逆強化学習法を用いた行動解析

Behavior Analysis by Inverse Reinforcement Learning

沖縄科学技術大学院大学

神経計算ユニット　Neural Computation Unit

教授　銅谷　賢治　Professor Kenji Doya

OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY
沖縄科学技術大学院大学

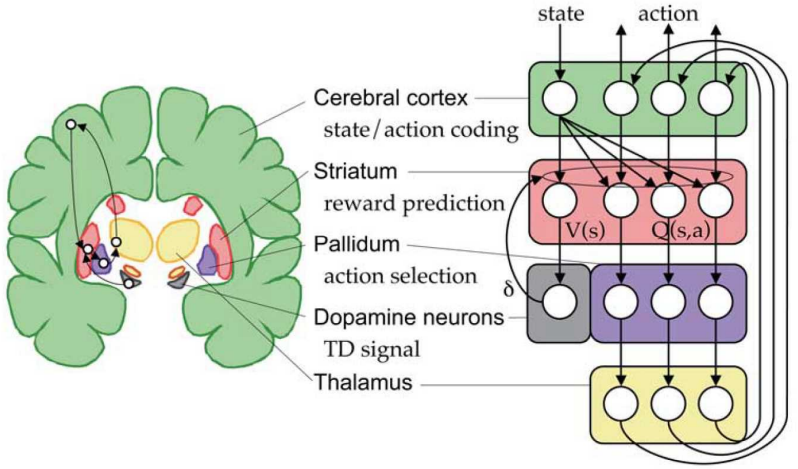
強化学習とは

- 試行錯誤を通して制御則（行動ルール）を学ぶ人工知能技術
- 囲碁のチャンピオンに勝利したアルファ碁は強化学習とディープラーニングの組み合わせ
  ➡ ロボットなどの制御へ応用
- ヒトや動物の意思決定のモデルとしても注目
  ➡ 脳科学の観点からの説明

[Nature Blog. The Go Files: AI computer wraps up 4-1 victory …]

強化学習

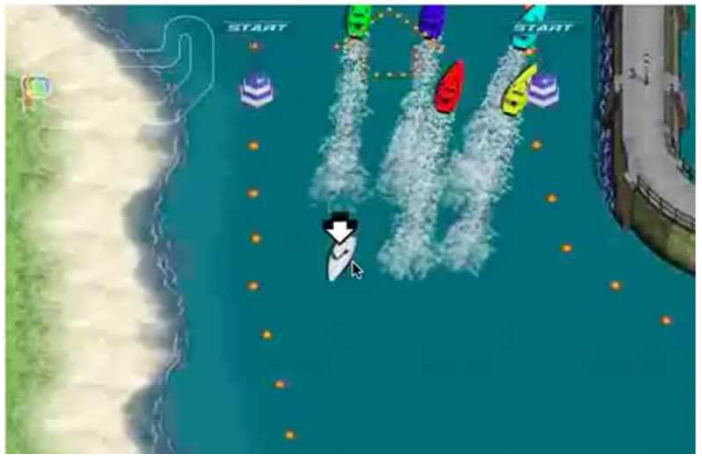
目的関数（報酬） → 制御則



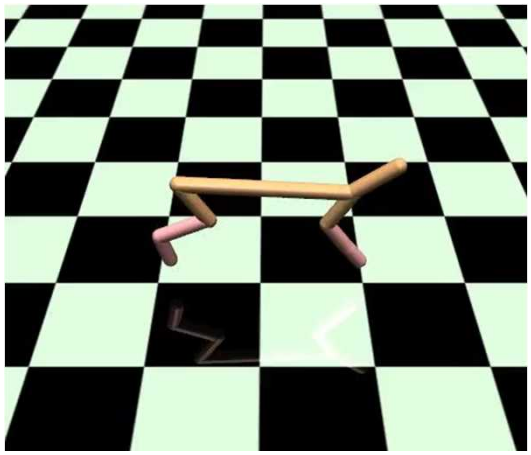
[Doya, 2007]

逆強化学習とは

- 単純な目的を使うと
  膨大な学習データと
  計算時間が必要
- 適切な目的を設計する
  ことは困難

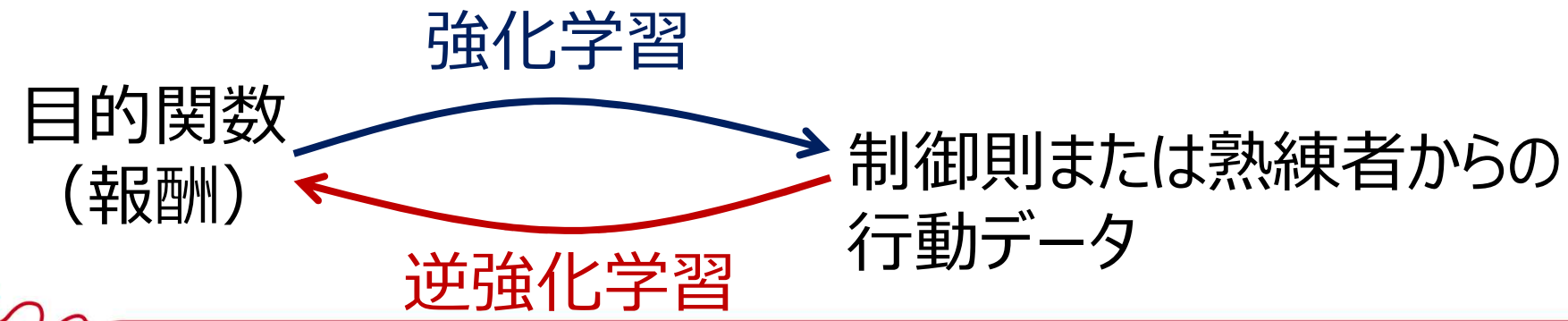


[OpenAI Blog. Faulty Reward …]

