

Baby Tetris First Part Report

ERGIN Seçkin Yağmur and GROUSSON Dylan

January 8, 2026

<https://github.com/diarca/baby-tetris/>

Contents

1 Question 1: Discounted MDP For Player	1
1.1 Finding Optimal Policy	2
1.2 Implementation	2
1.2.1 Main Architecture	2
1.2.2 Implementation Choices	3
1.2.3 Value Iteration Implementation For The Player	4
1.3 Results	4
1.3.1 Executions	5
1.3.2 Observations	6
1.4 Reachable States	6
2 Question 2: Adversarial MDP For Opponent	8
2.0.1 Choosing the Tromino Piece	8
2.1 Executions	9
2.1.1 Observations	9
3 Question 3: Robust Policy For Player	10
3.1 Executions	10
3.2 Observations	10

1 Question 1: Discounted MDP For Player

The objective of this part is to choose the optimal policy that maximizes the expected cumulative discounted reward over time for the "player".

States

The state is represented by the current configuration of the grid, and the next piece to be placed.

$S := (\text{current grid}, \text{next piece to place})$

Actions

The actions correspond to a position and orientation to place the piece on the grid.

$A := (\text{position}, \text{orientation})$

Transition Function

The transition function is defined as $P(s'|s, a) := 1/2$

There are two possible pieces (I or L) as the next piece to place, and considering our state representation, and our assumption of uniform distribution between the pieces, each transition has an equal probability of 1/2.

Reward Function and Discount Factor

The reward function is computed based on the number of lines cleared after placing the piece, and a discount factor λ is used to weigh future rewards.

$$R(s, a, s') := \text{coefficient} * \text{number of lines cleared with } \lambda \in [0, 1]$$

1.1 Finding Optimal Policy

We use the value iteration method to find the optimal policy that maximizes the expected rewards, iteratively updating the value function for each state based on the Bellman equation until it converges to the optimal value (margin ϵ).

$$V(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s, a) \cdot (R(s, a, s') + \lambda V(s'))$$

Value iteration is better suited for maximizing the expected discounted reward since it computes the optimal value then extracts the policy and it's easier to implement.

1.2 Implementation

1.2.1 Main Architecture

Based on the our design choices from Section 1, we decided to use an object oriented architecture (C++) for our implementation.

For our structure we have declared 7 classes:

1. *Point*: classical point (x,y) class used for grid coordinates
2. *Field*: contains a boolean grid s.t. $\text{grid}[i][j] = 1$ if the corresponding cell is filled in the field
3. *Tromino*: corresponds to the baby Tetris pieces (I Piece or L Piece)
4. *State*: contains both a field grid and an incoming Tromino
5. *Action*: represents a choosable action on a given state (combination for placement and rotation of an incoming Tromino)
6. *Game*: contains the configuration (completed lines scores), the current state and score of a baby Tetris game (the configuration is loaded from a file : *config.txt*)
7. *MDP*: encompasses all the necessary methods in order to compute the value iteration over a game

Now that we presented the overview of the global implementation architecture, we will describe our implementation choices.

1.2.2 Implementation Choices

- **Pieces teleported to their final position:** In order to reduce the number of intermediate states when choosing an action, we compute on a state all available actions, then according to the policy, we make a choice over these actions and we teleport the incoming piece in the current state directly to its final position (with the corresponding rotation)
- **Reward Function:** As defined in Section 1, the reward function takes both the current state s and an available action a_s . Since we can calculate the available actions, we know in which state s' we're going to arrive and we can compute directly in s' the number of completed lines before removing them from the game. Then we abstract the reward function by counting the number of complete lines in the resulting state
- **Game Ending:** According to the subject, the game should end whenever a piece exceeds the height limit of the grid. Since we compute all available actions on a given state, if a piece cannot be placed on the grid, there will be no available action for that state, the game will terminate when there is no available action for the current state.
- **Max Iteration:** Since the value iteration isn't an exact method, if we fix an ϵ too small, we might iterate for a very long time. For this reason, we fix a maximum number of iteration for the expected value improvement to have a value iteration method that computes in reasonable time, even if the expected value vector is less qualitative than the ϵ margin we fixed

