Preprints of the 2012 20th Mediterranean Conference on Control & Automation (MED), Barcelona, Spain, July 3-6, 2012

Global Control Lyapunov Function design based on decision by Q-learning

H. Akiba (Tokyo University of Science), H. Nakamura(Tokyo University of Science)

***Abstract*— In nonlinear control theory, multilayer minimum projection method is proposed for Control Lyapunov Function (CLF) design. The method generates a global control Lyapunov function from local control Lyapunov functions. However, the automatic generation method from local functions is not developed. In this paper, we consider the control problem including learning from exploring space. The learning from exploring is defined in discrete space, however the control problem is defined in continuous space. Thus, we need to relate the discrete space to the continuous space.**

This paper focuses on the CLF design including learning from exploring by Q-learning. Our goal is to develop a CLF design method.

I. Introduction

In nonlinear control theory, the multilayer minimum projection method is available for designing global Control Lyapunov Functions (global CLFs).

The paper considers a control problem including learning from exploration. Reinforcement learning is used for system defined in a discrete space. However the control problem in the paper is defined in continuous space. This implies that we need to relate the discrete space to the continuous space.

In this paper, we propose a global CLF design method with the multilayer minimum projection method and Q-learning. The advantages of the proposed methods are confirmed by an example.

II. Preliminaries

In this paper, we use the Q-learning and multilayer minimum projection methods to design a global CLF. In this section, we introduce Q-learning, multilayer minimum projection method, and control design with CLFs.

A. Finite Markov decision process[1]

If conditional provability on the future states depends only on the current state in a stochastic process, the property is called Markov property. If a reinforcement learning task has Markov property, the task is called Markov Decision Process (MDP). If state and action spaces are finite, the task is called finite Markov Decision Process (finite MDP).

B. Action-value function[1]

We can evaluate value of states and action by an action-value function in reinforcement learning.

The method to decide the probability of selection of the action and to select the action is called policy.

This work was supported by JSPS Grant-in-Aid for Scientific Research(B) (22360167)(23360185)

H. Akiba and H. Nakamura are with the Department of Electrical Engineering, Faculty of Science and Technology, Tokyo University of Science, Yamazaki 2641, Noda, Chiba, Japan nakamura@rs.tus.ac.jp

Let the state after time step t be s_t and the action after time step t be a_t. Then, the expected return under the policy π is updated by the following equation:

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1}\middle| s_t = s, a_t = a\right\}, \quad (1)$$

where γ satisfying $0 \leq \gamma \leq 1$ is a parameter called a discount rate.

C. Q-learning [1] [2]

In Q-learning, we evaluate the effectiveness of an action a in a state s by action-value function $Q(s,a)$.

Consider an agent with the state s_t at time step t. When the agent takes the action a_t and moves to state s_{t+1}, $Q(s_t,a_t)$ is expressed by the following equation[1] :

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a)\right], \quad (2)$$

where $\leftarrow$ denotes updating and α satisfying $0 \leq \alpha \leq 1$ is a constant called a learning rate. The learning rate weights past rewards and determine learning speed.

The procedure of Q-learning is summarized below.

1) Initialing all Q and time step $t = 0$.
2) Agent observes the state s_t.
3) The agent selects and takes the action a_t according to the policy π.
4) The agent observes the next state s_{t+1}.
5) The agent obtains the reward r_{t+1}.
6) $Q(s_t,a_t)$ is updated by (2).
7) Set $t = t+1$, and go to 2.

In this paper, the agent employs the ϵ-greedy method for the policy. We introduce the ϵ-greedy method in the following subsection.

D. ϵ-greedy method

In ϵ-greedy, parameter ϵ represents a possibility satisfying $0 \leq \epsilon \leq 1$. Selecting the action by ϵ-greedy is accomplished as follows.

- The agent randomly selects an available action in the present state s with probability ϵ.
- The agent selects the best action in the present state s with probability $1-\epsilon$. The best action in the state s is the action having $\max_a Q(s,a)$.

978-1-4673-2529-5/12/$31.00 ©2012 IEEE

1) Convergence condition of Q-learning: We assume that the learning rate α satisfies the following two equations regarding time t and the agent performs all actions at all states by a certain policy:

$$\sum_{t=0}^{\infty} \alpha(t) \to \infty, \ \sum_{t=0}^{\infty} \alpha(t)^2 < \infty. \tag{3}$$

Then , $Q(s_t, a_t)$ converges to the optimal $Q(s_t, a_t)$ in all states with possibility 1. When we achieve the optimal $Q(s_t, a_t)$, we can solve the reinforcement learning problem by greedy policy with $\epsilon = 0$.

