Modeling Neuromodulated Synaptic Plasticity in Spiking Neural Network (文献紹介)

数理生命情報学研究室 修士 1 年 高野 成章
指導教員：河野 崇 教授

数理情報学輪講 秋学期 第 6 回

2018 年 11 月 8 日

概要

神経系における学習はシナプスの可塑性により実現され，代表的な学習則にシナプス前細胞と後細胞の発火に同時性に着目した Hebb 則がある．近年神経細胞の発火の同時性に加え，大脳基底核回路などで広範囲に分布する神経修飾物質のシナプスへの投射のタイミングにより可塑性が変化することが明らかになっており，強化信号を担うと考えられている．本稿では神経修飾物質によるシナプス可塑性への影響を Spiking Neural Network(SNN) 上でモデル化した研究 [1] を紹介する．まず最初に連続マルコフ決定過程における TD 誤差学習からスパイキングニューロンにおける学習則 (TD-LTP) を導出し，Actor-Critic 法により Navigation タスクを行なった実験を紹介する．

1 はじめに

1.1 Hebb 則について

神経系における学習は図 1.1 のようなニューロン間のシナプスの可塑性により実現され，代表的な学習則に Hebb 則が知られている．Hebb 則は一般的にシナプス前細胞と後細胞の発火の同時性によりそのシナプスの結合荷重を変化させる学習則である．

$$\Delta w = f(x_{pre}, x_{post}) \tag{1.1}$$

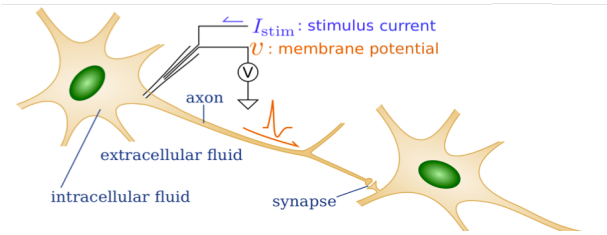


図 1.1: シナプス前細胞とシナプス後細胞の概要図．シナプス前細胞が発火した時シナプス小胞から神経伝達物質 (Neurotransmitter) が放出される．神経伝達物質はシナプス後細胞に到達した時にシナプス後細胞の電位に影響を与える．(図は [2] より)

Hebb 則のうち前細胞と後細胞の発火の前後関係に着目したものに STDP 則 (Spike Timing Dependent Plasticity) があり，代表的な STDP 則として次のような非対称型 STDP(図 1.2) が知られている：

- 前細胞が後細胞より先に発火した場合，シナプスの結合荷重は増加する (Long term potentiation, LTP).
- 一方で後細胞が発火してから前細胞が発火した時，結合荷重が減少する (Long term depression, LTD).

また STDP 則をはじめとした Hebb 則は (スパイク系列の) パターン認識など，ある種の教師なし学習を実現することが知られている．

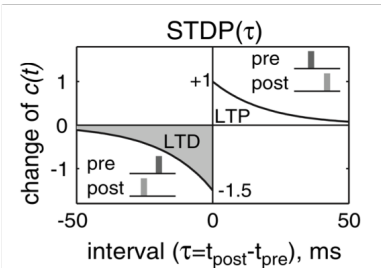


図 1.2: 非対称型 STDP 則の概要図．指数関数型の時間窓を採用している．t_{pre} はシナプス前細胞の最新の発火時刻．t_{post} はシナプス後細胞の最新の発火時刻に相当する．具体的な式は補遺 B.2 を参照．(図は [3] より)

近年シナプス前細胞と後細胞の発火の同時性に加えて，大脳基底核回路 (Basal Ganglia Circuit) や海馬 (Hippocampus) においてドーパミンなどの神経修飾物質 (Neuromodulator) がシナプス結合荷重に影響を与えることが実験的に知られている．神経修飾物質は一般に通常の神経伝達物質とは異なり周辺のシナプスに影響を与えることが知られており [4]，とくにドーパミンは報酬の情報を与えることで脳における強化学習を実現すると考えられている．したがって，これらの部位では神経修飾物質による影響を含んだシナプス学習則が考えられる．

$$\Delta w = f(x_{pre}, x_{post}, reward) \tag{1.2}$$

本稿では報酬学習において実験的に見られている性質を，強化学習のフレームワーク (TD 誤差学習) により定式化し，式 1.2 の具体的な形を導出する．導出された学習則には神経修飾物質による影響と考えられる項が含まれており，新たな生理学実験による検証項目を示唆することができると考えられる．

1.2 強化学習における基本的な用語

1.2.1 有限マルコフ決定過程 (finite MDP)

強化学習は機械学習の分野の一つであり, 最適制御の理論, Trial-Error の理論, Temporal Difference Error の理論などが統合され誕生した分野である．一般に強化学習の理論ではデータの生成源 (環境) に対して作用 (行動) することによって, 最適な行動規則 (方策) を学習する．正解は直接的には示されず, 報酬という形で行動の良し悪しの指標は環境から与えられる．具体的にはエージェントとの環境に対する作用によりエージェントは報酬を多くもらえるような行動を学習する．

強化学習の定式化については有限マルコフ決定過程 (MDP) を用いたフレームワークによるものが代表的である．MDP は次の集合により構成される．

- 状態 $s \in \mathcal{S}$
- 行動 $a \in \mathcal{A}(s)$
- 報酬 $r \in \mathcal{R}$

有限マルコフ決定過程 (MDP) とは，環境とエージェントの間のダイナミクスが次式で表されることを言う (式は図 1.3 のような設定)．

$$p\left(s', r|s, a\right) \coloneqq \Pr\left\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\right\} \tag{1.3}$$

