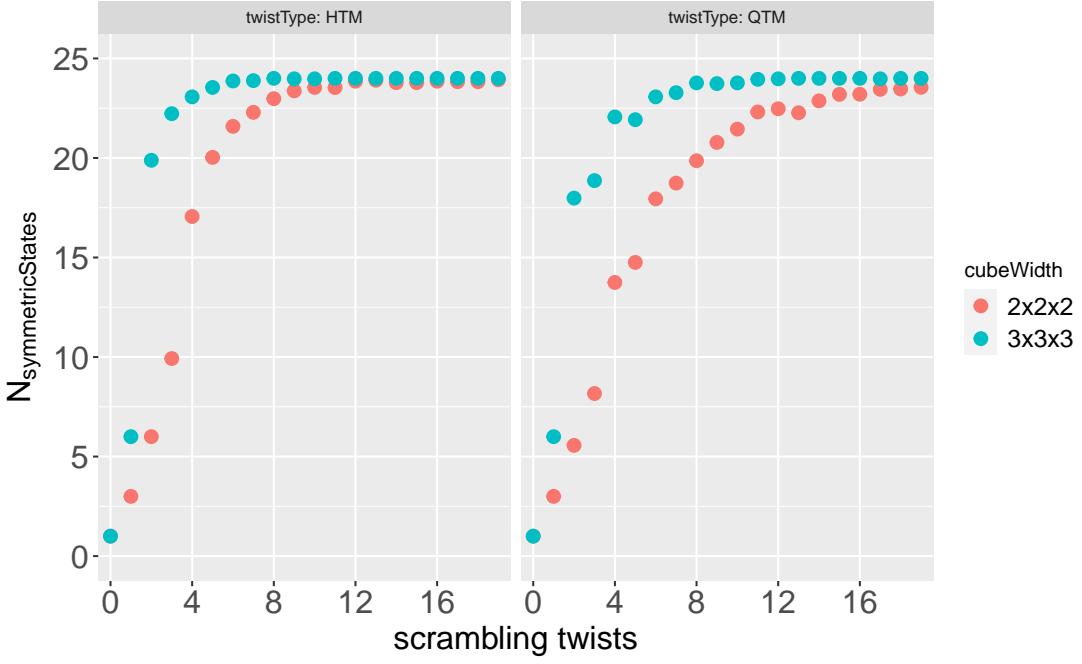
Scheiermann and Konen (2022) have shown, that the performance of agents, namely TD-N-tuple agents, is largely improved, if the trained agents are wrapped during test, play and evaluation by an MCTS wrapper. This holds for Rubik’s cube as well, as Fig. 12 shows: For the 2x2x2 cube, the non-wrapped agent *TCL-base* (red curve) is already quite good, but with wrapping it becomes almost perfect. For the 3x3x3 cube, the red curves are not satisfactorily: the solved-rates are below 20% for  $p = 9$  [13] in the **HTM** [**QTM**] case. But at least MCTS wrapping boosts the solved-rates by a factor of 3 [QTM: from 16% to 48%] or 4.5 [HTM: from 10% to 45%].

All these results are without incorporating symmetries. How symmetries affect the solved-rates will be investigated in Sec. 6.4. But before this, we look in the next section at the number of symmetries that effectively exist in a cube state.

## 6.3 Number of Symmetric States

Not every cube state has 24 truly different symmetric states ( $24 = \text{number of color symmetries}$ ). For example in the solved cube, all color-symmetric states are the same (after normalization). Thus, we have here only one truly different symmetric state.

However, we show in this section that for the majority of cube states the number of truly different symmetric states is close to 24. Two states are truly different if they are not the same after the **normalizing operation**. We generate a cube state by applying  $p$  random scrambling twists to the default cube. Now we apply all 24 color transformations (Sec. 2.4.3) to it and count the truly different states. The results are shown in Fig. 13 for



**Figure 13: Count of truly different symmetric states for cube states generated by  $p$  random scrambling twists. Each point is an average over 500 such states.**

both cube sizes and both twist types. For the 3x3x3 cube, the number of states quickly (for  $p > 5$ ) approaches the maximum  $N = 24$ , while for the 2x2x2 cube it is a bit slower:  $p > 4$  or  $p > 8$  is needed to surpass  $N = 20$ .

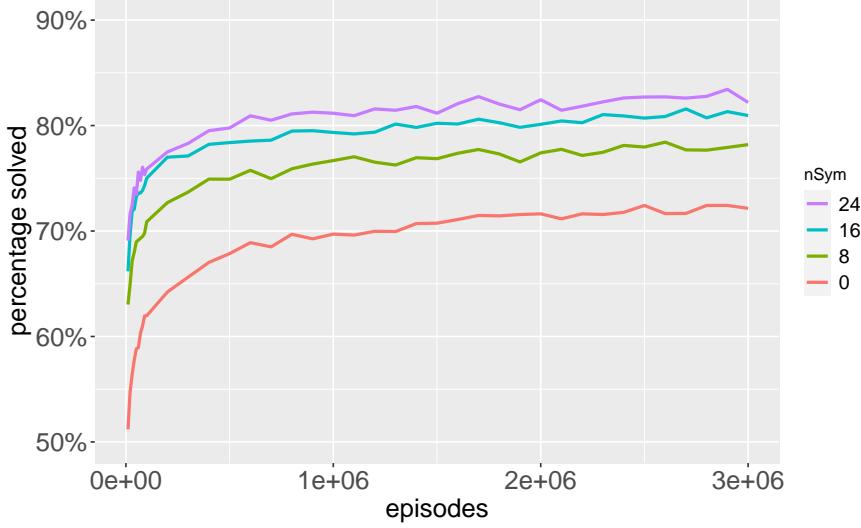
As a consequence, it makes sense to use 16 or even 24 symmetries when training and evaluating cube agents. Especially for scrambled states with higher  $p$ , the 24 color transformations used to construct symmetric states will usually lead to 24 different states.