E. Differentiable manifolds

In this paper, we consider the following nonlinear system on a two-dimensional arc connected C^1 differentiable manifold X:

$$\dot{x} = f(x, u), \tag{4}$$

where $x \in X, u \in U \subset F(\mathbb{R}, \mathbb{R}^2); t \mapsto u(t) \in \mathcal{U} \subset \mathbb{R}^2$, where $F(\mathbb{R}, \mathbb{R}^2)$ denotes a set of mappings from $\mathbb{R}$ to $\mathbb{R}^2$. The mapping $f : X \times \mathcal{U} \to T_x X$ is satisfied $f(0, 0) = 0$. $T_x X$ is a vector space called the tangent space to X at x.
We assume a chart (W, η) for X, where $W \subset X$, $\eta : W \to \Xi \subset \mathbb{R}^2$. Then $\eta(x)$ is called the local coordinate representation of x with the chart (W, η).
Consider the function $V : X \to \mathbb{R}$ and (W, η). The function $V_W : \Xi \to \mathbb{R}$ is defined by $V_W(x) = V \circ \eta^{-1}$. Note that V_W is defined on the subset of $\mathbb{R}^2$.
A function $V : X \to \mathbb{R}$ is called locally Lipschitz at $x \in X$ if there exist $K > 0$ and a neighborhood Ω of x such that

$$|V_W(\eta(x)) - V_W(\eta(y))| \leq K \|\eta(x) - \eta(y)\| \tag{5}$$

for all $x \in X$ and $y \in \Omega \subset W$ with local chart (W, η).

Local semiconcavity is defined as follows: a continuous function V is called locally semiconcave at $x \in X$ if there exist $C > 0$ and a neighborhood Ω of x such that

$$V(x) + V(y) - 2V_W\left(\frac{1}{2}(\eta(x) + \eta(y))\right) \leq C\|\eta(x) - \eta(y)\|^2 \tag{6}$$

for all $y \in \Omega \subset W$. When the function V is locally semiconcave, V is called locally semiconcave function on X.
In this paper, we use s-stability defined as follows, for the definition of stability in section V-B.

Definition 1: (Partition)[5][6][9] The infinite sequence $\pi = \{t_i \in \mathbb{R}_{\geq 0}\}_{i \in \mathbb{Z}_{\geq 0}}$ satisfied $0 = t_0 < t_1 < t_2 < \cdots$ and $\lim_{i \to +\infty} t_i = +\infty$ is called partition, and $d(\pi) := \sup_{i \in \mathbb{Z}_{\geq 0}}(t_{i+1} - t_i)$ is called the diameter of partition.

Definition 2: (Sample-hold solution)[5][6][9] Let $u = k(x)$ be feedback, π partition and $x \in X$ an initial state. Then, sample-hold solution $\psi(t, x, k(x)) : \mathbb{R}_{\geq 0} \times X \times U \to X$ is the continuous mapping obtained by repeatedly solving the following equation from the initial time t_i to the terminal time s_i:

$$\dot{x}(t) = f(x(t), k(x(t_i))), \tag{7}$$

where $x(0) = x$, and the terminal time s_i is defined by the following equation:

$$s_i = \max\{t_i, \sup\{s \in [t_i, t_{i+1}] | x(\cdot) \ is \ defined \ on \ [t_i, s)\}\}. \tag{8}$$

Definition 3: (S-stability)[5][6][9] Consider the system (4). A feedback control low $k : X \to \mathcal{U}$ is said to stabilize the origin if the following conditions are satisfied with $\mathcal{R}_1$, $\mathcal{R}_2 \in \mathfrak{P}$ such that $\mathcal{R}_1 \subset \mathcal{R}_2$. $\mathfrak{P}$ denotes the set of all open precompact subset of X containing the origin.

(1) There exists the set $\mathcal{M} \subset X$ depending only upon $\mathcal{R}_2$, the positive constant Ω depending on $\mathcal{R}_1$ and $\mathcal{R}_2$, and $T > 0$ exis. The sample-hold solution is satisfied the following conditions for an arbitrary initial value $x \in \mathcal{R}_2$ and any partition π whose diameter is smaller than Ω.
   (c1) For all $t \geq T$, $\psi(t, x, l(x)) \in \mathcal{R}_1$.
   (c2) For all $t \geq 0$, $\psi(t, x, l(x)) \in \mathcal{M}$.
(2) For each $\mathcal{O} \in \mathfrak{P}$ there exists a set $\mathcal{P} \in \mathfrak{P}$ such that if $\mathcal{R}_2 \subset \mathcal{P}$, $\mathcal{M}$ in (1) can be chosen satisfying $\mathcal{M} \subset \mathcal{O}$.