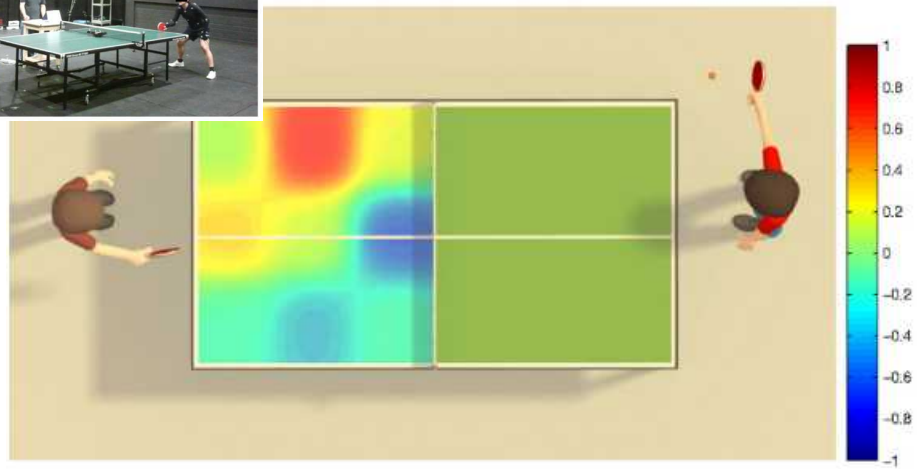


[Sorta Insightful (Blog)]

