Statistical performance of the value iteration in dynamic programming of Tetris game

Dongnam Ko

The Catholic University of Korea

Joint work with Jeongho Kim, Byungjoon Lee, and Johong Min.

IX Partial differential equations, optimal design and numerics,

Benasque, 29 August, 2022

Overview

Table of Contents

The game of Tetris

How can we survive as long as we can in the game of Tetris? Forever?

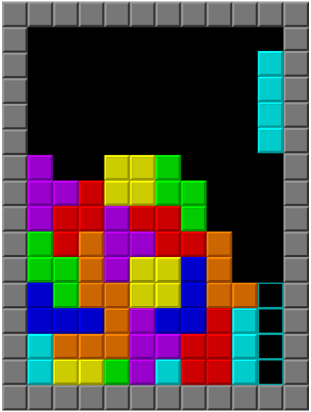

Figure: [Wikipedia; Tetris; a typical tetris screen]

The game of Tetris

For simplicity, we ignore 'operation' step of Tetris.

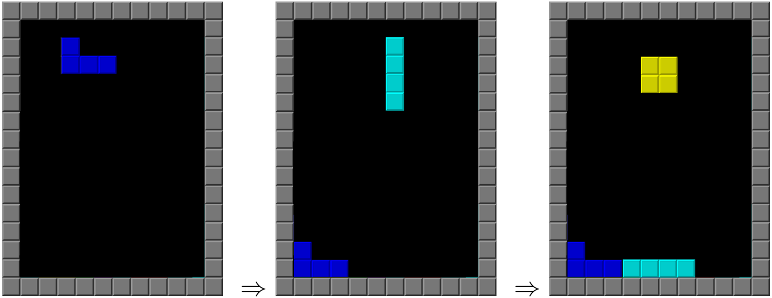

Figure: Time-discrete version of Tetris, $t = 0, 1, 2$.

Without consideration of falling speed, we just choose the place of the next block. Then, Tetris becomes **time-discrete** and **infinite horizon** optimal control problem.

Formulation of optimal control problem(OCP)

Optimal control problem: at a time instance, we have (1) stacked blocks in the screen and (2) the next block(shape) above. Our control is to choose a place for the next block.

State: a state i includes the stacked blocks (cardinality nearly $2^{\text{width} \times \text{height}}$) and the next block (all 7 shapes).

Control: a control μ maps a state i to the next stacked blocks. The shape is randomly given, therefore, $\mu(i)$ is a set of 7 states.

Next states: The state $j(t, i, \mu, \omega)$ after time t with control μ while the random element ω determines the next shapes. The set of possible $i(t, \mu, \omega)$ is denoted by $N(t, i, \mu)$.

Reward function for OCP

The running reward(cost) can be chosen as follows; Let i represent a state containing stacked blocks and the next block. Then,

$$r(i) = \begin{cases} 1 & \text{if } i \text{ is not at the terminal state (end of game)}, \\ 0 & \text{if } i \text{ is in the terminal}. \end{cases}$$

In order to make r integrable along time, we multiply $\alpha \in (0,1)$ for the next step (due to infinite-horizon).

$$\text{Reward}(i,\mu) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r(j(t,i,\mu))\right].$$

Then, the optimal control μ_* **want to survive as long as we can**.

Value and its control strategy

How can we compute the optimal control?

Dynamic Programming considers a value function V that maps a state i to a real number $V(i)$.

$V(i)$ gets 0 if i is the terminal (Boundary Condition), and $V(i)$ becomes bigger if it is a 'good' state to survive.

Then, the strategy μ corresponds V is the maximizer of

$$TV(i) := \max_{\mu} \left[r(i) + \alpha \mathbb{E} V(j(1, i, \mu)) \right].$$

The operator T is the Bellman operator, and μ is called the 1-step lookahead (control) of the value V.