F. Locally semiconcave practical control Lyapunov function[5][6]

Rifford introduced the semiconcave strict CLF for discontinuous feedback controller design.
However, strict CLFs are not appropriate for noncompact U. Hence, Nakamura introduced the locally semiconcave practical control Lyapunov function as follows[5][6]:

Definition 4: (Locally semiconcave practical control Lyapunov function) A Locally semiconcave practical control Lyapunov function for system (4) is a locally semiconcave function $V : X \to \mathbb{R}$ such that(a1), (a2) and the following condition hold.

(a1) V is proper; that is the set $\{x \in X | V(x) \leq L\}$ is compact for arbitrary L.
(a2) V is positive define; that is , $V(0) = 0$ and $V(x) > 0$ $\forall x \in X \backslash \{0\}$.
(a3) the compact set $\bar{\mathcal{U}} \subset \mathcal{U}$ satisfies the following equation, $Q > 0$, and the local chart (W, η) exist for arbitrary sets $R_1, R_2 \in \mathbb{R}_{>0}$, where $R_2 > R_1 > 0$.

$$\min_{u \in \bar{\mathcal{U}}} DV_W(\eta(x); f_W(\eta(x), u)) < -Q \tag{9}$$

$$\forall x \in \{x | R_1 \leq V(x) \leq R_2\} \tag{10}$$

where the mapping $f_W = f \circ \eta^{-1}$, the directional subderivative DV_W is defined as follows:

$$DV_W(\eta(x); v) = \lim_{t \to 0} \frac{V_W(\eta(x) + tv) - V(x)}{t}. \tag{11}$$

G. Multilayer minimum projection method

We can generate a global control Lyapunov function from local control Lyapunov functions by multilayer minimum projection method proposed by Nakamura[3][4]. We define the multilayered space coproduct of mappings as follows.

Definition 5: (Multilayered space) Consider a disjoint family of minifolds $\{\tilde{X}_i\}_{i \in L}$, manifold X and a family of mappings $\{\phi_i | \phi_i : \tilde{X}_i \to X\}_{i \in L}$, where $\{\tilde{X}_i\}_{i \in L}$ and X

are arc-connected manifolds. Then, the following manifold is said to be the coproduct of manifolds associated with the family of mappings$\{\phi_i\}_{i\in L}$:

$$\tilde{X} = \coprod_{i=1}^{l} \tilde{X}_i, \tag{12}$$

where $\coprod$ denotes a disjoint union of sets. In this paper, the coproduct manifold $\tilde{X}$ is said to be a multilayered space and each submanifold $\tilde{X}_i$ is called layer.

Definition 6: (Coproduct of mappings) Consider a disjoint family of minifolds $\{\tilde{X}_i\}_{i\in L}$ and manifold X. We assume there exists a family of mappings $\phi_i : \tilde{X}_i \to X$ Then, the following mapping from the coproduct of manifolds $\phi_i : \tilde{X}_i \to X$ is said to be the coproduct of $\{\phi_i\} i \in L$:

$$\phi(\tilde{x}) = \begin{cases} \phi_1(\tilde{x}) & (\tilde{x} \in \tilde{X}_1) \\ \vdots & \\ \phi_l(\tilde{x}) & (\tilde{x} \in \tilde{X}_l) \end{cases} \tag{13}$$

If $X = \mathbb{R}$, the coproduct of mappings ϕ is said to be a coproduct of functions.

The multilayer minimum projection method is summarized as follows:

(b1) Choose local diffeomorphisms $\{\phi_i | \phi : \tilde{X}_i \to X\}$ satisfying the following conditions on finite layers $\{\tilde{X}_i\}$, where $\tilde{X}_i$ is simple-structured two-dimensional C^1 differentiable manifold.
  (1) Continuous mapping $\phi_{i*} : T_{\tilde{x}_i}\tilde{X}_i \to T_{\phi(\tilde{x}_i)}X$ is bijection for $\tilde{x}_i \in \tilde{X}$.
  (2) $0_i \in \tilde{X}$ such as $\phi_i(0_i) = 0$ exist.

(b2) Choose $L = \{1, \cdots, l\}$ such as the coproduct of the family of mappings $\{\phi_i\}_{i\in L}$ is surjection

(b3) Design a CLF $\tilde{V}_i$ on each $\tilde{X}_i$ for globally asymptotic stabilization at the origin 0_i of $\dot{\tilde{x}}_i = \tilde{f}_i(\tilde{x}_i, u)$.

(b4) The following function is the CLF for asymptotic stabilization at the origin of X:

$$V(x) = \min_{\tilde{x}\in\phi^{-1}(x)} \tilde{V}(\tilde{x}), \tag{14}$$

where $\tilde{V} : \tilde{X} \to \mathbb{R}$ is the coproduct of mappings $\{\tilde{V}_{i\{i\in L\}}\}$.

Furthermore, the following theorem of multilayer minimum projection method holds.

Theorem 1: Consider finitely many layers $\tilde{X}_1, \cdots, \tilde{X}_l$ defined diffeomorphism and satisfied following conditions.

