

シナジーを利用した低品質な教示動作を含むデータセットからの模倣学習

Imitation learning from datasets containing low-quality demonstrations utilizing synergy

○田中裕人*, 沓澤京*, 大脇大*, 林部充宏*

○Yuto Tanaka*, Kyo Kutsuzawa*, Dai Owaki*, Mitsuhiro Hayashibe*

*東北大学

*Tohoku University

キーワード: 機械学習 (machine learning), 模倣学習 (imitation learning), シナジー (synergy)

連絡先: 〒980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-01 東北大学 青葉山キャンパス 機械系共同棟
503 田中裕人 Tel.:022-795-6970 Fax.:022-795-6971 E-mail: yuto.tanaka.r8@dc.tohoku.ac.jp

1. 緒言

模倣学習は、人間のデモンストレーションなどからその行動軌跡を模倣する学習アルゴリズムである。模倣学習は報酬設計を必要としない効果的な手法であり、強化学習におけるサンプリング効率の問題を解消できる手法として注目され、ロボット制御や自動運転システムなど幅広い分野で利用されている。

模倣学習はその性質上、模倣するデモンストレーションが最善のものであるという前提に基づいている。しかしながら、現実問題として、人間からデモンストレーションを取得する際、被験者間の熟練度の違いや疲労などから、模倣学習のデータには不完全なデータ（熟練度の低いデータ）が含まれることがある。この場合、模倣学習は不完全なデータをも模倣してしまうため、結果としてその性能を著しく低下させてしまうことがある。関連研究では、軌跡に対してランク付けや信用度の割り振りを行い、模倣

の優先度を定義することでこの問題に対処した¹⁾²⁾。一方で、それらの手法は複雑なアルゴリズムを要し、また人の手によるラベリングを必要とする。そこで本研究では、人間の運動に内在するシナジーという低次元化された構造に着目した、より簡潔かつ手動のラベリングを要しない手法でこの問題に対処することを試みた。

近年の神経科学の研究において、人や動物はシナジーと呼ばれる低次元化された機能群を用いて、冗長な手足の多数の関節を効率よく制御していることが示唆されている³⁾。シナジーはこれまでリハビリテーションやロボット工学の分野において利用されてきたが、先述の模倣学習における課題に対して利用されたことはない⁴⁾⁵⁾。一方で近年ではシナジーと運動の熟練度の関係が注目されており、Gentner ら⁶⁾は、熟練度の高いバイオリニストでは演奏動作において非熟練者よりもシナジーが発現していることを示した。また Chai らの研究⁷⁾では、ロコモーションにおける強化学習において学習が進むに

つれてエージェントの運動に低次元構造シナジーが発現することが明らかにされており、またシナジーと高いエネルギー効率との関係も示唆されている。同様に Han らの研究⁸⁾においても、リーチング運動における深層強化学習においてシナジーが発現することが明らかにされている。そこで、これらの研究結果および知見から、本研究ではシナジーを模倣学習の課題に利用できると考えた。シナジーを模倣学習に用いることで、学習が進んでいない不完全な軌道と、学習が進んだ最善な軌道を選択的に生成できると仮説を立て、検証を行った。検証の結果、本手法は高次元の連続制御タスクにおいて、ベースライン（通常の模倣学習）を上回った。

2. 手法

2.1 Spatial Synergy

本研究ではデータセットの軌跡における action (agent の joint torque) から Spatial Synergy を抽出し、評価した。Spatial Synergy は多次元の時空間データから低次元の時空間データとして抽出される。Spatial Synergy は次式のように表される。

$$x^l(t) = \sum_{p=1}^P c_p^l \cdot w_p(t) + residuals \quad (1)$$

$$X = W \cdot C \quad (2)$$

ここで $x^l(t)$ は1番目の試行の全関節への入力トルク、 w, c はそれぞれ抽出されるシナジーおよび可変パラメータである。また P は抽出されるシナジー数を表している。(2) 式は (1) 式の残差項を無視し、行列形式で書き換えたものである。 W は空間的なシナジー（どの関節を同時に動かすか）、 C は各時刻におけるシナジーの活性度を表す。この Spatial Synergy は PCA（主成分分析）を用いて *residuals* を最小化することにより、近似的に求めることができる。また

