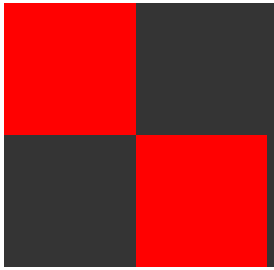
2003年3月18日



強化学習チュートリアル

松井藤五郎

名古屋工業大学 知能情報システム学科 世木研究室 博士後期課程 3 年
東京理科大学 理工学部 経営工学科 助手（4月から）

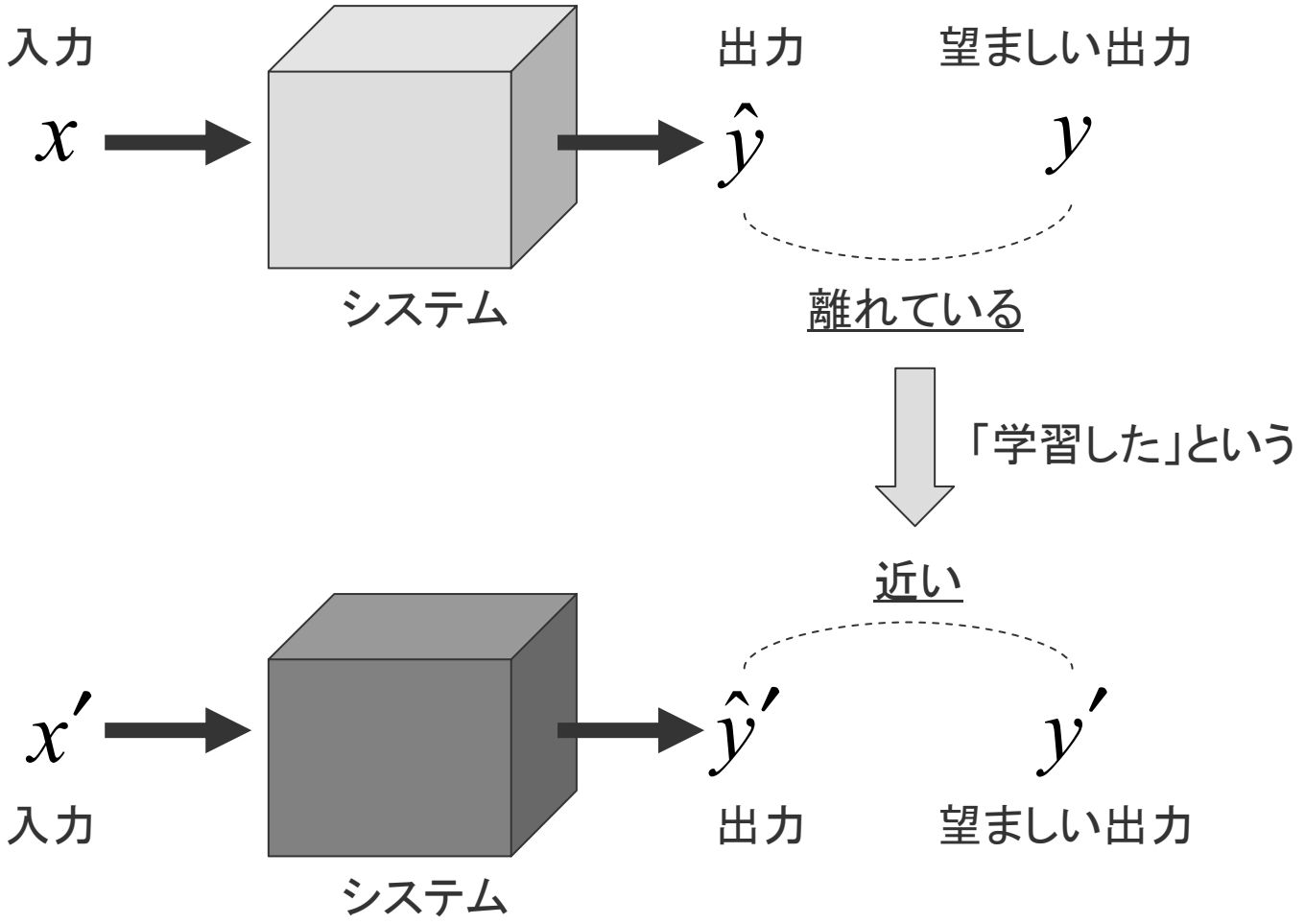
内容

- 強化学習とは
- 勉強の進め方
- これまでに提案された代表的な手法
- 私が提案した手法とそのプログラム
- 現在流行しているテーマ
- 実ロボットへの応用とその課題
- 強化学習に限らない研究の進め方

強化学習とは

機械学習

- 「経験から振る舞いを自動的に改善すること」



機械学習の例

PlayTennis: テニス大会の前日に,「翌日の大会が開催されるかどうか」を予測する問題

年度	予報	気温	湿度	風	開催
1989	晴れ	高い	高い	弱い	*No*
1990	晴れ	高い	高い	強い	*No*
1991	曇り	高い	高い	弱い	*Yes*
1992	雨	普通	高い	弱い	*Yes*
1993	雨	低い	普通	弱い	*Yes*
1994	雨	低い	普通	強い	*No*
1995	曇り	低い	普通	強い	*Yes*
1996	晴れ	普通	高い	弱い	*No*
1997	晴れ	低い	普通	弱い	*Yes*
1999	雨	普通	普通	強い	*Yes*
1999	晴れ	普通	普通	強い	*Yes*
2000	曇り	普通	高い	強い	*Yes*
2001	曇り	高い	普通	弱い	*Yes*
2002	雨	普通	高い	強い	*No*
2003	晴れ	低い	高い	弱い	?

分類

訓練事例＝経験

← 予測したい事例

入力 x

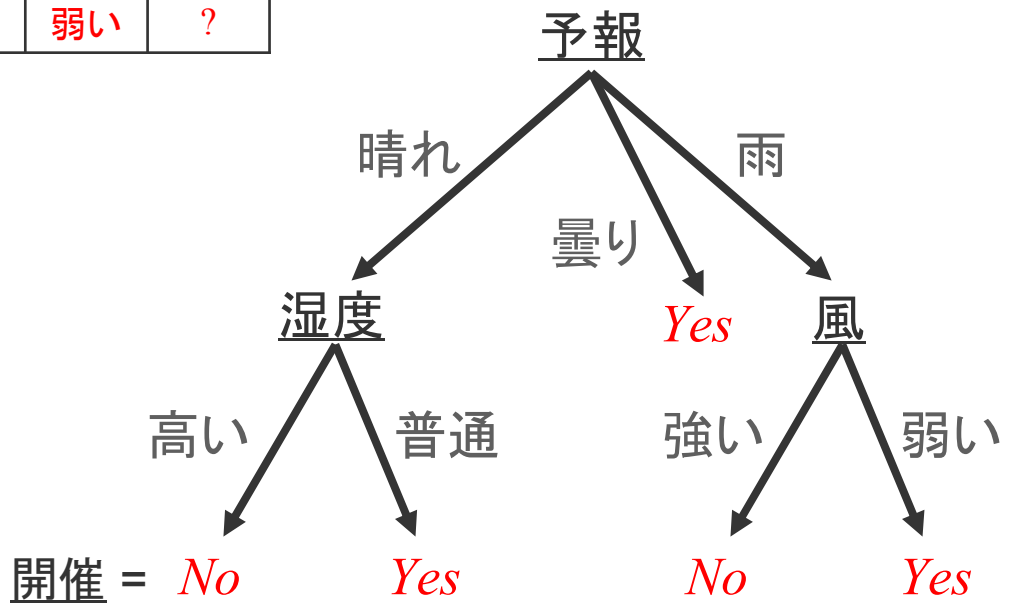
出力 y

学習する規則の例

- 決定木
  - 事例をいくつかの「カテゴリ」に分類する関数を表現

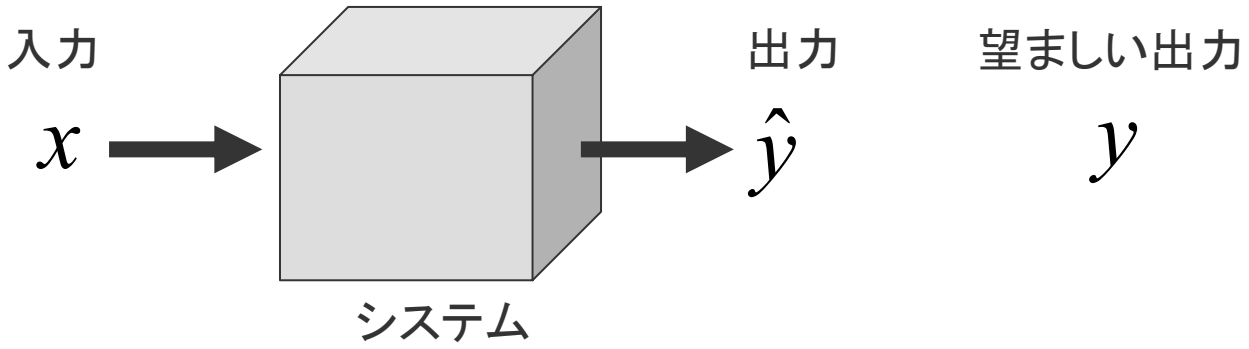
年度	予報	気温	湿度	風	開催
2003	晴れ	低い	高い	弱い	?

学習した決定木の例



教師つき学習と教師なし学習

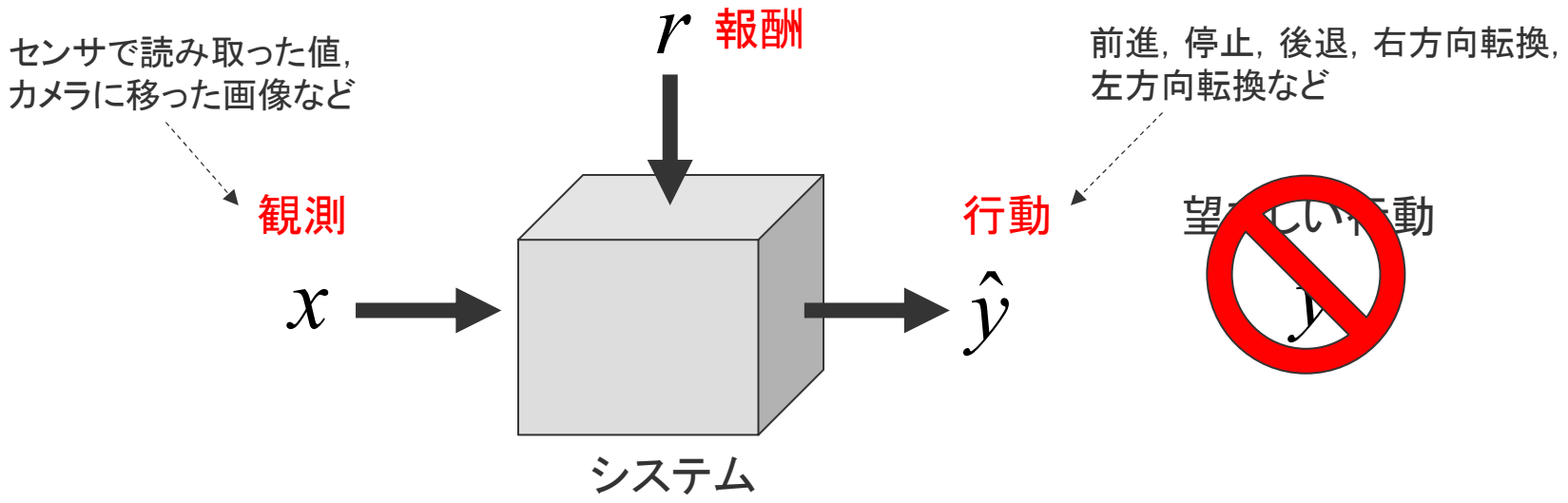
- 教師つき学習
  - 教師がいて，入力 x に対する望ましい出力 y を教えてくれる
  - 決定木学習，ILP，ニューラルネットワーク など



- 教師なし学習
  - 入力 x に対する望ましい出力 y を誰も教えてくれない
  - 強化学習，クラスタリング，GA など

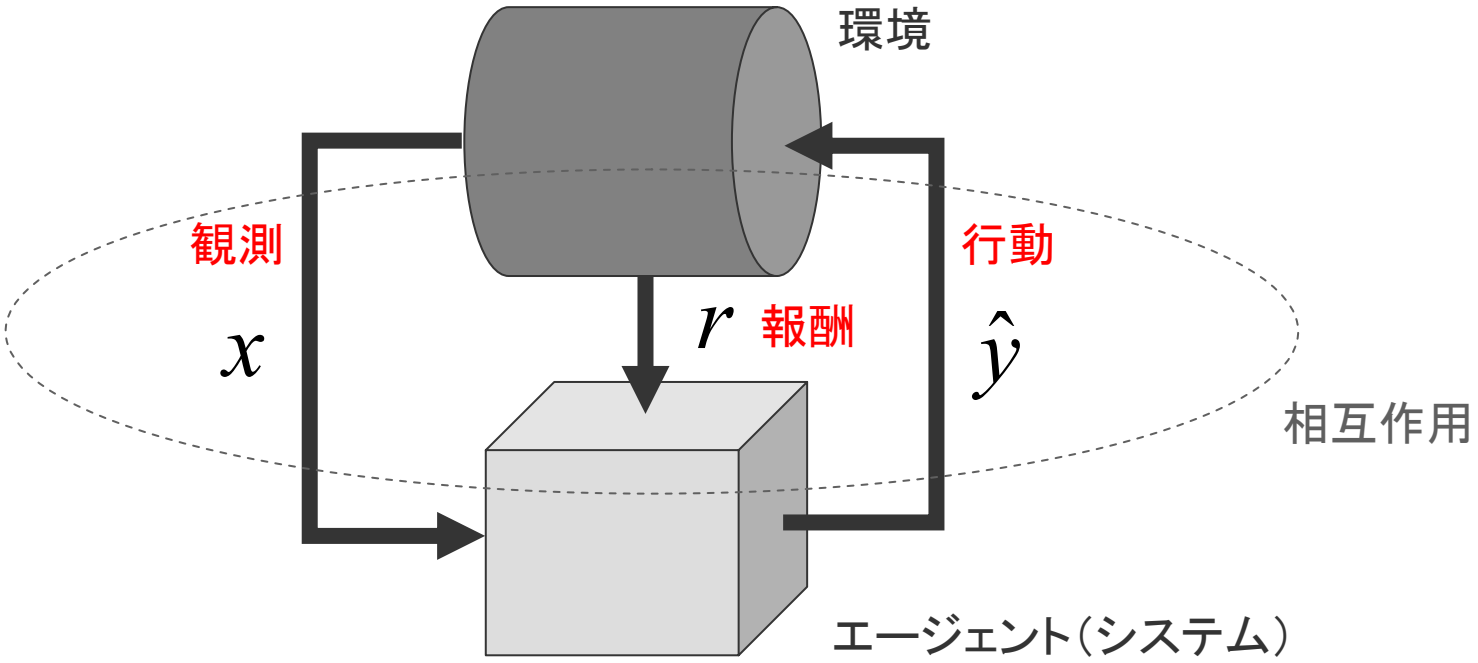
強化学習

- 「観測」が入力され，「行動」を出力
- 教師なし学習
  - 「望ましい行動」はわからない
- 「報酬」が与えられる
  - 「どの時点の行動に対する報酬か」はわからない



エージェントと環境

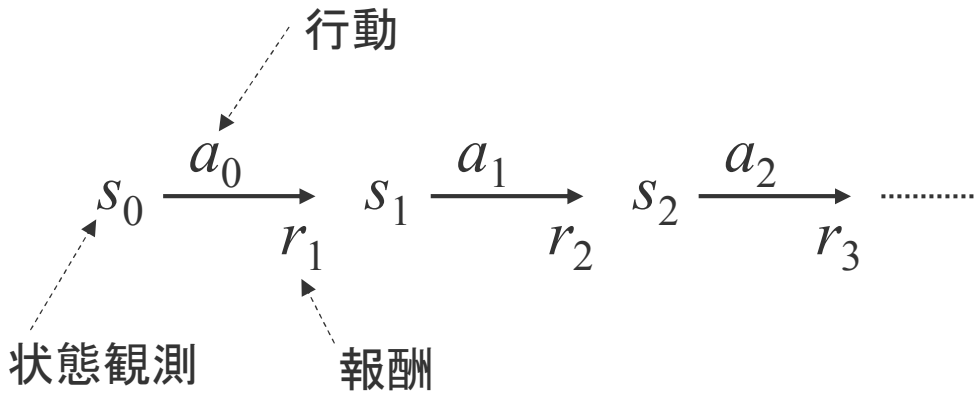
- エージェントは，「環境」と「相互作用」を行なう



- エージェントが自由に変更できないものを「環境」と呼ぶ

強化学習の枠組み

- 環境との相互作用を繰り返し行なう



- 目標
  - ある評価尺度 M を最大化する行動選択戦略の学習
  - (Sutton & Barto, 1998) では，M は期待割引収益

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

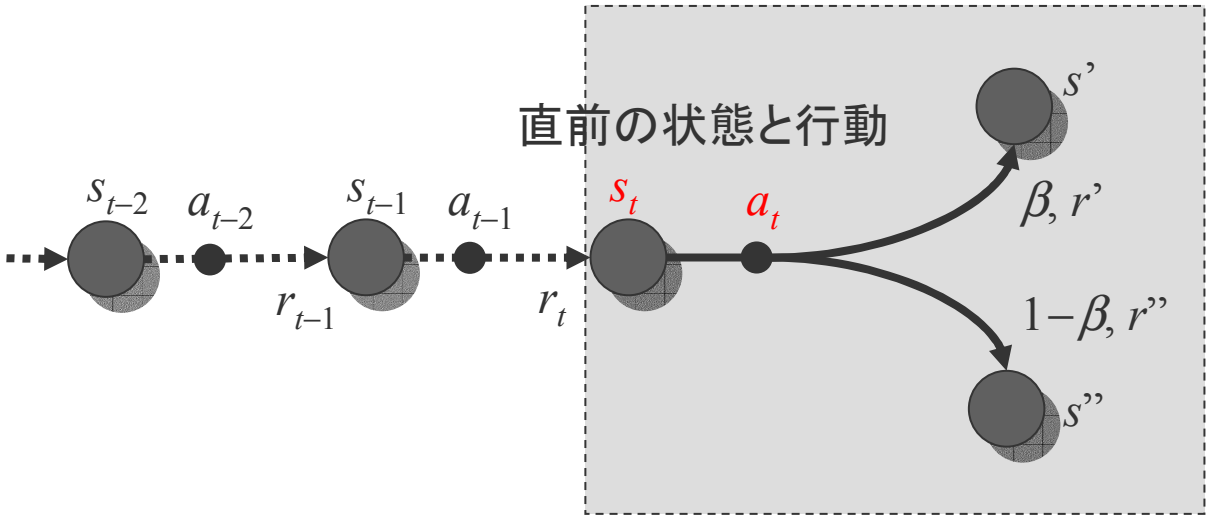
割引収益

割引率 $(0 \leq \gamma \leq 1)$

マルコフ性

- 次「状態」と「報酬」が直前の「状態」と「行動」にのみ依存

$$\Pr(s_{t+1} = s', r_{t+1} = r' \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0) =$$
$$\Pr(s_{t+1} = s', r_{t+1} = r' \mid s_t, a_t)$$