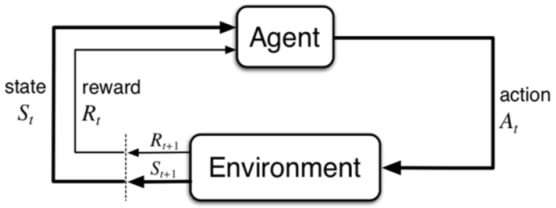


図 1.3: エージェントは環境に対し行動を行なった結果，エージェントは報酬を獲得し状態が遷移する．(図は [5] より)

MDP を仮定した時，状態と行動に対する報酬の期待値の関数は次のように定義することができる．

$$r(s, a) \coloneqq \mathbb{E}\left[R_t|S_{t-1} = s, A_{t-1} = a\right] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p\left(s', r|s, a\right) \tag{1.4}$$

ここで時刻 t 以降の累積報酬を G_t を考える．一般に未来に得られる報酬より現在の得られる報酬の方が優先されるため，割引率 $0 < \gamma \leqq 1$ を用いて，G_t を次のように定義する．

$$G_t \coloneqq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1.5}$$

次にエージェントの行動規則を表す方策について説明する．方策 π とはある状態で取りうる行動の確率である．例えば時刻 t で, $\pi(a(t), s(t)) := Pr(A(s) = a|s)$ とは状態 $S_t = s$ で行動 $A_t = a$ をとる確率を示す．

状態 s の状態価値関数とは, 状態 s からある方策 π を取った時に得られる累積報酬の期待値ことであり, 次のように表すことができる．

$$\begin{aligned} V^{\pi}(s) &\equiv \mathbb{E}_{\pi}\left[G_t|S_t = s\right] \\ &= \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\right] \\ &= R_{t+1} + \mathbb{E}_{\pi}\left[\gamma G_{t+1}|S_t = s\right] \\ &= R_{t+1} + \gamma V^{\pi}(s+1) \end{aligned} \tag{1.6}$$

同様に状態 s に加えてとる行動 a セットとしてまとめ，その価値を定義したものとして行動価値関数 q が定義できるが，本稿では明示的に使用しないこととする．

$$q_{\pi}(s, a) \coloneqq \mathbb{E}_{\pi}\left[G_t|S_t = s, A_t = a\right] = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right] \tag{1.7}$$

強化学習の目的は環境から得られる報酬を最大化することにある．そのために，状態の価値 (状態価値関数) を行動を通して学習し，またその状態に対する最適な行動 (方策) の学習を行う．

1.2.2 TD 誤差学習について

式 1.6 より真の状態価値関数では $V^\pi(s) = R_{t+1} + \gamma V^\pi(s+1)$ が成立し，この条件式を Self-Consistency equation と呼ぶ (現在の状態の価値関数は，遷移した先の状態の価値関数と遷移したことにより得られる報酬の和と等しい).

しかしエージェントは環境のダイナミクスに完全に知ることはできないため状態価値関数を近似的に与え，学習を行う必要がある．ここでパラメータ w を用いた関数 $V(x, \mathrm{w})$ により近似した時の Self-Consistency Equation に生じる誤差を TD 誤差 (Temporal Difference error) と呼ぶ.

$$\delta(t) \coloneqq R_{t+1} + \gamma V(s+1, \mathrm{w}) - V(s, \mathrm{w}) \tag{1.8}$$

学習を達成するために，Sarsa[6][7] や Q-Learning[8] など様々なアルゴリズムが知られているが，本稿では以降の節で述べるように Actor-Critic 法による学習を採用する[*1].

2 連続マルコフ決定過程 (MDP)

前述した STDP 則のように神経系における情報処理においてタイムスケールは重要な概念である．特に学習のメカニズムを理解するためにはニューロンの発火の時刻と報酬が与えられるまでの時間差を生物学的に妥当な値を用いたい．例えば神経細胞の発火はミリ秒の単位で起こるのに対し，その発火に対し報酬が得られるのは数秒後であることが一般的である．この時なぜ神経系は報酬が得られるような神経活動を強化することができるのかといった問題が存在する．このような理由により時間・状態を連続化し (連続 MDP の定式化 [9])，スパイキングニューロンによる学習を考える．

また，ここでは具体的に MDP の状態として二次元の物理空間を想定する．すなわち状態 S の要素は, 時刻 t における位置 $\mathbf{x}(t) \in \mathbb{R}^2$ である．同様に行動 A の要素も $\mathbf{a}(t) \in \mathbb{R}^2$ であるとする．環境のダイナミクスを表す関数 f を導入し次のように状態の遷移に関する定式化が行える．

$$\dot{\mathbf{x}}(t) = f(\mathbf{a}(t), \mathbf{x}(t)) \tag{2.1}$$

連続 MDP における方策 π は次のように定義する．

$$\pi(\mathbf{a}(t), \mathbf{x}(t)) := p(\mathbf{a}(t)|\mathbf{x}(t)) \tag{2.2}$$

エージェントの目標は次の価値関数 $V^\pi(\mathbf{x}(t))$ を最大化することである.

$$V^\pi(\mathbf{x}(t)) := \left\langle \int_t^\infty r\left(\mathbf{x}^\pi(s), \mathbf{a}^\pi(s)\right) e^{-\frac{(s-t)}{\tau_r}} \mathrm{d}s \right\rangle_{\mathbf{x}^\pi, \mathbf{a}^\pi} \tag{2.3}$$

ここで τ_r は報酬率 r の時定数に相当し，MDP における割引率と対応する.