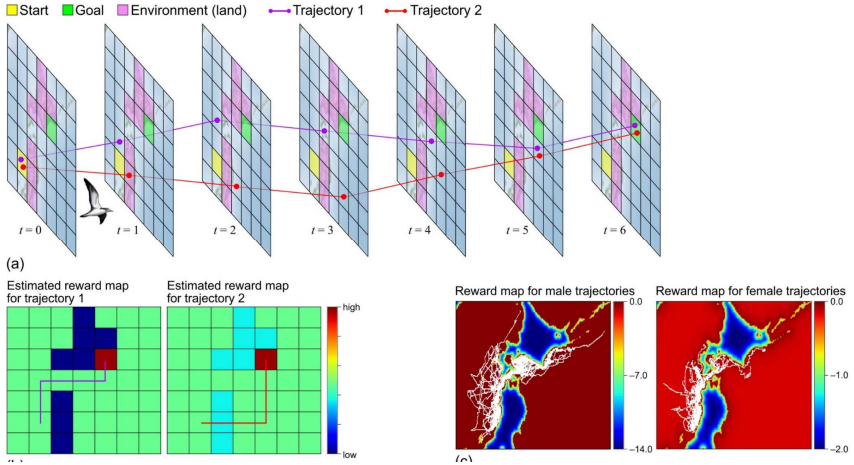
- 熟練者の行動データ
  をもとに目的を推定する技術が逆強化学習



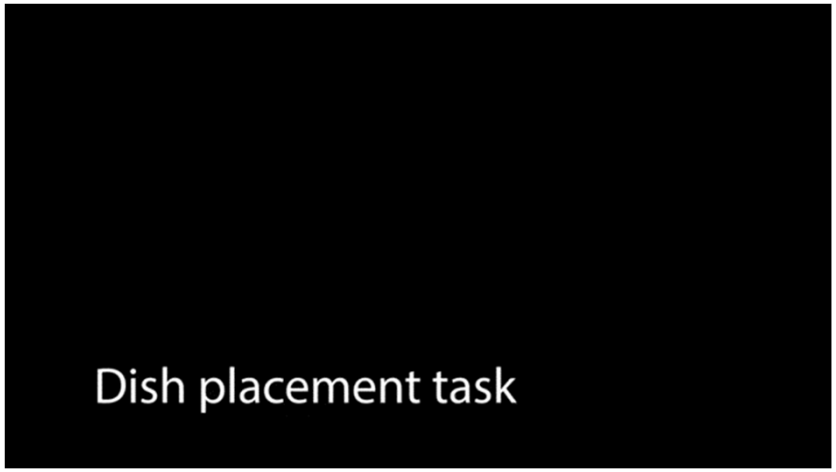
逆強化学習の適用例



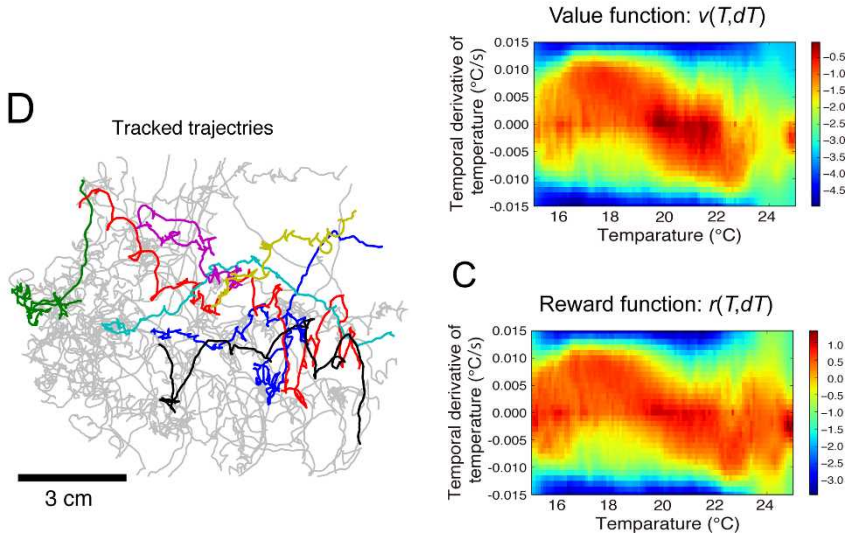
卓球の戦略 [Mueling et al., 2014]



海鳥の飛行経路予測 [Hirakawa et al., 2018]



ロボットの模倣学習 [Finn et al., 2016]

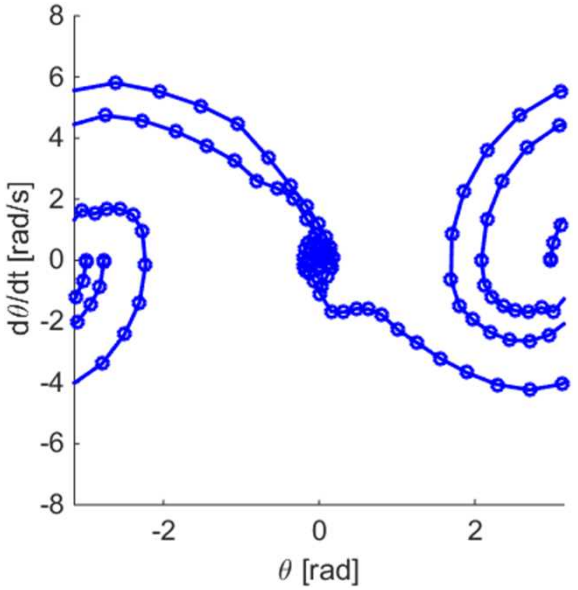


線虫の行動 [Yamaguchi et al., 2018]

従来技術とその問題点

- 現在主流となっている方法は
  最大エントロピ法を利用した逆強化学習

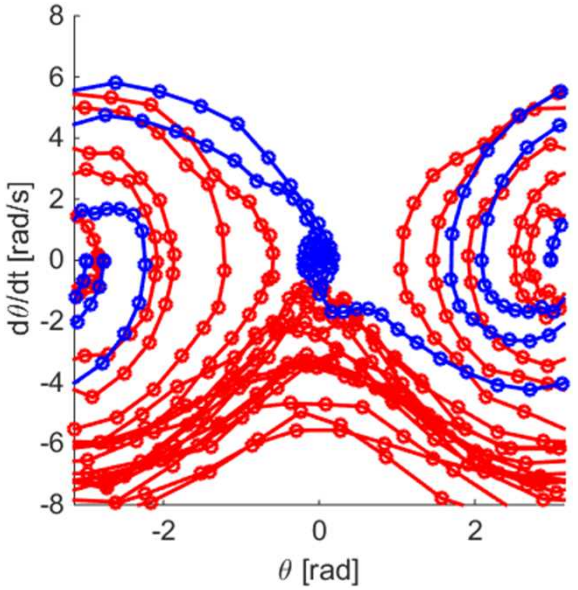
- 問題点
  - モデルベース手法
    ➡ 制御対象の数学モデルが必要
  - データの単位が状態の系列 ($x_0, x_1, \ldots, x_T$)
    ➡ データの収集コストが高い
  - 解析したい成功行動データだけから推定
    ➡ 確率モデルの学習に相当し、複雑な問題を解くことになる



