

階層強化学習から 重層強化学習へ —大脳基底核学習モデル

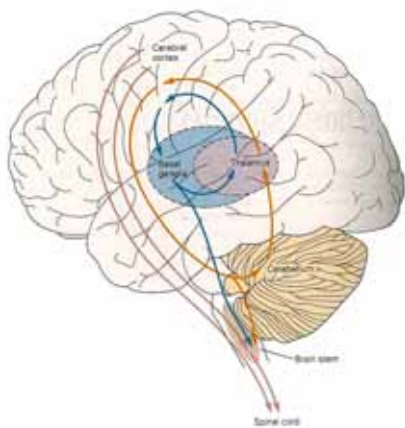
ATR脳情報研究所

川人光男

ヒト強化学習は大脳基底核で行われている

- 「報酬」を最大化するような行動を、探索により学習
 - 環境に応じて異なる最適行動を獲得
 - 目標出力がわからない問題に適用可
 - 人間や動物の行動学習のモデル
- 応用例; バックギャモン、ロボット
- 長期間にわたる報酬の累積を最大化するように行動を学習する: 現在と未来の報酬をどう評価するか

Figure 43-1 The relationships of the basal ganglia to the major components of the motor system. The basal ganglia and the cerebellum may be viewed as key elements in two parallel reentrant systems that receive input from and return their influences to the cerebral cortex through discrete and separate portions of the ventrolateral thalamus. They also influence the brain stem and, ultimately, spinal mechanisms.



Basal Ganglia

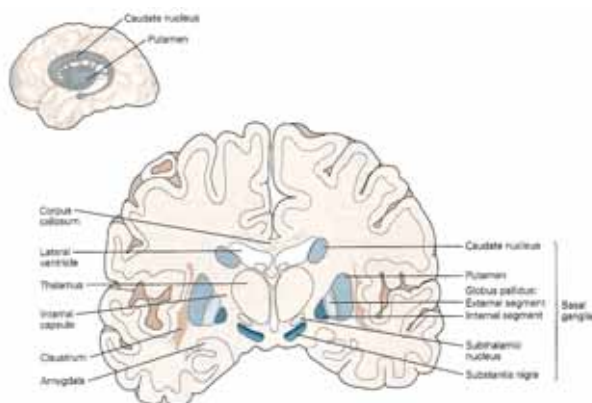
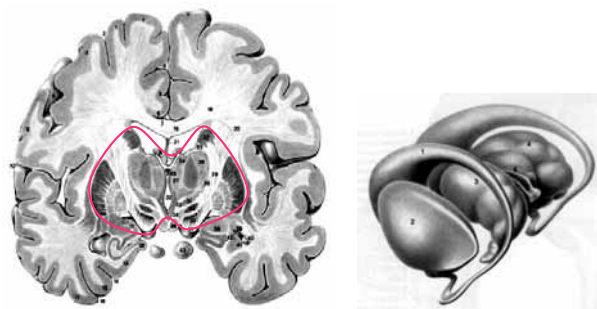
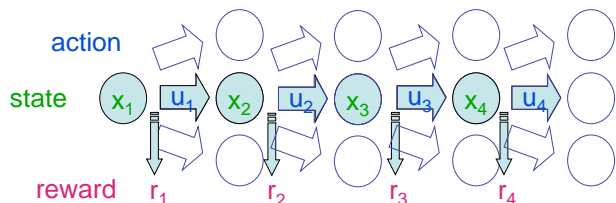


Figure 43-2 This coronal section shows the basal ganglia in relation to surrounding structures. (Adapted from Susswirth et al. 1981.)

「強化学習」 reinforcement learning

- 「報酬」を最大化するような行動を、探索により学習
 - 環境に応じて異なる最適行動を獲得
 - 目標出力がわからない問題に適用可
 - 人間や動物の行動学習のモデル
- 応用例
 - ゲームプログラム: バックギャモン, オセロ, ...
 - ロボット制御: 移動ロボット, サッカー, ...
 - 動的資源配分: 携帯電話チャンネル割り当て, ...