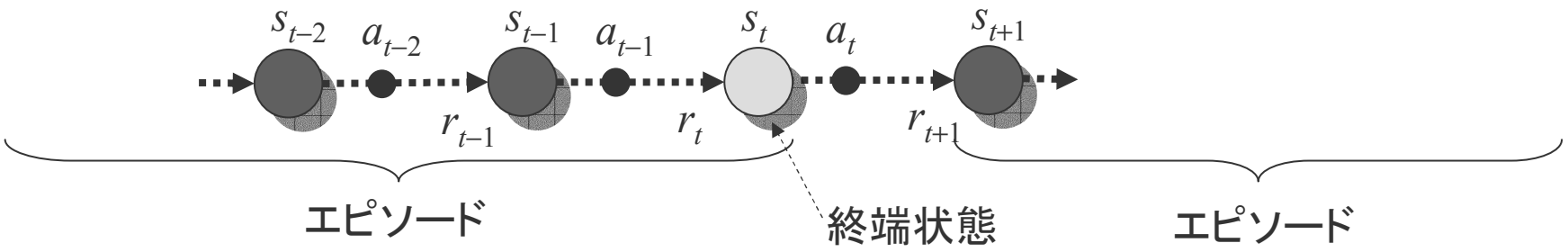
マルコフ決定過程（MDPs）

- 強化学習のタスクを表す

- MDP を構成するもの
  - 取りうる状態の集合 S
  - 取りうる行動の集合 A
  - 状態遷移確率関数 P
    - 状態 s で行動 a を取ったときに次状態が s' となる確率
  - 報酬関数 R
    - 状態 s で行動 a を取ったときに得られる報酬の期待値

エピソード型タスクと連続型タスク

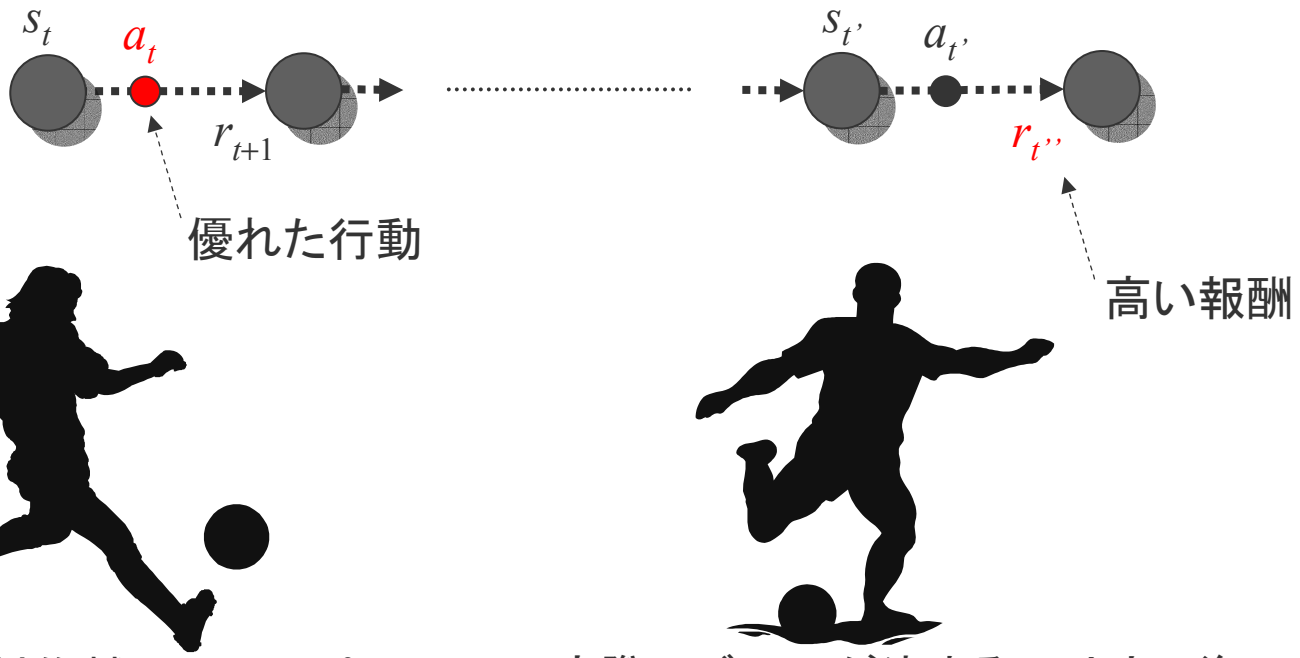
- エピソード型タスク
  - 相互作用の系列がある状態（終端状態）で分割できる
    - 例：チェスなどのゲーム，ゴールのある迷路など
  - MDP に終端状態の集合 S^+ を加える



- 連続型タスク
  - エピソード型でないタスク

報酬の遅れ

- 報酬は，すぐにもらえるとは限らない



たとえば，中村俊輔のキラーパス

実際にゴールが決まるのは少し後

これを「遅延報酬」という

まとめ：強化学習

- 機械学習
  - 経験から振る舞いを自動的に改善
- 強化学習の特徴
  - 試行錯誤
  - 教師なし学習
  - 遅延報酬

ニューラルネットワークの結合重みを強化する研究もあれば
論理的なルールベースの構築に強化学習を使う研究もある

- 強化学習のタスク
  - MDP で表現
  - エピソード型タスクと連続型タスク

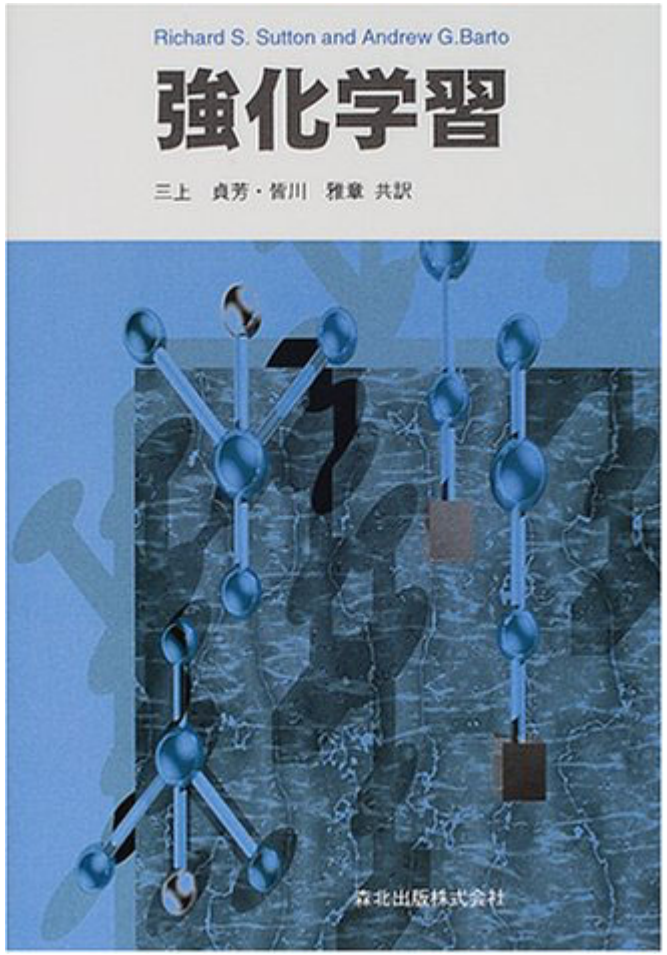
勉強の進め方

文献を読もう

- 読み物
  - 基本的なアルゴリズムなどが紹介されている
- 論文誌
  - 新しいアルゴリズムが提案されている
- 国際会議
  - 最新の研究成果が報告されている
  - 流行しているテーマを確認できる
- Web サイト
  - 最新情報を入手可能

読み物

- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press. 1998. 三上貞芳・皆川雅章共訳. *強化学習*. 森北出版
  - 強化学習関係者必携の書
  - 強化学習分野における（世界の）常識



「強化学習」の内容と読み方

I 強化学習問題
- 1 序章
- 2 評価フィードバック
- 3 強化学習問題

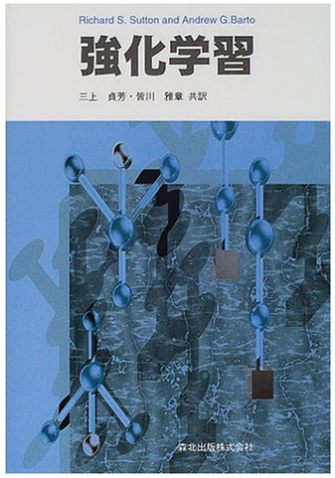
II 基本的な解法群
- 4 動的計画法
- 5 モンテカルロ法
- 6 TD 学習

III 統一された見方
- 7 適格度トレース
- 8 一般化と関数近似
- 9 プランニングと学習
- 10 強化学習の特徴軸
- 11 ケーススタディ

強化学習とは何か知りたい人

強化学習を使って研究したい人

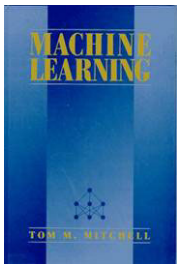
強化学習について研究したい人



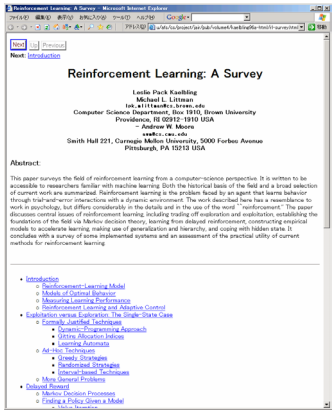
そのほかの読み物

- S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall. 1995. 古川康一 監訳. *エージェントアプローチ 人工知能*. 共立出版. 1997. 第18章および第20章.
  - エージェントの視点から強化学習を解説

- T. M. Mitchell. *Machine Learning*. McGraw-Hill. 1995. 第1章および第13章.
  - 機械学習の視点から強化学習を解説

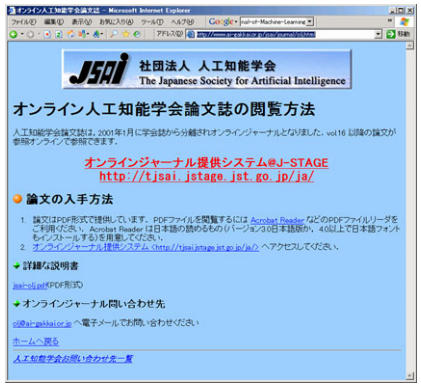


- L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research,* Vol. 4, pp. 237-285, 1996.
  - サーベイ論文. 入力や出力の一般化についても言及
  - http://www-2.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/rl-survey.html

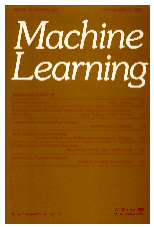


論文誌

- 人工知能学会論文誌
  - オンラインジャーナルとしても公開
  - http://www.ai-gakkai.or.jp/jsai/journal/olj.html



- *Machine Learning*
  - 図書館で借りることができる
  - *The Journal of Machine Learning Research* と間違えないように ☺



- 他多数

関連する学会

- 人工知能学会
- 情報処理学会
- 電子情報通信学会
- 日本ロボット学会
- 計測自動制御学会
- システム制御情報学会

国際会議

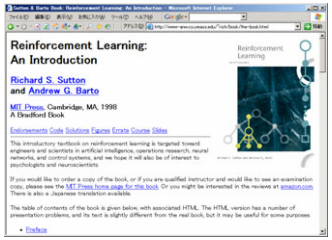
- International Conference on Machine Learning (ICML)
  - 機械学習に関する国際会議
  - 毎年開催，次回はワシントン DC
- International Joint Conference on Artificial Intelligence (IJCAI)
  - 人工知能に関する国際合同会議
  - 2年に1度（奇数年）開催，次回はアカプルコ（メキシコ）
- European Conference on Machine Learning (ECML)
  - 機械学習に関するヨーロッパ国際会議
  - 毎年開催，次回はクロアチア
- Pacific-Rim International Conference on Artificial Intelligence (PRICAI)
  - 人工知能に関する環太平洋国際会議
  - 2年に1度開催（偶数年）開催，次回はニュージーランド
- Conference on Uncertainty in Artificial Intelligence (UAI)
  - 人工知能における不確実性に関する会議
  - 毎年開催，次回はアカプルコ，メキシコ

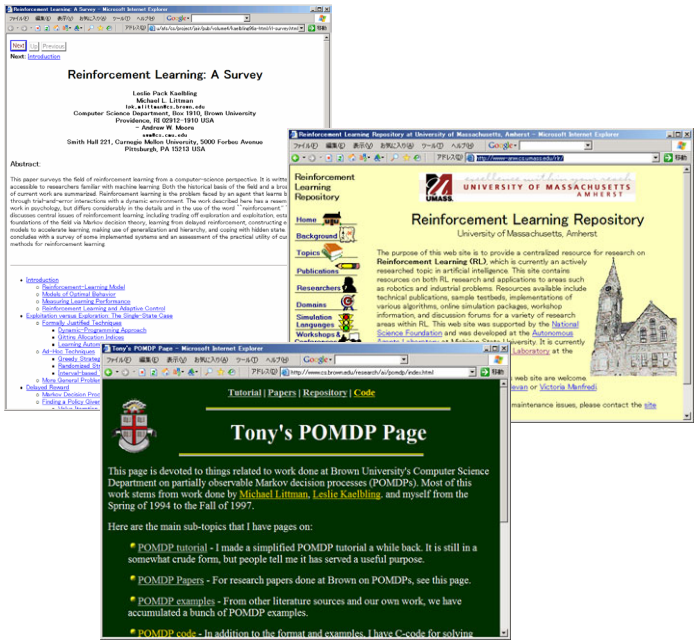
Proceedings の例

* 国際会議で発表された論文集のことを Proceedings（プロシーディングス）といいます．
** 国際会議の一覧が見れます．http://www-sekilab.ics.nitech.ac.jp/~tohgoroh/conferences.html

Web サイト

- Sutton & Barto Book
  - http://www-anw.cs.umass.edu/~rich/book/the-book.html
  - 「強化学習」の教科書
- Kaelbling et al. Survey
  - http://www-2.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/rl-survey.html
  - 強化学習についてのサーベイ論文 (1996)
- Reinforcement Learning Repository
  - http://www-anw.cs.umass.edu/rlr/
  - 強化学習についての情報源
- Tony's POMDP Page
  - http://www.cs.brown.edu/research/ai/pomdp/
  - 部分観測環境について
- 強化学習 FAQ
  - http://www.k2.t.u-tokyo.ac.jp/members/naoko/RL-FAQ-j.html
  - Sutton の強化学習 FAQ の日本語版





主な研究者（海外）1/3

- Richard S. Sutton
  - Massachusetts 大学特任教授
  - 教科書「強化学習」
  - http://www-anw.cs.umass.edu/~rich/sutton.html

- Andrew G. Barto
  - Massachusetts 大学教授
  - 教科書「強化学習」
  - http://envy.cs.umass.edu/People/barto/barto.html

- Leslie Pack Kaelbling
  - MIT 教授
  - サーベイ論文「強化学習」
  - http://www.ai.mit.edu/people/lpk/

（注）このリストは主観的であり，完全ではありません．

* 写真はそれぞれのサイトから引用

主な研究者（海外）2/3

- Michael L. Littman
  - Duke 大学助教授
  - サーベイ論文「強化学習」, POMDPs
  - http://www.cs.duke.edu/~mlittman/

- Andrew W. Moore
  - CMU 助教授
  - サーベイ論文「強化学習」, データマイニング
  - http://www-2.cs.cmu.edu/~awm/