シナジー $W \cdot C$ が元データ X をどれほど正確に再構成できるかを次式に示す R^2 指標にて評価することができる。

$$R^2 = 1 - \frac{\|X - \hat{X}\|_F^2}{\|X - W \cdot C\|_F^2} \quad (3)$$

ここで $\|\cdot\|_F$ はフロベニウスノルムを表している。 R^2 の値が大きいほど元のデータの再現率が高い。しかしながら R^2 指標は抽出するシナジー数 P によって値が変化してしまう。そのため本研究では低次元構造の指標として各 P における R^2 の平均をとった *Synergy Level Index (SLI)* を用いる。

$$SLI = \frac{1}{N} \sum_{i=1}^N R_i^2 \quad (4)$$

SLI を用いることで P に依らないシナジーの評価が可能となる。

2.2 模倣学習

2.2.1 Behavior cloning および提案手法

本研究では単に suboptimal な軌道が含まれるデータセットから模倣学習することを目的とする。そのため過去の行動履歴から方策を復元するシンプルな模倣学習アルゴリズムである Behavior Cloning (BC) によって学習する。BC の目的関数は次のように表すことができる。

$$\min_{\pi} E_{(s,a) \sim \mathcal{D}} [-\log \pi(a|s)] \quad (5)$$

s, a はそれぞれデータセット $\mathcal{D}$ から得られる状態、行動である。本手法を次式に示す。本手法では後述の SLI と報酬和の関係から、状態 s に加え、標準化した SLI, β で条件付けすることで SLI に応じた軌道を生成できるよう模倣学習を行う。つまり本手法を用いることで optimal と suboptimal な軌道が混在したデータであっても

optimal な軌道 (SLI が高い軌道) を選択的に生成することができる。

$$\min_{\pi} E_{(s,\beta,a)\sim\mathcal{D}}[-\log \pi(a|s,\beta)] \quad (6)$$

2.3 評価方法

本研究では以下の二つのデータセットで本手法を評価した。また本手法の概要図とアルゴリズムを Fig.3, Algo.1 に示す。

2.3.1 D4RL

本研究では模倣学習のデータセットとして D4RL⁹⁾ のロコモーションデータセット medium-expert データセット (halfcheetah, Walker2d, Hopper) を用いた。medium-expert は medium(学習を途中で打ち切った方策) と expert (ほぼ完全に学習が完了した方策) の混合データセット (suboptimal と optimal を含む) であり、本研究が想定している問題設定に最適であるといえる。初めにそれぞれのデータセットにおいて長さ 1000step(10 秒間) の軌跡を取り出し、それぞれの軌跡に PCA を行うことで各軌跡の SLI を算出した。次に SLI と報酬和に相関があることを示し、 SLI で条件付けされた提案手法で模倣学習を行い、シミュレーション環境でその性能を評価した。また β 値を変化させることで性能に生じる変化を確認した。

2.3.2 Robomimic

より実用的なタスクにおいても検証するため、模倣学習データセット Robomimic¹⁰⁾ における Pick and Place Can タスクにおけるデータセットを用いた。このデータセットは多数の人間から集められたデータセットであり、それぞれのデモンストレーションの品質が異なる。2.3.1 と同様、一回のデモンストレーションを取り出し、それぞれの軌跡に PCA を行うことで各軌跡の

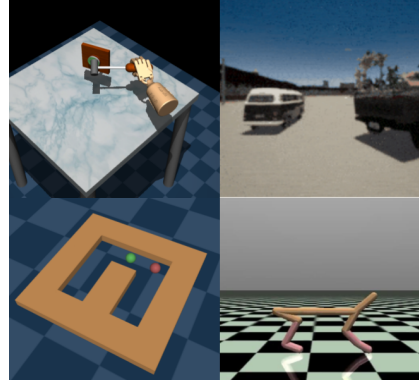


Fig. 1: ロコモーションや迷路など様々なタスクを含む模倣学習やオフライン強化学習用データセット D4RL⁹⁾。

SLI を算出した。またシミュレーション環境でタスクの成功率およびタスクの実行にかかった時間を評価した。

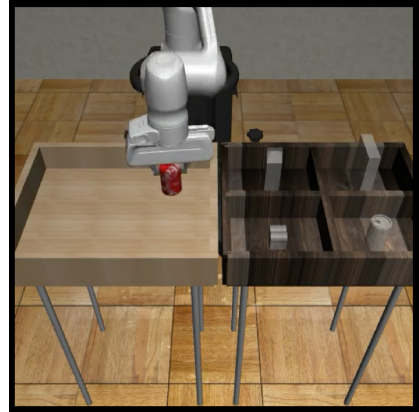


Fig. 2: 多関節ロボット Panda を人間が操作し、持っているオブジェクト (缶) を決められた枠に入れるタスクの行動軌跡を集めたデータセット Pick and Place Can¹⁰⁾