新技術の特徴・従来技術との比較

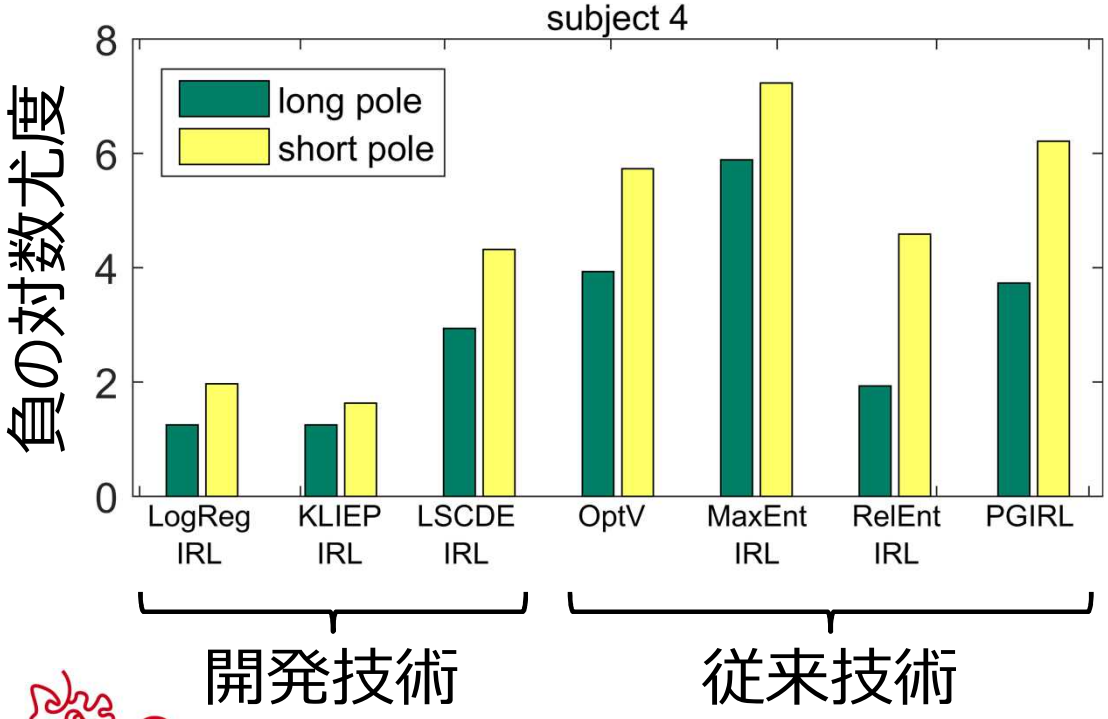
- 開発した技術はエントロピ正則化と
  密度比推定を利用した逆強化学習

- 問題点の解決
  - モデルフリー手法
    ➡ 制御対象の数学モデルが不必要
  - データの単位が状態の遷移 (x_t, x_{t+1})
    ➡ データの収集コストが低い
  - 成功行動データと基準行動データから推定
    ➡ 事例の識別問題に相当し、簡単な最適化問題になる