式 2.3 を時間 t に関して微分すると次の等式が得られる.

[*1] Q-Learning をスパイキングニューロンモデルを用いて実装するにはより多くの仮定が必要であるため.

$$\dot{V}^{\pi}(\mathbf{x}(t)) - \frac{1}{\tau_r}V^{\pi}(\mathbf{x}(t)) + r(\mathbf{x}(t), \mathbf{a}(t)) = 0 \tag{2.4}$$

前節と同様にエージェントは環境の関する情報である $r(\mathbf{x}, \mathbf{a})$ と f について知らないため，$V^{\pi}(\mathbf{x}(t))$ を直接計算することはできない．したがって，$V^{\pi}(\mathbf{x}(t))$ を近似するための関数 $V(\mathbf{x}(t), \mathbf{w})$ を与え，パラメータ $\mathbf{w}$ を学習するようなアルゴリズムを考える．近似により生じる誤差は,

$$\delta(t) := \dot{V}(\mathbf{x}(t), \mathbf{w}) - \frac{1}{\tau_r}V(\mathbf{x}(t), \mathbf{w}) + r(\mathbf{x}(t), \mathbf{a}(t)) \tag{2.5}$$

であり，これを連続 MDP における TD 誤差と呼ぶ[*2] ．

また本稿では $\delta(t)$ の最小化を直接行わず，状態価値関数の損失関数 E の最小化を考える．

$$E(t) = \left[V^{\pi}(\mathbf{x}(t)) - V(\mathbf{x}(t), \mathbf{w})\right]^2 \tag{2.6}$$

この損失関数を減少させる方向の $\mathbf{w}$ の更新は勾配法により次のようにして与える．

$$\dot{\mathbf{w}} = \eta \left[V^{\pi}(\mathbf{x}(t)) - V(\mathbf{x}(t), \mathbf{w})\right] \nabla_{\mathbf{w}} V(\mathbf{x}(t), \mathbf{w}) \tag{2.7}$$

η を学習率と呼ぶ．
ここで式 2.4 を用いて,

$$\dot{\mathbf{w}} = \eta \left[\dot{V}^{\pi}(\mathbf{x}(t)) + r(\mathbf{x}(t), \mathbf{a}(t)) - \frac{1}{\tau_r}V(\mathbf{x}(t), \mathbf{w})\right] \nabla_{\mathbf{w}} V(\mathbf{x}(t), \mathbf{w}) \tag{2.8}$$

なお計算途中で τ_r をくくり出し $\tau_r\eta$ を改めて η と書き直した．

最後に,

$$\dot{V}^{\pi}(\mathbf{x}(t)) \approx \dot{V}(\mathbf{x}(t), \mathbf{w}) \tag{2.9}$$

を仮定し式 2.5 を用いることで次の更新則が得られる．

$$\dot{\mathbf{w}} \approx \eta\delta(t)\nabla_{\mathbf{w}} V(\mathbf{x}(t), \mathbf{w}) \tag{2.10}$$

式 2.9 については理論的な保証はないが，導かれた式 2.10 は $\delta(t)$ が減少する方向に更新するため結果的に有効な更新則だと考えることにする[*3].

3　スパイキングニューロンによる TD 誤差学習の実現

学習のフレームワークとして，アクター・クリティックネットワークを考える．アクター・クリティックネットワークでは エージェントが方策と状態価値関数を分けて学習する．本稿ではアクター・クリティックはそれぞれ神経細胞の集合により構成され, それぞれがスパイキングニューロンモデル (補遺 A) である (=Spiking Neural Network(SNN) の一種)．

[*2] 連続 MDP における $\delta(t)$ は時刻 t のみに対応する項であり, 時刻 $t'(\neq t)$ とは無関係である．時刻 t での情報が正しく時刻 t' に伝播するためには,V は連続かつ微分可能である必要がある．

[*3] $\delta(t) > 0$ の場合を考える．$\nabla_{\mathbf{w}} V(\mathbf{x}(t), \mathbf{w}) > 0$ の時 $\mathbf{w}$ は増加する方向に進む．V が増加すると式 2.5 により,$\delta(t)$ は減少する．$\delta(t)$ が減少すること次の $\mathbf{w}$ の更新の大きさが小さくなる．$\delta(t) < 0$ の場合も同様に更新の大きさが小さくなり，各パラメータを適切に設定することで収束すると考えられる．

- アクターニューロン：方策 π を学習する
- クリティックニューロン：状態価値関数 $V(\mathbf{x})$ を学習する

具体的な構成は図 3.1 のようなものを考える．

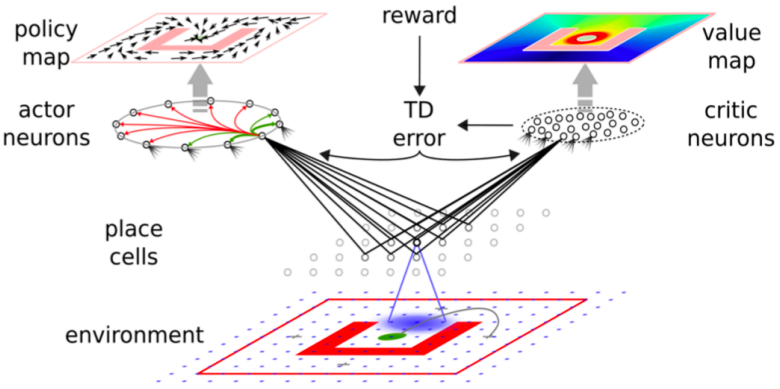


図 3.1: Actor-Critic Network の概要．場所細胞からアクターニューロン，クリティックニューロンそれぞれにフィードフォーワードで繋がっている．後述 (節 3.1.2) で説明するようにエージェントの位置情報は場所細胞 (Place Cell) の発火頻度によりエンコードされる．クリティクニューロンは入力された位置に対する価値を学習し (Value map の作成), アクターニューロンはエージェントの動く方向を学習する (Policy map の作成).