#### 6.4 The Benefit of Symmetries

In order to investigate the benefits of symmetries, we first train a TCL agent with different numbers of symmetries. As described in Sec. 2.5, we select in each step  $n_{\text{Sym}} = 0, 8, 16, 24$  symmetric states. Which symmetric states are chosen is selected randomly. Symmetries are used (a) to update the weights for each symmetric state and (b) to build with Eq. (18) a smoothed value function which is used to decide about the next action during training. For 0, 8, 16, 24 symmetries, we train 3 agents each (3x3x3 cube, STICKER2, QTM). The 3 agents differ due to their differently created random-walk n-tuple sets.

Fig. 14 shows the learning curves for different  $n_{\text{Sym}} = 0, 8, 16, 24$ . It is found that agents with  $n_{\text{Sym}} > 0$  learn faster and achieve a higher asymptotic solved rate.

Next, we evaluate each of the trained agents by trying to solve for each  $p \in \{1, \dots, 15\}$  (scrambling twists) 200 different scrambled cubes. During evaluation, we use again the same  $n_{\text{Sym}}$  as in training to form a smoothed value function. We compare in Fig. 15 different



**Figure 14: Learning curves for different numbers  $n_{\text{Sym}} = 0, 8, 16, 24$  of symmetries.** Shown is the solved rate of (3x3x3, QTM) cubes. The solved rate is the average over all twist numbers  $p = 1, \dots, 13$  with 200 testing cubes for each  $p$  and over 3 agents with different random-walk  $n$ -tuple sets.

symmetry results, both without wrapping (*TCL-base*, red curves) and with MCTS-wrapped agents using 100 (green) or 800 (blue) iterations. It is clearly visible that MCTS wrapping has a large effect, as it was also the case in Fig 12. But in addition to that, the use of symmetries leads for each agent, wrapped or not, to a substantial increase in solved-rates (a surplus of 10-20%). It is remarkable, that even for  $p=14$  or  $15$  a solved rate above or near 50% can be reached<sup>19</sup> by the combination ( $n_{\text{Sym}}=16$ , 800 MCTS iterations).

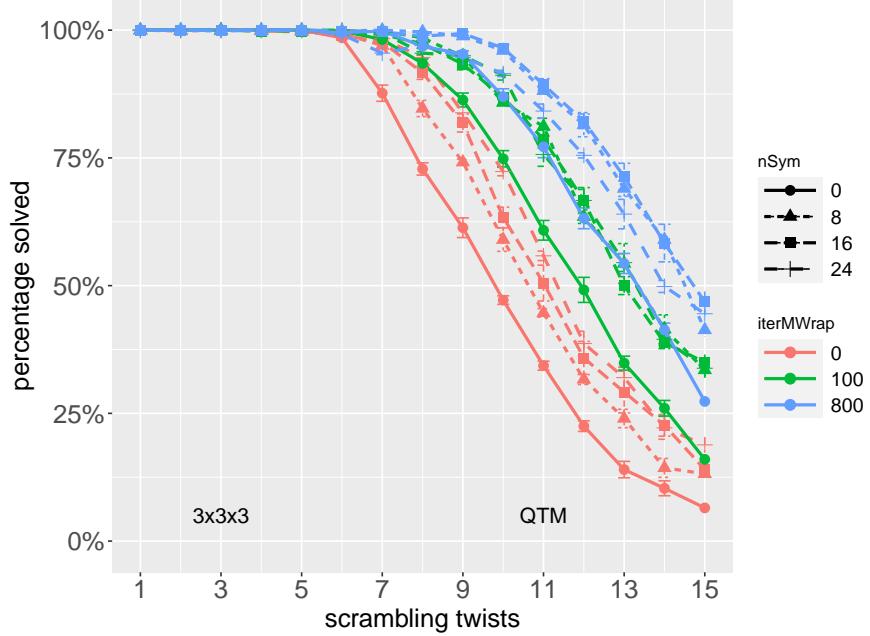
Surprisingly, it seems that *with wrapping* it is only important whether we use symmetries, not how many, since the difference between  $n_{\text{Sym}} = 8, 16, 24$  is only marginal. For 800 MCTS iterations, the solved rate for  $n_{\text{Sym}} = 24$  is in most cases even smaller than that for  $n_{\text{Sym}} = 8, 16$ . This is surprising because it would have been expected that also with wrapping a larger  $n_{\text{Sym}}$  should lead to a smoother value function and thus should in theory produce larger solved rates. – Note that this is not a contradiction to Fig. 14, because the learning curves were obtained *without wrapping* and the red *TCL-base* curves in Fig. 15 (again without wrapping) show the same positive trend with increasing  $n_{\text{Sym}}$ <sup>20</sup>. The red curves in Fig. 15 show approximately the same average solved rates as the asymptotic values in Fig. 14.