(1) For all $i \in L := \{1, 2, \cdots, l\}$ and $\tilde{x}_i \in \tilde{X}_i$, continuous mappings$\phi_{i*} : T_{\tilde{x}_i}\tilde{X}_i \to T_{\phi_i(\tilde{x}_i)}X$ are bijections.
(2) $0_1 \in \tilde{X}_1$ satisfied $\phi_1(0_1) = 0$ exist.
(3) The coproduct of mapping $\phi : \tilde{X} \to \tilde{X}$ associated $\{\phi_i\}_{i\in L}$ is surjection.
(4) Consider $\tilde{X}_i$ existed $k \in \{1, \cdots, i-1\}$ satisfied $\phi_i(0_i) \in Im(\phi_k)$. $0_i \in \tilde{X}_i$ exist for all $i \in L\backslash\{1\}$.

Consider the family of CLF for asymptotic stabilization at the origin $0_i \in \tilde{X}$ and following functions.

$$\hat{V}_1 = \tilde{V}_1 \tag{15}$$

$$\hat{V}_i = \tilde{V}_i + c_i (i \in L\backslash\{1\}) \tag{16}$$

where offset c_i is defined below.

$$c_i > \min_{1,\cdots,i-1} \min_{\tilde{x}_k \in {\phi_k}^{-1}(\phi_i(0_i))} \hat{V}_k(\tilde{x}_k) \tag{17}$$

Then, the following function $V(x)$ is the CLF on X:

$$V(x) = \min_{\tilde{x}\in\phi^{-1}(x)} \hat{V}(\tilde{x}), \tag{18}$$

where $\hat{V}$ is the coproduct of family of functions.

H. CLF design with multilayer minimum projection method and local semiconcave practical control Lyapunov function[6]

Consider the following discontinuous state feed back control with locally semiconcave control Lyapunov function V designed by the multilayer minimum projection method:

$$u_i = \begin{cases} -\frac{\langle p,f\rangle + \sqrt{\langle p,f\rangle^2 + (\sum_{i=1}^2 \langle p,g_i\rangle^2)^2}}{\sum_{i=1}^2 \langle p,g_i\rangle^2} \langle p, g_i\rangle \\ \qquad\qquad \left(\sum_{i=1}^2 \langle p, g_i\rangle^2 \neq 0\right) \\ 0 \\ \qquad\qquad \left(\sum_{i=1}^2 \langle p, g_i\rangle^2 = 0\right), \end{cases} \tag{19}$$

where $p : X \to T_x^*X$ is a discontinuous mapping such that $p(x) \in \tilde{D}V(x)$ for all $x \in X$. Then, the input (19) stabilizes the origin of (4). $\tilde{D}V(x)$ is called the disassembled differential defined as follows:

Definition 7: (Disassembled differential) Consider a local semiconcave function on an two-dimensional C^1 differentiable manifold X.

An arbitrary compact set $S \subset \mathbb{R}^n$ and an indexed family of C^2 function $(\tilde{V}_s)_{s\in S}$ satisfy the following function for an arbitrary compact set on the local chart (W, η):

$$V(x) = \min_{s\in S} \tilde{V}_s(x), \forall x \in M. \tag{20}$$

Then the set-valued mapping $\tilde{D}V : X \to 2^{T_x X}$ defined by the following function is called disassembled differential of V.

$$\begin{aligned} \tilde{D}V(x) = & \left\{ d\tilde{V}_s(x) | s \in \{s \in S | V(x) = \tilde{V}_s(x)\} \right\} \\ = & \left\{ \tfrac{\partial \tilde{V}_s W}{\partial \xi}(\eta(x)) d\xi | s \in \{s \in S | V(x) = \tilde{V}_s(x)\} \right\} \end{aligned} \tag{21}$$

We can easily calculate the inner product $\langle p, f\rangle$, $\langle p, g_i\rangle$ in (19) as follows:

$$\langle p, f\rangle = L_{\tilde{f}}\tilde{V}(x) = \frac{\partial \tilde{V}}{\partial x}\tilde{f}(x), \tag{22}$$

$$\langle p, g_i\rangle = L_{\tilde{g}_i}\tilde{V}(x) = \frac{\partial \tilde{V}}{\partial x}\tilde{g}_i(x). \tag{23}$$

III. Problem statement

We consider a problem that a holonomic mobile robot autonomously moves to an unknown goal in a maze; the robot needs to explore the space, and learn the optimal path to the destination. In this paper, a lattice maze defined on a two-dimensional arc connected C^1 differentiable manifold $X \subset \mathbb{R}^2$ is considered.

The objective of the paper is to design the global CLF on the continuous space obtained by Q-learning in the discrete space.