3.1 クリティックニューロンのモデル化

3.1.1 学習則の導出

クリティックニューロンの実装にあたり，まず最初に状態価値関数をスパイキングニューロンの発火率により表現すると仮定する．したがって，多くの累積報酬が得られる状態であるほど, クリティックニューロンはたくさん発火すると考える．

$$V(\mathbf{x}(t)) := \nu\rho(t) + V_0 \tag{3.1}$$

また，ここでの発火率 $\rho(t)$ の定義として，時間平均およびニューロン数による平均を考える．時間平均 $\rho(t)$ は次のようにして定義する．まず, 各々のニューロンの発火率を ρ_i として次のようにして定義する．

$$\rho_i(t) = \int_{-\infty}^{\infty} Y_i(s)\kappa(t-s)\mathrm{d}s \equiv (Y_i \circ \kappa)(t) \tag{3.2}$$

ただし $Y_i(t)$ はニューロン i のスパイク列であり, 発火時刻 $t_i^{(f)}$ を用いて $Y_i(t) = \sum_{t_i^{(f)} \in \mathcal{F}} \delta_D\left(t - t_i^{(f)}\right)$

κ はスパイク数の平均を取る窓関数 (フィルター，カーネル）であり $\kappa(t) := \frac{e^{\frac{-t}{\tau_k}} - e^{\frac{-t}{v_k}}}{\tau_\kappa - v_\kappa}$ として与える．

式 3.2, クリティックに相当するニューロン数を N_{critic} としてニューロン数の平均を合わせる．

$$V(\mathbf{x}(t)) := \frac{\nu}{N_{critic}} \sum_{i=1}^{N_{critic}} \rho_i(t) + V_0 \tag{3.3}$$

したがって TD 誤差を次のように表すことができる.

$$\delta(t) = \frac{\nu}{N_{critic}} \sum_{i=1}^{N_{critic}} \left(Y_{i^\circ} \left[\dot{\kappa} - \frac{\kappa}{\tau_r} \right] \right)(t) - \frac{V_0}{\tau_r} + r(\mathbf{x}(t), \mathbf{a}(t)) \tag{3.4}$$

(途中で $\dot{\rho}_i(t) = (Y_i \circ \dot{\kappa})(t)$ を用いた.)

次に上記の TD 誤差の最小化を式 2.9 により行う. ここでのパラメータ $\mathbf{w}$ は場所細胞ニューロン j からクリティックニューロン i へのシナプス強度 w_{ij} である.

$$\dot{w}_{ij} = \eta \delta(t) \frac{\partial V(\mathbf{x}(t))}{\partial w_{ij}} \tag{3.5}$$

ここで,

$$\frac{\partial V(\mathbf{x}(t))}{\partial w_{ij}} = \frac{\nu}{N_{critic}} \frac{\partial \rho_i(t)}{\partial w_{ij}} = \frac{\nu}{N_{critic}} \left(\frac{\partial Y_i}{\partial w_{ij}} \circ \kappa \right)(t) \tag{3.6}$$

が成立する.

次に $\frac{\partial Y_i(t)}{\partial w_{ij}}$ は直接計算できないため, $\frac{\partial \langle Y_i(t) \rangle_{Y|X,\hat{t}_i}}{\partial w_{ij}}$ の計算を行う. すなわちニューロン i が発したスパイクを, ニューロン i への入力スパイクとニューロン i の最後の発火時刻 $\hat{t}_i$ で条件づけた平均を計算する.

$$\begin{aligned} &\frac{\partial \langle Y_i(t) \rangle_{Y_i|\mathbf{X},\hat{t}_i}}{\partial w_{ij}} \approx \frac{\partial \tilde{\rho}_i(t)}{\partial w_{ij}} {}^{*4} = \frac{\partial g\left(u_i(t), \hat{t}_i\right)}{\partial w_{ij}} \\ &= \frac{1}{\Delta u} g\left(u_i(t), \hat{t}_i\right) \left(X_j^{\hat{t}_i} \circ \varepsilon \right) \\ &= \frac{1}{\Delta u} \left\langle Y_i \left(X_j^{\hat{t}_i} \circ \varepsilon \right) \right\rangle_{Y_i|X,\hat{t}_i} \end{aligned} \tag{3.7}$$

以上より,

$$\frac{\partial V(\mathbf{x}(t))}{\partial w_{ij}} = \frac{\nu}{N_{critic} \Delta u} \left(\left[Y_i \left(X_j^{\hat{t}_i} \circ \varepsilon \right) \right] \circ \kappa \right)(t) \tag{3.8}$$

が導出された.