- Satinder Singh
  - Colorado 大学助教授
  - POMDPs, 適格度トレース
  - http://www.eecs.umich.edu/~baveja/

* 写真はそれぞれのサイトから引用

主な研究者（海外）3/3

- Anthony R. Cassandra
  - St. Brown 大学助教授
  - POMDPs に関する論文多数，Tony’s POMDP Page の作者
  - http://www.cassandra.org/arc/cv.html
- C. J. C. H. Watkins
  - Q学習の提案者
- G. A. Rummery
  - Sarsa の提案者
- G. J. Tesauro
  - プロ並みのバックギャモンプログラム TD-Gammon の作者
- P. Stone
  - RoboCup サッカー優勝チーム CMUnited-99 の作者
- S. Russell
  - 「エージェントアプローチ 人工知能」の著者

主な研究者（日本）

- 浅田稔
  - 大阪大学教授
  - 実ロボットへの応用
  - http://www.er.ams.eng.osaka-u.ac.jp/
- 宮崎和光
  - 大学評価機構助教授
  - Profit Sharing の合理性定理，罰回避政策
  - http://svrrd2.niad.ac.jp/faculty/teru/
- 荒井幸代
  - CMU 研究員
  - マルチエージェント強化学習
  - http://www.ri.cmu.edu/people/arai_sachiyo.html
- 木村元
  - 東京工業大学助手
  - POMDPs, 確率的政策，実ロボットへの応用
  - http://www.fe.dis.titech.ac.jp/~gen/indexj.html

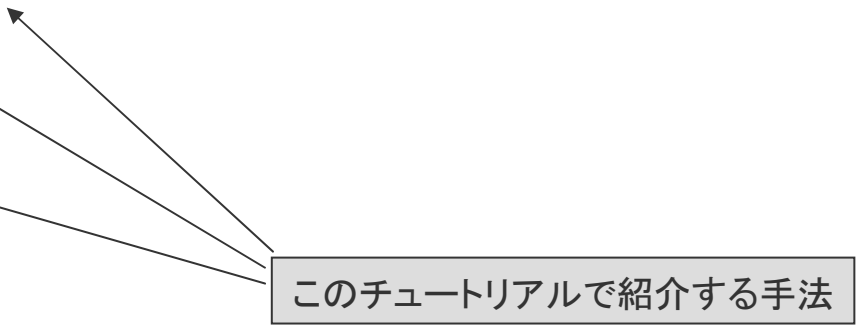
* 写真はそれぞれのサイトから引用

これまでに提案された代表的な手法

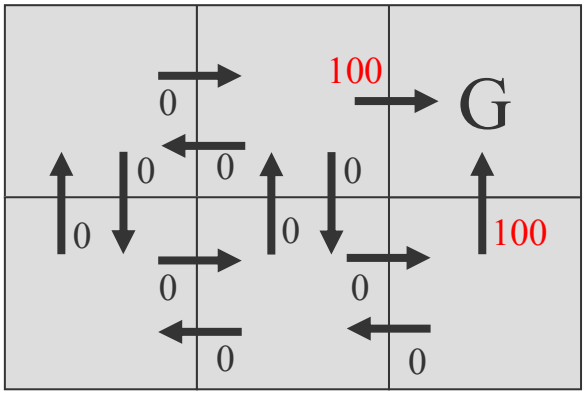
代表的な手法と技法

- 手法
  - Monte Carlo 法（第5章）
    - Monte Carlo Control 法（5.3節）
  - TD法（第6章）
    - Q学習（6.5節）
    - Sarsa（6.4節）
  - Profit Sharing
- 技法
  - 適格度トレース（第7章）
  - 関数近似（第8章）

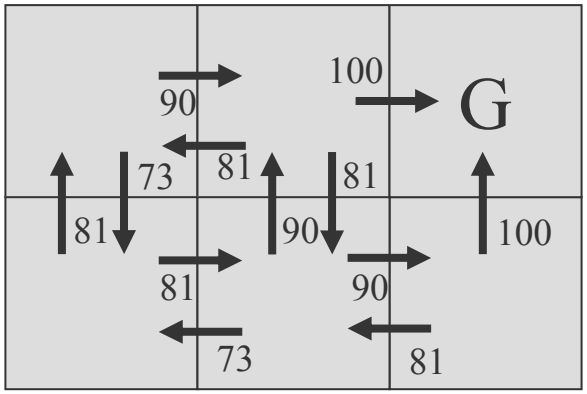
TD = Temporal Difference



Q学習



報酬 r



$\gamma = 0.9$

（最大）割引収益 $Q(s,a)$

「推定値」を意味する記号

$$\hat{Q}(s,a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s',a')$$

1 ステップ後の状態

学習率 α を導入 $(0 \le \alpha \le 1)$

$$\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha\left(r + \gamma \max_{a'} \hat{Q}(s',a') - \hat{Q}(s,a)\right)$$

Q学習のアルゴリズム

$Q(s,a)$ を任意に初期化

各エピソードに対して繰り返し:

s を初期化

エピソードの各ステップに対して繰り返し:

Q から導かれる政策（後述）を使って，s での行動 a を決定する

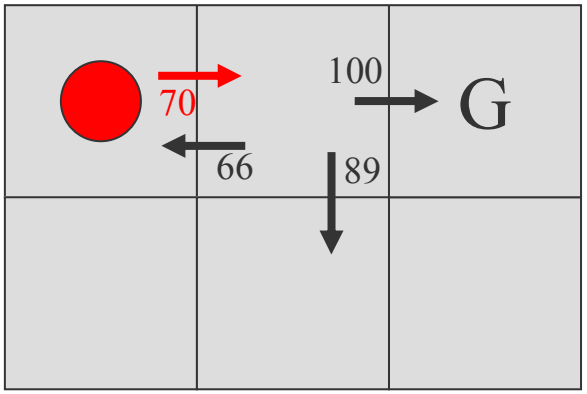
行動 a を取り，r, s' を観測する

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$$

$s \leftarrow s'$;

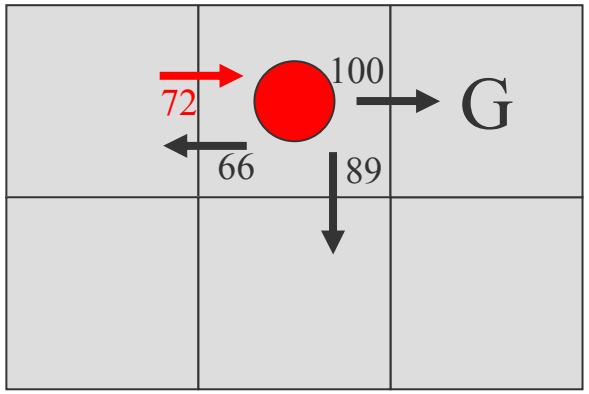
s が終端状態ならば繰り返しを終了

Q学習の実行例



現在の予測値

行動 a_{right} を選択



1 ステップ後の予測値

$$\hat{Q}(s,a_{\text{right}}) \leftarrow \hat{Q}(s,a)+\alpha\left(r+\gamma \max_{a'} \hat{Q}(s',a')-\hat{Q}(s,a)\right)$$
$$=70+0.1\times\left(0+0.9\max\{66,89,100\}-70\right)$$
$$=70+0.1\times(90-70)$$
$$=70+2$$
$$=72$$

学習率 α が 1 に近いほど，値の変化が大きい．
ただし，報酬や予測値の誤差の影響も大きい．

Sarsa

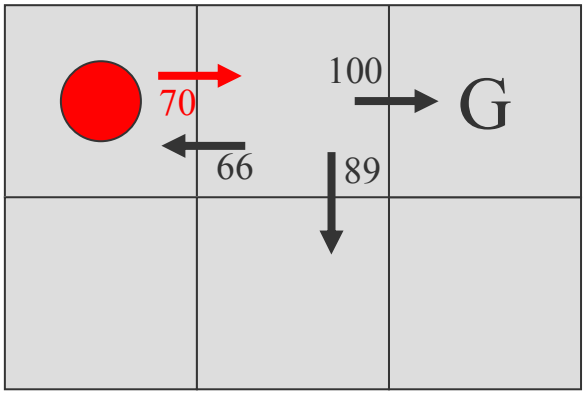
- Q学習での更新に使われる $Q(s',a')$ は，政策が選んだものとは無関係＝政策オフ型

- Sarsa は，政策が選んだものを使う＝政策オン型

$$\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha\left(r + \gamma\hat{Q}(s',a') - \hat{Q}(s,a)\right)$$

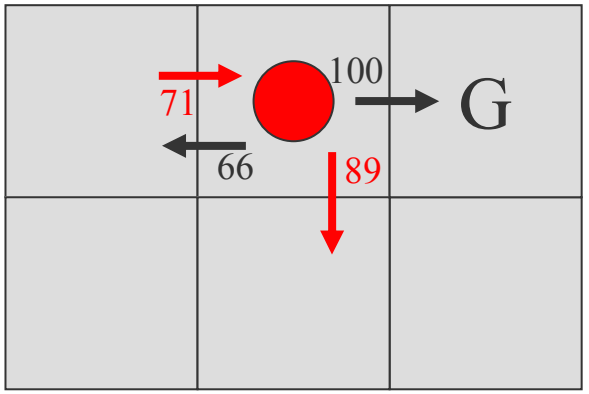
次の状態で実際に選択した行動

Sarsa の実行例



現在の予測値

行動 a_{right} を選択



1 ステップ後の予測値

行動 a_{down} を選択

$$\hat{Q}(s,a_{\text{right}}) \leftarrow \hat{Q}(s,a) + \alpha\left(r + \gamma\hat{Q}(s',a_{\text{down}}) - \hat{Q}(s,a)\right)$$
$$= 70 + 0.1\times\left(0 + 0.9\times 89 - 70\right)$$
$$= 70 + 0.1\times(80 - 70)$$
$$= 70 + 1$$
$$= 71$$

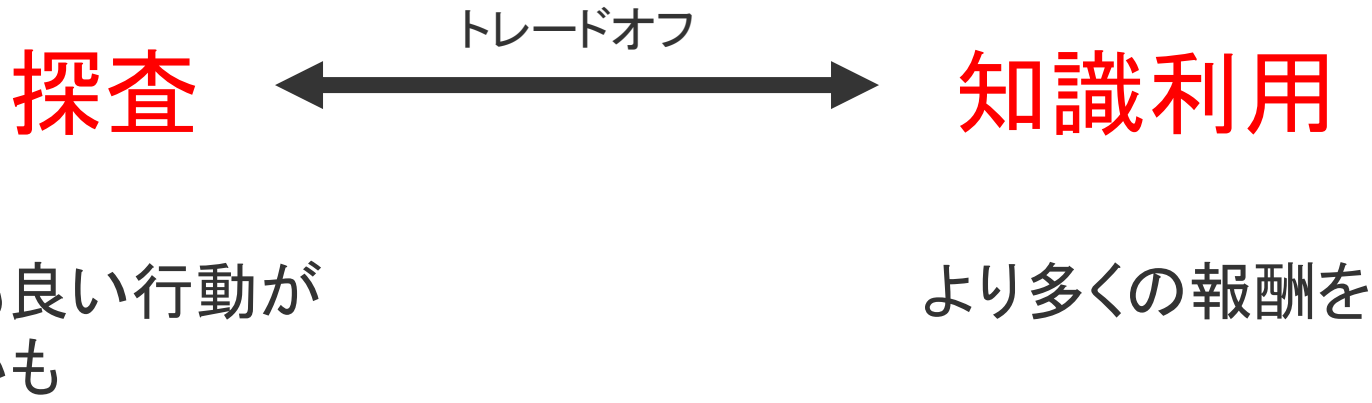
政策（行動選択）

- グリーディ選択
  - 行動価値が最大のものを選択
- ε グリーディ選択
  - 確率 ε で一様なランダム選択
  - 確率 1–ε でグリーディ選択
- ソフトマックス選択
  - 行動価値に応じて選択確率を変える

$$\Pr(a \mid s) = \frac{e^{Q(s,a)/\tau}}{\sum_{a'} e^{Q(s,a')/\tau}}$$

探査か知識利用かのジレンマ

- 知識利用＝グリーディ選択
- 探査＝それ以外の選択

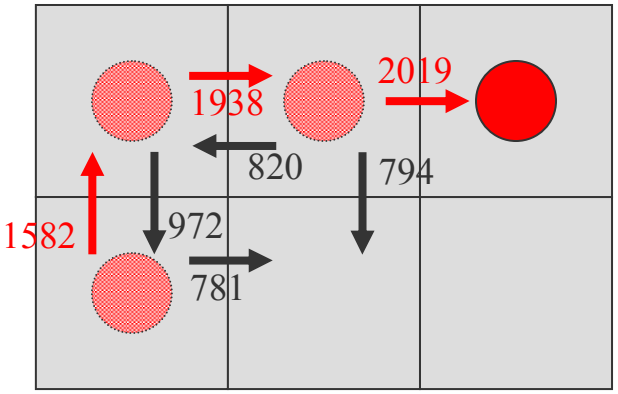


Q学習・Sarsa のまとめ

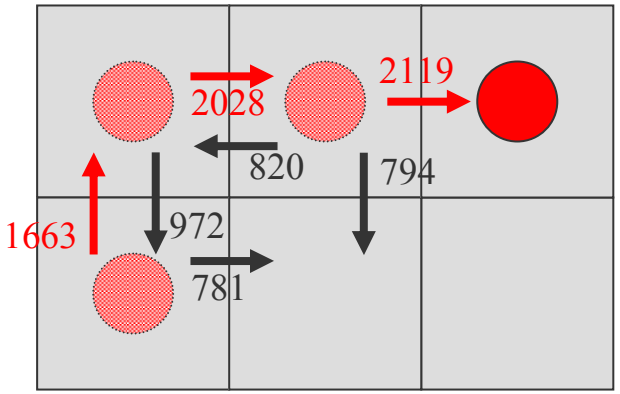
- 行動価値を推定する手法
  - 行動価値＝期待割引収益
- 真の行動価値への収束性が証明されている
  - ただし，ランダムな行動選択で
- 探査か知識利用かのジレンマ

Profit Sharing

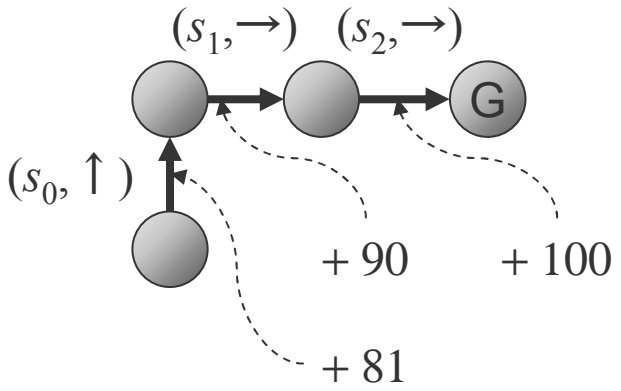
エピソード開始時の優先度



ゴールした後で



ゴールに近いものほど優先度を増やす



エピソード

* 状態行動対の系列

$(\gamma = 0.9$ のとき$)$

エピソード終了後,

エピソードに含まれるすべての (s_t, a_t) に対し

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + \gamma^{T-t-1} r_T$$

s_t での a_t の優先度

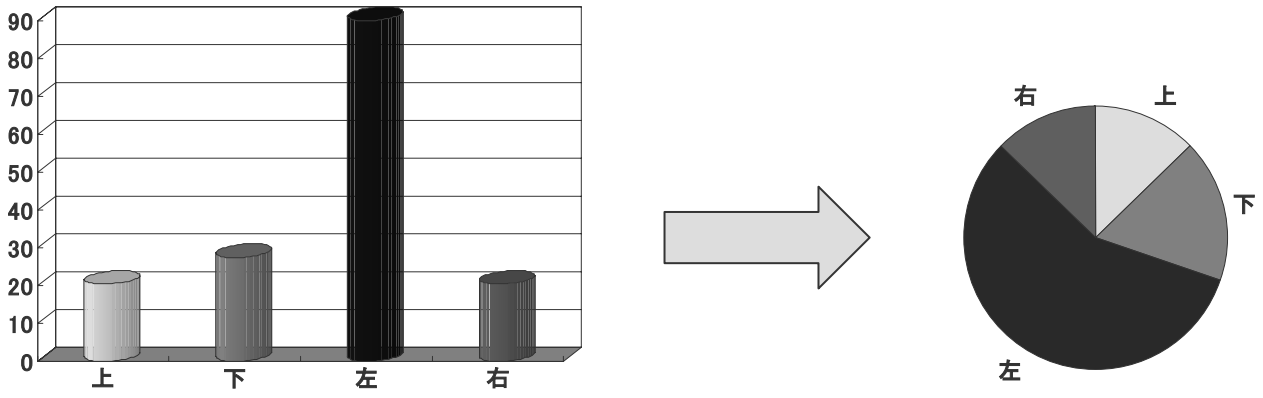
割引率 $(0 \le \gamma \le 1)$

報酬

Profit Sharing における行動選択

- 重み付きルーレット選択
  - 優先度に比例した確率分布に従って選択

$$\Pr(a \mid s) = \frac{P(s,a)}{\sum_{a'} P(s,a')}$$



優先度　　選択確率

Profit Sharing のアルゴリズム

すべての s, a に対し $P(s,a) \leftarrow C$（小さい定数）

各エピソードに対して繰り返し:

s を初期化

エピソードの各ステップに対して繰り返し:

P に比例した確率分布に従って，s での行動 a を決定する

行動 a を取り，r, s' を観測する

$s \leftarrow s'$;

s が終端状態ならば繰り返しを終了

エピソードに含まれるすべての s_t, a_t に対して:

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + \gamma^{T-t-1} r_T$$

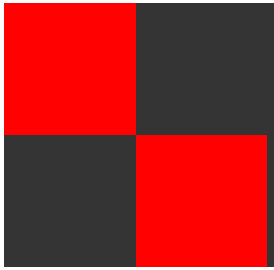
Profit Sharing における期待値合理性定理

$$\gamma \leq \frac{1}{\max_{s \in \mathcal{S}} |\mathcal{A}(s)|}$$

とりうる行動の数

$$\gamma \leq 1$$

- 宮崎の合理性定理 (1994)
  - ○合理的な解を学習することを保証
  - ×条件が厳しく，学習が遅い
- 期待値合理性条件 (松井, 2003)
  - ○期待値の意味で合理的な解を学習できることを証明
  - ○宮崎の条件より緩く，学習が速い



現在流行しているテーマ

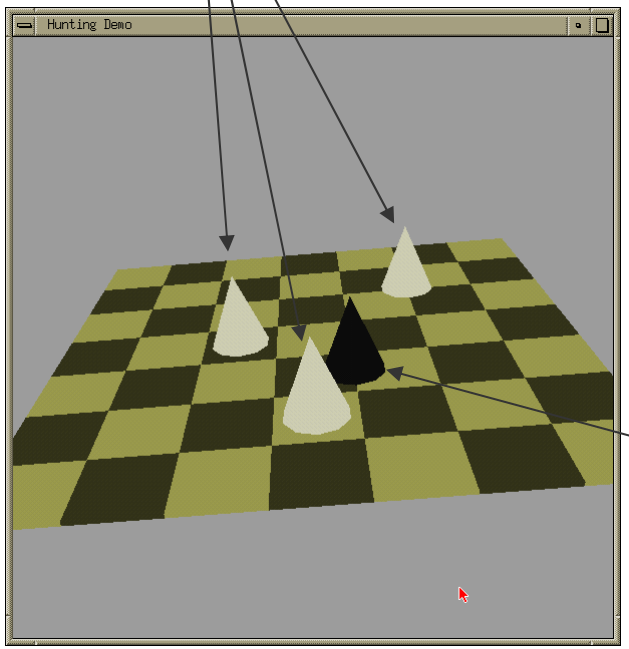
国際会議 ICML-2002 からの分析

現在流行しているテーマ

- マルチエージェント強化学習
- 部分観測マルコフ決定過程（POMDPs）
- 階層型強化学習

マルチエージェント強化学習

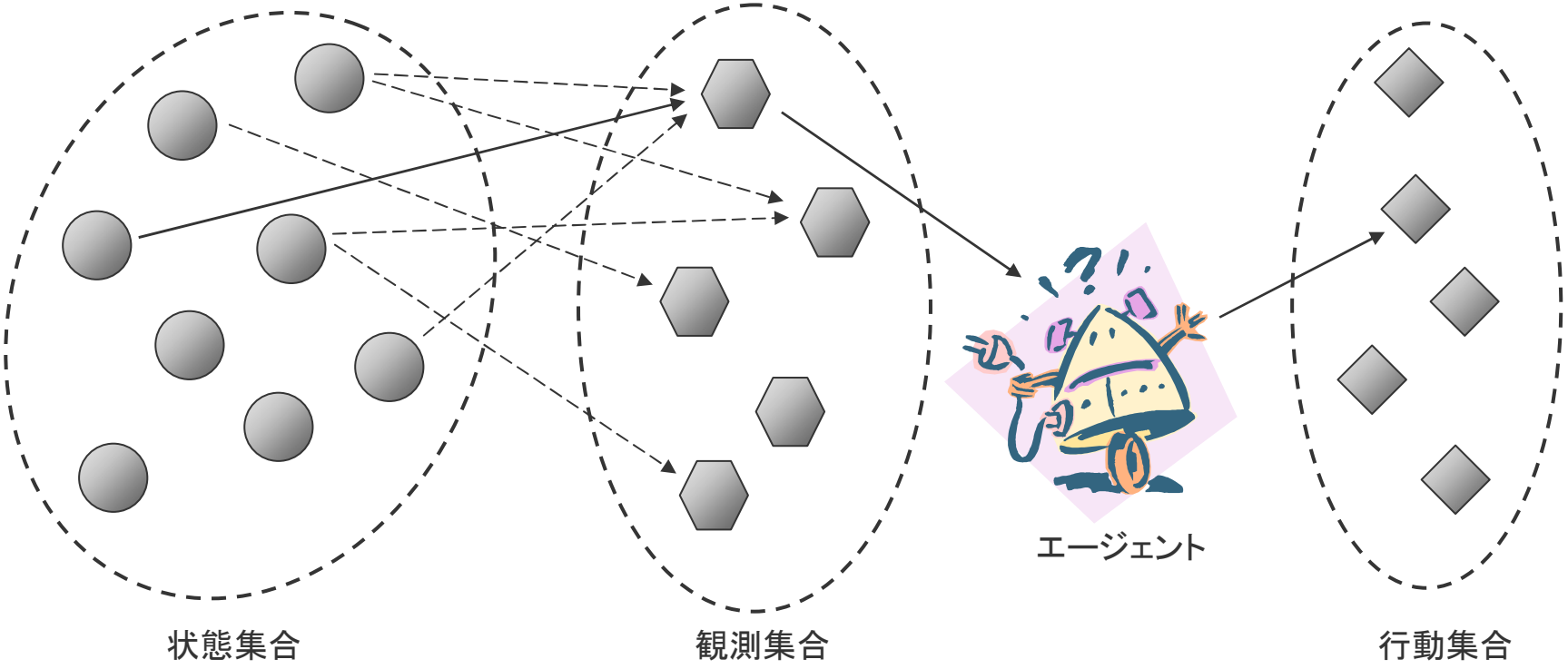
- 複数のエージェントがいる環境での強化学習
  - 協調
  - 敵対
  - 同時学習
  - 部分観測



例：追跡問題
（2体のエージェントで周りを囲む問題）

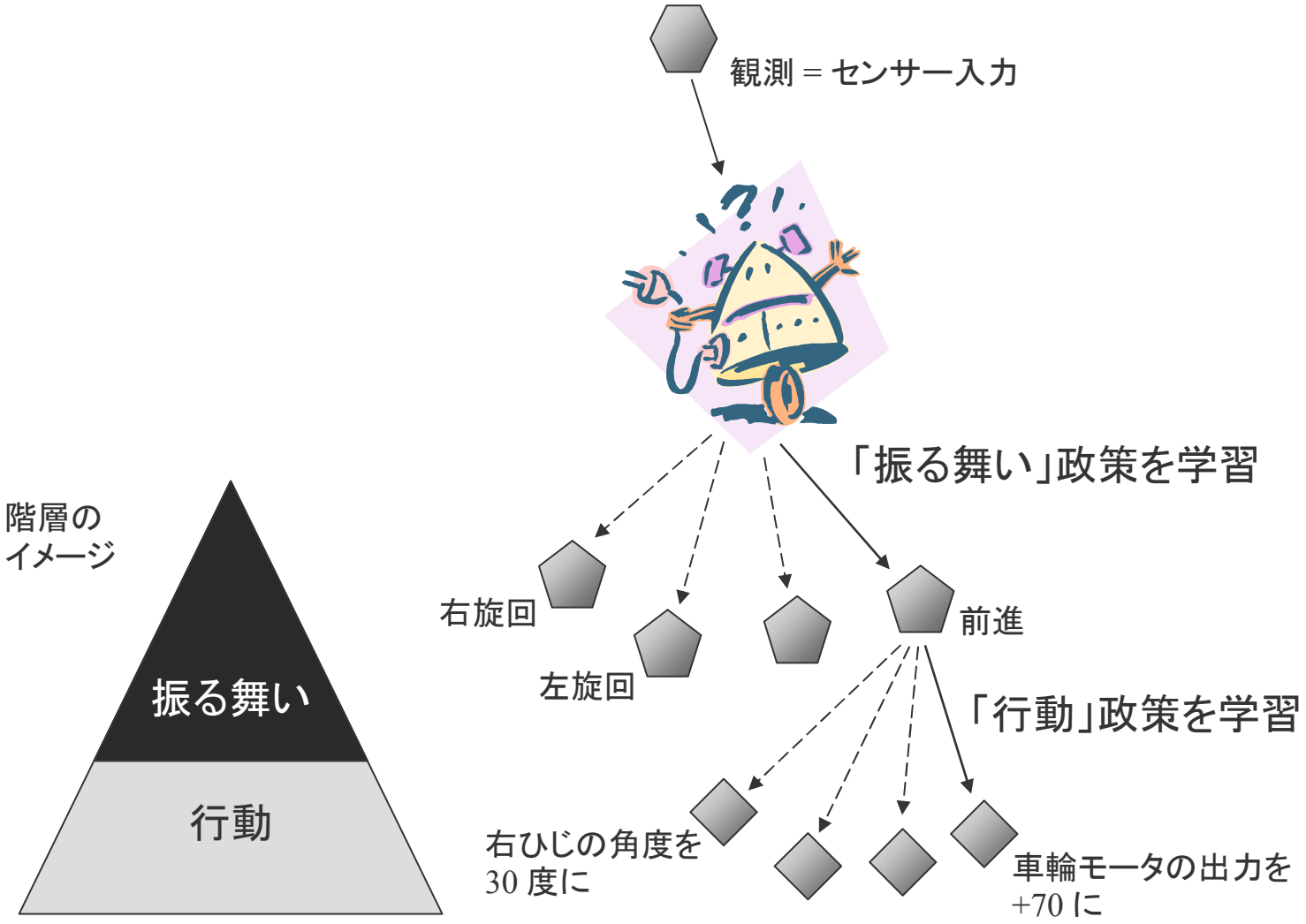
部分観測マルコフ決定過程（POMDPs）

- POMDPs = Partially Observable MDPs
- エージェントが状態を混同する
  - センサーの情報が不足しているとき
  - サッカーで周りの選手の区別がつかないとき



階層型強化学習

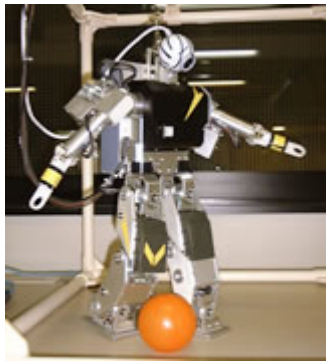
- 低レベルの学習と高レベルの学習を分離



実ロボットへの応用とその課題

実ロボットへの応用例

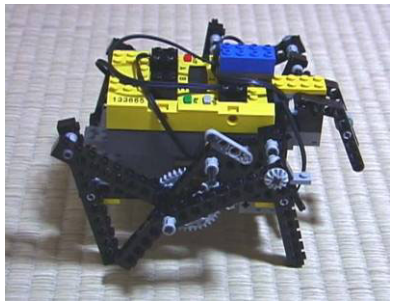
- サッカーロボット
  - 大阪大学浅田研究室が有名
- AIBO
  - Sony
  - 日本全国多数の研究室
- その他
  - 起き上がりロボット
    - 川人学習動態脳プロジェクト
  - MindStorms
    - 宮崎（大学評価機構）



HOAP-1

AIBO

2関節3リンクロボット（合成）



MindStorms

* 写真はそれぞれのサイトから引用

実ロボットに応用する上での課題 1/2

- 状態定義
  - 実世界の状態を完全に記述することはできない
- 観測・行動の不確実性
  - センサーからの入力にはノイズが含まれる
  - モーターの調子によって結果が異なる
- 連続値の扱い
  - 状態や行動が連続的な値をとる

実ロボットに応用する上での課題 2/2

- 動的環境
  - 自分以外にも移動するものが存在する
- マルチエージェント
  - 協力・敵対関係にある他のロボットが存在する
  - 他のロボットも学習するとき，環境が安定しない

もし，実ロボットで強化学習研究をするなら

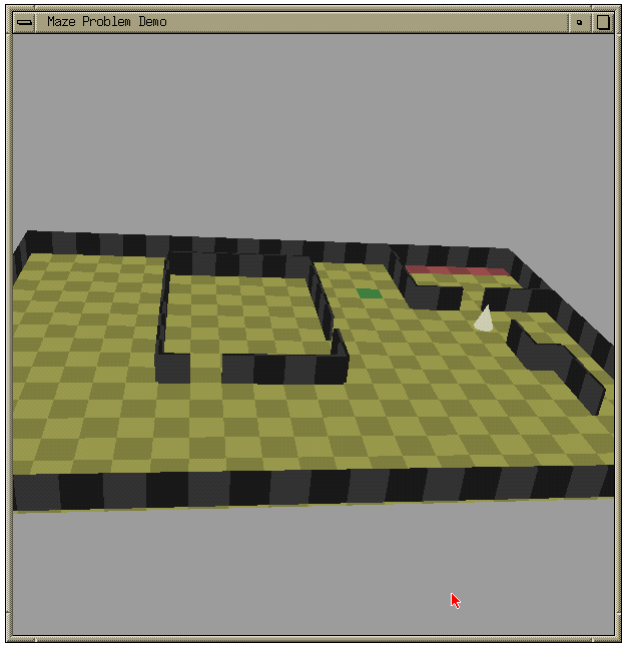
- 既存アルゴリズム・技術の応用「だけ」ではだめ
  - すでに多数の応用例が報告されているから
- 特殊なロボットに特化した研究は役に立たない
  - 他のロボットへの応用ができないから
  - したがって，理論的な議論もするべき
    - ヒューマノイドロボットに共通する問題点など
- 関連研究を挙げ，研究の位置付けを明確にする
  - 他で発表された研究とどこが同じでどこが違うのか

私が提案した手法と
そのプログラム

概要

- 提案した強化学習アルゴリズム
  - On-Line Profit Sharing (OnPS)
  - Last-Visit Profit Sharing (LVPS)

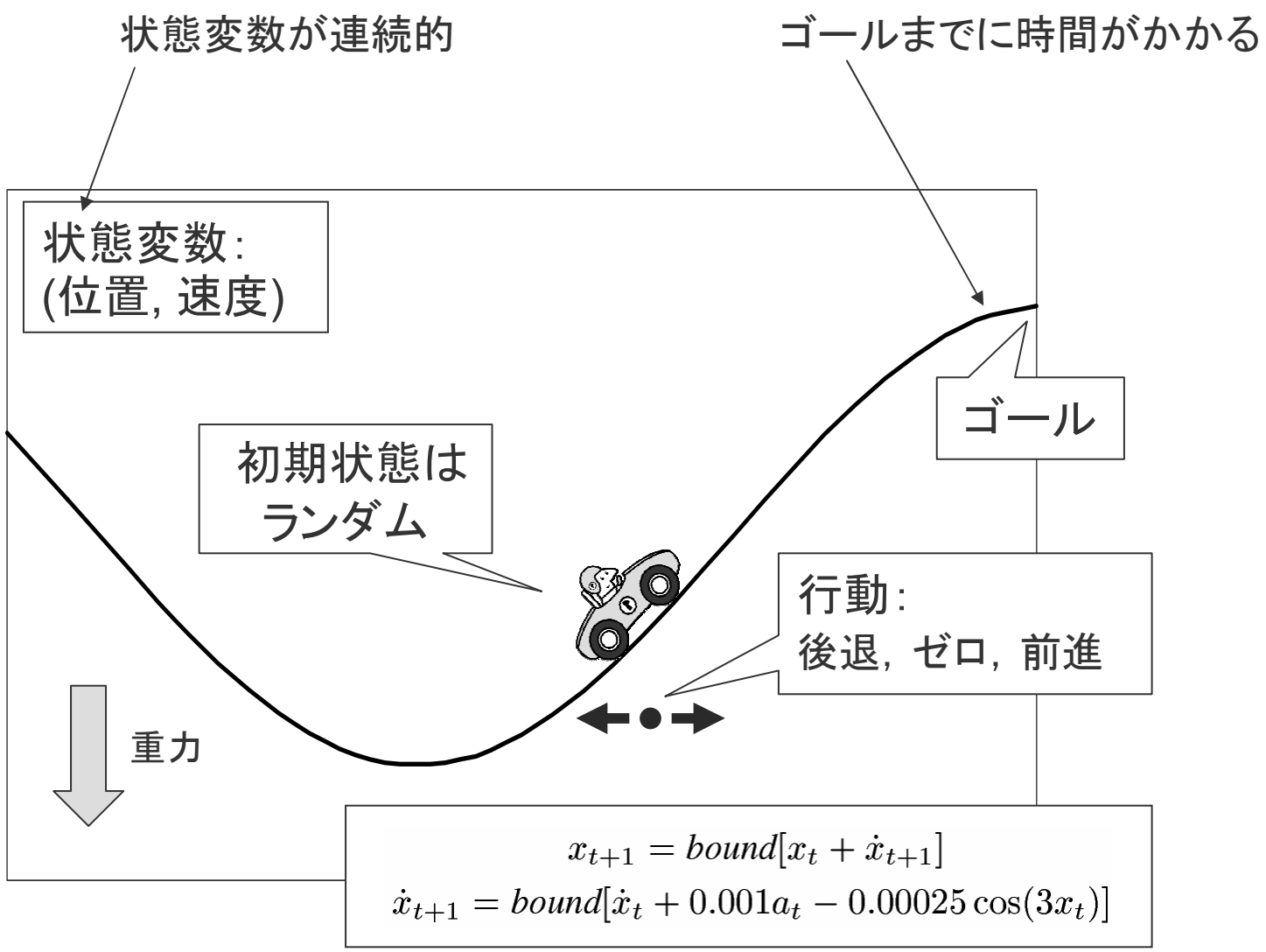
- 作成した強化学習プログラム
  - 迷路の問題
    - OpenGL を使って 3D 表示可能
  - 車の山登り問題
    - 視覚効果なし
  - POMDPs
    - http://www-sekilab.ics.nitech.ac.jp/~tohgoroh/POMDPs/
  - RoboCup Soccer
    - プロトタイプのみ（未完成）