Bellman operators and the optimal value

Since there are 7 shapes of blocks, the expectation is on the 7 shapes;

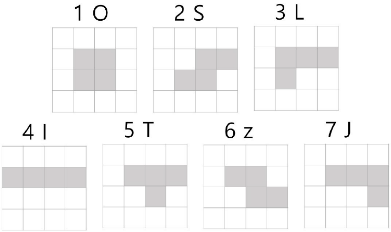

Figure: 7 shapes of Tetris

$$TV(i) = \max_{\mu} \left[1 + \frac{\alpha}{7} \sum_{j \in N(1,i,\mu)} V(j) \right].$$

Tetris is a typical example of a 'survival game' since we can just count time steps for scores and in every step there is randomness on shapes. (We may consider snake game, pacman, or flappy birds for comparison.)

Table of Contents

lookahead and the optimal reward

How we can **compute the optimal value** function?

For example, the 2-step lookahead is the maximizer of

$$\begin{aligned} T^2 V(i) &= \max_{\mu} [r(i) + \alpha \mathbb{E} [r(j(1,i,\mu)) + \alpha \mathbb{E} V(j(2,i,\mu))]] \\ &= \max_{\mu} \mathbb{E} \left[r(j(0,i,\mu)) + \alpha r(j(1,i,\mu)) + \alpha^2 V(j(2,i,\mu)) \right] \end{aligned}$$

which searches the values at states after two discrete time steps. ℓ-step lookahead follows

$$T^{\ell} V(i) = \max_{\mu} \mathbb{E} \left[\sum_{t=0}^{\ell-1} \alpha^t r(j(t,i,\mu)) + \alpha^{\ell} V(j(\ell,i,\mu)) \right].$$

As $\ell \to \infty$, it will converge to the **optimal reward function**

$$T^{\ell} V(i) \to \text{Reward}(i) = \max_{\mu} \mathbb{E} \left[\sum_{t=0}^{\infty} \alpha^t r(j(t,i,\mu)) \right]$$

with the exponential rate (at least) α in ℓ^{∞}-norm.

The reward of a given control

Computing $T^\ell V$ ($\ell \to \infty$) to find V^* is called the **value iteration(VI)**.

Note that, for each μ, we may define $T_\mu^\ell V$ and it also converges to the unique equilibrium V_μ, the expected reward using μ (Performance):

$$T_\mu^\ell V(i) = \mathbb{E}\left[\sum_{t=0}^{\ell-1} \alpha^t r(j(t,i,\mu)) + \alpha^\ell V(j(\ell,i,\mu))\right] \to \mathrm{Reward}(i,\mu) = V_\mu(i).$$

This is called the **policy evaluation**.

Performance estimate [D. P. Bertsekas, Book, 2019]

Let μ be the ℓ-step lookahead of a value V. Then, the policy evaluation V_μ from the strategy μ satisfies

$$\|V_\mu - V^*\|_\infty \leq \frac{2\alpha^\ell}{1-\alpha}\|V - V^*\|_\infty.$$

The meaning of value

What exactly is the meaning of the value $V_\mu(i)$?

- Case 1; let a control μ always make Tetris terminates after t step. Then, the total reward becomes

$$V_\mu(i) = 1 + \alpha + \alpha^2 + \cdots + \alpha^{t-1} = \frac{1-\alpha^t}{1-\alpha}.$$

- If the optimal control could keep Tetris forever, then,

$$V_\mu(i) - V^*(i) = V_\mu(i) - \frac{1}{1-\alpha} = -\frac{\alpha^t}{1-\alpha}.$$

In general, we will get

$$V_\mu(i) - V^*(i) = -\frac{\alpha^\tau}{1-\alpha},$$

where τ is an weighted expected survival time with the strategy μ. Therefore, the performance estimate says that

$$\tau \geq \ell + \text{Constant}.$$

Numerical simulation 1; standard Tetris

We never see this performance in practice.

Consider Tetris with the standard scale; width 10 and height 12. The value function is set to be (the remaining height - number of holes).