式 3.5, 3.8 より,

$$\dot{w}_{ij} = \tilde{\eta} \delta(t) \left(\left[Y_i \cdot \left(X_j^{\hat{t}_i} \circ \varepsilon \right) \right] \circ \frac{\kappa}{\tau_r} \right)(t), \quad \tilde{\eta} = \frac{\eta \nu}{N_{critic} \Delta u} \tag{3.9}$$

式 3.9 は, ニューロンの接続に依存しない項 $\delta(t)$ と, ニューロンの接続に関係する項 $\left(\left[Y_i \cdot \left(X_j^{\hat{t}_i} \circ \varepsilon \right) \right] \circ \frac{\kappa}{\tau_r} \right)(t)$ で分けることができる ($\tilde{\eta}$ は新しく定義した学習率). ニューロンの接続に依存しない $\delta(t)$ は, 広範囲に分布する神経修飾物質 (Neuromodulator) との対応が考えられる.

ニューロンの接続に関係する $\left[Y_i \cdot \left(X_j^{\hat{t}_i} \circ \varepsilon \right) \right] \circ \frac{\kappa}{\tau_r}(t)$ は, シナプス前細胞の発火により生じた興奮性シナプス後電位 (EPSP)$(X_j^{\hat{t}_i} \circ \varepsilon)$ とシナプス後細胞の発火のタイミングによる関数であり Hebb 則に対応する.

---

*4 $\left\langle \int_t^{t+\Delta t} Y_i(t')\,\mathrm{d}t' \right\rangle_{\mathrm{Y}_i|\mathrm{X}} = \int_t^{t+\Delta t} \langle Y_i(t') \rangle_{\mathrm{Y}_i|\mathrm{X}}\,\mathrm{d}t' = \int_t^{t+\Delta t} \tilde{\rho}_i(t')\,\mathrm{d}t' \approx \tilde{\rho}_i(t)\Delta t$
よって $\Delta t \to 0$ で $\langle Y_i(t) \rangle_{Y_i|\mathbf{x}} = \tilde{\rho}_i(t)$ が成り立つ [10]. ここでの発火率は単純化のため, そのニューロンの過去の発火の履歴を無視する. またここでの導出は一部 [1] とは異なる.

実際に,

$$Y_i \cdot (X_j^{\hat{t}_i} \circ \varepsilon) = \sum_{t_i^{(f)} \in \mathcal{F}_i} \delta_D \left(t - t_i^{(f)} \right) \sum_{t_j \in \mathcal{G}_j^{\hat{t}_i}} \epsilon(t - t_j) \tag{3.10}$$

であり ($\mathcal{G}_j^{\hat{t}_i}$ は $\hat{t}_i$ 以降の X の発火時刻の集合とする), シナプス後細胞 Y が発火した時に正の値をとることから LTP であることがわかる. 最後に κ は過去の Hebb 則の値を畳み込むことから, Hebb 則が起こったシナプスの履歴 (Eligibility Trace) の項として機能すると考えることができる. 以上の性質を踏まえ, 式 3.9 を TD-LTP と呼ぶ.

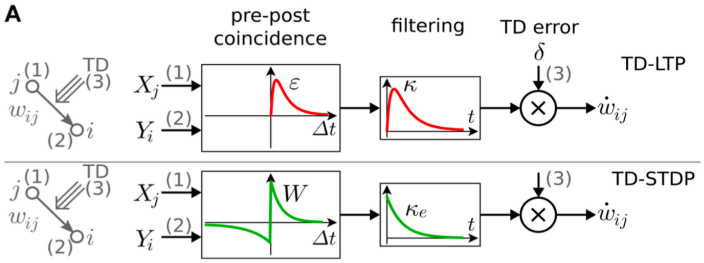


図 3.2: (上)TD-LTP(式 3.9) の概略図. (下)TD-STDP の概略図. どちらも, シナプス前細胞の発火による EPSP を調べ, そのあとにシナプス後細胞の発火時刻との近さを計算している. 計算結果は κ により, 状態価値関数フィルターされる

3.1.2 Linear Track Simulation

クリティックニューロンの振る舞いを調べるため，アクターニューロンを実装する前に次のような実験を行う. なお本文で紹介する実験の詳細な設定, およびその他紹介しきれなかった数値実験については [1] を参照にされたい.

まず最初に場所細胞のモデル化を行う. j 番目の場所細胞 (place cell) は特定の座標 $\mathbf{x}_j$ で強く発火すると仮定する. さらに場所細胞の発火はポアソン過程に従うと仮定すると j 番目の場所細胞の発火率を式 3.11 として表すことができる.

$$\rho_j(\mathbf{x}(t)) = \rho_{PC} \exp\left(-\frac{\|\mathbf{x}(t) - \mathbf{x}_j\|^2}{\sigma_{PC}^2} \right) \tag{3.11}$$

ただし $\mathbf{x}(t)$ は時刻 t におけるエージェントの位置, ρ_{pc}, σ_{pc} は場所細胞の発火率を特徴付ける定数である.

ここで次のようなケースを考える. エージェントは原点から直線上に一定の速度で進み, ある地点で報酬を得る. 報酬を得た後は原点に戻り, シナプスの結合荷重は変えずにこの試行を数回繰り返す. ここでのケースは方策がない状況，あるいは方策が固定されている状況と捉えることができる.

報酬獲得時刻を t_r とすると, $t < t_r$ で $\dot{V}(\mathbf{x}(t), \mathbf{w}) - \frac{1}{\tau_r} V(\mathbf{x}(t), \mathbf{w}) = 0$ であり, 実際に図 3.3 の左図に見られるように $V(\mathbf{x}(t)) \propto e^{-\frac{t - t_r}{\tau_r}}$ に収束していることが見られる. 一方で右図の TD 誤差の時間変化はノイズによる影響を大きく受けていることが見られるものの，最終的に 0 に収束することは確認された. 以上より，上記の設定においてクリティックニューロンは状態価値関数を正しく学習することが確認された.