迷路の問題

OnPS と LVPS

Mountain-Car タスク (Moore, 1991)



必要なメモリ量が無制限

- 実世界のタスクでは，ゴールするのに多くのステップが必要

⇒ 多くの状態行動対を記憶しなければならない

⇒ Profit Sharing のオンライン更新化

提案：On-Line Profit Sharing (OnPS)

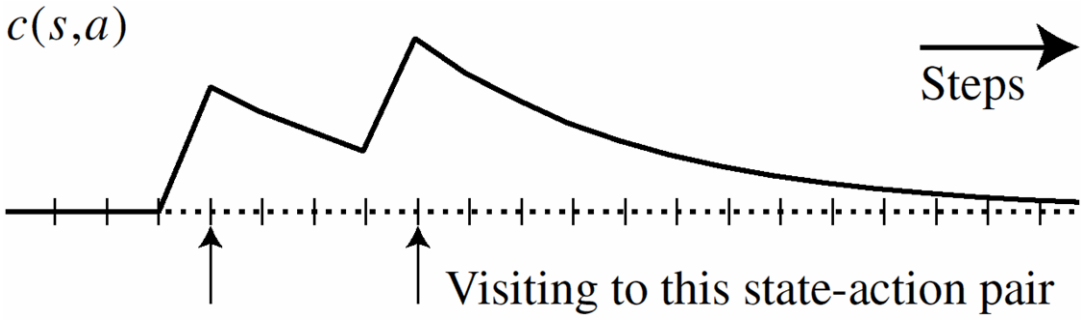
- ステップごとに優先度 P を更新
- 状態行動対を憶えなくていい

$$\Delta P_t(s,a) = r_{t+1} c_t(s,a) \quad \text{for all } s, a$$

そのときの報酬（中間報酬）

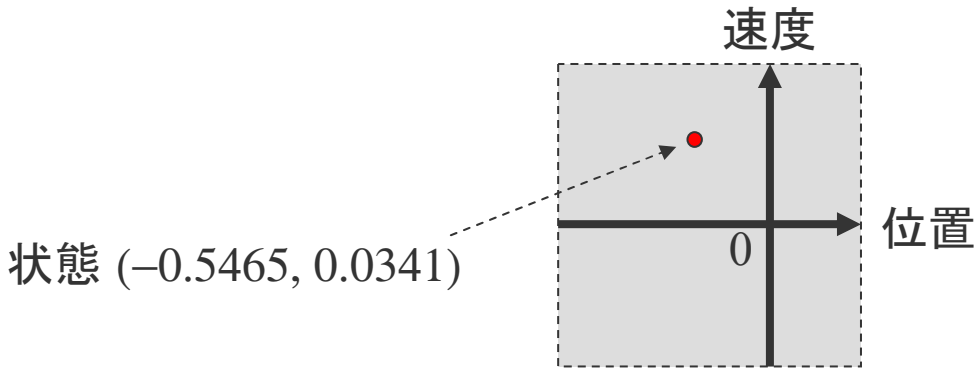
$$c_t(s,a) = \begin{cases} \gamma c_{t-1}(s,a) + 1 & s = s_t \text{ かつ } a = a_t \text{ のとき} \\ \gamma c_{t-1}(s,a) & \text{そうでないとき} \end{cases}$$

信用トレース
s, a を更新したのが
どのくらい前か



連続的な状態空間

- 従来は状態空間を離散的としていたため，そのまま扱えない



  - 他の強化学習法では，関数近似により解決

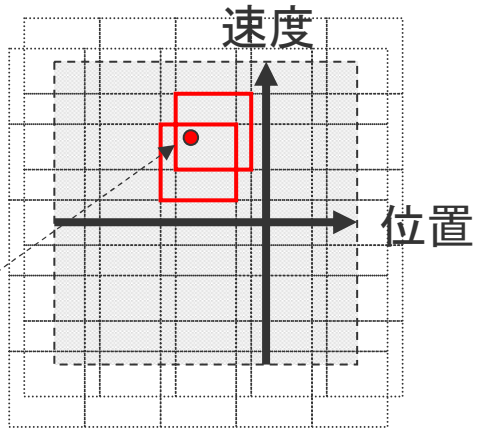
⇒ Profit Sharing にも関数近似を導入

線形関数近似つき On-Line Profit Sharing

- P をパラメータベクトル $\vec{\theta}$ の線形和で表す

$$P(s,a) \approx \vec{\theta}^{\mathrm{T}} \vec{\phi}_{sa} = \sum_{i=1}^{n} \theta(i)\phi_{sa}(i)$$

特徴ベクトル
（含まれていれば1，そうでないとき0）



状態
(–0.5, 0.03)

CMACs
(Albus, 1971;
Sutton, 1998)

更新式

$$\vec{\theta} \leftarrow \vec{\theta} + r_{t+1}\vec{c}_t$$
$$\vec{c}_t = \gamma\vec{c}_{t-1} + \alpha_t \nabla_{\vec{\theta}_t} P(s_t, a_t)$$
$$= \gamma\vec{c}_{t-1} + \alpha_t \vec{\phi}_{s_t a_t}$$

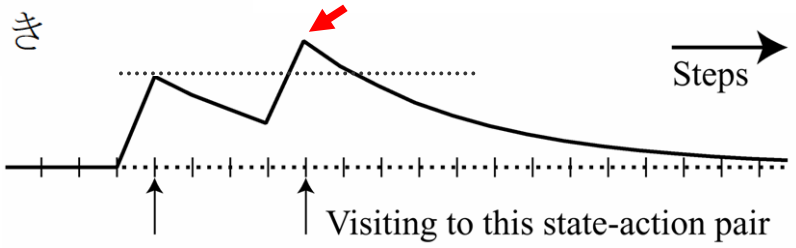
勾配法に基づく更新

提案：Last-Visit Profit Sharing (LVPS)

オンライン型 Profit Sharing

$$c_t(s,a) = \begin{cases} \gamma c_{t-1}(s,a) + 1 & s = s_t \text{ かつ } a = a_t \text{ のとき} \\ \gamma c_{t-1}(s,a) & \text{そうでないとき} \end{cases}$$

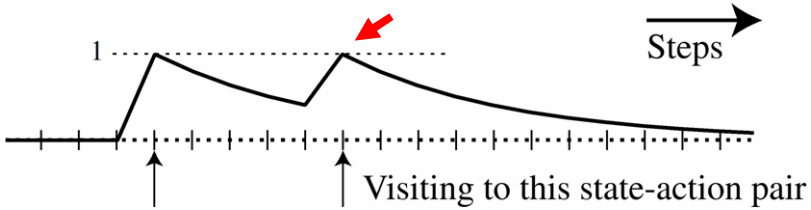
（選択すると，トレースの値を 1 増やす）



Last-Visit Profit Sharing

$$c_t(s,a) = \begin{cases} 1 & s = s_t \text{ かつ } a = a_t \text{ のとき} \\ \gamma c_{t-1}(s,a) & \text{そうでないとき} \end{cases}$$

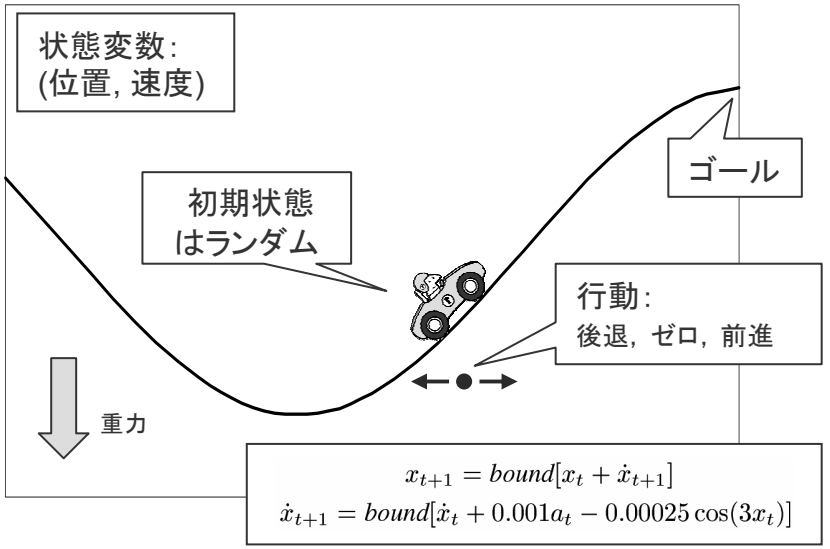
（選択すると，トレースの値を 1 に再設定）



同じ (s,a) を再び選択すると，前の時刻の情報が消える

実験

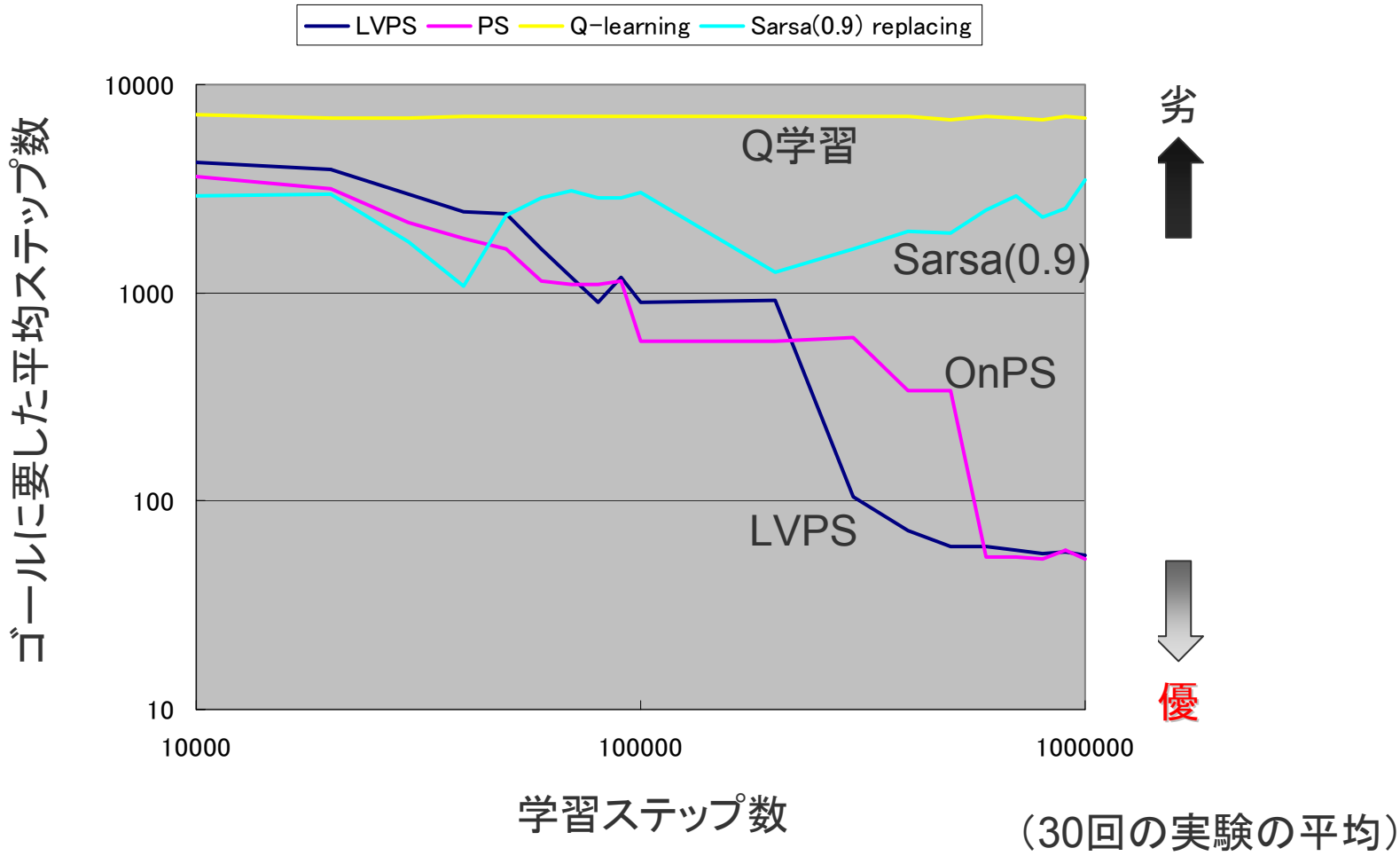
- Mountain-Car タスク
  - 9x9 の格子を10枚
  - $\alpha = 0.1, \gamma = 0.9$



- 報酬
  - ゴールに着いたとき 1
  - そのほか 0

- 評価
  - グリーディ方策がゴールに要する平均ステップ数

実験結果



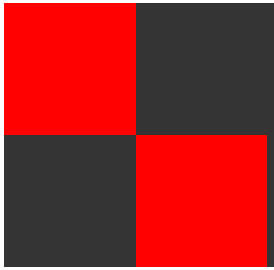
（30回の実験の平均）

OnPS と LVPS のまとめ

- オンライン更新化
  - ○必要なメモリ量が有限
  - ○中間報酬から学習可能
  - ○冗長な強化を削除可能
    (Last-Visit Profit Sharing)
  - ○オフライン更新と等しい
  - ×ステップあたりの計算量が増える

- 関数近似の導入
  - ○連続的な状態空間を扱うことができる
  - ×合理性が証明できない

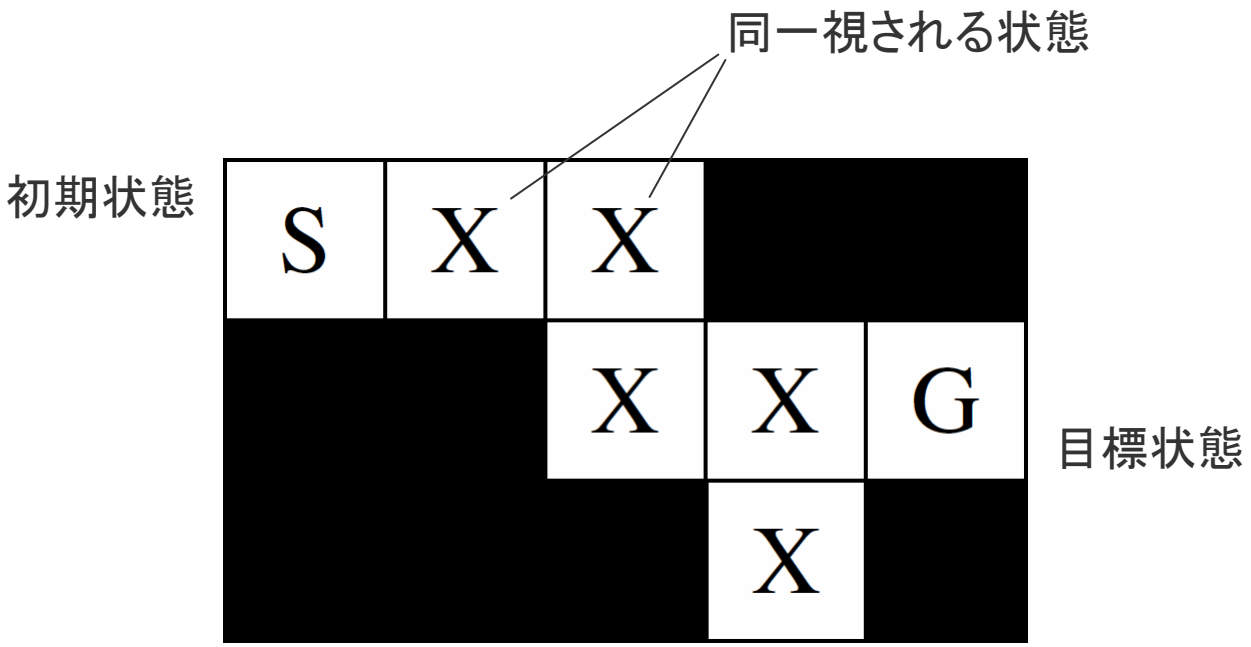
⇒ 適格度トレースでの技法 (Sutton & Barto, 1998) により軽減可能



POMDPs への応用

POMDPs では確率的政策が有効

- 決定的政策では目標に到達できないことがある
  - 目標に到達できない例



行動価値推定は POMDPs で破綻

- 行動価値として
  - $Q(s,a) = \mathrm{E}_\pi\{R_t | s_t{=}s, a_t{=}a\}$ の替わりに
  - $Q(x,a) = \mathrm{E}_\pi\{R_t | x_t{=}x, a_t{=}a\}$ を推定