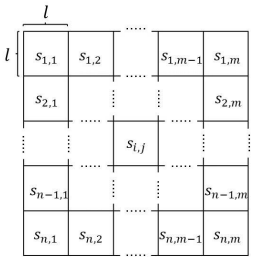

Fig. 1. Defined lattice maze

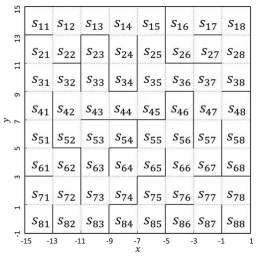

Fig. 2. Lattice Maze

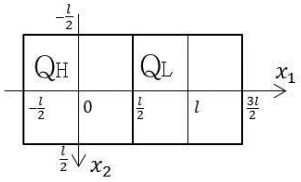

Fig. 3. Two lattices maze

A. Agent

Consider the holonomic mobile robot (agent) modeled by the following equation:

$$\dot{x} = u, \tag{24}$$

where $x \in X \subset \mathbb{R}^2$ is the state and $u \in U \subset \mathbb{R}^2$ is the input.

B. Maze definition

The maze is composed with some square l on a side sections (Fig.1). The $n \times m$ lattice maze composed n sections in column and m sections in row; the maze has sufficiently thin walls and the walls are set along each side of any square section. Assume the agent never touch the walls, then the lattice maze defined above is a two-dimensional C^1 differentiable manifold. The agent can know the movable direction by sensing the walls and knows only the initial state where the agent start to explore the lattice maze.

C. Q-learning setting

To apply Q-learning, we need to relate the lattice maze in the continuous space to the set S in the discrete space. Relate each section of the lattice maze in the continuous space to state $s_{ij} \in S$, and the lattice maze can be modeled as follows in the discrete space:

$$S := \{s_{ij} | i \in \{1, \cdots, n\}, j \in \{1, \cdots, m\}\} \tag{25}$$

The discrete space representing the lattice maze has Markov property and is a finite space. Let a set $A_{s_{ij}}$ be a set of selectable actions at state s_{ij}. The set $A_{s_{ij}}$ consists of all of the motions to the neighboring lattices (up, down, right and left).

We suppose that an initial value of $Q = 0$ for all states, we do not update the value of Q at the goal state. If the agent selects the action to move toward the wall, the agent obtains the reward $r = -1$. If the agent reaches the goal, a trial is finished, and the agent goes back to the initial state. After the exploring, we use the goal reward r_g as the value of Q at the goal.

IV. Locally semiconcave control Lyapunov function design with $Q(s, a)$

In this section, we design a locally semiconcave control Lyapunov function with the value of Q. We set the locally semiconcave control Lyapunov function that the value of the CLF at the middle of the each section of the lattice maze is $V_q(s_{ij})$ used $\max_a Q(s_{ij}, a)$.

A. Use of the value of Q

The converged value of Q has the following relation according to (2):

$$\max_a Q(s_{ij}, a) = R_t = \sum_{k=0}^{d-1} \gamma^k r_{t+k+1}, \tag{26}$$

where d is the number of the states that are including the shortest path from the initial state to the goal state; that is, if the agent is at the goal state, $d = 1$. If the agent is at the neighbor state of the goal, $d = 2$. $\max_a Q(s_{ij}, a)$ of the each state s_{ij} can be calculated from the discount return R_t. To simplify the discussion, the agent earns the positive rewards at the only goal state. Let the reward when the agent selects the action to reach the goal state be γr_g. Then, the value of Q is represented by the following equation with d defined above:

$$\max_a Q(s_{ij}, a) = r_g \gamma^{d-1}. \tag{27}$$

To design the global CLF with $V_q(s_{ij})$, the CLF need to have an unique equilibrium at the goal and $V_q(s_{ij}) = 0$ at the goal. $V_q(s_{ij})$ is set so that the value of the global CLF at the middle of the each section of the lattice maze changes as a quadratic function, and $V_q(s_{ij})$ is defined as follows:

$$V_q(s_{ij}) = (\log \max_a Q(s_{ij}, a)) - \log r_g)^2 \tag{28}$$

$$= \{(d-1) \log \gamma\}^2, \tag{29}$$

where $\log$ is a natural logarithm.

B. Lyapunov function design by the minimum projection method

In this subsection, we design a local Lyapunov function for two neighboring and movable sections of the lattice maze for the agent. Then, we design a global Lyapunov function by the minimum projection method.