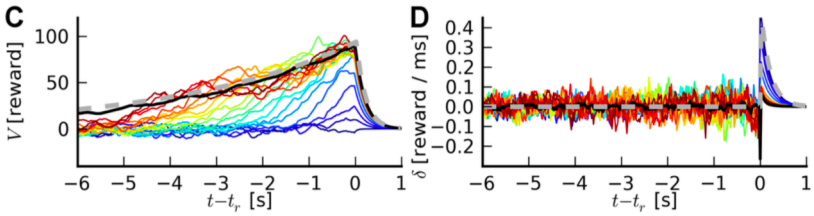


図 3.3: Linear Track Simulation を行なった結果． (左図) 状態価値関数． 各試行回数は濃い青から始まり，赤色までで合計 20 回，黒線は 30 回から 50 回を平均した結果である．グレーの点線は TD 誤差が 0 に収束した時の理論線である．(右図)TD の値の時間変化．各試行回数は濃い青から赤色の線に対応する．試行回数が増えるにつれて TD 誤差が 0 に近づいていることが確認できる．

3.2 アクターニューロンのモデル化

次にアクターニューロンを実装する．アクターニューロンは前節のクリティックニューロンと同様の学習規則に従うとを仮定する．すなわち，

$$\dot{w}_{ij} = \tilde{\eta}\delta(t)\left(\left[Y_i \cdot \left(X_j^{\hat{t}_i} \circ \varepsilon\right)\right] \circ \frac{\kappa}{\tau_r}\right)(t), \quad \tilde{\eta} = \frac{\eta\nu}{N_{actor}\Delta u} \tag{3.12}$$

ただしここでの Y はアクターニューロンの発火系列に対応する．また $\delta(t)$ はクリティックにより計算されたものを利用する (式 3.4)．アクターニューロンとクリティックニューロンは直接つながっていないが, TD 誤差により状態の価値が間接的に共有される．まず $\delta(t) > 0$ の時を考える．このときクリティックニューロンは報酬を予測していないのにかかわらず, 報酬を獲得した時を状況を表す．この時クリティックニューロンと同様に $\delta(t)$ により場所細胞とアクターニューロンの結合が増加する．すなわち, 今後同じ状態で同じアクターニューロンが活動するようになる．逆に $\delta(t) < 0$ ではシナプスの結合荷重が減少し, 他のアクターニューロンが活動する機会が与えられる．

図 3.4 のような Navigation タスクにおいては $\mathbf{a}_k$ は各アクターニューロン k がエンコードする方向ベクトルに相当する．

$$\mathbf{a}_k = a_0\left(\sin\left(\theta_k\right), \cos\left(\theta_k\right)\right)^T \quad \left(\theta_k = \frac{2k\pi}{N_{actor}},\ \text{for } k = 1, \ldots, N_{actor}\right) \tag{3.13}$$

時刻 t におけるエージェントの行動 $a(t)$ はアクターニューロンの発火率 ρ_k で重みづけたものによって計算されると考える．

$$\mathbf{a}(t) = \frac{1}{Z_a(t)}\sum_k \rho_k(t)\mathbf{a}_k \tag{3.14}$$

なお発火率はクリティックニューロンと同様にエンコードする．

$$\rho_k(t) := \left(Y_k \circ \gamma\right)(t) \quad \gamma(t) := \frac{e^{\frac{-t}{\tau_\gamma}} - e^{\frac{-t}{v_\gamma}}}{\tau_\gamma - v_\gamma}\Theta(t) \tag{3.15}$$

またアクターニューロンでは N-Winner-take-all の機構を考える．N-winner-take-all とは発火したニューロンと類似するニューロンには興奮性刺激により発火させ，それ以外のニューロンは抑制する仕組みを有するネットワークである．具体的には，あるアクターニューロンが発火した時そのニューロンはそれがエンコード

する方向と近い方向をエンコードするアクターニューロンに興奮性刺激に与え，異なる方向には抑制性の刺激を与えることで，アクターニューロンの集団として一つの似たような方向を与えることになる．

これを実現するアクターニューロン間のシナプス w_{kk} の実装は次のようにして行う．

$$w_{kk'} = \frac{w_-}{N_{actor}} + w_+ \frac{f_{nav}\left(k, k'\right)}{Z_k^f} \tag{3.16}$$

(Z_k^f は規格化定数である．$w_+ > 0$，$w_- < 0$(定数))

$$f_{nav}\left(k, k'\right) = \left(1 - \delta_{kk'}\right) \exp\left(\zeta \cos\left(\theta_k - \theta_{k'}\right)\right) \tag{3.17}$$

以上の手法によりエンコードされたアクター・クリティックネットワークは正しく Navigation タスクを行う（図 3.4 はタスクを行うメカニズムの概要，図 3.5 は実際のシミュレーションを行った結果).