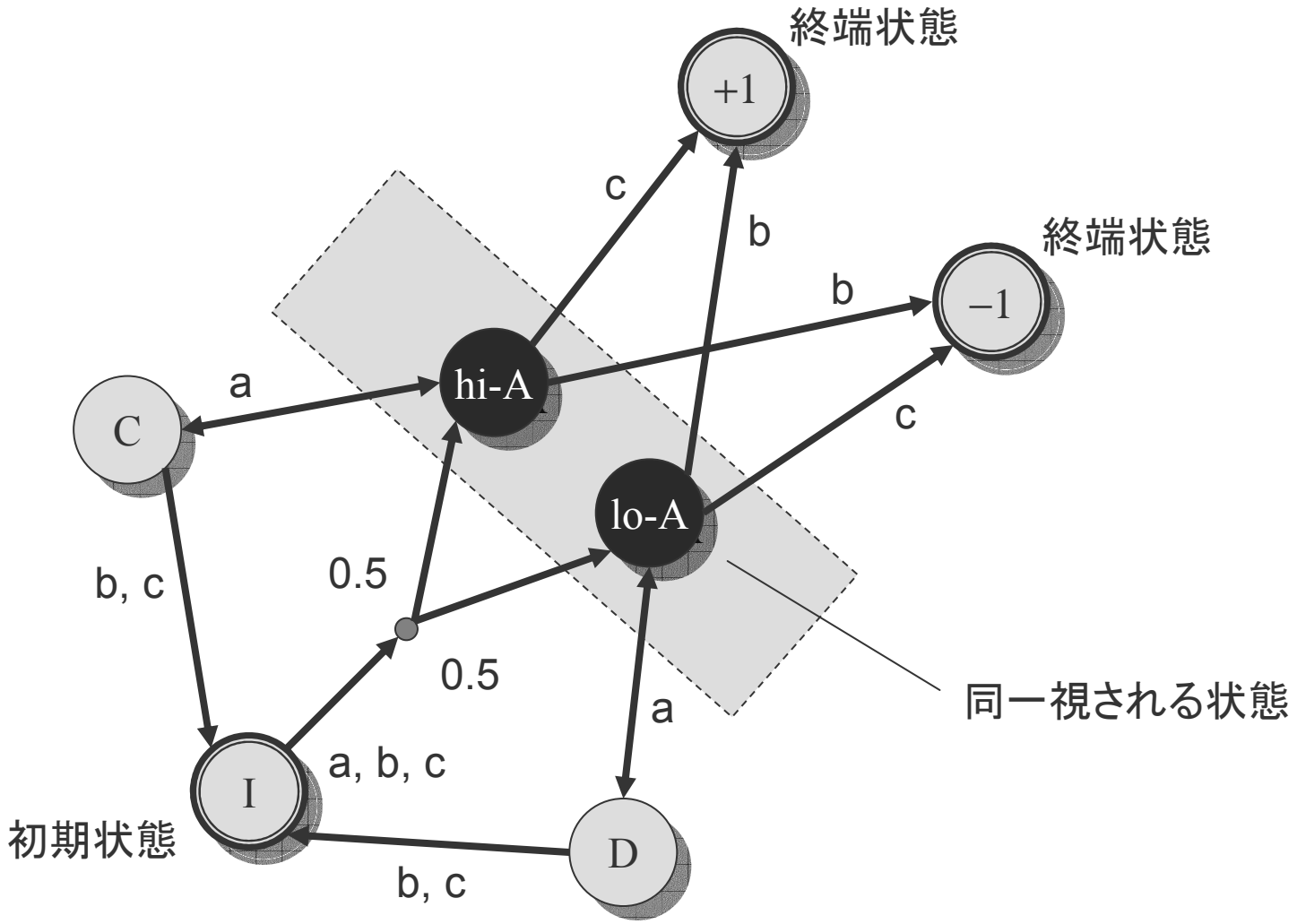
- しかし，$Q(x,a)$ から，最適な確率的政策を導くことができない
  - よって，行動価値は使えない

行動優先度学習型の有効性を調べる

- Profit-sharing methods:
  - OnPS by (松井, 2003)
  - FVPS by (Arai & Sycara, 2001)
  - LVPS by (松井, 2003)
- 比較として
  - 入れ替え更新トレースを用いた Sarsa(0.9) with Gibbs 分布ソフトマックス選択, $\tau = 0.2$
  - パラメータは (Loch & Singh, 1998) と同じ

Parr 95 (Parr & Russell, 1995)

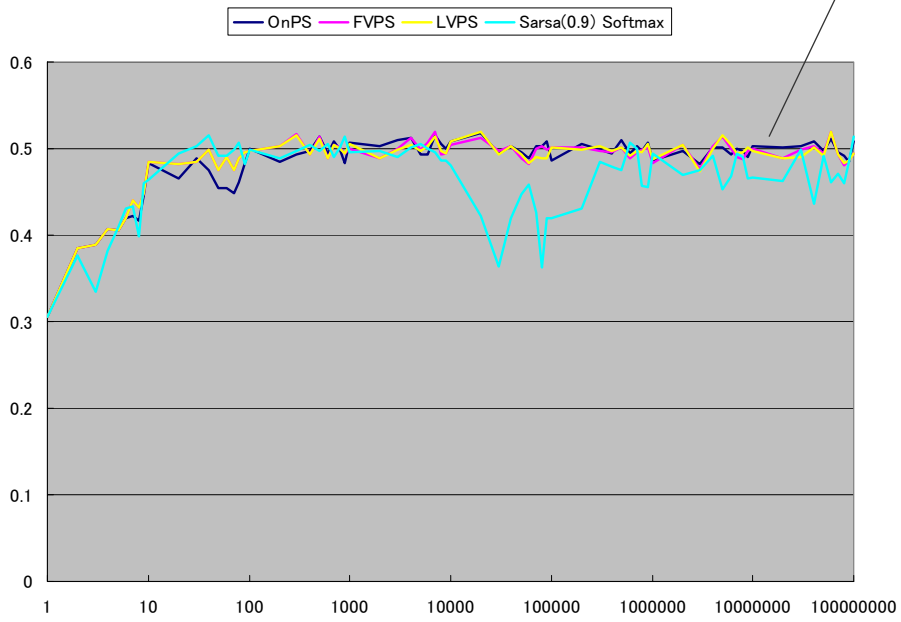
実験のやり方は (Loch & Singh, 1998) と同じ



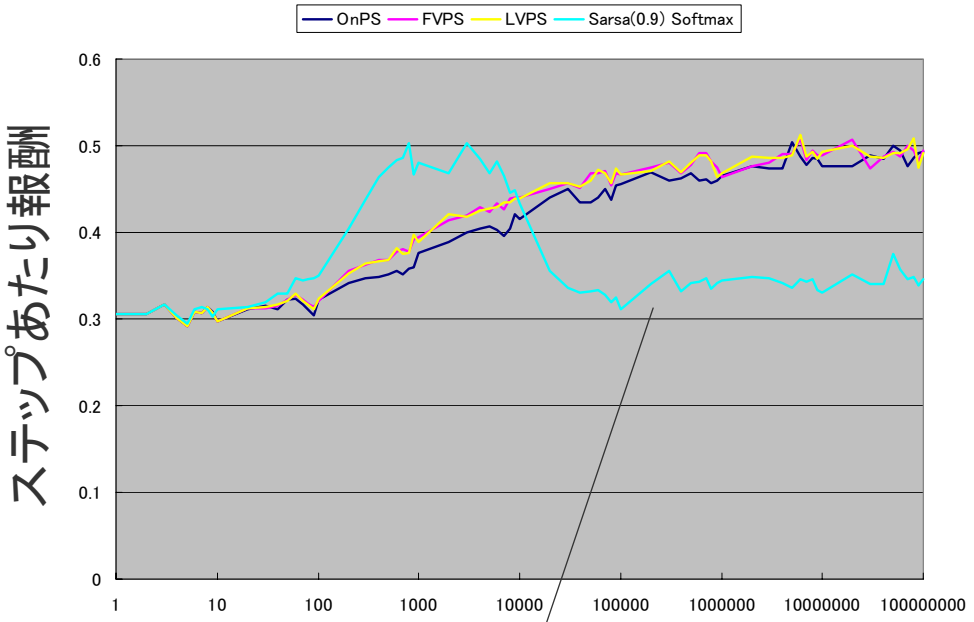
Parr 95 の結果

グリーディ政策の性能

ほぼ同等



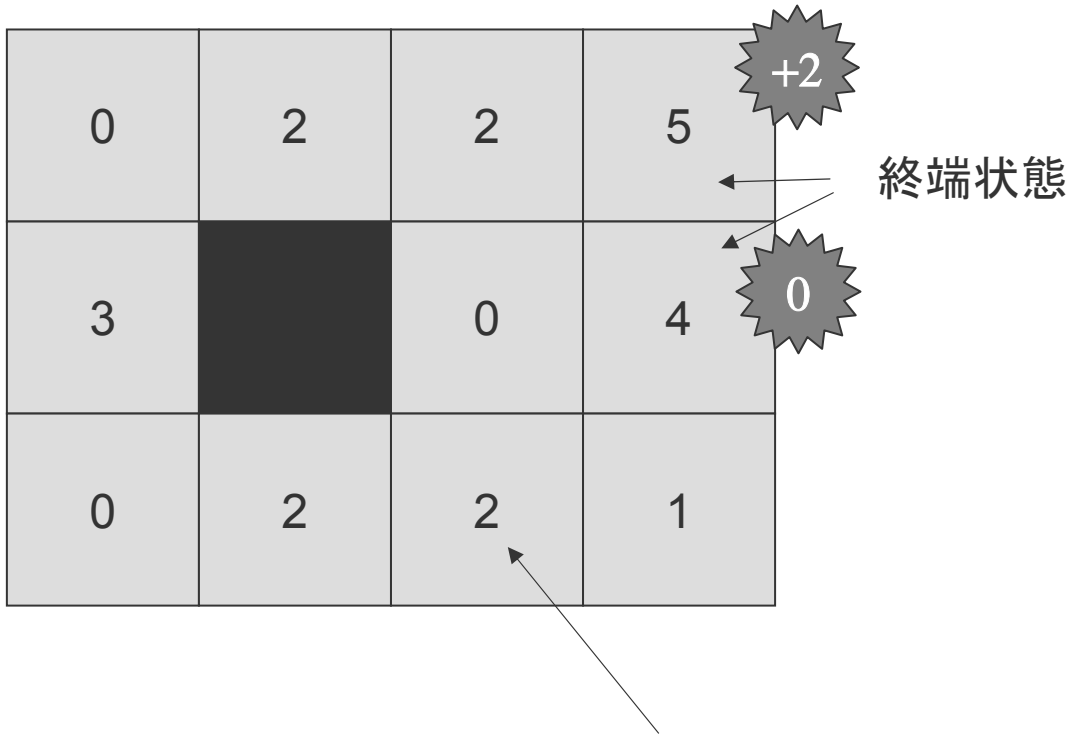
確率的政策の性能（学習中の性能）



Sarsa(0.9) だけ劣る

4x3 Maze (Parr & Russell, 1995)

実験のやり方は (Loch & Singh, 1998) と同じ



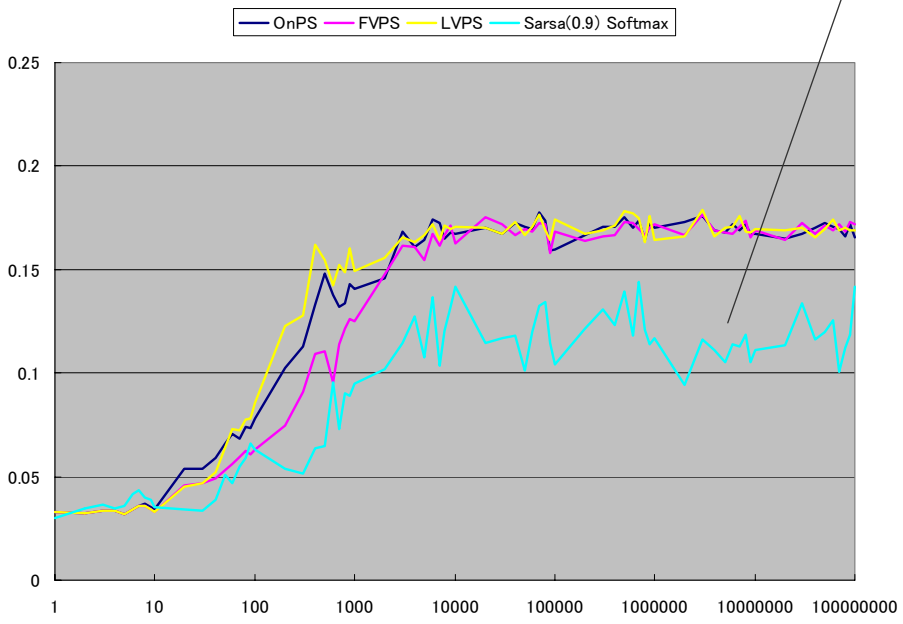
エージェントは，左隣および右隣のマスに
移動可能かどうか—だけを観測できる.

観測番号:
（同じ番号の状態は同一視される）

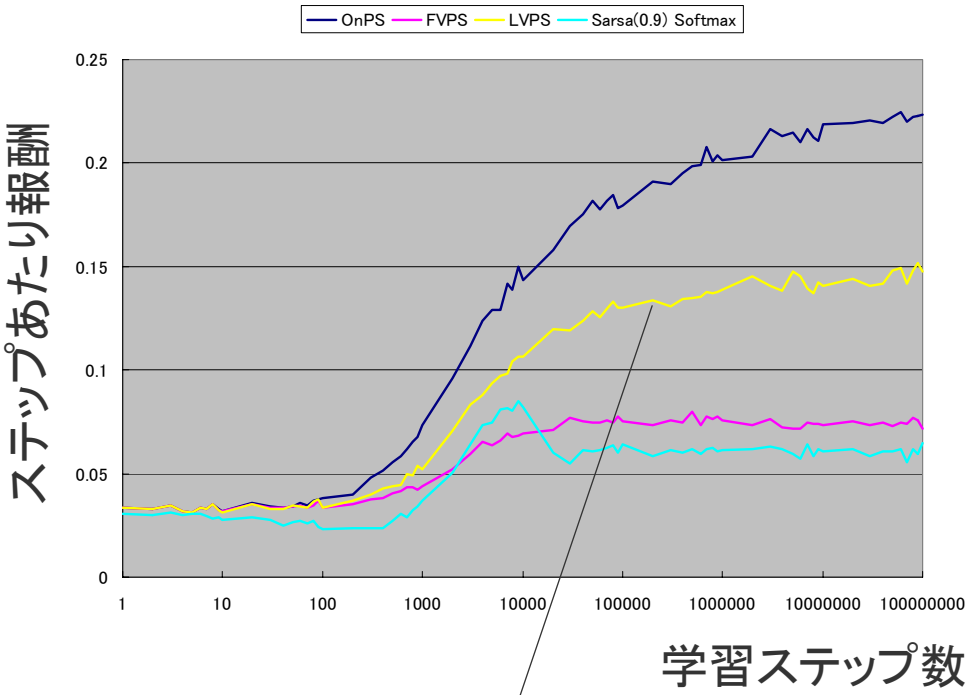
Russell 4x3 の結果

グリーディ政策の性能

Sarsa(0.9) だけ劣る

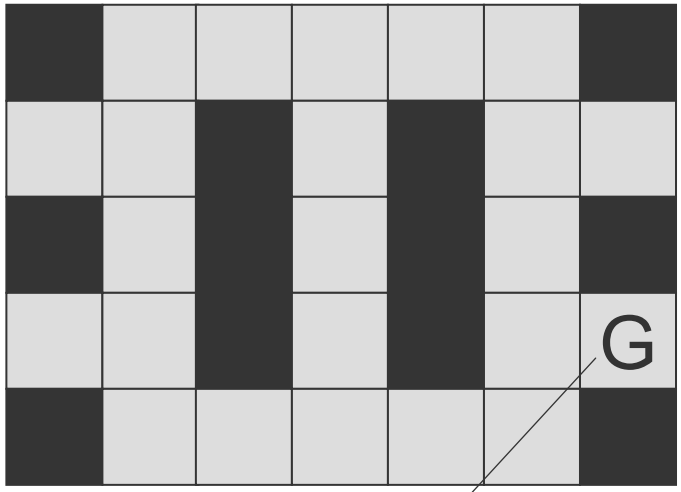


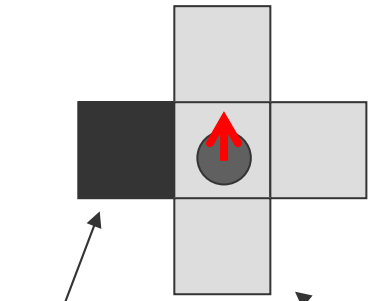
確率的政策の性能（学習中の性能）



1. OnPS, 2. LVPS, 3. FVPS, 4. Sarsa(0.9) の順に良い

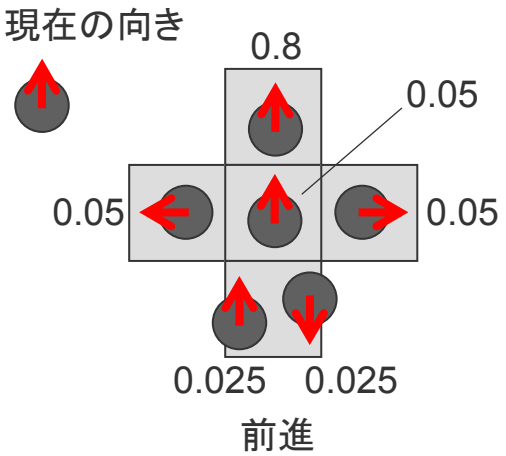
89 state office (Littman, Kassandora, Kaelbling, 1995)