Consider the rectangle $D := (-l/2, -l/2)\ (-l/2, l/2)\ (3l/2, -l/2)\ (3l/2, l/2)$ that is combined the two sections of the 1 on-a-side lattice maze. (Fig. 3) Let two $\max Q$ of the sections be Q_H and $Q_L (Q_H > Q_L)$, and consider the mapping $\phi : \mathbb{R}^2 \to D$:

$$\phi : \begin{cases} x_1 = \begin{cases} \frac{l}{\pi} \tan^{-1}(\frac{3}{V_{q1}} \tilde{x}_1) & (\tilde{x}_1 < 0) \\ \frac{3l}{\pi} \tan^{-1}(\frac{1}{V_{q1}} \tilde{x}_1) & (\tilde{x}_1 \geq 0) \end{cases} \\ x_2 = \frac{l}{\pi} \tan^{-1}(\frac{1}{V_{q1}} \tilde{x}_2), \end{cases} \tag{30}$$

where $\tilde{x}_1, \tilde{x}_2 \in \tilde{X} = \mathbb{R}^2$ and a constant V_{q1} is defined as follows:

$$V_{q1} = \frac{\log \gamma}{\tan \frac{3}{\pi}}. \tag{31}$$

Then, the following function is a control Lyapunov function on $\tilde{X} = \mathbb{R}^2$.

$$\tilde{V}(\tilde{x}_1, \tilde{x}_2) = \tilde{x}_1^2 + \tilde{x}_2^2 \tag{32}$$

Proof: Let the function $\tilde{X} = \tilde{x}_1^2 + \tilde{x}_2^2$ be a CLF candidate. $\tilde{V}$ is proper and positive definite. The system (4) on X can uniquely determine the corresponding system on $\tilde{X}$ as follows[8]:

$$\dot{\tilde{x}} = \left(\frac{\partial \phi}{\partial \tilde{x}}\right)^{-1} f\left(\phi(\tilde{x}), u\right). \tag{33}$$

Then,

$$\dot{\tilde{x}}_1 = \frac{\pi V_q}{l}\left\{1 + \left(\frac{\tilde{x}_2}{V_q}\right)^2\right\} u_1 \tag{34}$$

$$\dot{\tilde{x}}_2 = \begin{cases} \frac{\pi V_q}{3l}\left\{1 + \left(\frac{3\tilde{x}_1}{V_q}\right)^2\right\} u_2 & (\tilde{x}_1 < 0), \\ \frac{\pi V_q}{3l}\left\{1 + \left(\frac{\tilde{x}_1}{V_q}\right)^2\right\} u_2 & (0 \le \tilde{x}_1). \end{cases} \tag{35}$$

Consider the following input obtained by (19):

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{cases} \begin{bmatrix} -\frac{4\pi V_{q1}^2}{9l} \frac{\sin \frac{2\pi}{l} x_1}{(1+\cos \frac{2\pi}{l} x_1)^2} \\ -\frac{4\pi V_{q1}^2}{l} \frac{\sin \frac{2\pi}{l} x_2}{(1+\cos \frac{2\pi}{l} x_2)^2} \end{bmatrix} & (-\frac{l}{2} \le x_1 < 0), \\ \begin{bmatrix} -\frac{4\pi V_{q1}^2}{3l} \frac{\sin \frac{2\pi}{3l} x_1}{(1+\cos \frac{2\pi}{3l} x_1)^2} \\ -\frac{4\pi V_{q1}^2}{l} \frac{\sin \frac{2\pi}{l} x_2}{(1+\cos \frac{2\pi}{l} x_2)^2} \end{bmatrix} & (0 \le x_1 \le \frac{3l}{2}). \end{cases} \tag{36}$$

Then,

$$\frac{d\tilde{V}}{dt} = 2\tilde{x}_1 \dot{\tilde{x}}_1 + 2\tilde{x}_2 \dot{\tilde{x}}_2 < 0. \tag{37}$$

Therefore, $\tilde{V} = \tilde{x}_1^2 + \tilde{x}_2^2$ is a CLF on $\tilde{X}$. ■

We design a local Lyapunov function on D by substituting (30) into (32) as follows:

$$\tilde{V}(\phi^{-1}(x_1, x_2)) = \begin{cases} \frac{V_{q1}^2}{9} \tan^2(\frac{\pi}{l} x_1) + V_{q1}^2 \tan^2(\frac{\pi}{l} x_2) \\ \qquad (x_1 < 0) \\ V_{q1}^2 \tan^2(\frac{\pi}{3l} x_1) + V_{q1}^2 \tan^2(\frac{\pi}{l} x_2) \\ \qquad (x_1 \ge 0). \end{cases} \tag{38}$$

We extend the discussion from the neighboring, movable two lattice sections to the defined maze. We design the local Lyapunov functions among the each neighboring, movable two sections; then, we design the offset c_s for the each local Lyapunov function as follows:

$$c_s = (\log Q_L - \log r_g)^2 - (\log \gamma)^2. \tag{39}$$

The global Lyapunov function is designed as the following equation by the multilayer minimum projection method:

$$V(x) = \min_{\tilde{x} \in \phi^{-1}(x)} \hat{V}\left(\tilde{x}\right). \tag{40}$$