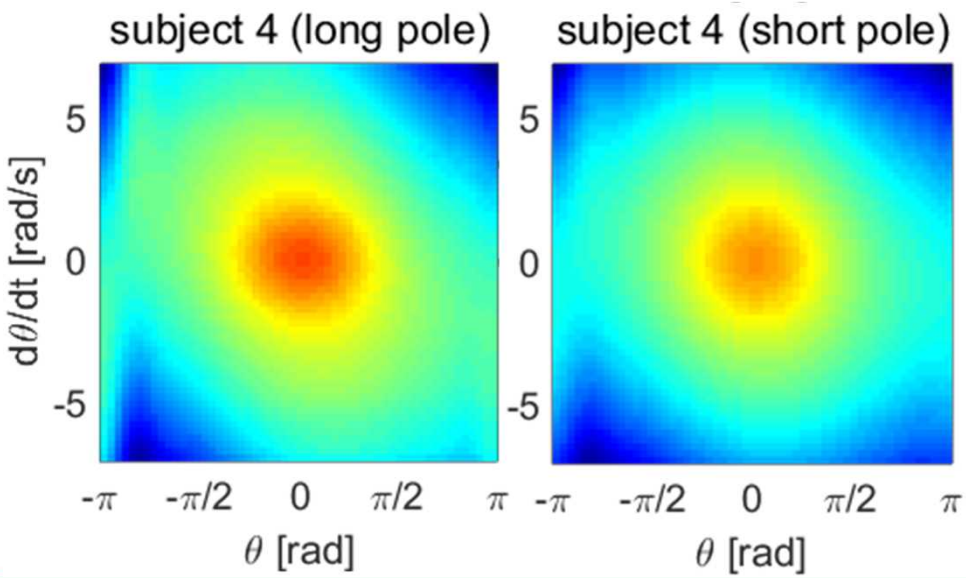


従来技術との比較（人の運動の解析）

- 振り子の振り上げ・安定化課題に対して行動データを計測
  - ➡ 開発技術は従来技術よりも被験者の動作を復元できている

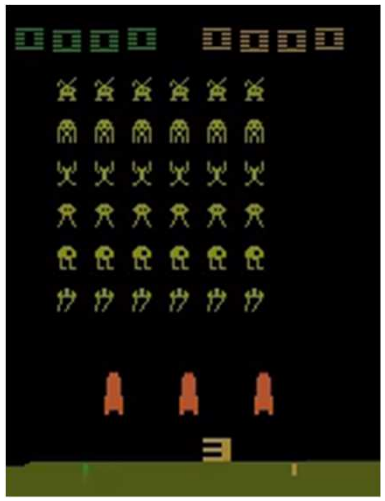


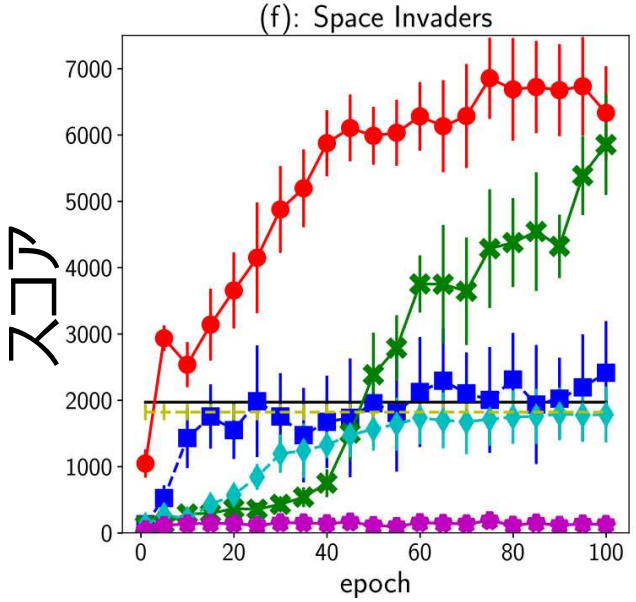
推定された目的関数

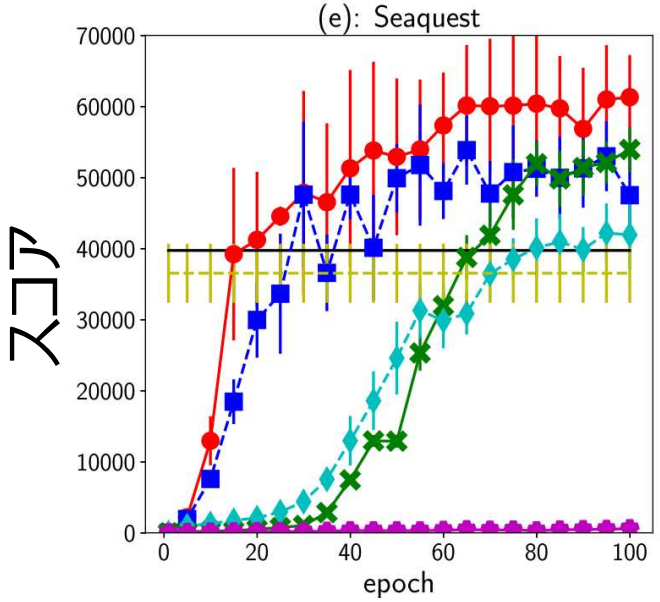


従来技術との比較（ゲームプレイの解析）

- 人のゲームプレイデータからゲームをうまくプレイするための目的を推定し、人工エージェントにゲームを学習（模倣学習）
  - ➡ 少ないデータから効率よく制御則を学習