報酬が遅れを持って与えられる場合



—報酬の直前の行動だけを考えたのでは不十分

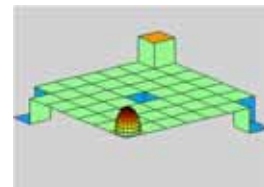
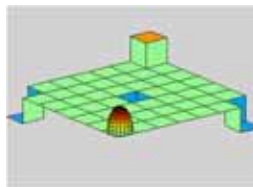
- 途中の状態の評価が必要

—状態の評価関数

$$V(x(t)) = E[r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \dots | u = g(x)]$$

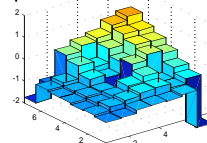
Example: Navigation

— Reward field

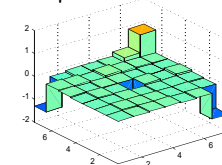


Value function

$\gamma=0.95$



$\gamma=0.3$



動的計画法と強化学習

- 動的計画法: 既知の環境, オフライン
 - 最適行動の条件: Bellman 方程式

$$V^*(x) = \max_u E[r(x(t), u) + \gamma V^*(x(t+1))]$$

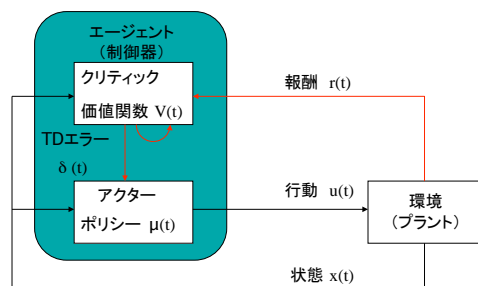
- 強化学習: 未知の環境, オンライン
 - 学習信号: TD誤差

$$\delta(t) = r(t) + \gamma V(x(t+1)) - V(x(t))$$

$$V(x(t)) := (1 - \alpha)V(x(t)) + \alpha \delta(t)$$

$$Q(x(t), u(t)) := (1 - \alpha)Q(x(t), u(t)) + \alpha \delta(t)$$

単純な強化学習



— 将来の累積報酬を最大化するようなポリシーを見つける

- 状態と行動の関数としての価値関数を学習する

TD誤差による報酬予測と行動の学習

— 状態価値関数

$$V(x(t)) = E[r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \dots]$$

— 行動価値関数

- 行動選択 $Q(x(t), u) = E[r(x(t), u) + \gamma V(x(t+1))]$

$$u(t) = \arg \max_u Q(x(t), u)$$

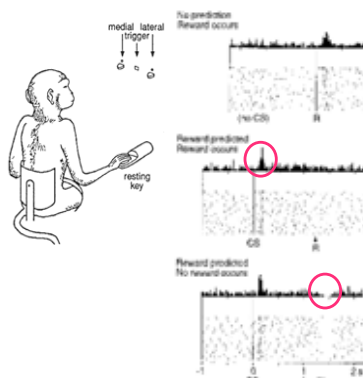
— TD誤差

- 予測の更新 $\delta(t) = r(t) + \gamma V(x(t+1)) - V(x(t))$

$$\Delta V(x(t)) = \alpha \delta(t)$$

$$\Delta Q(x(t), u(t)) = \alpha \delta(t)$$

ドーパミンニューロンが報酬予測を表現 (Schultz et al. 1993)

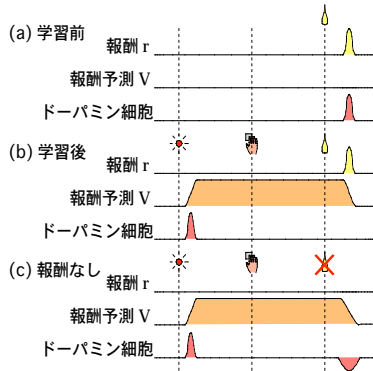


予期しない報酬

報酬を予告する刺激

報酬の欠如

ドーパミン細胞の報酬予測応答 (Schultz et al. 1993)