C. Confirmation of the validity of the offset c_s for the maze having one goal state

In this subsection, we confirm that the offset c_s satisfies condition (17).
The offset c_s is given by (39). Note that the difference of a local control Lyapunov function between the middle of two lattice sections is $(\log \gamma)^2$. Then, to satisfy (17), c_s should satisfy the following inequality:

$$c_s - V_q > 0. \tag{41}$$

(41) is calculated as follows:

$$\begin{aligned} c_s - V_q &= (\log Q_L - \log r_g)^2 - (\log \gamma)^2 - \{(p-1)\log \gamma\}^2 \\ &= \{(q-1)^2 - 1 - (p-1)^2)\}(\log \gamma)^2, \end{aligned} \tag{42}$$

where p, q is the number of the states that are included in the shortest path from the state having Q_H, Q_L to the goal state, $q \ge p+1$ and $p \ge 2$. Then, c_s satisfies (41) by (42). Thus, the offset c_s satisfies the condition (17).

V. Exploring simulation in 8×8 lattice maze

A. Lyapunov function design for 8×8 lattice maze having one goal state by minimum projection method

In this subsection, we show the effectiveness of the proposed method by computer simulation with an 8×8 lattice maze with $l = 2$(Fig. 2). Consider the following holonomic moving robot:

$$\begin{cases} \dot{x} = u_1 \\ \dot{y} = u_2 \end{cases}, \tag{43}$$

where $(x, y) \in \mathbb{R}^2$ is a state, $(u_1, u_2) \in \mathbb{R}^2$ is an input.

The parameters needed for updating the action-value function is collected as below.

- Learning rate α: 0.1
- Discount rateγ: 0.9
- Policy: ϵ-greedy (ϵ: 0.7)

The purpose of this subsection is to design the CLF affected by the learning result of Q-learning. First, we relate each lattice section in the 8×8 lattice maze to a state in the discrete space as Fig.2. Then, we apply the Q-learning as follows; the agent start to explore at the initial state $s_{11}(x = -14, y = 14)$, the goal state is $s_{88}(x = 0, y = 0)$, and the goal reward $r_g = 30$. Then, we obtained $\max Q$ as follows:

$$Q_{ij} = \begin{bmatrix} 5.559 & 6.177 & 6.863 & 7.626 & 8.473 & 7.626 & 8.473 & 9.414 \\ 5.003 & 5.559 & 6.177 & 6.863 & 9.414 & 8.473 & 9.414 & 10.46 \\ 5.559 & 6.177 & 6.863 & 6.177 & 10.46 & 11.62 & 12.91 & 11.62 \\ 5.003 & 6.863 & 7.626 & 8.473 & 9.414 & 15.94 & 14.35 & 12.91 \\ 5.559 & 7.626 & 8.473 & 7.626 & 15.94 & 17.72 & 15.94 & 11.62 \\ 6.177 & 6.863 & 9.414 & 12.91 & 14.35 & 19.68 & 17.72 & 10.46 \\ 6.863 & 7.626 & 10.46 & 11.62 & 19.68 & 21.87 & 24.30 & 27.00 \\ 6.177 & 8.473 & 9.414 & 15.94 & 17.72 & 19.68 & 21.87 & 30.00 \end{bmatrix}$$

Note that there exists no positive reward without the goal state. Then, the value of the Lyapunov function at each section of the lattice maze is as follows:

$$V_{s_{ij}} = \begin{bmatrix} 2.842 & 2.498 & 2.176 & 1.876 & 1.599 & 1.876 & 1.599 & 1.343 \\ 3.208 & 2.842 & 2.498 & 2.176 & 1.343 & 1.599 & 1.343 & 1.110 \\ 2.842 & 2.498 & 2.176 & 2.498 & 1.110 & 0.899 & 0.711 & 0.899 \\ 3.208 & 2.176 & 1.876 & 1.599 & 1.343 & 0.400 & 0.544 & 0.711 \\ 2.842 & 1.876 & 1.599 & 1.876 & 0.400 & 0.278 & 0.400 & 0.899 \\ 2.498 & 2.176 & 1.343 & 0.711 & 0.544 & 0.178 & 0.278 & 1.110 \\ 2.176 & 1.876 & 1.110 & 0.899 & 0.178 & 0.100 & 0.044 & 0.011 \\ 2.498 & 1.599 & 1.343 & 0.400 & 0.278 & 0.178 & 0.100 & 0 \end{bmatrix}.$$

The Lyapunov function for the lattice maze is designed as Fig.4.