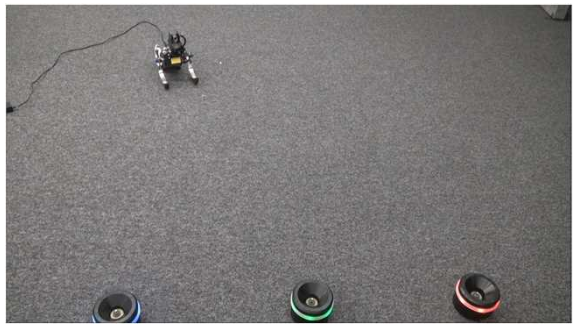


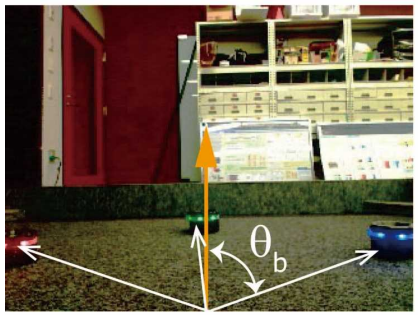




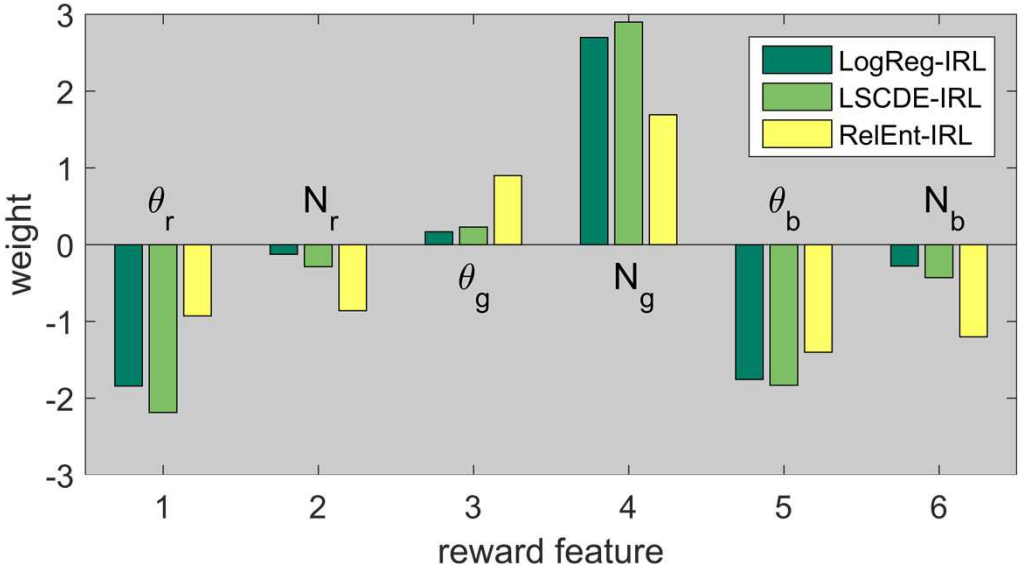
従来技術との比較（ロボット制御）

- 人の操作データからロボットの行動を学習するための目的と制御則を学習



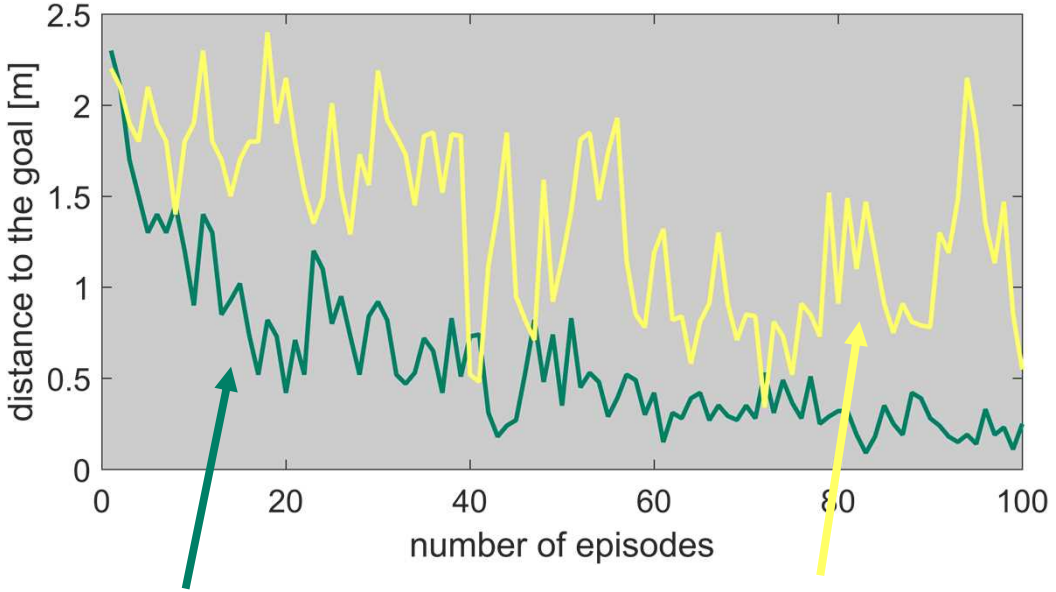


推定報酬のパラメータ



目標の大きさNgへの正の評価と
ランドマークとの角度θr,θbへの負の評価が重要

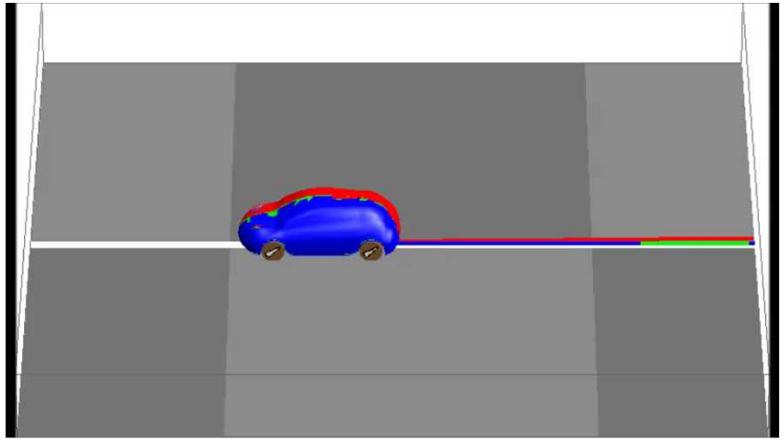
推定報酬を使って制御則を学習した時の比較



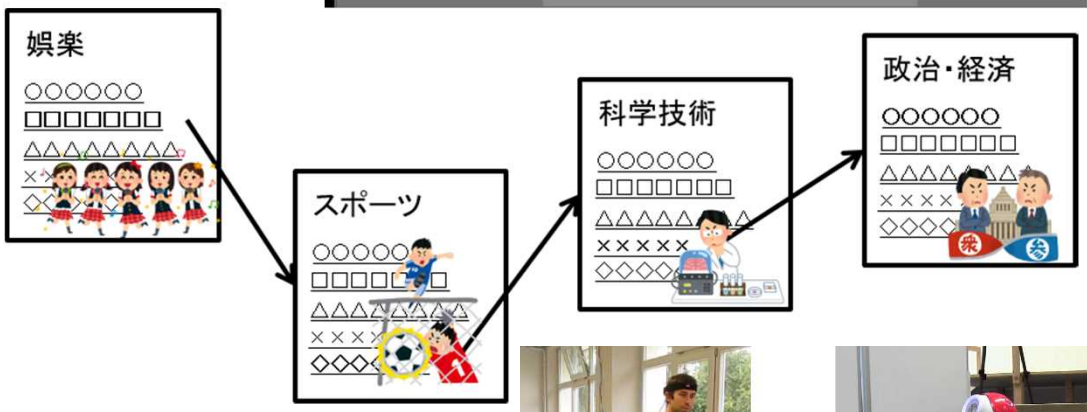
新技術: 早く収束　　従来技術: 不安定

想定される用途

- ドライビング技術の解析