## 6.5 Computational Costs

Table 9 shows the computational costs when training and testing with symmetries. All computations were done on a single CPU Intel i7-9850H @ 2.60GHz. If we subtract the

<sup>19</sup>  $p$  is above  $p_{\max}=13$ , the maximum twist number used during training.

<sup>20</sup> i.e.  $n_{\text{Sym}}= 24$  is for every  $p$  clearly better than  $n_{\text{Sym}}= 16$



**Figure 15: With symmetries:** Percentage of solved cubes (3x3x3, QTM) as a function of scrambling twists  $p$  for TD-N-tuple agents trained and evaluated with different numbers of symmetries  $n_{\text{Sym}}$  and wrapped by MCTS wrappers with different iterations. The red curves are *TCL-base* (without wrapper), the other colors show different forms of *TCL-wrap*. The solved rates are the average over 200 testing cubes for each  $p$  and over 3 agents with different random-walk n-tuple sets.

computational costs for  $n_{\text{sym}} = 0$ , computation time increases more or less linearly with  $\text{iter}$  and roughly linearly with  $n_{\text{Sym}}$ . Computation times for  $n_{\text{Sym}} = 24$  are approximately 10x larger than those for  $n_{\text{Sym}} = 0$ .

Computation times are dependent on the solved rate: If a cube with  $p = 13$  is solved, the episode takes normally 12-15 steps. If the cube is not solved, the episode needs 50 steps, i.e. a factor of 3-4 more. Thus, the numbers in Table 9 should be taken only as rough indication of the trend.

Bottom line: Training time through symmetries increases by a factor of  $13/0.5 = 26$  ( $n_{\text{Sym}} = 24$ ) and testing time increases through 800 MCTS iterations by a factor of about  $3130/8 \approx 400$ .

Training with symmetries takes between 5.4h and 13h on a normal CPU, depending on the number of symmetries. This is much less than the 44h on a 32-core server with 3 GPUs that were used by McAleer et al. (2019). But it also does not reach the same quality as McAleer et al. (2019).

**Table 9: Computation times with symmetries.** All numbers are for 3x3x3 cube, STICKER2 and QTM. Training: 3 million self-play episodes, w/o MCTS in the training loop. Testing: 200 scrambled cubes with  $p = 13$ , agents wrapped by MCTS wrapper with  $\text{iter}$  iterations.

nSym	training [hours]	testing [seconds]				
		iter	0	100	400	800
0	0.5		0.5	48	196	390
8	5.4		4.0	241	877	1400
16	9.5		7.3	464	1380	2330
24	13.0		8.0	550	1760	3130

## 7 Related Work

Ernö Rubik invented Rubik’s cube in 1974. Rubik’s cube has gained worldwide popularity with many human-oriented algorithms being developed to solve the cube from arbitrary scrambled start states. By ‘human-oriented’ we mean algorithms that are simple to memorize for humans. They usually will find long, suboptimal solutions. For a long time it was an unsolved question what is the minimal number of moves (**God’s Number**) needed to solve any given cube state. The early work of Thistlethwaite (1981) put an upper bound on this number with his 52-move algorithm. This was one of the first works to systematically use group theory as an aid to solve Rubik’s cube. Later, several authors have gradually reduced the upper bound 52 (Joyner, 2014), until Rokicki et al. (2014) could prove in 2014 for the 3x3x3 cube that **God’s Number** is 20 in HTM and 26 in QTM.

Computer algorithms to solve Rubik’s cube rely often on hand-engineered features and group theory. One popular solver for Rubik’s cube is the two-phase algorithm of Kociemba (2015). A variant of A\* heuristic search was used by Korf (1991), along with a pattern database heuristic, to find the shortest possible solutions.

The problem of letting a computer *learn* to solve Rubik’s cube turned out to be much harder: Irpan (2016) experimented with different neural net baseline architectures (LSTM gave for him reportedly best results) and tried to boost them with AdaBoost. However, he had only for scrambling twist  $\leq 7$  solved rates of better than 50% and the baseline turned out to be better than the boosted variants. Brunetto and Trunda (2017) found somewhat better results with a DNN, they could solve cube states with 18 twists with a rate above 50%. But they did not learn from scratch because they used an optimal solver based on Kociemba (2015) to generate training examples for the DNN. Smith et al. (2016) tried to learn Rubik’s cube by genetic programming. However, their learned solver could only reliably solve cubes with up to 5 scrambling twists.

A breakthrough in learning to solve Rubik’s cube are the works of McAleer et al. (2018, 2019) and Agostinelli et al. (2019): With Autodidactic Iteration (ADI) and Deep Approximate Value Iteration (DAVI) they were able to *learn from scratch* to solve Rubik’s cube in QTM for arbitrary scrambling twists. Their method has been explained in detail already in Sec. 5.1, so we highlight here only their important findings: McAleer et al. (2019) only needs to inspect less than 4000 cubes with its trained network DeepCube when solving

for a particular cube, while the optimal solver of Korf (1991) inspects 122 billion different nodes, so Korf’s method is much slower.