3. 結果

3.1 D4RL における本手法の性能

Fig.4- 6 に Half Cheetah, Walker2d, Hopper の medium-expert データセットにおける報酬和を横軸、標準化した SLI, β を縦軸にとしたときの散布図を示す。

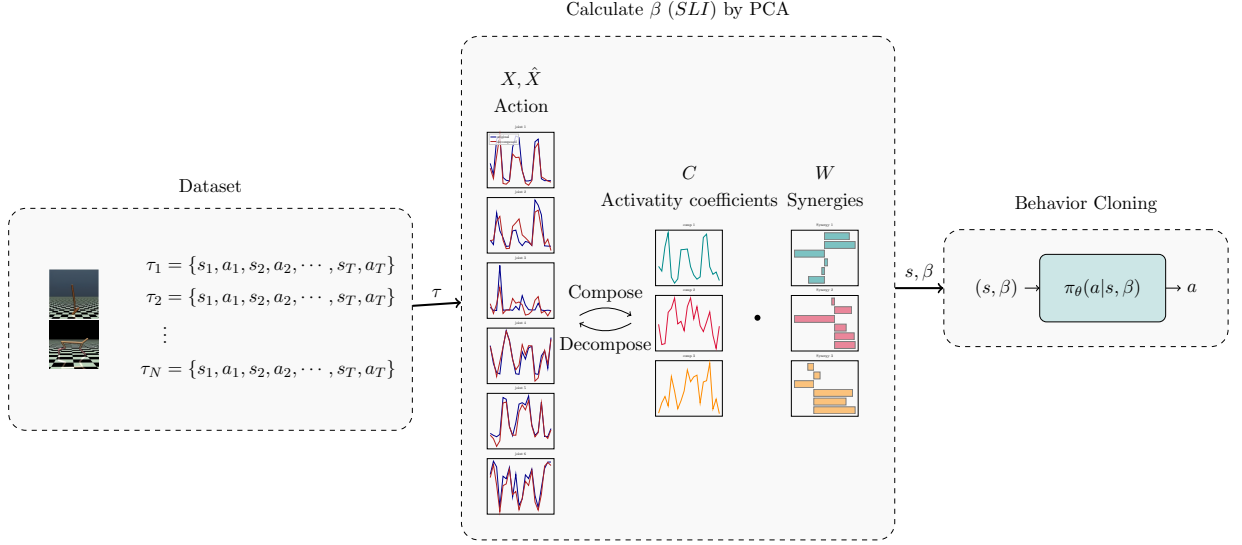


Fig. 3: 本手法の概要図

Algorithm 1 Imitation Learning with Synergy Level Index (SLI)

- 1: **Input:** Dataset $\mathcal{D}$ with trajectories $\tau_i = \{s_1, a_1, s_2, a_2, \dots, s_T, a_T\}$
 - 2: **Output:** Policy $\pi(a|s, \beta)$
 - 3: **procedure** CALCULATE SLI
 - 4: **for** each trajectory τ_i in $\mathcal{D}$ **do**
 - 5: Extract actions a_i from τ_i
 - 6: Perform PCA on a_i to get synergy W and coefficients C
 - 7: Compute $SLI_i = \frac{1}{N} \sum_{i=1}^N R_i^2$ where $R^2 = 1 - \frac{\|X - \hat{X}\|_F^2}{\|X - W \cdot C\|_F^2}$
 - 8: **end for**
 - 9: **end procedure**
 - 10: **procedure** BEHAVIOR CLONING WITH SLI
 - 11: Normalize SLI to get β
 - 12: Learn policy $\pi(a|s, \beta)$ by minimizing $E_{(s, \beta, a) \sim \mathcal{D}} [-\log \pi(a|s, \beta)]$
 - 13: **end procedure**
 - 14: **Return:** Policy $\pi(a|s, \beta)$
-