  - ➡ 熟練ドライバと初心者を比較し、どのような特徴が「うまい」運転に重要かを調査



- ウェブサイトのアクセス解析
  - ➡ ユーザの閲覧履歴から行動の意図を推定

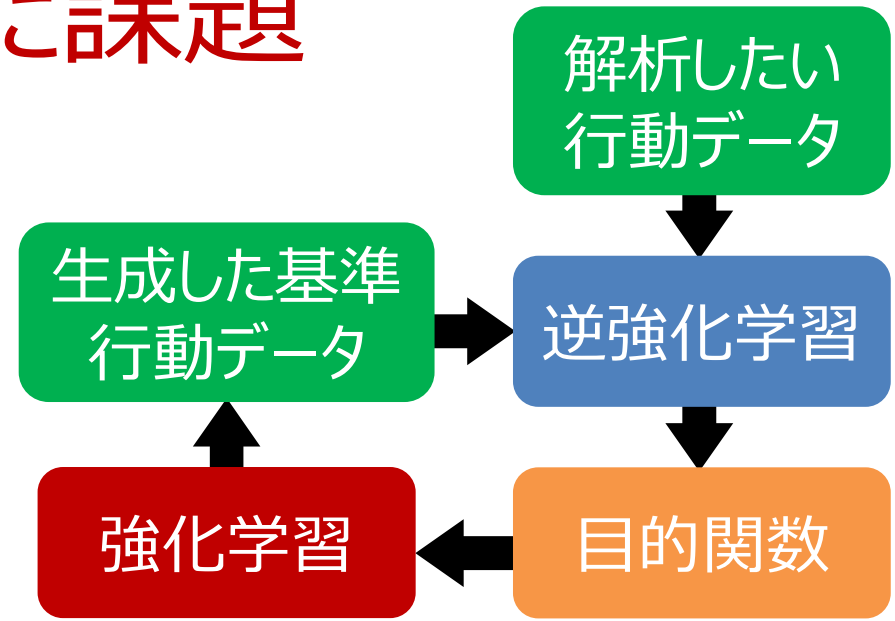


- ロボットの模倣学習
  - ➡ 人の動作データを使ってロボットの制御則を作成

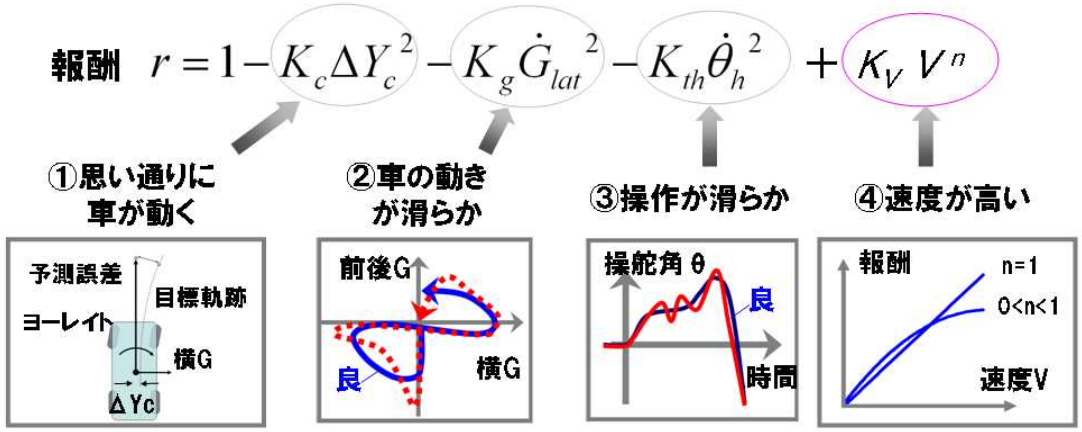
[Koenemann et al., 2014]

実用化に向けた課題

- 解析したい行動データと比較する
  基準行動データが必要
  ➡ 推定した目的関数と強化学習で
  基準データを生成
  ただしシミュレータが必要



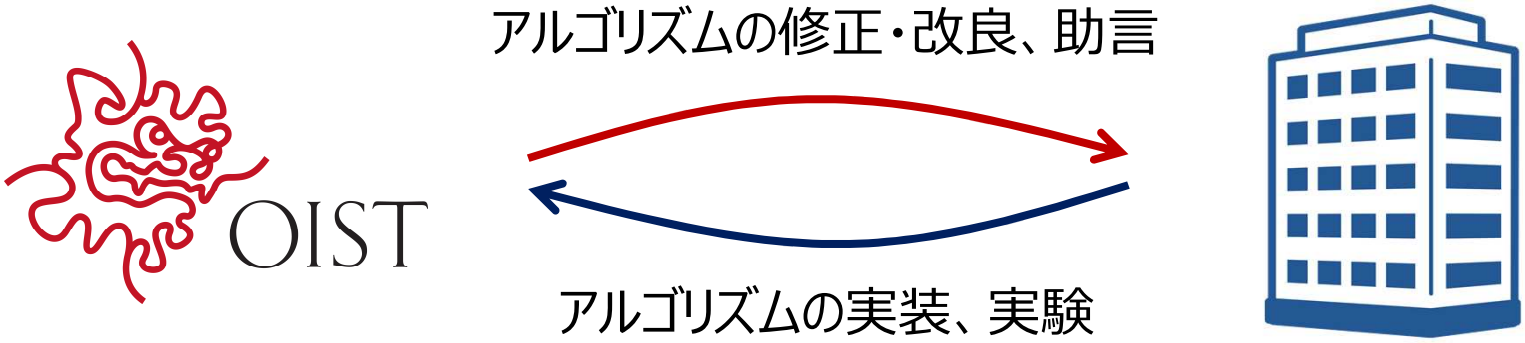
- 目的関数は特徴となる要素の重みつき
  線形和であることを仮定
  ➡ ディープラーニングを使った
  特徴の同時学習
  ただしデータ数は増加