1.2.3 Value Iteration Implementation For The Player

The value iteration algorithm is implemented for the player by maintaining a value function for each state and updating it based on the expected rewards from taking actions in that state. The output is a policy that maps each state to the action that maximizes the expected cumulative discounted reward.

```
// ... (within the loop over all reachable states)
// for (auto& [currState, currValue] : V)
    Action& action : currState.getAvailableActions()
    for (int k = 0; k < nbActions; k++)
    {
        // rewards for action index k
        rewards[k] = 0.0;
        for (State& placedState : currState.genAllStatesFromAction(action))
        {
            State afterState = placedState.completeLines();
            double vAfter = V.at(afterState);
            double reward[k] = 0.5 * (placedState.evaluate(config_) +
                lambda * vAfter);
        }
    }
    vPrime = *std::max_element(rewards.begin(), rewards.end());

    for (int k = 0; k < nbActions; k++)
    {
        if (rewards[k] == vPrime)
        {
            // If we find a matching action, we add it to the policy
            A.insert_or_assign(currState.clone(),
                std::make_unique<Tromino>(k));
            break;
        }
    }
// ... (the rest of the loop updates delta and V[currState] with vPrime)
```

Listing 1: Lowest Maximal Expected Reward

1.3 Results

Here, we will present multiple examples of execution of our implementation. In these executions we will use fixed values for: $\text{width} = 4$, $\epsilon = 0.00000001$, $\text{maxIteration} = 100$, $\text{probaIPiece} = 0.5$ and $\text{maxGameAction} = 10000$. Note that we fixed a small ϵ in order to approach the optimal policy, keeping in mind that the algorithm is bounded in iteration by the *maxIteration*. And the variant variables will be: the height $\in \{4, 5\}$ and the discount factor $\lambda \in [0, 1]$.

For each execution we will measure:

1. the number of iterations for the value iteration algorithm
2. the average expected gain of the optimal policy
3. the final score either at the end of the game or when we reach the maximum number of actions
4. the execution time of the said execution

1.3.1 Executions

```
All constants:
width = 4, height = 4, probaIPiece
= 0.5, maxGameAction = 10000
epsilon = 1e-08, maxIteration =
100, lambda = 0.1

i = 0 and delta = 6
i = 1 and delta = 0.6
i = 2 and delta = 0.06
i = 3 and delta = 0.00525
i = 4 and delta = 0.00045
i = 5 and delta = 0

average over final V 0.412926
...
...
**..
**..

Game Over! Global score: 7849 in
10000 actions
make run 3.02s user 0.08s system
92% cpu 3.340 total
```

height: 4 and λ : 0.1

```
All constants:
width = 4, height = 4, probaIPiece
= 0.5, maxGameAction = 10000
epsilon = 1e-08, maxIteration =
100, lambda = 0.4

i = 0 and delta = 6
i = 1 and delta = 2.4
i = 2 and delta = 0.96
i = 3 and delta = 0.336
i = 4 and delta = 0.1152
i = 5 and delta = 0

average over final V 0.473608
***.
***.
***.
***.

Game Over! Global score: 225 in
285 actions
make run 2.93s user 0.08s system
93% cpu 3.237 total
```

height: 4 and λ : 0.4

```
All constants:
width = 4, height = 5, probaIPiece
= 0.5, maxGameAction = 10000
epsilon = 1e-08, maxIteration =
100, lambda = 0.1

i = 0 and delta = 6
i = 1 and delta = 0.6
i = 2 and delta = 0.06
i = 3 and delta = 0.006
i = 4 and delta = 0.00058125
i = 5 and delta = 3.1875e-05
i = 6 and delta = 0

average over final V 0.462044
...
...
...
...
...*
.***

Game Over! Global score: 7864 in
10000 actions
make run 58.26s user 0.32s system
99% cpu 58.837 total
```