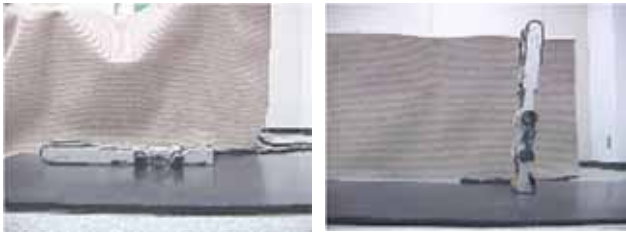


階層モジュール強化学習

- 単純な強化学習は現実的な問題ではとても遅いので、内部モデル、モジュール性、階層性などが必要（例、モデルに基づく階層強化学習、強化学習モザイクなど）
- 強化学習の行動の変容を引き起こすシナプス可塑性は脳基底核線状体で起きているのか？

階層強化学習による ロボットの起き上がり

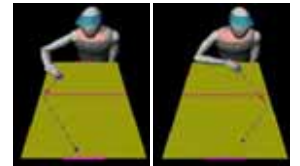
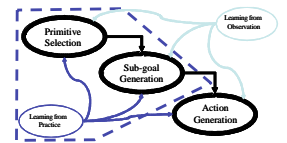
森本淳、銅谷賢治



Morimoto J. and Doya K.: Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, 36, 37-51 (2001)

強化学習・見まね・熟練学習 エアホッケー

- Learn appropriate actions and sub-goals for the observed situation.
 - Database initialized with supervised data; observes human player.
 - Actions: Right bank shot, left bank shot, etc.
- Learn by adjusting the distance to the query point within the database.
 - Data is retrieved using locally weighted learning (LWL) techniques.
 - Weights are updated using Q learning techniques.
 - Agent receives feedback (reward and penalty) while playing.

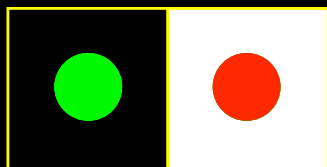


エアホッケー

Darrin C. Bentivegna (darrin@atr.jp)

Third Trial

-5yen

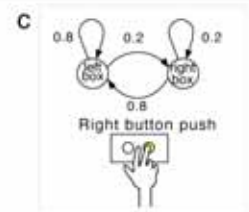
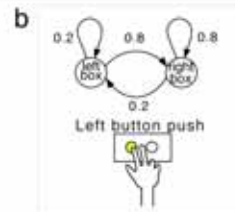
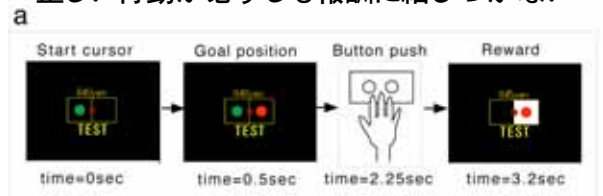