Agostinelli et al. (2019) extended the work of McAleer et al. (2019) by replacing the MCTS solver with a batch-weighted A\* solver which is found to produce shorter solution paths and have shorter run times. At the same time, Agostinelli et al. (2019) applied their agent DeepCubeA successfully to other puzzles like LightsOut, Sokoban, and the 15-, 24-, 35- and 48-puzzle<sup>21</sup>. DeepCubeA could solve all of them.

The deep network used by McAleer et al. (2019) and Agostinelli et al. (2019) were trained without human knowledge or supervised input from computerized solvers. The network of McAleer et al. (2019) had over 12 million weights and was trained for 44 hours on a 32-core server with 3 GPUs. The network of McAleer et al. (2019) has seen 8 billion cubes during training. – Our approach started from scratch as well. It required much less computational effort (e.g. 5.4h training time on a single standard CPU for nSym=8, see Table 9). It can solve the 2x2x2 cube completely, but the 3x3x3 cube only partly (up to 15 scrambling twists). Each trained agent for the 3x3x3 cube has seen 48 million scrambled cubes<sup>22</sup> during training.

## 8 Summary and Outlook

We have presented new work on how to solve Rubik’s cube with n-tuple systems, reinforcement learning and an MCTS solver. The main ideas were already presented in Scheiermann and Konen (2022) but only for HTM and up to  $p = 9$  twists. Here we extended this work to QTM as well and presented all the details of cube representation and n-tuple learning algorithms necessary to reproduce our Rubik’s cube results. As a new aspect, we added cube symmetries and studied their effect on solution quality. We found that the use of symmetries boosts the solved rates by 10-20%. Based on this, we could increase for QTM the number of scrambling twists where at least 45% of the cubes are solved from  $p = 13$  without symmetries to  $p = 15$  with symmetries.

We cannot solve the 3x3x3 cube completely, as McAleer et al. (2019) and Agostinelli et al. (2019) do. But our solution is much less computational demanding than their approach.

Further work might be to look into larger or differently structured n-tuple systems, perhaps utilizing the staging principle that Jaśkowski (2018) used to produce world-record results in the game 2048.

---

<sup>21</sup>a set of 15, 24, ... numbers has to be ordered on a  $4 \times 4$ ,  $5 \times 5$ , ... square with one empty field

<sup>22</sup> $3 \cdot 10^6 \times 16 =$  training episodes  $\times$  episode length  $E_{train}$ . This is an upper bound: some episodes may have shorter length, but each unsolved episode has length  $E_{train}$ .

## References

- F. Agostinelli, S. McAleer, A. Shmakov, and P. Baldi. Solving the Rubik’s cube with deep reinforcement learning and search. *Nature Machine Intelligence*, 1(8):356–363, 2019. [1](#), [4](#), [5](#), [21](#), [22](#), [32](#), [33](#)
- S. Bagheri, M. Thill, P. Koch, and W. Konen. Online adaptable learning rates for the game Connect-4. *IEEE Transactions on Computational Intelligence and AI in Games*, 8(1):33–42, 2015. [25](#)
- D. F. Beal and M. C. Smith. Temporal coherence and prediction decay in TD learning. In T. Dean, editor, *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 564–569. Morgan Kaufmann, 1999. ISBN 1-55860-613-0. [25](#)
- W. W. Bledsoe and I. Browning. Pattern recognition and reading by machine. In *Proceedings of the Eastern Joint Computer Conference*, pages 225–232, 1959. [16](#)
- C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012. [25](#)
- R. Brunetto and O. Trunda. Deep heuristic-learning in the Rubik’s cube domain: An experimental evaluation. In *ITAT*, pages 57–64, 2017. URL <http://ceur-ws.org/Vol-1885/57.pdf>. [32](#)
- A. Irpan. Exploring boosted neural nets for Rubik’s cube solving. Technical report, University of California, 2016. URL [https://www.alexirpan.com/public/research/nips\\_2016.pdf](https://www.alexirpan.com/public/research/nips_2016.pdf). [32](#)
- W. Jaśkowski. Mastering 2048 with delayed temporal coherence learning, multistage weight promotion, redundant encoding, and carousel shaping. *IEEE Transactions on Games*, 10(1):3–14, 2018. [33](#)
- D. Joyner. The man who found God’s number. *The College Mathematics Journal*, 45(4):258–266, 2014. [32](#)
- H. Kociemba. The two-phase-algorithm, 2015. URL <http://kociemba.org/twophase.htm>. Details in <http://kociemba.org/math/imptwophase.htm>, retrieved Sep-01-2022. [32](#)
- W. Konen. General board game playing for education and research in generic AI game learning. In D. Perez, S. Mostaghim, and S. Lucas, editors, *Conference on Games (London)*, pages 1–8, 2019. URL <https://arxiv.org/pdf/1907.06508.pdf>. [5](#)
- W. Konen. The GBG class interface tutorial V2.3: General board game playing and learning. Technical report, TH Köln, 2022. URL <http://www.gm.fh-koeln.de/ciopwebpub/Konen22a.d/TR-GBG.pdf>. [5](#), [42](#)