height: 5 and λ : 0.1

```
All constants:
width = 4, height = 5, probaIPiece
= 0.5, maxGameAction = 10000
epsilon = 1e-08, maxIteration =
100, lambda = 0.9

i = 0 and delta = 6
i = 1 and delta = 5.4
i = 2 and delta = 4.86
i = 3 and delta = 4.374
i = 4 and delta = 3.81358
i = 5 and delta = 1.88219
i = 6 and delta = 0

average over final V 0.676855
...
...
...
...
...*
.***

Game Over! Global score: 8513 in
10000 actions
make run 58.20s user 0.34s system
99% cpu 58.809 total
```

height: 5 and λ : 0.9

1.3.2 Observations

Over our executions, we remark that the λ has a huge impact on the optimal policy computed. On the smaller grid (4×4), it is really difficult to reach the *maxGameAction* with $\lambda = 0.4$, since the model will prioritize short term decisions in order to increase its score faster. In comparison, with $\lambda = 0.1$, the model will not gain that much per action, prioritizing safer actions that allow a higher score on the long term.

That result was as expected, since in a 4×4 grid, the space for the risky moves is limited. The study of the game with the 4×5 grid shows that bigger grid provides a larger space to take risks in order to increase the global score. The game with $height = 5$ and $\lambda = 0.9$ illustrates this perfectly because it scores higher than the game with $height = 5$ and $\lambda = 0.1$, since the risky actions can be better exploited on a bigger grid.

In both cases, the total computational time is pretty acceptable (max 1 minute) even if it can be largely optimized (see Section ??). About the exact same improvement, the average expected gain V cannot be really discussed here since there is a huge part of states that aren't reachable in practice.

Finally, we notice that the value iteration algorithm choice described in Section 1 was pretty relevant because we can observe the quick convergence to the optimal policy (at the last iteration, Δ is always equal to 0).

1.4 Reachable States

Instead of computing the value iteration on all possible states, we focus on the reachable states from a given starting state s_0 . To achieve this, we implement a Breadth First Search (BFS) algorithm that explores all reachable states iteratively from the current state.:

Algorithm 1 Generate Reachable States

```

1:  $Q$ : a queue of states
2:  $s$ : a state
3:  $H$ : a hashMap of states
4:  $s \leftarrow s_0$ 
5:  $Q.push(s)$ 
6: while  $Q.isNotEmpty()$  do
7:    $s \leftarrow Q.pop()$ 
8:   for  $a$ : availableActions( $s$ ) do
9:      $s_{Placed} \leftarrow applyAction(s, a)$ 
10:     $s_{After} \leftarrow s_{Placed}.completeLines()$ 
11:    if  $Q.doesNotContain(s_{After})$  and  $H.doesNotContain(s_{After})$  then
12:       $Q.push(s_{After})$ 
13:    end if
14:   end for
15:    $H.push(s)$ 
16: end while
17: return  $H$ 

```

Then we can repeat our experiments that produces results that are way better:

```
All constants:
width = 4, height = 4, probaIPiece
= 0.5, maxGameAction = 10000,
epsilon = 1e-08, maxIteration =
1000, lambda = 0.1

i = 0 and delta = 6.10288
i = 1 and delta = 0.352625
i = 2 and delta = 0.0250445
i = 3 and delta = 0.0010013
i = 4 and delta = 4.28143e-05
i = 5 and delta = 1.79071e-06
i = 6 and delta = 4.8676e-08
i = 7 and delta = 1.07541e-09

average over final V 0.565842
....
....
...**
...**

Game Over! Global score: 7632 in
10000 actions
-----
Executed in 7.69 secs
```