TEST



確率遷移規則

正しい行動が必ずしも報酬に結びつかない



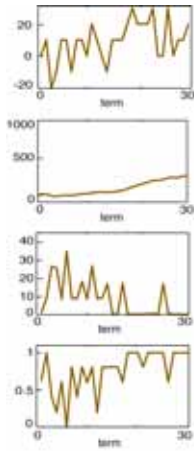
•短期報酬 (SR): 6回のボタン押し(1ターム)で得られた報酬の額

•累積報酬 (AR): それぞれの遷移規則に対して学習のはじめから今まで得られた累計の報酬額

•学習速度インデックス (LRI): 2つの隣接するターム間での行動変化の指数

•学習収束インデックス (LCI): 現在と最後の行動の間の距離から測った、最適な行動にどれほど近いかの指数

Dominant probability=0.8



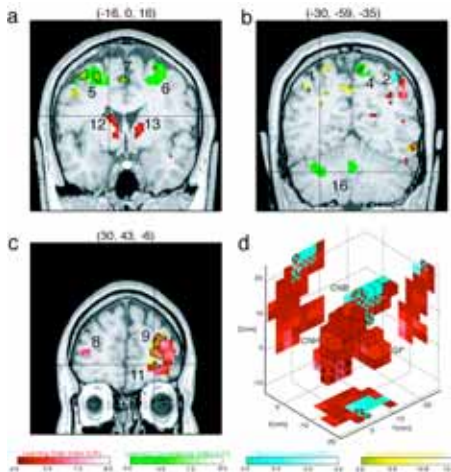
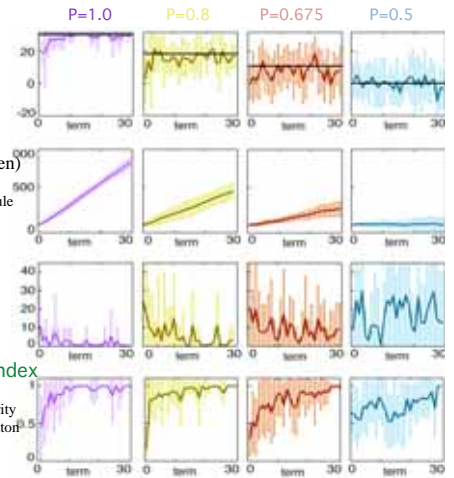
被験者平均

Short-term reward (yen)
the amount of money obtained in 6 button pushes (one term)

Accumulated Reward (yen)
the amount of money accumulated from the start of each transition rule to the current term

Learning rate index
the amount of behavioral changes between two adjacent terms

Learning convergence index
the degree of learning convergence measured as the normalized similarity between the current and final button push strategy



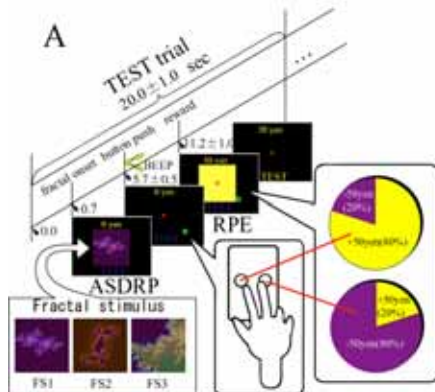
明らかになったモジュール構造

- 尾状核
強化学習の主要な座で、
短期報酬に導かれる (LRI and SR)
- 前頭眼窩野と前頭前野
長期に渡る累積報酬 (AR)
- 小脳、運動前野、SMA
内部モデルの格納、最適行動の記憶と実行 (LCI)

Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, Imamizu H, Kawato M:
A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *Journal of Neuroscience* 24, 1660-1665 (2004)

強化学習の脳活動

選択した行動とフラクタル図形に応じた報酬の予測
予測した報酬の狂い(誤差)



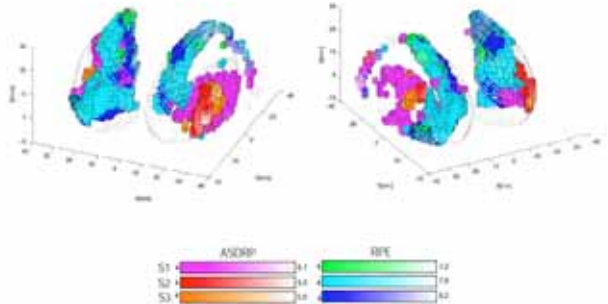
Action-State Value Function estimated by Q-algorithm, reproducing Behavioral Decisions

- Action-stimulus dependent reward prediction (ASDRP) estimated by Q-learning algorithm from subjects decisions and real rewards
- Behaviors were well reproduced by Q-model: 92% (SD21), 85% (SD32), 73% (SD42)
- Learning coefficient decayed with trials (RLS)

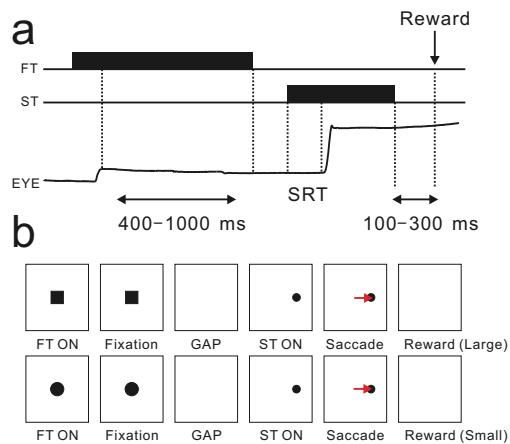
$$Q(fs, bp) = Q(fs, bp) + \alpha_t^{fs} (r - Q(fs, bp))$$