- W. Konen and S. Bagheri. Reinforcement learning for n-player games: The importance of final adaptation. In *9th International Conference on Bioinspired Optimisation Methods and Their Applications (BIOMA)*, Nov. 2020. URL <http://www.gm.fh-koeln.de/ciopwebpub/Konen20b.d/bioma20-TDNTuple.pdf>. 5
- W. Konen and S. Bagheri. Final adaptation reinforcement learning for n-player games. *arXiv preprint arXiv:2111.14375*, 2021. 24, 40, 41
- R. E. Korf. Multi-player alpha-beta pruning. *Artificial Intelligence*, 48(1):99–111, 1991. 32, 33
- S. M. Lucas. Learning to play Othello with n-tuple systems. *Australian Journal of Intelligent Information Processing*, 4:1–20, 2008. 16
- S. McAleer, F. Agostinelli, A. Shmakov, and P. Baldi. Solving the rubik’s cube without human knowledge. *arXiv preprint arXiv:1805.07470*, 2018. 1, 4, 21, 32
- S. McAleer, F. Agostinelli, A. Shmakov, and P. Baldi. Solving the Rubik’s cube with approximate policy iteration. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/pdf?id=Hyfn2jCcKm>. 1, 4, 5, 18, 19, 21, 22, 25, 31, 32, 33, 39
- V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. 4
- T. Rokicki, H. Kociemba, M. Davidson, and J. Dethridge. The diameter of the Rubik’s Cube group is twenty. *siam REVIEW*, 56(4):645–670, 2014. 7, 27, 32
- J. Scheiermann and W. Konen. AlphaZero-inspired game learning: Faster training by using MCTS only at test time. *IEEE Transactions on Games*, 2022. doi: 10.1109/TG.2022.3206733. URL <https://ieeexplore.ieee.org/document/9893320>. 5, 26, 28, 33, 41
- D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 4
- D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017. 22, 25
- R. J. Smith, S. Kelly, and M. I. Heywood. Discovering Rubik’s cube subgroups using coevolutionary GP: A five twist experiment. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 789–796, 2016. 32
- R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998. 17

M. Thistlethwaite. Thistlethwaites's 52-move algorithm, 1981. URL <https://www.jaapsch.net/puzzles/thistle.htm>. Reconstructed by Jaap Scherphuis, retrieved Sep-01-2022. 32

Wikipedia. Pocket Cube, 2022a. URL [https://en.wikipedia.org/wiki/Pocket\\_Cube](https://en.wikipedia.org/wiki/Pocket_Cube). retrieved Aug-17-2022. 7

Wikipedia. Rubik's Cube, 2022b. URL [https://en.wikipedia.org/wiki/Rubik's\\_Cube](https://en.wikipedia.org/wiki/Rubik's_Cube). retrieved Aug-17-2022. 7

# Appendix

## A Calculating $s_{\text{loc}}$ from $f_{\text{col}}$

Given the face colors  $f_c$  (Eq. (3)) of a transformed cube, how can we calculate the transformed sticker locations  $s_\ell$  (Eq. (4))?

This problem seems ill-posed at first sight, because a certain face color, e.g. *white*, appears multiple times in  $f_c$  and it is not possible to tell from the appearance of *white* alone to which sticker location  $s_\ell$  it corresponds. But with a little more effort, i.e. by looking at the neighbors of the *white* sticker, we can solve the problem, as we show in the following.

### A.1 2x2x2 cube

All **cubies** of the 2x2x2 cube are corner cubies. We track for each cubie exactly one **sticker**. This can be for example the set

$$\mathfrak{B} = \{0, 1, 2, 3, 12, 13, 14, 15\}$$

of 8 stickers, which is the same as the set of tracked stickers shown in Fig. 11.

For each  $s \in \mathfrak{B}$ :

1. Build the cubie that contains  $s$  as the first sticker.<sup>23</sup>
2. Locate the cubie in  $f_c$ . That is, find a location in  $f_c$  with the same color as the 1<sup>st</sup> cubie face. If found, check if the neighbor to the right<sup>24</sup> has the color of the 2<sup>nd</sup> cubie face. If yes, check if its neighbor to the right has the color of the 3<sup>rd</sup> cubie face. If yes, we have located the cubie in  $f_c$  and we return it, i.e. its three sticker locations  $C = [a, b, c]$ .
3. Having located the cubie, we can infer three elements of  $s_\ell$ :

$$s_\ell[s] = C[0] \tag{26}$$

$$s_\ell[R[s]] = C[1] \tag{27}$$

$$s_\ell[R[R[s]]] = C[2] \tag{28}$$

Here  $R[s]$  is the right neighbor of sticker  $s$ .  $R[R[s]]$  is the left neighbor.

In total, we have located  $8 \times 3 = 24$  stickers, i.e. the whole transformation for  $s_\ell$ .<sup>25</sup>

---

<sup>23</sup>We know for example from looking at the default cube in Fig. 11 that sticker  $s = 0$  is part of the 0-8-4-cubie.

<sup>24</sup>By *neighbor to the right* we mean the next sticker when we march in clockwise orientation around the actual cubie.

<sup>25</sup>The relevant GBG source code is in CubeState.locate and CubeState2x2.apply\_sloc\_slow.