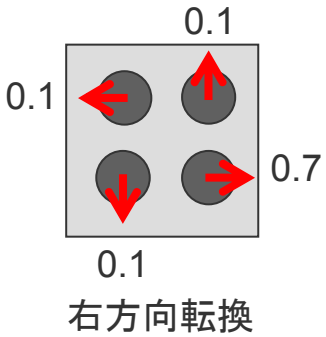


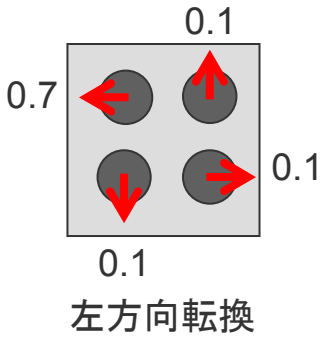


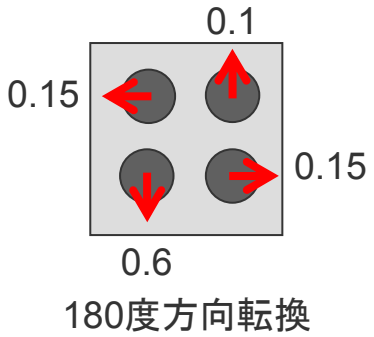
「方向」が考慮される
隣接するマスを独立に観測
行動:
- 前進
- 右方向転換
- 左方向転換
- 180度方向転換

「状態遷移」と「観測」は確率的



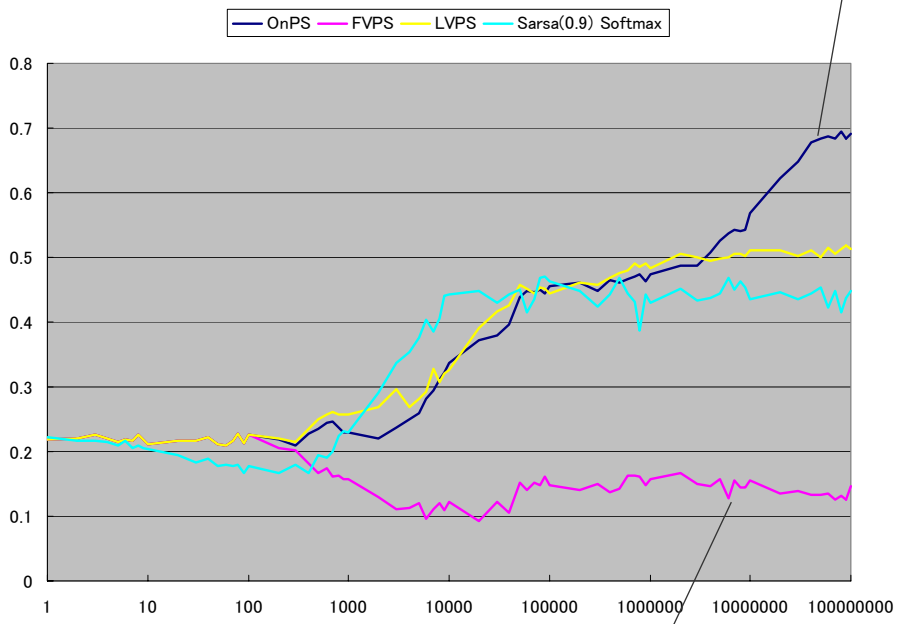






89 state office の結果（ゴール到達率）

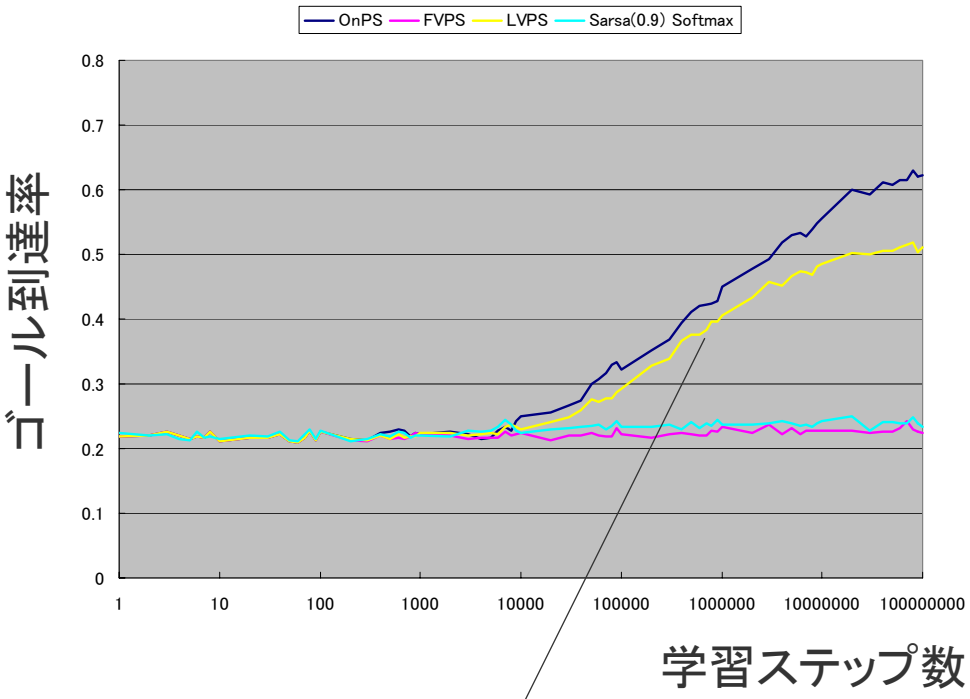
グリーディ政策の性能



OnPS は良い

FVPS は劣る

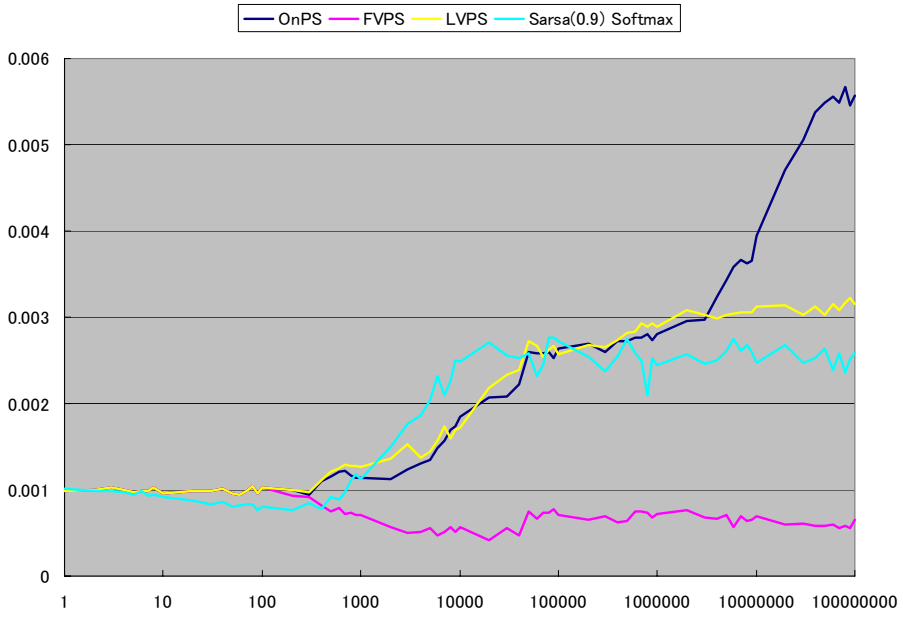
確率的政策の性能（学習中の性能）



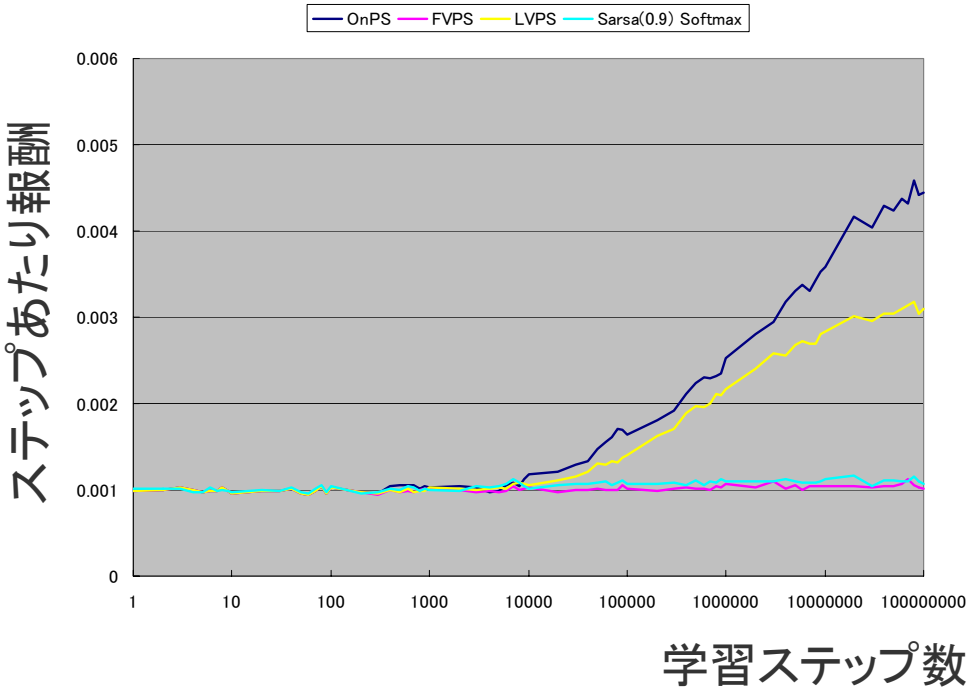
1. OnPS, 2. LVPS, 3. Sarsa(0.9), FVPS の順に良い

89 state office の結果（ステップあたり報酬）

グリーディ政策の性能

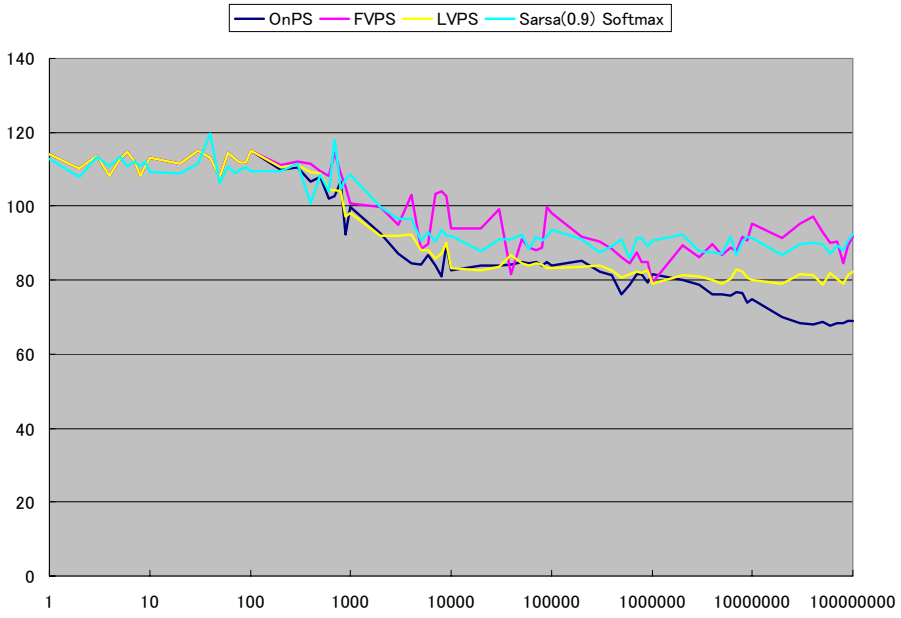


確率的政策の性能（学習中の性能）

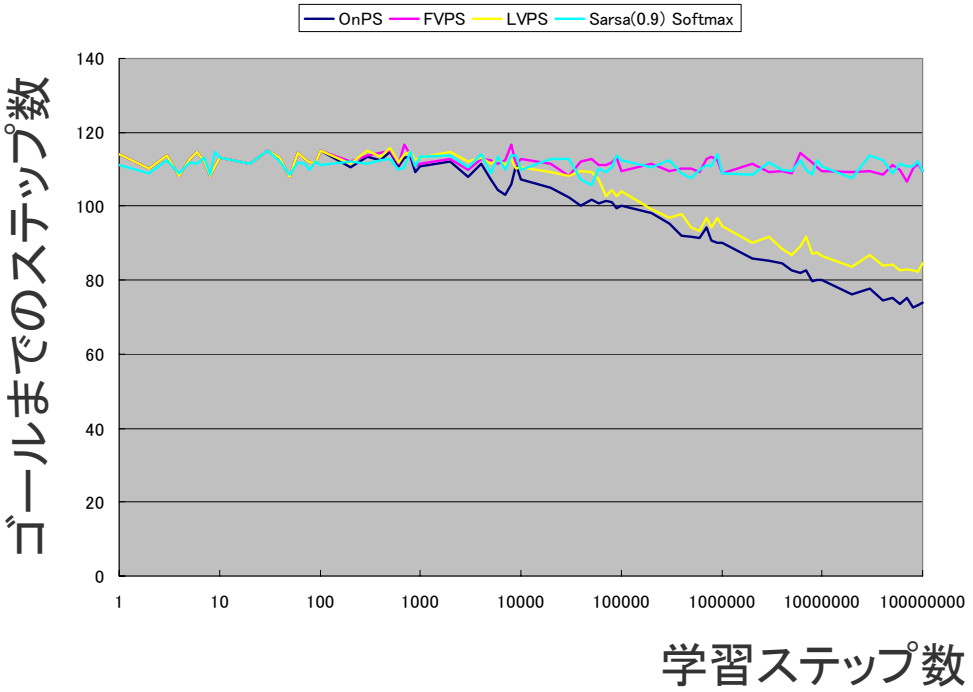


89 state office の結果（ゴールまでのステップ数）

グリーディ政策の性能



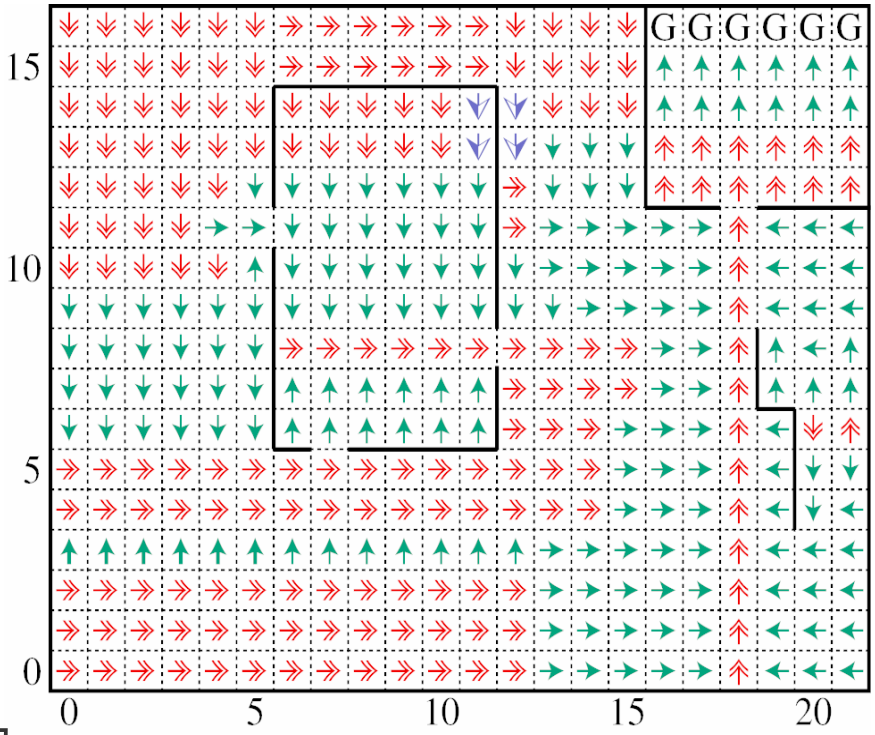
確率的政策の性能（学習中の性能）



環境の変化への適応

例題: 迷路の問題

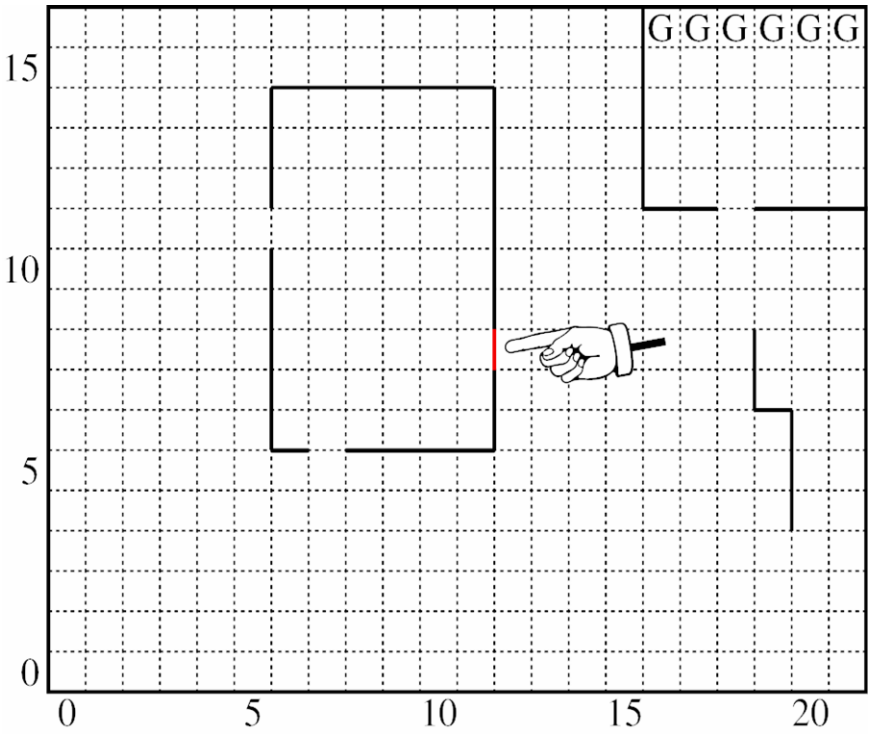
- ランダムな場所からスタート
- ゴール（G）へたどり着くと100点
- 利用可能なオペレータ
  - *Simple move*（コスト 1）
    - 隣のマスへ移動
  - *To-wall*（コスト 3）
    - 壁に突き当たるまで移動
  - *Wall-following*（コスト 5）
    - 壁に沿って，壁がなくなるまで移動