Then, the survival time $\tau(\ell)$ from ℓ-step lookahead scores

$$\tau(1) = 120, \quad \tau(2) = 370, \quad \tau(3) = 1070, \quad \text{and} \quad \tau(4) = 2810.$$

(averaged over 20 simulations)

The score grows, not just linearly, but may exponentially.

Numerical simulation 2; small Tetris

If the scale gets smaller, then the cardinality of states becomes few enough to **compute the optimal value. For** $4 * 5$ **Tetris,** the cardinality is less than $7 \times 2^{20} \sim 1MB$.

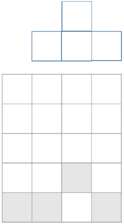

Figure: Tetris with width 4 and height 5

Here a 1-step lookahead control can survive around 20 time steps in expectation.

Numerical simulations; small Tetris

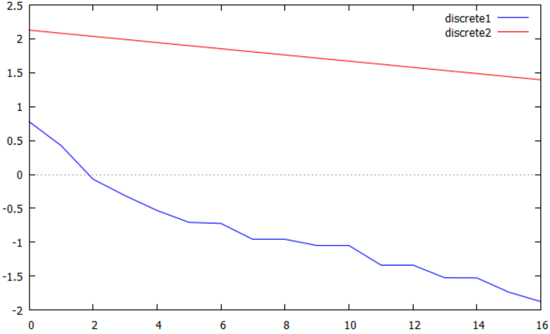

Figure: $\log(V_\mu(i) - V^*(i))$ along ℓ (blue), compared with α decay (red).

As ℓ grows, the ratio seems much stiff, at least $(\alpha/2)$. This shows that, in score,

$$\tau \gtrapprox 22\ell + \text{Constant}.$$

In the remaining of the talk, we want to explain this $(\alpha/2)$ factor.

Table of Contents

The performance estimates of VI

From now on, we introduce the results of [Kim-K.-Lee-Min, preprint]:

The performance estimate follows

$$\|V_\mu - V^*\|_\infty \le \frac{\alpha^\ell}{1-\alpha}\|V - V^*\|_\infty.$$

The proof is elementary, and here we reorganizing it with Bellman operators,

$$T_\mu V(i) := 1 + \frac{\alpha}{7}\sum_{j\in N(i,\mu)} V(j), \quad TV(i) := 1 + \frac{\alpha}{7}\max_u \left(\sum_{j\in N(i,u)} V(j)\right),$$

and its zero-reward versions,

$$L_\mu V(i) := 0 + \frac{\alpha}{7}\sum_{j\in N(i,\mu)} \frac{1}{7}V(j), \quad LV(i) := 0 + \frac{\alpha}{7}\max_u \left(\sum_{j\in N(i,u)} \frac{1}{7}V(j)\right).$$

The performance equality

Theorem 1; operator expression of performance

Given a value function $\tilde{V}$ of V^*, let μ be the ℓ-step lookahead policy generated by $\tilde{V}$, i.e., $T_\mu T^{\ell-1}\tilde{V} = T^\ell \tilde{V}$. Then, we have

$$V_\mu - V^* = (I + L_\mu + L_\mu^2 + \cdots)\left[\left(T^\ell V^* - T^\ell \tilde{V}\right) + L_\mu\left(T^{\ell-1}\tilde{V} - T^{\ell-1}V^*\right)\right].$$

Given this result, since all the operators are contraction with ratio α,

$$\begin{aligned}
&\|V_\mu - V^*\|_\infty \\
&\quad\le (1 + \alpha + \alpha^2 + \cdots)(\|T^\ell \tilde{V} - T^\ell V^*\|_\infty + \alpha \cdot \|T^{\ell-1}\tilde{V} - T^{\ell-1}V^*\|_\infty) \\
&\quad\le (1 + \alpha + \alpha^2 + \cdots)(\alpha^\ell \|\tilde{V} - V^*\|_\infty + \alpha \cdot \alpha^{\ell-1}\|\tilde{V} - V^*\|_\infty) \\
&\quad\le \frac{2\alpha^\ell}{1-\alpha}\|\tilde{V} - V^*\|_\infty.
\end{aligned}$$