## A.2 3x3x3 cube

The 3x3x3 cube has 8 corner **cubies** and 12 edge cubies. We track for each cubie exactly one **sticker**. This can be for the corners the set

$$\mathfrak{B} = \{0, 2, 4, 6, 24, 26, 28, 30\}$$

and for the edges the set

$$\mathfrak{E} = \{1, 3, 5, 7, 25, 27, 29, 31, 11, 15, 21, 33\}.$$

We do for the corner set  $\mathfrak{B}$  the same as we did for the 2x2x2 cube.

For each element  $s \in \mathfrak{E}$  of the edge set:

1. Build the edge cubie  $c_E$  that contains  $s$  as the first sticker.
2. Locate the cubie in  $f_c$ . That is, find an edge location in  $f_c$  with the same color as the 1<sup>st</sup> cubie face. If found, check if the other sticker of that cubie has the same color as the other sticker of  $c_E$ . If yes, we have located the edge cubie in  $f_c$  and we return it, i.e. its two stickers  $C = [a, b]$ .
3. Having located the cubie, we can infer two elements of  $s_\ell$ :

$$s_\ell[s] = C[0] \tag{29}$$

$$s_\ell[O[s]] = C[1] \tag{30}$$

Here  $O[s]$  is the other sticker of the edge cubie that has sticker  $s$  as first sticker.

In total, we have located

$$8 \times 3 + 12 \times 2 = 48$$

stickers, i.e. the whole transformation for  $s_\ell$ .<sup>26</sup>

## B N-Tuple Representations for the 3x3x3 Cube

In this appendix we describe the n-tuple representations of the cube, analogously to the 2x2x2 cube Sec. 4, but now for the 3x3x3 cube.

### B.1 CUBESTATE

A natural way to translate the cube state into a board is to use the flattened representation of Fig. 4 as the board and extract from it the 48-element vector  $\mathbf{b}$ , according to the given numbering. The  $k$ th element  $b_k$  represents a certain cubie face location and gets a number from  $\{0, \dots, 5\}$  according to its current face color  $f_c$ . The solved cube is for example represented by  $\mathbf{b} = [00000000 11111111 \dots 55555555]$ .

This representation CUBESTATE is what the BoardVecType CUBESTATE in our GBG-implementation means: Each board vector is a copy of  $\mathbf{f}_{\text{col}}$ , the face colors of all cubie faces. An upper bound of possible combinations for  $\mathbf{b}$  is  $6^{48} = 2.2 \cdot 10^{32}$ . This is much larger than the true number of distinct states (Sec. 2.2.2) which is  $4.3 \cdot 10^{19}$ .

---

<sup>26</sup>The relevant GBG source code is in `CubeState.locate`, `CubeState3x3.locate_edge` and `CubeState3x3.apply_sloc_slow`.

**Table 10: The correspondence edge location  $\leftrightarrow$  STICKER2 for the solved cube. The yellow colored cells show the location of the 12 edge stickers that we track.**

3x3x3	location	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	37	39	41	43	45	47
STICKER2	edge	A	B	C	D	D	G	K	E	E	J	F	A	I	J	K	L	H	B	F	I	L	G	C	H
	face ID	1	1	1	1	2	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	1	1	1	1

## B.2 STICKER

McAleer et al. (2019) had the interesting idea for the 3x3x3 cube that 20 stickers (cubie faces) are enough. To characterize the 3x3x3 cube, we need according to McAleer et al. (2019) only one (not 2 or 3) sticker for every of the 20 cubies, as shown in Fig. 10. This is because the location of one sticker uniquely defines the location and orientation of that cubie. We name this representation STICKER in GBG.

We track the 4 top corner stickers 0,2,4,6 plus the 4 bottom corner stickers 24,26,28,30 plus one sticker for each of the 12 edge stickers as shown in Fig. 10, in total 20 stickers and ignore the 28 other stickers.

How to lay out this representation as a board? – McAleer et al. (2019) create a rectangular one-hot-encoding board with  $20 \times 24 = 480$  cells (20 rows for the stickers and 24 columns for the locations<sup>27</sup>) carrying only 0's and 1's. This is fine for the approach of McAleer et al. (2019), where they use this board as input for a DNN, but not so nice for n-tuples. Without constraints, such a board amounts to  $2^{480} \approx 10^{145}$  combinations, which is unpleasantly large (much larger than in CUBESTATE).<sup>28</sup>

Another possibility to lay out the board: Specify 20 board cells (the stickers) with 24 position values each. This amounts to  $24^{20} = 4.0 \cdot 10^{27}$  combinations.

## B.3 STICKER2

Analogously to Sec. 4.3, we represent the 24 corner locations and 24 edge locations as:

$$\text{corner location} = (\text{corner cubie}, \text{face ID}),$$

$$\text{edge location} = (\text{edge cubie}, \text{face ID}).$$

That is, each corner location is represented by a corner cubie a,b,c,d,e,f,g,h and by a face ID 1,2,3. Table 7 shows the explicit numbering in this new representation. Additionally, each edge location is represented by an edge cubie A,B,C,D,E,F,G,H,I,J,K,L<sup>29</sup> and by a face ID 1,2. Convention for face ID numbering of edge cubies: For top- and bottom-layer edge cubies, it is 1 for U and D stickers, 2 else. The face ID for middle-layer edge cubies is 1 for F and B stickers, 2 else. Table 10 shows the explicit numbering in this representation.