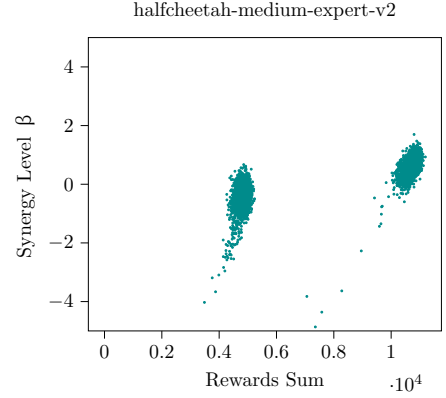


Fig. 4: Half Cheetah

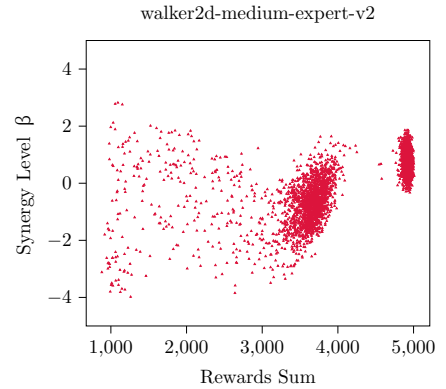


Fig. 5: Walker2d

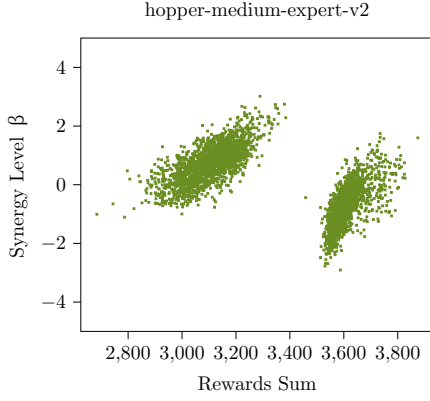


Fig. 6: Hopper

図から optimal な軌道 (報酬和の高い軌道) は SLI が高く suboptimal な軌道 (報酬和の低い軌道) は相対的に SLI が低いことがわかる。また SLI で条件付けを行い、BC を行ったときのスコア (10 エピソードの累積報酬の平均) を図3に示す。ベースラインとして SLI を用いた条件付けのない通常の BC のスコアを示している。

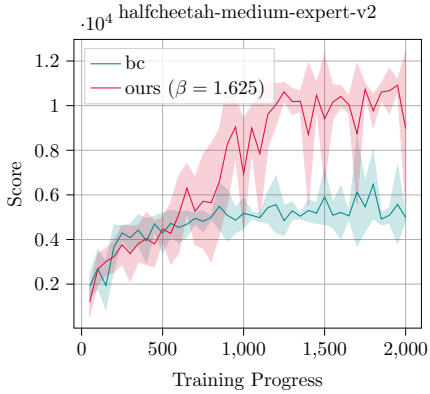


Fig. 7: Half Cheetah

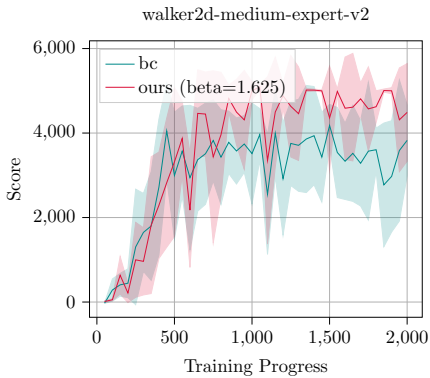


Fig. 8: Walker2d

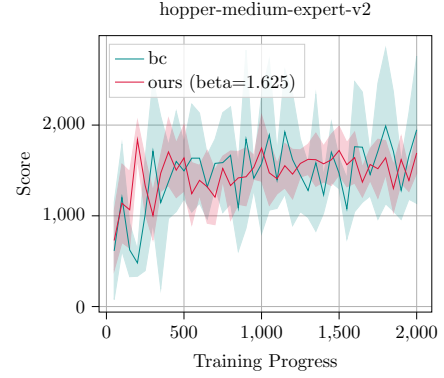


Fig. 9: Hopper

Fig.7, 8 に各 Agent のスコア (縦軸はスコア (10 エピソードの累積報酬の平均、影部は標準偏差)、横軸は学習ステップを示す) いずれの agent においても本手法がベースラインを上回っていることがわかる。Half Cheetah においては通常の BC の約 2 倍、Walker2d においては約 1.3 倍の性能を示した。一方で Hopper は性能が向上しなかった。また通常の BC では suboptimal の軌道も模倣してしまい、スコアが伸びていない。また Fig.7, 8 の結果は行動軌跡のみから算出した SLI を用いることで、一般の模倣学習で想定される報酬ラベルのないケースを想定しても、高い性能を達成できることを示している。次に標準化した SLI , β の値を変化させたときのスコアを Table.1-3 に示す。