The performance equality

Theorem 2; scaling of values in performance

Suppose that the running reward $r(i)$ is constant. Given a state i, a value function $\tilde{V}$, and a series of constants $a_m > 0$, we have

$$(V_\mu - V^*)(i) = \sum_{m=0}^{\infty} L_\mu^m \left(T^\ell V^* - T^\ell(a_m \tilde{V}) \right)(i) + \sum_{m=1}^{\infty} L_\mu^m \left(T^{\ell-1} V^* - T^{\ell-1}(a_{m-1} \tilde{V}) \right)(i).$$

Moreover, each term is affine to $a_1, a_2, \cdots$.

This implies that, with carefully determined $a_m > 0$, we have

$$|(V_\mu - V^*)(i)| = \sum_{m=0}^{\infty} \left| L_\mu^m (T^\ell V^* - T^\ell(a_m \tilde{V}))(i) \right|.$$

Therefore we only need to analyze $|T^\ell V^*(j) - T^\ell(a_m \tilde{V})(j)|$.

Proof of the operator expression

The proof needs two ingredients. One is the convergence result,

$$\lim_{m\to\infty} T_\mu^m V = V_\mu \quad \text{for any } V.$$

The other is on the linear Bellman operators,

$$T_\mu V_1 - T_\mu V_2 = L_\mu V_1 - L_\mu V_2 = L_\mu(V_1 - V_2).$$

Therefore, we have

$$\begin{aligned} V_\mu - V^* &= \sum_{m=1}^{\infty} (T_\mu^m V^* - T_\mu^{m-1} V^*) \\ &= \sum_{m=1}^{\infty} (T_\mu^m V^* - T_\mu^m T^{\ell-1} \tilde{V}) + \sum_{m=1}^{\infty} (T_\mu^{m-1} T^\ell \tilde{V} - T_\mu^{m-1} V^*) \\ &= \sum_{m=1}^{\infty} L_\mu^m (T^{\ell-1} V^* - T^{\ell-1} \tilde{V}) + \sum_{m=1}^{\infty} L_\mu^{m-1} (T^\ell \tilde{V} - T^\ell V^*) \end{aligned}$$

Proof of the scaling values

The proof needs two ingredients. One is the control conservation (from the constant reward assumption),

$$T_\mu \tilde{V} = T\tilde{V} \;\Leftrightarrow\; T_\mu(a\tilde{V}) = T(a\tilde{V})$$

The other is the affine properties of Bellman operators,

$$L_\mu^k T^\ell(aV) - L_\mu^k L^\ell(aV) = \text{Constant} \quad \text{for any } a > 0.$$

It also guarantees that every term is affine on a. Therefore, we have

$$\begin{aligned}(V_\mu - V^*)(i) = &\sum_{m=0}^{\infty} L_\mu^m \left(T^\ell V^* - T^\ell(a_m\tilde{V})\right)(i)\\ &+ \sum_{m=0}^{\infty} L_\mu^m L_\mu \left(T^{\ell-1}(a_m\tilde{V}) - T^{\ell-1}V^*\right)(i).\end{aligned}$$

Table of Contents

Effects of lookaheads

The performance equality now becomes

$$|(V_\mu - V^*)(i)| = \sum_{m=0}^{\infty} \left| L_\mu^m (T^\ell V^* - T^\ell (a_m \tilde{V}))(i) \right|.$$

In a rough estimation, we have

$$\frac{\left\| T^\ell V^* - T^\ell \tilde{V} \right\|_\infty}{\| V^* - \tilde{V} \|_\infty} \leq \alpha^\ell.$$

From a viewpoint of Hamilton-Jacobi-Bellman equation, it corresponds to a kind of 'local' contraction rate near the initial state (in spatial position) and near the optimal value (in solution function space).

Argument of choosing maxima

Suppose that our problem is deterministic (no random blocks).