The corresponding board consists of  $8 + 8 + 12 + 12 = 40$  cells shown in Table 11. The 8 cell pairs in the first two rows code the locations of the tracked corner stickers

<sup>27</sup>8 · 3 for the corner stickers and 12 · 2 for the edge stickers

<sup>28</sup>McAleer et al. (2019) do not need a weight for every of the  $2^{480}$  possible states, as the n-tuple network would need. Instead they need only  $480 \cdot 4096 = 2 \cdot 10^6$  weights to the first hidden layer having 4096 neurons.

<sup>29</sup>4 U-stickers, 4 D-sticker, 4 middle-layer stickers (2F, 2B)

0,2,4,6,24,26,28,30, see Table 7 in Sec. 4.3. The 12 cell pairs in the last two rows code the location of the tracked edge stickers 1,3,5,7,17,21,43,47,25,27,29,31, see Table 10. This n-tuple coding requires tuple cells with varying number of position values and leads to

$$8^8 \cdot 3^8 \cdot 12^{12} \cdot 2^{12} = 4.0 \cdot 10^{27}$$

combinations in representation STICKER2.<sup>30</sup>

**Table 11: STICKER2 board representation for the default 3x3x3 cube. For the `BoardVector`, cells are numbered row-by-row from 0 to 39.**

corner	a	b	c	d	e	f	g	h		8 positions
face ID	1	1	1	1	1	1	1	1		3 positions
edge	A	B	C	D	E	F	G	H	I	12 positions
face ID	1	1	1	1	1	1	1	1	1	2 positions

## B.4 Adjacency Sets

To create n-tuples by random walk, we need to define adjacency sets (sets of neighbors) for every board cell  $k$ .

For CUBESTATE, the board is the flattened representation of the 3x3x3 cube (Fig. 4). The adjacency set is defined as the 4-point neighborhood, where two stickers are neighbors if they are neighbors (share a common edge) on the cube.

For STICKER2, the board consists of 40 cells shown in Table 11. Since it matters for the corner stickers mostly where the other corner stickers are and for the edge stickers mostly where the other edge stickers are, it is reasonable to form two adjacency subsets  $S_1 = \{00, \dots, 15\}$  and  $S_2 = \{16, \dots, 39\}$  and to define the adjacency set

$$\text{Adj}(k) = S_i \setminus \{k\}$$

for each  $k \in S_i$ ,  $i = 1, 2$ .

## C Hyperparameters

In this appendix we list all parameter settings for the GBG agents used in this paper. Parameters were manually tuned with two goals in mind: (a) to reach high-quality results and (b) to reach stable (robust) performance when conducting multiple training runs with different random seeds. The agents listed further down are the best-so-far agents found (best among all agents that learn from scratch by self-play).

The detailed meaning of RL parameters is explained in [Konen and Bagheri \(2021\)](#):

<sup>30</sup>This is, by the way, identical to  $(8 \cdot 3)^8 \cdot (12 \cdot 2)^{12} = 24^{(8+12)} = 24^{20} = 4.0 \cdot 10^{27}$ , the same number we had above in the second mode of STICKER. But STICKER2 has the advantage that the combinations are spread over more board cells (40) than in STICKER (20). By having more board cells with fewer position values, the n-tuples can better represent the relationships between cube states.

- Algorithms 2, 5 and 7 in [Konen and Bagheri \(2021\)](#) explain parameters  $\alpha$  (learning rate),  $\gamma$  (discount factor),  $\epsilon$  (exploration rate) and output sigmoid  $\sigma$  (either identity or tanh).
- Appendix A.3 explains our eligibility method, parameters are: eligibility trace factor  $\lambda$ , horizon cut  $c_h$ , eligibility trace type ET (normal) or RESET (reset on random move). If not otherwise stated, we use in this paper  $\lambda = 0$  (no eligibility traces). For  $\lambda > 0$ , horizon cut  $c_h$  and eligibility trace type are irrelevant. If  $\lambda > 0$ , their defaults  $c_h = 0.1$  and trace type ET apply.
- Appendix A.5 explains our TCL method (also summarized in Sec. [5.2.1](#)). Parameters of TCL are: TC-Init (initialization constant for counters), TC transfer function (TC-id or TC-EXP),  $\beta$  (exponential factor in case of TC-EXP), TC accumulation type (delta or recommended weight-change).

Another branch of our algorithm is the MCTS wrapper, which can be used to wrap TD-N-tuple agents during evaluation and testing. MCTS wrapping is briefly explained in Sec. [5.2.2](#). The precise algorithm for MCTS wrapping is explained in detail in ([Scheiermann and Konen, 2022](#), Sec. II-B).<sup>[31](#)</sup> Parameters of MCTS are:

- $c_{PUCT}$ : relative weight for the prior probabilities of the wrapped agent in relation to the value that the wrapper estimates
- $d_{max}$ : maximum depth of the MCTS tree, if -1: no maximum depth
- UseSoftMax: boolean, whether to use SoftMax normalization for the priors or not
- UseLastMCTS: boolean, whether to re-use the MCTS from the previous move within an episode or not

Further parameter explanations:

- Sec. [4](#) in this document explains n-tuples, parameters are: number of n-tuples, length of n-tuples, and n-tuple creation mode (fixed, random walk, random points).
- Sec. [2.5](#) in this document explains symmetries. If parameter  $nSym = 0$ , do not use symmetries. If  $nSym > 0$ , use this number  $nSym$  of symmetries. In the Rubik's cube case,  $nSym$  is a number between 0 and 24.
- LearnFromRM: whether to learn from random moves or not. (Does not apply here, because we use in Rubik's cube always  $\epsilon = 0$ , i.e. we have no random moves.)
- ChooseStart-01: whether to start episodes from different 1-ply start states or always from the default start state. (Does not apply here, because we start in Rubik's cube never from the default cube, but always from the  $p$ -twisted cube.)