height: 4 and λ : 0.1

```
All constants:
width = 4, height = 5, probaIPiece
= 0.5, maxGameAction = 10000,
epsilon = 1e-08, maxIteration =
1000, lambda = 0.1

i = 0 and delta = 6.20552
i = 1 and delta = 0.477762
i = 2 and delta = 0.0334845
i = 3 and delta = 0.00181086
i = 4 and delta = 6.29689e-05
i = 5 and delta = 2.6382e-06
i = 6 and delta = 1.32337e-07
i = 7 and delta = 6.63914e-09

average over final V 0.559813
....
....
...
*...
***.

Game Over! Global score: 7620 in
10000 actions
-----
Executed in 46.73 secs
```

height: 5 and λ : 0.1

```
All constants:
width = 4, height = 4, probaIPiece
= 0.5, maxGameAction = 10000,
epsilon = 1e-08, maxIteration =
1000, lambda = 0.9

i = 0 and delta = 8.36976
i = 1 and delta = 5.32365
i = 2 and delta = 3.54512
i = 3 and delta = 2.10366
...
i = 87 and delta = 1.33224e-08
i = 88 and delta = 1.0733e-08
i = 89 and delta = 8.64692e-09

average over final V 4.15109
*...
*...
*...
***.

Game Over! Global score: 11118 in
10000 actions
-----
Executed in 26.53 secs
```

height: 4 and λ : 0.9

```
All constants:
width = 4, height = 5, probaIPiece
= 0.5, maxGameAction = 10000,
epsilon = 1e-08, maxIteration =
1000, lambda = 0.9

i = 0 and delta = 9.7563
i = 1 and delta = 6.7369
i = 2 and delta = 4.43912
i = 3 and delta = 2.70431
...
i = 88 and delta = 1.44415e-08
i = 89 and delta = 1.16674e-08
i = 90 and delta = 9.42616e-09

average over final V 4.54237
*...
*...
***.
***.
***.

Game Over! Global score: 13305 in
10000 actions
-----
Executed in 402.82 secs
```

height: 5 and λ : 0.9

We can then remark that with this new implementation, the value iteration algorithm is quite slower both in terms of computation (per iteration) and in number of iteration. For instance, the 4×4 game with $\lambda = 0.4$ wasn't able to play all the 10000 actions with being blocked while we can now play it with a way more "risky" policy ($\lambda = 0.9$).

2 Question 2: Adversarial MDP For Opponent

The goal of this part is to design an adversary policy that selects tromino pieces for the game with the intention of minimizing the players score. For this policy we use once again value iteration, but this time we minimize the expected cumulative discounted reward over time from the available actions. The policy chooses the tromino piece that leads to the lowest expected score for the player. We have the same states (reachable states) and reward function as in Section 1.

Actions

The actions for the adversary correspond to the choice of the next tromino piece for the player to place on the grid.

$$A := \{I, L\}$$

Transition Function

The transition function for the adversary is defined as :

$$P(s'|s, a, t) := 1/|A_t(s)|$$

where $A_t(s)$ is the set of available actions for the player given the current state s and the chosen tromino piece t , with the assumption of a uniform distribution over the player's possible actions for the given piece.

The player will still use the optimal policy computed in Section 1 to play the game.

2.0.1 Choosing the Tromino Piece

We considered two different heuristic approaches for choosing the tromino piece, their value iterations are defined as:

- **Lowest Maximal Expected Reward:** The policy chooses the tromino piece that leads to the lowest maximal expected reward for the player.

$$V(s) = \min_{t \in \{I, L\}} \max_{a \in A_t(s)} \sum_{s'} P(s'|s, a, t) \cdot (R(s, a, s') + \lambda V(s'))$$

- **Lowest Average Expected Reward:** The policy chooses the tromino piece that leads to the lowest average expected reward for the player.

$$V(s) = \min_{t \in \{I, L\}} \frac{1}{|A_t(s)|} \sum_{a \in A_t(s)} \sum_{s'} P(s'|s, a, t) \cdot (R(s, a, s') + \lambda V(s'))$$

Both approaches produce similar results, with no remarkable difference. However, compared to the random adversary policy, where the tromino pieces are chosen randomly, the tromino value iteration, for both approaches, successfully decreases the player's score.