企業への期待

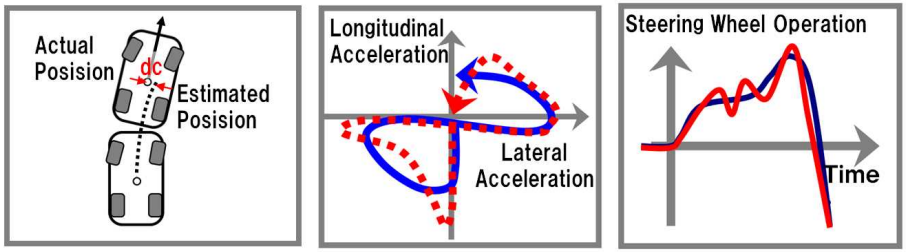
- ヒトの行動データの解析や、ヒトの作業のロボットへの移植に興味を持ち、実際にデータを所有している企業との共同研究を希望

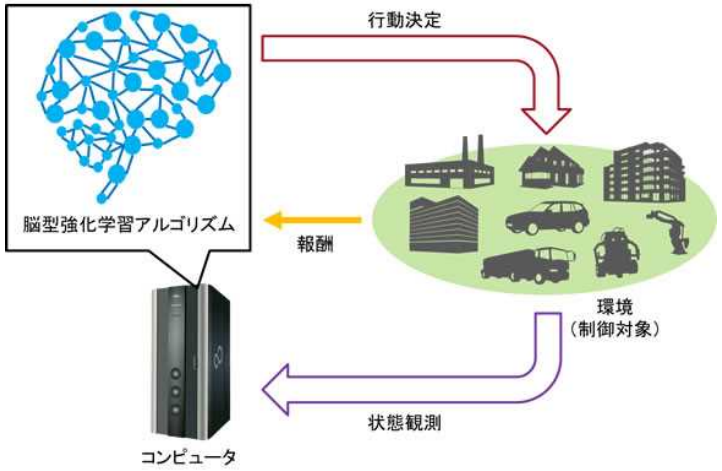
- 当研究室ではアルゴリズムの開発などの基礎研究が主体



産学連携の経歴

- 2016年-2018年
  富士通研究所と脳科学の知見から、
  強化学習アルゴリズムの共同研究
  [Sasaki et al., 2017a; 2017b]





- その他、メーカー企業との共同研究経験
  多数あり

特許

発明の名称：　(1) 逆強化学習の方法、逆強化学習用アルゴリズムをプロセッサに実行させる指示を記憶する記憶媒体、逆強化学習用システム、及び逆強化学習用システムを含む予測システム
(2) 密度比推定による直接逆強化学習

出願番号：　(1) 日本 登録　特許6417629
米、欧、中、韓へ出願済み
(2) 日本 特願2018-546050
米、欧、中、韓へ出願済み

出願人：　沖縄科学技術大学院大学（単独）

発明者：　銅谷　賢治　（教授）
内部　英治　（グループリーダー研究員）

お問い合わせ先

沖縄科学技術大学院大学（OIST）
技術移転セクション

TEL　：　098-966-8937
FAX　：　098-982-3424
E-mail　：　tls@oist.jp

参考文献

- Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. HFSP Journal, 1(1): 30-40.
- Finn, C., Levine, S., and Abbeel, P. (2016). Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization. In Proc. of ICML 33, 49-58.
- Hirakawa, T., Yamashita, T., Tamaki, T., Fujiyoshi, H., Umezu, Y., Takeuchi, I., Matsumoto, S., and Yoda, K. (2018). Can AI predict animal movements? Filling gaps in animal trajectories using inverse reinforcement learning. Ecosphere.
- Koenemann, J., Burget, F., Bennewitz, M. (2014). Real-time imitation of human whole-body motions by humanoids. In Proc. of ICRA, pp. 2806-2812.
- Muelling, K., Boularias, A., Mohler, B., Schölkopf, B., and Peters, J. (2014). Learning strategies in table tennis using inverse reinforcement learning. Biological Cybernetics, 108(5): 603-619.
- Nature Blog (2016). The Go Files: AI computer wraps up 4-1 victory against human champion.
- OpenAI Blog (2016). Faulty Reward Functions in the Wild.
- Sakuma, T., Shimizu, T., Miki, Y., Doya, K., and Uchibe, E. (2013). Computation of driving pleasure based on driver's learning process simulation by reinforcement learning. In Proc. of Asia Pacific Automotive Engineering Conference.

参考文献

- Sasaki, T., Uchibe, E., Iwane, H. Yanami, H., Anai, H., and Doya, K. (2017a). Policy gradient reinforcement learning method for discrete-time linear quadratic regulation problem using estimated state value function. In Proc. of SICE Conference.
- Sasaki, T., Uchibe, E., Iwane, H., Yanami, H., Anai, H., and Doya, K. (2017b). Derivation of integrated state equation for combined outputs-inputs vector of discrete-time linear time-invariant system and its application to reinforcement learning. In Proc. of SICE Conference.
- Sorta Insightful (2018). Deep Reinforcement Learning Doesn't Work Yet.
- Uchibe, E. and Doya, K. Inverse reinforcement learning using dynamic policy programming. In Proc. of ICDL and Epirob, 2014.
- Uchibe, E. Model-Free Deep Inverse Reinforcement Learning by Logistic Regression. Neural Processing Letters, 47(3): 891-905, 2018.
- Yamaguchi, S., Honda, N., Ikeda, M., Tsukada, Y., Nakano, S., Mori, I., and Ishii, S. (2018). Identification of animal behavioral strategies by inverse reinforcement learning. PLoS Computational Biology.
- Ziebart, B.D., Maas, A., Bagnell, J.A., and Dey, A. (2008). Maximum entropy inverse reinforcement learning. In Proc. of AAAI, 1433-1438.