最適な解の一つ

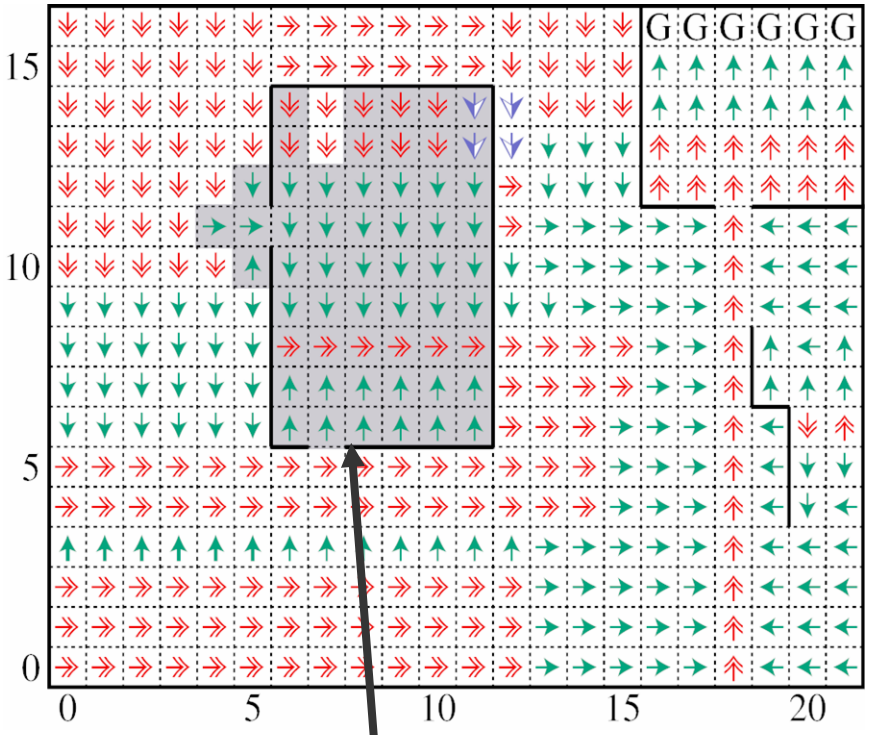
課題：環境の変化にどう適応するか？

- 環境が変化したからといって，学習をやりなおしたくない



環境変化の例
壁（矢印）が追加された.

知識を再利用できる部分・できない部分



矢印のとおりに行動しても
ゴールへたどり着けない

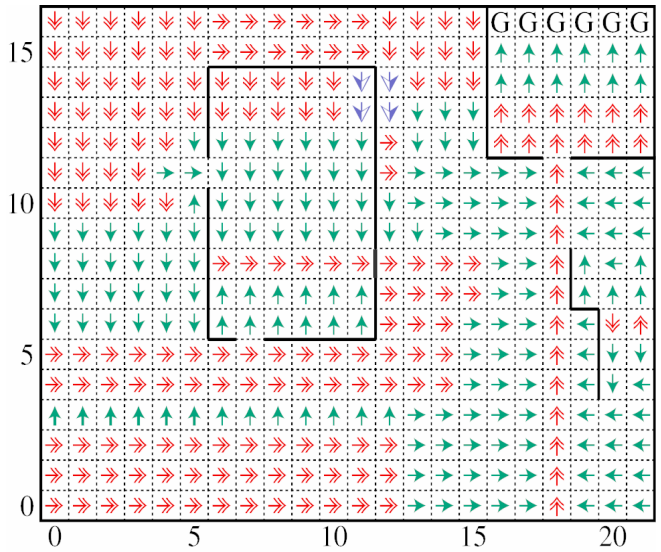
- 再利用できない部分だけを学習しなおせばいい

Q この部分をどうやって見つけるか？

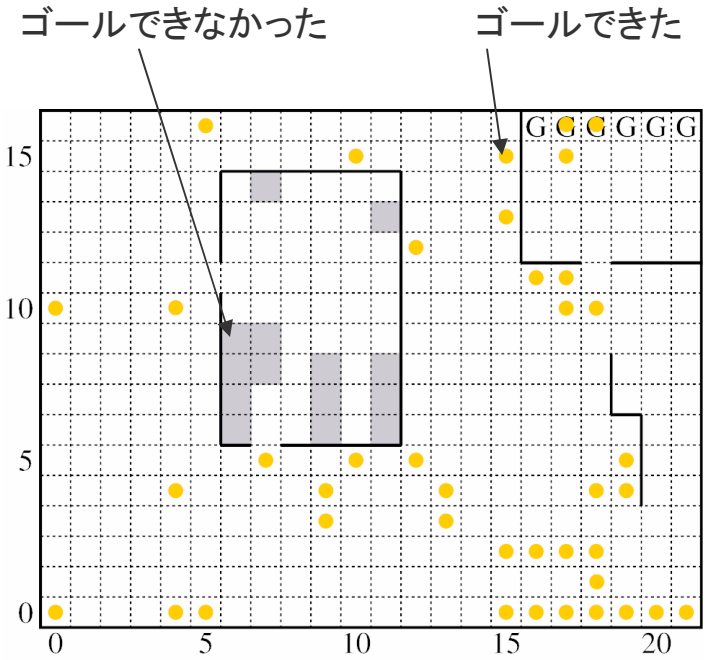
提案

概念学習を使って
帰納的にこの部分を
見つける

知識が利用できる範囲を概念学習で見つける



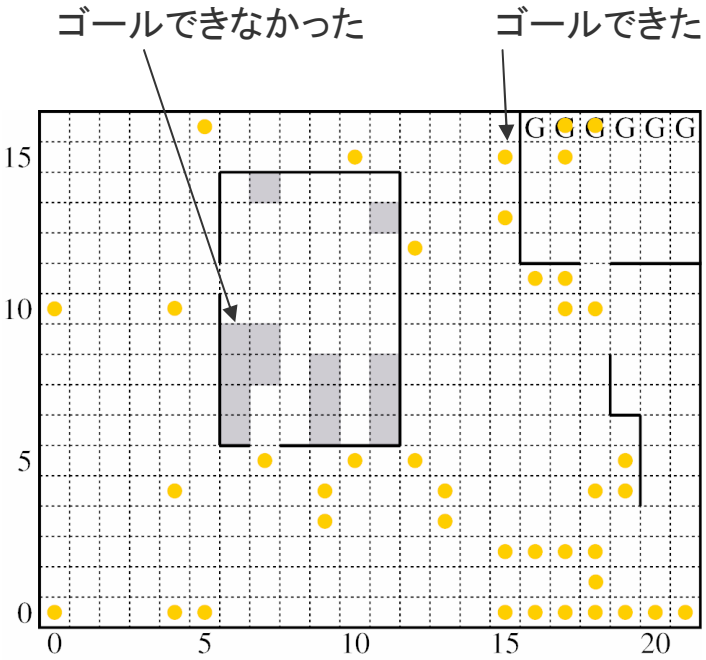
強化学習で学習した知識
（実際は，実験ごとに異なる知識を学習する）



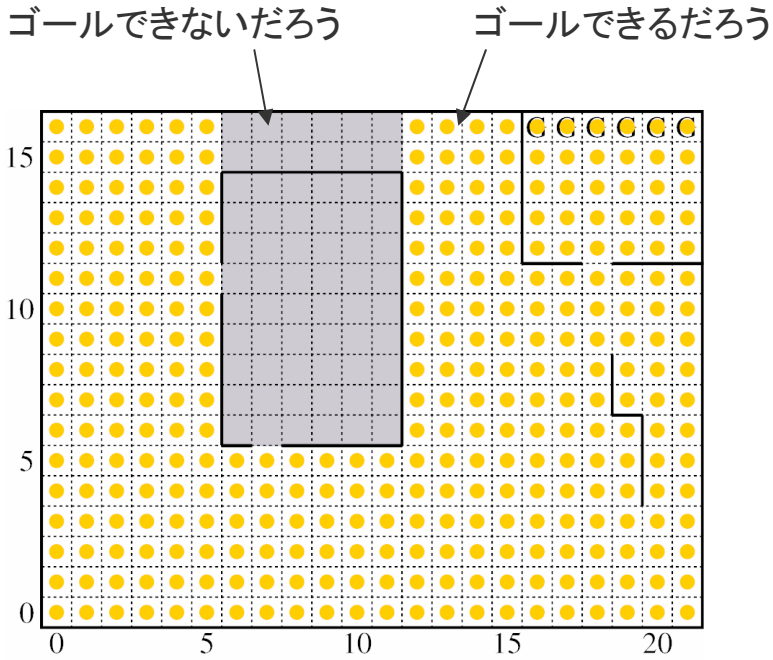
知識を使って実際に経験した状態

- 「経験した状態」を事例として，「失敗の概念」を学習

ゴールできないと予測される部分だけを再学習



実際に経験した状態



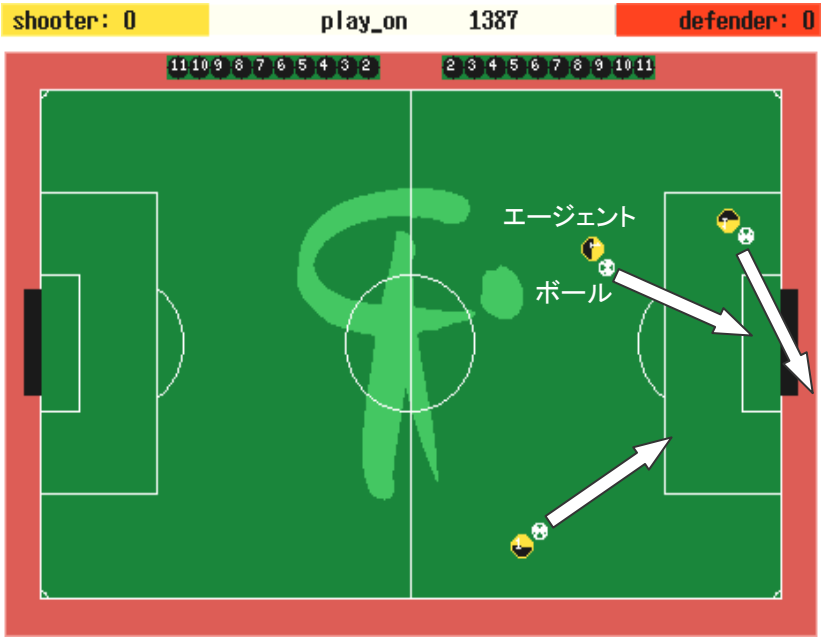
学習した決定木による分類

- 知識を再利用することにより，新しい環境（迷路）に早く適応できる

ロボットサッカーへの応用

- 単純なシュート政策の事前条件を学習

- 事例
  - 成功例　80
  - 失敗例　141
- 背景知識
  - サッカーについての知識
  - 大小比較



ロボットサッカーシミュレータのイメージ

* 実際には，エージェントとボールはひとつずつ

結果

- 学習した規則

can_kick_to_goal(A) :- my_speed(A,B), goal_distance(A,C), gteq(B,0.080), lteq(C,18.260).

can_kick_to_goal(A) :- my_speed(A,B), goal_distance(A,C), lteq(B,0.030), lteq(C,8.180).

* 自分の速度が 0.080 以上で，ゴールまでの距離が 18.260 以下ならば成功する
* 自分の速度が 0.030 以下で，ゴールまでの距離が 8.180 以下ならば成功する

（状態数）	結果は成功	結果は失敗	計
成功と予測	62.5±10.4	6.7±2.0	69.2±11.3
失敗と予測	18.5±10.3	143.4±5.7	161.9±12.8
計	81.0±1.3	150.1±5.8	231.1±6.7

成功率
36.2% → 90.3±2.4%

* 失敗と予測した場合に，行動を「ドリブル」にしたとき

まとめ

- 政策事前条件
  - 前の環境で学習した政策が，今の環境でも使えるかどうかを表す
  - 真のところでは，政策を再利用可能
  - 偽のところだけを学習しなおすことによって，環境の変化に早く適応できる
  - 成功例・失敗例から概念学習によって獲得できる
- 課題
  - 最適解を得るにはどうしたらよいか？
    - 今の条件は目標に到達することしか考えていない

強化学習に限らない話題

* 書籍の写真は Amazon.co.jp から引用しています

研究の進め方

1. 調査・勉強
   - 研究対象の基本技術や従来手法を調べる

2. 考察・提案
   - 従来手法の問題点を洗い出し，改善方法を考える

3. 実装・検証
   - 考えた方法を実装し，実験により効果を確認する
   - あるいは，考えた方法の有効性を理論的に示す

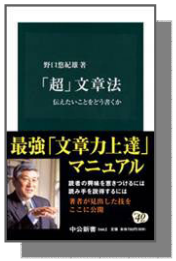
4. 発表
   - 結果を論文にまとめ，発表する

学会に入りましょう

- 人工知能学会
  - 人工知能学会誌がもらえる
- 情報処理学会
  - 情報処理学会誌「情報処理」がもらえる
- 電子情報通信学会
  - 電子情報通信学会誌と論文誌がもらえるらしい
- その他
  - 自分の研究に関連する学会ならどこでも O.K.

論文を書く前に読むべき5冊の本

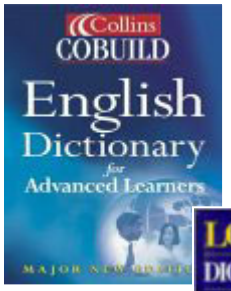
- 木下是雄「理科系の作文技術」中央新書
  - 理科系の学生必携の書.
- 野口悠紀雄「『超』文章法」中央新書
  - 日本語の文章をいかにわかりやすく書くか.
- 藤沢晃治「『分かりやすい説明』の技術」ブルーバックス
  - 「分かりやすい説明」と「分かりにくい説明」の違い.
- 酒井聡樹「これから論文を書く若者のために」共立出版
  - 「論文」には何を書いたらいいのか.
- 野矢茂樹「論理トレーニング101題」産業図書
  - 「したがって」など接続詞の使い方について.

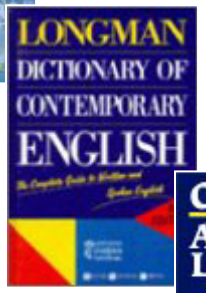


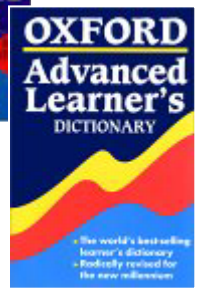
研究するために必要な3つの辞典

- 研究社「リーダーズ英和辞典」第2版
  - 古いものや大学受験用のものは捨てる

- 英英辞典
  - 英語の論文を読むとき，書くときに使う
  - 次のいずれか
    - Collins COBUILD English Dictionary
    - Longman Dictionary of Contemporary English
    - Oxford Advanced Learner's Dictionary



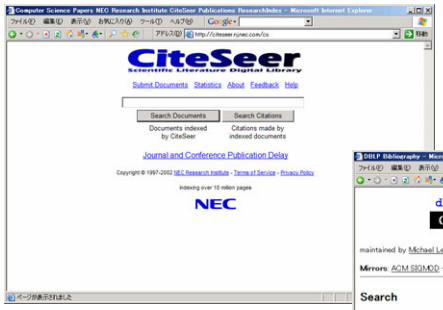


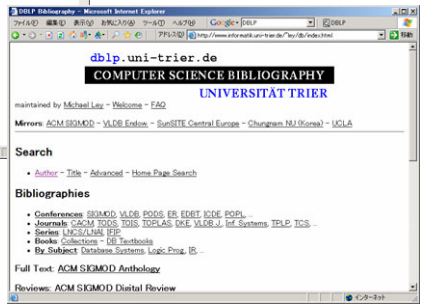


- シソーラス（英語の類義語辞典）
  - 英語の論文を書くときに使う
  - 書店で手に取り，自分が使いやすいものも選ぶ

研究するのに欠かせない3つのサイト

- CiteSeer
  - http://citeseer.nj.nec.com/
  - 論文を検索・入手できる





- DBLP Bibliography
  - http://www.informatik.uni-trier.de/~ley/db/
  - 論文を検索できる

- Amazon.co.jp
  - http://www.amazon.co.jp/
  - 本の評価や関連する本を調べられる
  - ユーザー登録すると本を薦めてくれる

その他

- リストマニア in Amazon.co.jp
  - http://www-sekilab.ics.nitech.ac.jp/~tohgoroh/index_j.html
  - 私が購入した本とその短評

連絡先

- E-mail:
  - tohgoroh@ics.nitech.ac.jp
- WWW
  - http://www-sekilab.ics.nitech.ac.jp/~tohgoroh/