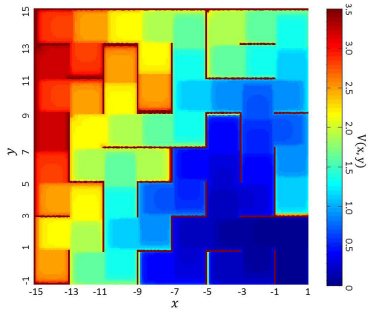

Fig. 4. Lyapunov Function

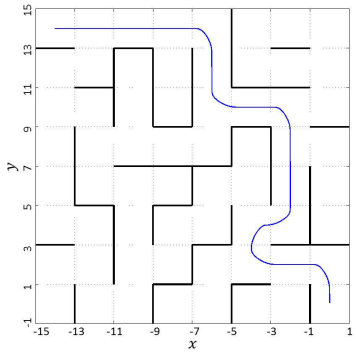

Fig. 5. Simulation

B. Determining the control input and the simulation result

In this subsection, we design the control input for s-stability at the goal state with the local semiconcave control Lyapunov function designed with the value of Q, and show the result of the simulation. The agent is modeled by (24). The control input for s-stabilization at the goal state is obtained as the following equation by (19):

$$u = \begin{cases} -\langle p, g_i \rangle & \left(\sum_{i=1}^{2} \langle p, g_i \rangle^2 \neq 0 \right) \\ 0 & \left(\sum_{i=1}^{2} \langle p, g_i \rangle^2 = 0 \right) \end{cases}, \tag{44}$$

where g_i is a unit matrix $E \in \mathbb{R}^{2\times 2}$, p is randomly selected in disassemble differentiable $\tilde{D}V$.

The simulation result with the control input (44) and the initial state (-14,14) is Fig.5.

The proposed method navigates the agent smoothly without hitting a wall from the start state to the goal state. The path of the simulation coincides with the path selected by the learning result with Q-learning. Thus, we can confirm that the learning result in the discrete space reflects the CLF in the continuous space.

In this paper, the agent does not obtain any positive reward without the goal state. The reason is no equilibrium except at the goal state as the discussion of the subsection IV-C.

VI. Conclusion

In this paper, we proposed the CLF design method with the learning result applying Q-learning on-line for the problem of the exploring a lattice maze. We confirm that it is possible to design the global CLF based on the learning result in the discrete space by multilayer minimum projection method. The learning result is included in the optimal path and goal selected by the agent with Q-learning.

Furthermore, we apply Q-learning on-line in this paper. However, convergence of Q is very slow. Thus, it is difficult to apply the proposed method on-line at present.

In the proposed method, there is not a positive reward without the goal state. It is because there is a possibility of stationary points on the path from the start state to the goal state if we consider the some rewards. we can not discuss the problem in this paper. If we solve that problem, we will be able to design the global CLF that is affected by more complex decision by the agent.

References

[1] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning An Introduction," The MIT Press Cambridge, Massachusetts London, England, 1998.

[2] Christopher J. C. H. Watkins and Peter Dayan, "Technical Note: Q-Learning, Machine Learning," 1992: 279 292.

[3] Hisakazu Nakamura, Yoshiro Fukui, Nami Nakamura and Hirokazu Nishitani, "Multilayer Minimum Projection Method with Singular Point Assignment for Nonsmooth Control Lyapunov Function Design," 49th IEEE Conference on Decision and Control(2010).

[4] Hisakazu Nakamura, Yoshiro Fukui, Nami Nakamura, Hirokazu Nishitani, "Multilayer minimum projection method for nonsmooth strict control Lyapunov function design," Systems and Control Letters 59 (2010) 563-570.

[5] Hisakazu Nakamura, Takayuki Tsuzuki, Yoshiro Fukui and Nami Nakamura, "Asymptotic Stabilization with Locally Semiconcave Control Lyapunov Function on General Manifold," Systems and Control Letters, submittfed.

[6] Hisakazu Nakamura, Takayuki Tsuzuki, "S-stabilization on Manifolds Locally Semiconcave Practical Control Lyapunov Function," 40th Control Theorem Symposium(2011) (in Japanese)

[7] Yoshiro Fukui, Hisakazu Nakamura, Hirokazu Nishitani, "search control on unknown field of two-wheeled mobile via minimum projection method," 39th Control Theorem Symposium, 141/144 (2010) (in Japanese).

[8] Hisakazu Nakamura, Yuh Yamashita, Hirokazu Nishitani "Minimum Projection Method for Nonsmooth Control Lyapunov Function Design on General Manifolds," Systems & Control Letters, Vol. 58, No. 10-11, pp. 716-723, 2009.

[9] M.Malisoff, M. Krichman and E. Sontag, "Global stabilization for systems evolving on manifolds," Journal of Dynamical and Control Systems, 12(2006)pp. 161-184.

[10] L. Rifford, Semiconcave control-Lyapunov functions and stabilizing feedbacks, SIAM Journal on Control and Optimization, 41(2002) pp.659-681.