Then, the situation becomes simple; for two value functions V_1 and V_2 defined on a set N, we need to estimate

$$\frac{|T^\ell V_1(i_0) - T^\ell V_2(i_0)|}{\max_{i_\ell \in N} |V_1(i_\ell) - V_2(i_\ell)|} = \frac{\alpha^\ell | \max_{i_\ell \in N} V_1(i_\ell) - \max_{i_\ell \in N} V_2(i_\ell)|}{\max_{i_\ell \in N} |V_1(i_\ell) - V_2(i_\ell)|},$$

which is a simple maximum argument among $|N|$ elements.

When this ratio becomes α^ℓ? From general values V_1 and V_2, it occurs when both V_1 and V_2 have their maxima at the same point!

Note also that if they always have the same maximizers, then two corresponding controls are idential, and it implies that $V_1 = V_2$.

Choosing maxima - closely corellated values

For example, suppose that V_1 is V^* and V_2 follows $V^* + \varepsilon$ where $\varepsilon \sim U[0, \sigma]$ is a small uniform error.

The ratio now becomes

$$\frac{\alpha^{\ell} |\max_{i_\ell \in N} V^*(i_\ell) - \max_{i_\ell \in N} (V^* + \varepsilon)(i_\ell)|}{\max_{i_\ell \in N} \|\varepsilon(i_\ell)\|},$$

From simulations, we can get this is around from 0.5 to 0.8.

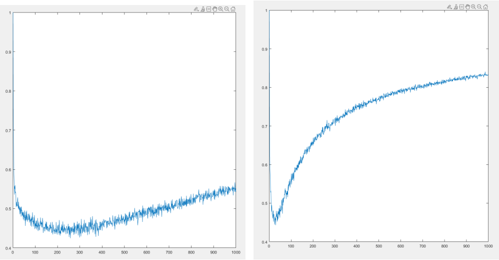

Figure: Left: with $E \sim U[0, 0.01]$, Right: with $E \sim U[0, 0.1]$, drawn over $|N|$.

Lookahead is a critical gain for performance

The maximum and the averaging arguments exclusively affects the ratio. For large ℓ, $T^\ell \tilde{V}$ and V^* are closely correlated; The performance improves around $\alpha/2$, and it fits the data.

Numerical simulations; [Kim-K.-Lee-Min, preprint]

Consider the small Tetris game with width 4 and height 5. Suppose that the initial value function $\tilde{V}$ follows $V^* + \varepsilon$ where $\varepsilon \sim U[0, 0.01]$. Then, the convergence ratio of $T^\ell(a\tilde{V})(i_0)$ to $V^*(i_0)$ becomes $(\alpha/2)$ in expectation, for proper $a > 0$ on each i_0. The expectation is over $\tilde{V}$.

Note: the convergence of values and of controls are different problems.

Simulation revisited

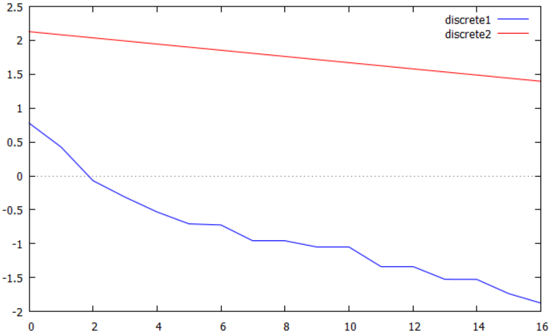

Figure: Graph shows $\log(V_\mu(i) - V^*(i))$ along ℓ (blue), compared with α decay (red).

We can check that with small ℓ the decay is nearly $\alpha/5$ and for large ℓ it becomes $\alpha/2$.

Table of Contents

Summary and remarks

Summary

- Value iteration near good initial state and value exceeds the performance estimate which uses ℓ^∞-norm.

Remark

- The main point is on the local contraction ratio of the Bellman operator, near the initial state, near the optimal value.

- Unfortunately, there seems not many results on this, even for Hamilton-Jacobi-Bellman equation with infinite-horizon.

THANK YOU FOR YOUR ATTENTION.