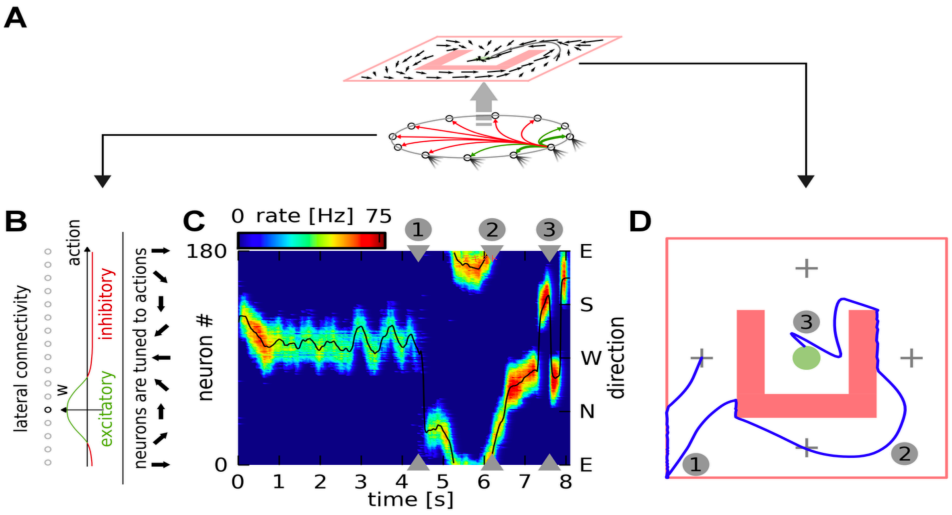


図 3.4: アクターニューロンネットワークを追加した上で行なった Navigation タスク．(A) アクターニューロンが学習した Policy map, (B) アクターニューロン間で N-winnter-take all 結合がされている．(C) 活動 (発火) しているアクターニューロン．(D) 図 (C) に対応するエージェントの動き．

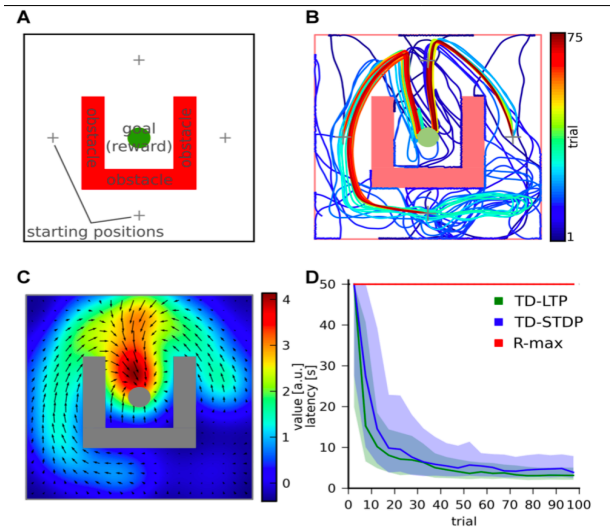


図 3.5: (A) Navigation タスク (具体的には Morris の water maze 実験を模した設定)．Trial 毎にエージェントは異なる初期位置 (4 点) からスタートする．(B) シミュレーション結果．75 回試行を繰り返した時のエージェントの軌跡を表す (C) シナプス結合強度を用いて作成した Value Map と Policy Map (D) ゴールに到達するまでの時間 (TD-STDP, R-MAX による学習則と比較)．

4　結論

本稿で紹介した論文の新規性はスパイキングニューロンによる TD 学習のフレームワークを構築し, 連続状態・時間におけるシナプスの学習則を導出したことにある．過去に神経修飾物質による TD 誤差の実現は指摘されてきたが今回の研究により，神経修飾物質が関わっている際の Hebb 則を予想することができた．したがって今後粒度の高い生理学的な検証実験を示唆することが可能になると考えられる．

本稿では神経修飾物質によるシナプス荷重への作用のメカニズムを明らかにすることに重点が置かれており, 提案されたアクター・クリティックネットワーク自体については脳における具体的な部位との直接的な対応は目的とされていない．したがって，場所細胞については海馬がエンコードしていることが明らかになっているように，今後は脳の具体的な部位とマッピングした詳細なネットワークに拡張できると考えられる．

また論文で導出されたアルゴリズムの理論解析は不十分だと考えられる．今後は学習率などのハイパーパラメータがどのように学習に影響を及ぼすか, そしてそれが脳においてどのように実現されるかについて深める必要があると考える．

補遺 A　Spike Response Model (SRM および SRM_0)

SRM[11] ではニューロン i の状態は膜電位 u のみで表される．ニューロンの入力に対して, ϵ は

$$\begin{aligned} u_{\mathrm{i}}(t) =& \eta\left(t-\hat{t}_i\right)+\sum_j w_{\mathrm{ij}} \sum_j \epsilon_{ij}\left(t-\hat{t}_i, t-t_{\mathrm{j}}^{(f)}\right) \\ &+\int_0^{\infty} \kappa_{ij}\left(t-\hat{t}_i, s\right) I_i^{ext}(t-s) \mathrm{d}s \end{aligned} \tag{補遺 A.1}$$

w_{ij} はニューロン j からニューロン i へのシナプス結合荷重を示す．ϵ は興奮性シナプス後電位 (EPSP)[*5]のカーネルである．すなわち, 入力に対する応答を表す項である．η はニューロンの膜電位が閾値を超えた時のふるまいを表すカーネルである．最後に，κ は外部入力の電流 I_i^{ext} に対する応答を表すカーネルである．$\hat{t}_i$, $\hat{t}_j^f$ はニューロン i およびニューロン j の最後に発火した時刻である．

本冊子では SRM を単純化したモデル SRM_0 を使用する．SRM_0 ではニューロンの i の膜電位は次の式で表される．

$$u_i(t)=\sum_j w_{ij} \sum_{t_j^{(f)} \in \mathcal{F}_j, t_j^{(f)}>\hat{t}_i} \varepsilon\left(t-t_j^{(f)}\right)+\chi \Theta\left(t-\hat{t}_i\right) \exp\left(\frac{\hat{t}_i-t}{\tau_m}\right) \tag{補遺 A.2}$$

$$\varepsilon(s)=\frac{\varepsilon_0}{\tau_m-\tau_s}\left(e^{\frac{-s}{\tau_m}}-e^{\frac{-s}{\tau_s}}\right) \Theta(s) \tag{補遺 A.3}$$

また SRM_0 の発火は膜電位 u に依存する Inhomogenous ポアソン過程でありニューロンの平均発火率 $\tilde{\rho}$ は,

$$\tilde{\rho}_i(t)=g\left(u_i(t)\right)=\rho_0 \exp\left(\frac{u_i(t)-\theta}{\Delta u}\right) \tag{補遺 A.4}$$

として定義する ($\Delta u \to 0$ で LIF モデルと等価).