大脳基底核の中で行動に依存した報酬の予測と 予測誤差が別々に表現されている

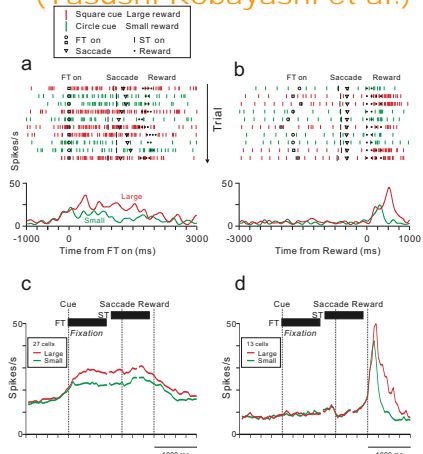
被核 = 行動状態価値関数
尾状核 = 報酬予測誤差



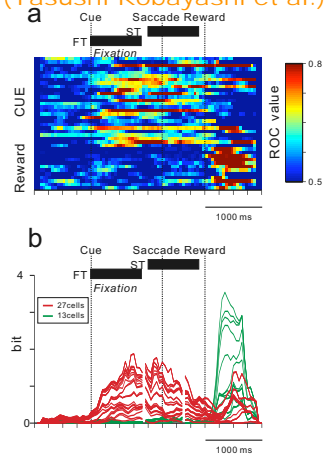
Task (Yasushi Kobayashi et al.)



Two Populations of Neurons (Yasushi Kobayashi et al.)

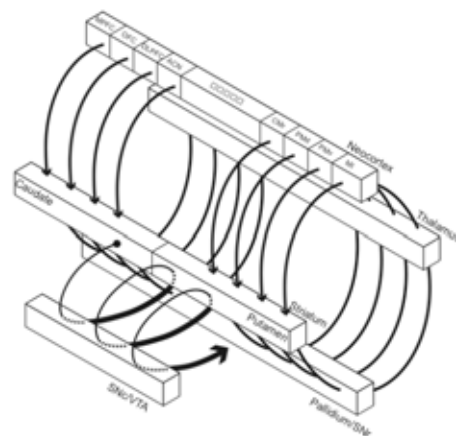
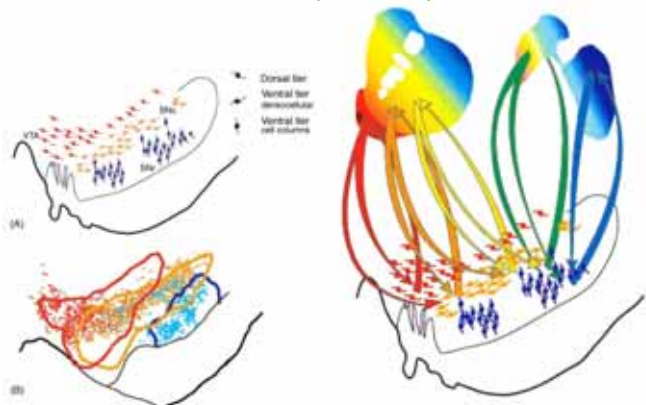


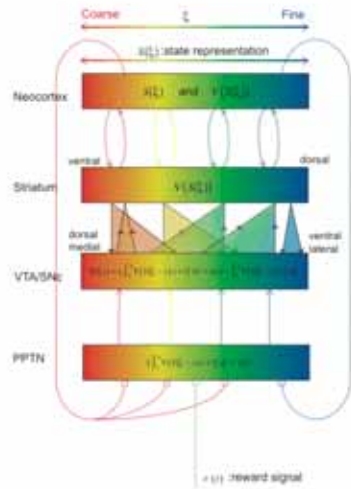
ROC and Information Quantity (Yasushi Kobayashi et al.)