---

<sup>[31](#)</sup>As ([Scheiermann and Konen, 2022](#), Sec. IV-E) shows, the MCTS wrapper may be used as well during training, but due to large computation times needed for this, we do not follow that route in this paper.

- $E_{train}$ : maximum episode length during training, if -1: no maximum length.
- $E_{eval}$ : maximum episode length during evaluation and play, if -1: no maximum length.

All agents were trained with no MCTS wrapper inside the training loop. The hyper-parameters of the agent for each cube variant were found by manual fine-tuning. See also (Konen, 2022).

In the following, we list the precise settings for all agents used in this paper. If not stated otherwise, these common settings apply to all agents: sigmoid  $\sigma = id$ , LearnFromRM = false, ChooseStart-01 = false. Wrapper settings during test and evaluation: MCTS wrapper with  $c_{PUCT} = 1.0$ ,  $d_{max} = 50$ , UseSoftMax = true, UseLastMCTS = true.

Common parameters of Algorithm 2 in Sec. 5.2 are: cost-to-go  $c = -0.1$  and positive reward  $R_{pos} = 1.0$ .

The parameters for training without symmetries ( $nSym = 0$ ) in Sec. 6.2 are:

- **2x2x2 cube, HTM**:  $\alpha = 0.25$ ,  $\gamma = 1.0$ ,  $\epsilon = 0.0$ ,  $\lambda = 0.0$ , no output sigmoid. N-tuples: 60 7-tuples created by random walk. TCL activated with transfer function TC-id, TC-Init=  $10^{-4}$  and rec-weight-change accumulation. 3,000,000 training episodes.  $p_{max} = 13$ ,  $E_{train} = 16$ ,  $E_{eval} = 50$ .  
Agent filename in GBG: 2x2x2\_STICKER2\_AT/TCL4-p13-ET16-3000k-60-7t-stub.agt.zip
- **2x2x2 cube, QTM**: same as 2x2x2 cube, HTM, but with  $p_{max} = 16$ ,  $E_{train} = 20$ .  
Agent filename in GBG: 2x2x2\_STICKER2\_QT/TCL4-p16-ET20-3000k-60-7t-stub.agt.zip
- **3x3x3 cube, HTM**: same as 2x2x2 cube, HTM, but with 120 7-tuples created by random walk,  $p_{max} = 9$ ,  $E_{train} = 13$ .  
Agent filename in GBG: 3x3x3\_STICKER2\_AT/TCL4-p9-ET13-3000k-120-7t-stub.agt.zip
- **3x3x3 cube, QTM**: same as 3x3x3 cube, HTM, but with  $p_{max} = 13$ ,  $E_{train} = 16$ .  
Agent filename in GBG: 3x3x3\_STICKER2\_QT/TCL4-p13-ET16-3000k-120-7t-stub.agt.zip

The agent files given in the list above are just stubs, i.e. agents that are initialized with the correct parameters but not yet trained. This is because a trained agent can require up to 80 MB disk space, which is too much for GitHub. Instead, a user of GBG may load such a stub agent, train it (takes between 10-40 minutes) and save it to local disk.

When evaluating in Sec. 6.2 the trained agents with different MCTS wrappers, we test in each case whether  $c_{PUCT} = 1.0$  or 10 is better. In most cases,  $c_{PUCT} = 1.0$  is better, but for (2x2x2, QTM, 800 iterations) and for (3x3x3, HTM, 100 iterations)  $c_{PUCT} = 10.0$  is the better choice.

The parameters for training with symmetries ( $nSym = 8, 16, 24$ ) in Sec. 6.4 are:

- **3x3x3 cube, QTM**: same as 3x3x3 cube, QTM in Sec. 6.2, but with  $n_{sym} = 8, 16, 24$ .  
Agent filename in GBG: 3x3x3\_STICKER2\_QT/TCL4-p13-ET16-3000k-120-7t-nsym08-stub.agt.zip,  
3x3x3\_STICKER2\_QT/TCL4-p13-ET16-3000k-120-7t-nsym16-stub.agt.zip,  
3x3x3\_STICKER2\_QT/TCL4-p13-ET16-3000k-120-7t-nsym24-stub.agt.zip.

Again, the agent filenames are just stubs, i.e. agents that are initialized with the correct parameters but not yet trained. As above, a user of GBG may load such a stub agent, train it (which takes in the symmetry case between 5.4h and 13h, see Table 9) and save it to local disk.

For further details and experiment shell scripts, see also the associated **Papers-with-Code repository** <https://github.com/WolfgangKonen/PapersWithCodeRubiks>.