2.1 Executions

According to our implementation, we made the same executions in order to compare our adversary policies. Here we play the value iteration policy (from Section /refsec:question1) for the player on 3 adversary policies: random (based on *PROBA_I_PIECE*), Min-Max value iteration and Min-Avg iteration:

```
All constants:  
width = 4, height = 4, probaIPiece  
= 0.5, maxGameAction = 10000,  
epsilon = 1e-08, maxIteration =  
1000, action policy lambda = 0.9,  
tromino policy lambda = 0.1 and  
number of simulation = 50
```

```
Average results:  
Random Adversary moves: 11086.3  
Min Max Adversary Moves: 9999  
Min Avg Adversary Moves: 10000
```

Executed in 35.04 secs

height: 4 and $\lambda_{tromino}$: 0.1

```
All constants:  
width = 4, height = 5, probaIPiece  
= 0.5, maxGameAction = 10000,  
epsilon = 1e-08, maxIteration =  
1000, action policy lambda = 0.9,  
tromino policy lambda = 0.9 and  
number of simulation = 50
```

```
Average results:  
Random Adversary moves: 13285  
Min Max Adversary Moves: 9996  
Min Avg Adversary Moves: 9996
```

Executed in 630.78 secs

height: 5 and $\lambda_{tromino}$: 0.1

```
All constants:  
width = 4, height = 4, probaIPiece  
= 0.5, maxGameAction = 10000,  
epsilon = 1e-08, maxIteration =  
1000, action policy lambda = 0.9,  
tromino policy lambda = 0.9 and  
number of simulation = 50
```

```
Average results:  
Random Adversary moves: 11091.7  
Min Max Adversary Moves: 9999  
Min Avg Adversary Moves: 10000
```

Executed in 40.04 secs

height: 4 and $\lambda_{tromino}$: 0.9

```
All constants:  
width = 4, height = 5, probaIPiece  
= 0.5, maxGameAction = 10000,  
epsilon = 1e-08, maxIteration =  
1000, action policy lambda = 0.9,  
tromino policy lambda = 0.1 and  
number of simulation = 50
```

```
Average results:  
Random Adversary moves: 13292.5  
Min Max Adversary Moves: 9996  
Min Avg Adversary Moves: 9996
```

Executed in 537.57 secs

height: 5 and $\lambda_{tromino}$: 0.9

2.1.1 Observations

We can first observe that the score is not that much reduced by the adversary policies even if it is more noticeable in the 4×5 case. This is caused by the lack of possibilities, we can only choose between 2 pieces where the player may have to choose between way more actions.

Contrarily to the player value iteration we can remark that the λ parameter doesn't affect the score of the game. This is due to the fact that the score is minimized in that case and then the factor with which we multiply the $V(s)$ will not change anything.

3 Question 3: Robust Policy For Player

In this section, we study the case where the player doesn't make any assumptions on the adversary policy, searching for a generally suitable policy instead of optimizing against a specific one. We accept a lower maximum score if it elevates the minimal score across all adversary policies.

To this end, the first idea we had was to implement a Max-Min value iteration. Player policy always considers the worst-case scenario against all adversary policies. This is achieved by modifying the value iteration to maximizing the minimum expected reward over all possible states resulting from the adversary's actions (tromino choices).

$$V(s) = \max_{a \in A} \min_{t \in \{I, L\}} P(s'|s, a, t) \cdot (R(s, a, s') + \lambda V(s'))$$

The second approach we had was to parameterize the base player value iteration. The idea is to use these parameters as weight in the reward function to adjust the importance of each factor.

$$R(s, a, s') := lw \cdot \text{completed_lines}_{s'} - hw \cdot \text{max_cell_height}_{s'} + sw \cdot \text{score}_{s'} - gr \cdot \text{gaps}_{s'}$$

Where lw , hw , sw and gr are respectively the weights for completed lines, the highest occupied cell's height, the produced score and the number of gaps obtained after the tromino placement.