Spiral Connections between Striatum and VTA/SNc

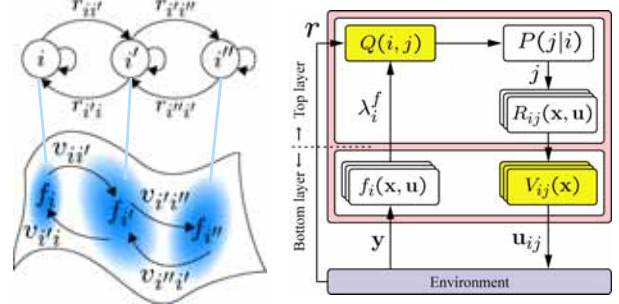
Haber et al. (2000,2003)



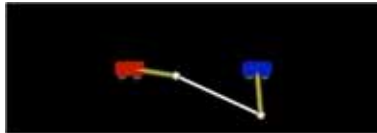


HMM-RL MOSAIC

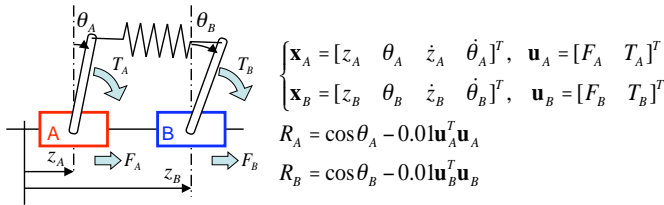
Upper layer: Q-learning in semi-Markov dynamics of symbols which correspond to forward models
 Bottom layer: continuous RL with value-function approximation derived from sub-goals, which are upper-layer actions



Collaboration with/without Symbol

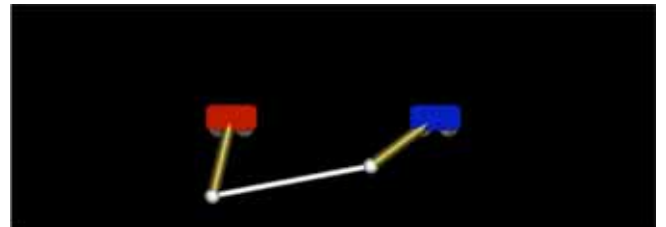


- Two agents (cart-poles) need to collaborate because pole tips are connected by a spring
- Inference of intension of other agent is necessary
- With and without direct symbol communication



Collaboration when Two Agents possess Different Symbols

- Because of different brain structures, symbol sets are different between two agents.
- From previous inference-learning through movement observations, two symbols sets were corresponded.



Prisoner's Dilemma with Computer Agents

individual differences and forward-model based RL

Subject chose to cooperate or defect by a button press and obtained rewards based on the payoff matrix.

Subject's task is to learn the strategy for agent A and B to maximize their rewards.

Explicit use of agent's strategy is essential for A, while B is identical to probabilistic instrumental conditioning.

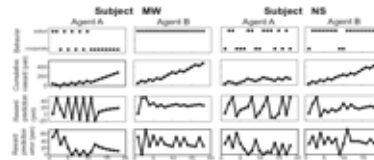


Haruno and Kawato (in preparation)

Subject Behaviors

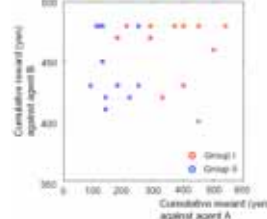
Behaviors and reward prediction

Subject groups (from n=32)



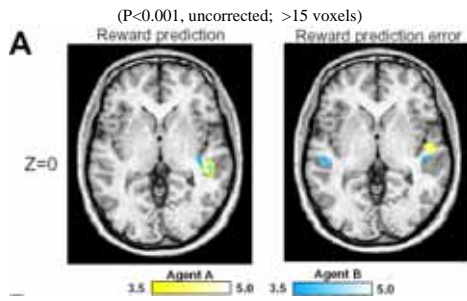
Group1 (8 males and 4 females) learned the optima for both A and B.
 Group2 (8 males and 4 females) learned the optima for only B.

Cumulative reward



Group 1 and Group 2 are different only against agent A ($P < 0.001$), but not against B.

Differential Correlation with Reward Prediction and Reward Prediction Error



- The STS shows significant difference in correlation with reward prediction and reward prediction error.
- This difference persisted against B as well as A.