補遺 B　その他のモデル (Neuromodulated STDP[3])

Izhikevich は Eligibility Trace を明示的にモデル化したメカニズムを提案し，大脳皮質の一部を再現したネットワークモデル上でシミュレーションを行った．

シナプス結合荷重の更新則は次のように設定されている．

$$\dot{C}=-C/\tau_c+\mathrm{STDP}(\Delta\tau)\delta_D\left(t-t_{\mathrm{pre/post}}\right) \tag{補遺 B.1}$$

ここでの STDP は非対称型を仮定する．(図 1.2 を参照)

$$STDP(\Delta\tau)=\begin{cases} a^+ \cdot \exp\left(\frac{-\Delta\tau}{\tau^+}\right) & \text{if } \Delta\tau>0 \quad \text{(LTP)} \\ -a^- \cdot \exp\left(\frac{\Delta\tau}{\tau^-}\right) & \text{if} \quad \Delta\tau<0 \quad \text{(LTD)} \end{cases} \tag{補遺 B.2}$$

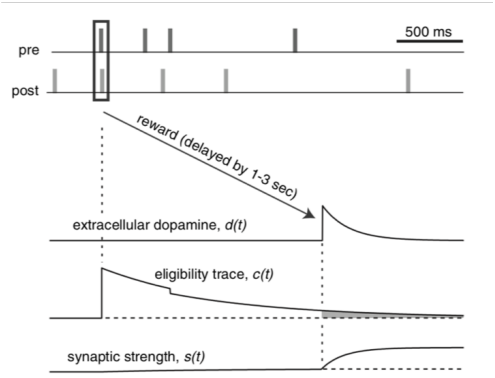
一方で，広範囲に投射されるドーパミン (DA) は次の式に従うとする．

$$\dot{D}=-D/\tau_d+\mathrm{DA}(t) \tag{補遺 B.3}$$

[*5] 抑制性シナプス電位 (IPSP) でも同様に表されるがここでは考慮しない．

$$\dot{w} = CD \tag{補遺 B.4}$$

以上の式を解釈すると，STDP が起こった時直接シナプス結合を変化させるのではなく，まず Eligibility Trace としてその情報をマークする．そして指数関数的に減少する Eligibility Trace に対し，その時間窓の中でドーパミンが到達すればそのシナプス結合が増強することになる．



図補遺 B.1: [3] で提案されているシナプス結合荷重の更新則．図中の s は本稿での w と対応し，c と d はそれぞれ C, D に対応する．([3] より引用)

[3] では，このメカニズムを仮定したネットワークモデルにより，古典的条件付け (Classical Conditioning), 道具的条件付け (Instrumental Conditioning)，無条件刺激 (Unconditioned Stimulus) から条件刺激 (Conditioned Stimulus) への反応のシフトなどが再現できることが明らかにされている．以下に発表資料で紹介した実験におけるパラメータを載せる．

表補遺 B.1: 共通のパラメータ

各ニューロンの発火率	τ_d	τ_c	DA(t) (tonic)	Δ DA(t) (per DA spike)
1Hz	0.2 s	1.0 s	0.01 $\mu M/s$	0.5 μM

表補遺 B.2: Shift of DA Response from US to Reward-Predicting CS in Classical Conditioning(実験 4)

DA(t) (per VTA_p spike)
$0.004\mu M/s$

参考文献

[1] N. Frémaux, H. Sprekeler, and W. Gerstner, "Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons," *PLoS Computational Biology*, vol. 9, no. 4, 2013.

[2] T. Kohno, "SIGNAL TRANSMITTION IN NEURONS," 2011.

[3] E. M. Izhikevich, "Solving the distal reward problem through linkage of STDP and dopamine signaling," *Cerebral Cortex*, vol. 17, no. 10, pp. 2443–2452, 2007.

[4] E. Marder and V. Thirumalai, "Cellular, synaptic and network effects of neuromodulation," *Neural Networks*, vol. 15, no. 4-6, pp. 479–493, 2002.

[5] R. S. Sutton and A. G. Barto, "Reinforcement Learning : An Introduction; Second edition, in progress," p. 426, 2018.

[6] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems.* PhD thesis, Cambridge University, 1994.

[7] Richard S. Sutton, "Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding," *Advances in Neural Information Processing Systems 8 (NIPS 1995)*, 1996.

[8] C. J. C. H. Watkins, *Learning from Delayed Rewards.* Ph.d. thesis, University of Cambridge, 1989.

[9] K. Doya, "Reinforcement learning in continuous time and space," *Neural Computation*, vol. 12, no. 1, pp. 219–245, 2000.

[10] N. Fremaux, H. Sprekeler, and W. Gerstner, "Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity," *Journal of Neuroscience*, vol. 30, no. 40, pp. 13326–13337, 2010.

[11] W. Gerstner and W. M. Kistler, *Spiking Neuron Models Single Neurons, Populations, Plasticity.* Cambridge University Press, 2002.