3.1 Executions

The following table presents selected results from our grid search experiments for the parameterized policy and the performance of the dedicated Max-Min Value Iteration (VI) policy.

L.W.	H.W.	S.W.	G.R.	Rand	MinMax	MinAvg	GapAvg	Min
Max-Min VI (Fixed Parameters)				10685	9999	9999	10356	9999
0	0	3	0	11090.70	9999	10000	10622	9999
0	0	1	1	10837.25	9999	10000	9999	9999
10	0	4	0	10611.33	9998	10000	7500	7500
10	0	3	1	9833.09	9998	11248	7500	7500
20	0	4	1	9242.76	7499	10000	7499	7499
0	1	4	0	10046.42	9999	10000	8749	8749
0	1	4	1.5	9810.65	9999	10000	9999	9810.65
10	2	1	0	7856.08	7499	8750	7499	7499
20	2	3	1	7934.26	9996	8750	7499	7499
0	2	1	1	0.05	0	0	0	0

Table 1: Selected Results from Parameter Grid Search and Max-Min VI

3.2 Observations

The grid search across the four parameters, alongside the dedicated Max-Min Value Iteration, revealed several key insights into what constitutes a robust and high-performing policy.

- **Height Weight is Critical:** The most dramatic results are seen with the *height_weight*. Configurations where this parameter is greater than zero perform exceptionally poorly, often failing to achieve any significant score. This indicates that a strong penalty for increasing the board's height is counterproductive, as the agent becomes too conservative to clear lines effectively. The best-performing policies consistently had a *height_weight* of 0.

- **Score and Line Weights Matter:** High scores were achieved when the *score_weight* was a dominant factor, especially when both *line_weight* and *height_weight* were zero. For instance, configuration $\{0, 0, 3, 0\}$ showed strong performance, suggesting that directly optimizing for the game's intrinsic score (which is based on line clears) is a very effective strategy. A high *line_weight* also performs well but can be less consistent against different adversaries if not balanced.
- **Max-Min VI Performance:** The *robustActionValueIterationMaxMin* achieved a commendable minimal score of 9999, matching some of the best grid-searched configurations. Its performance against the Random adversary was 10685, which is solid, though some parameter-tuned policies (e.g., $\{0, 0, 3, 0\}$) slightly surpassed it. Against MinMax and MinAvg adversaries, it consistently achieved 9999, indicating good inherent robustness. This confirms the effectiveness of the Max-Min approach as a baseline for robust play.
- **Robustness vs. High Score:** Comparing the "Most Robust" policy from the parameter search (configuration #0 $\{0, 0, 1, 0\}$ with a min score of 9999) with the explicit Max-Min VI, both achieve the same minimal score against specific adversaries. However, the Max-Min VI exhibits a slightly higher average performance against the random adversary, making it a strong contender for overall robustness without requiring extensive parameter tuning.
- **Variability:** Configurations that were highly variable (e.g., configuration $\{10, 0, 3, 1\}$) demonstrated policies specialized for certain adversaries (e.g., MinAvg) but with lower scores against others. This highlights the trade-off between a specialized, high-reward policy and a generally robust one.
- **Gap Reduction:** The *gap_reduction* parameter had a more subtle effect. While not as impactful as the other weights, a small penalty for gaps (e.g., a weight of 1.0) often helped stabilize the policy. However, a weight of 0 for this parameter was part of some of the best overall configurations, suggesting its importance is secondary to a strong line-clearing and scoring incentive, or its optimal value is very low.

In conclusion, the experiments show that a robust policy for this game should primarily focus on maximizing the score derived from line completions while completely avoiding penalties for board height. The explicit Max-Min Value Iteration provides a strong robust baseline. Parameter tuning allows for fine-grained control and can yield slightly higher average scores against random adversaries, but the core principle of robustness against the worst-case remains paramount.