β	3	2	1	0	-1	-2	-3
Score	1450	6835	9952	5484	4940	4908	4824

Table 1: Half Cheetah

β	3	2	1	0	-1	-2	-3
Score	4981	4579	3917	3667	2906	3351	2835

Table 2: Walker2d

β	3	2	1	0	-1	-2	-3
Score	1570	1629	1427	1750	3382	3468	3154

Table 3: Hopper

表より β を適切な値に設定 (学習し直す必要

はない) することで optimal に近い軌道を生成できている。

3.2 Robomimic における本手法の性能

Fig.10, 11 に Robomimic の Pick and Place Can タスクにおけるモデルアーキテクチャを MLP、RNN、横軸を学習プロセスとしたときのタスクの成功率 (50 Rollout) を示す。

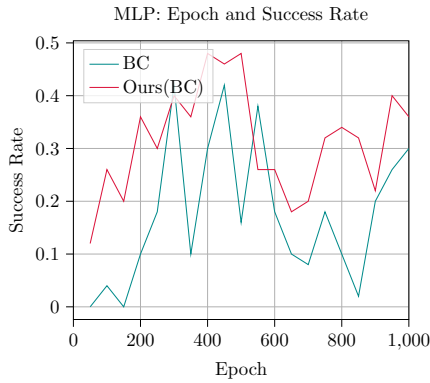


Fig. 10: 成功率 (MLP)

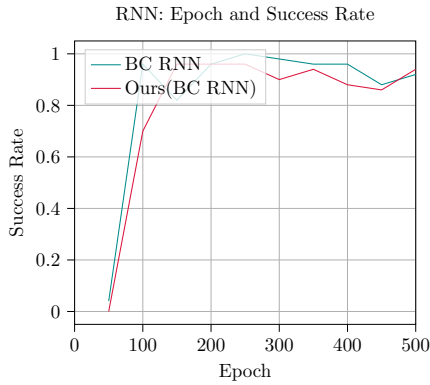


Fig. 11: 成功率 (RNN)

次に Table.4 に BC(MLP) および RNN における 500epoch 間の平均成功率と平均消費時間を示す。

	BC	Ours(BC)	RNN	Ours(RNN)
Success Rate	0.18	0.31	0.81	0.85
Time Consumption	8.6	7.7	5.1	4.6

Table 4: 500epoch 間の平均成功率および平均消費時間

MLP ではタスクの成功率が本手法を用いたことにより上昇した。タスクの成功率は RNN および提案手法 (RNN) でほぼ同じ値だが、本手法の方がタスクにかかる平均消費時間が短い。

4. 考察

先行研究の通り、学習が進んだ、熟練した expert はよりシナジーが発現しやすいと考えられる。とくに Fig.4, 5 の散布図において expert データが多く分布する SLI の値を用いることで optimal な軌道が生成されることが考えられる。また β の値も高すぎても元のデータが分布していないため、性能が低下すると考えられる。Walker2d では比較的スコアがあまり伸びていないが、これは Fig. 5 の報酬和の散布図からわかるように Walker2d は expert と medium の性能差が小さいためである。一方で Hopper では SLI の高い領域では性能が低下していた。これは Hopper は 3 関節しかないため、そもそもシナジーが発現しにくかったと考えられる。そのため本手法は高次元の空間を持つタスクがより適していると考えられる。Table.4 では、本手法はベースラインと同様、あるいはそれ以上の成功率であることに加え、タスクを完了するまでの時間がより短くなっていた。このことから、シナジーは関節や筋肉の冗長性を評価することができるが、本研究ではタスクにおける冗長性 (どのように動かしたら缶を効率よく運べるかなど) も評価することができるようになったと考えられる。そのためより無駄のない動き、素早い動きを選択的に生成することができたと考えられる。

5. 結言

本研究では行動軌跡のみから算出される SLI が optimal と suboptimal で異なることを示し、 SLI を用いることで報酬や状態の情報なしに、BC の性能が向上することを示した。今後の展

望として、より高難易度なタスクについても検証したい。また他のアルゴリズムとの性能の比較も行いたい。

参考文献

- 1) Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.
- 2) Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. Confidence-aware imitation learning from demonstrations with varying optimality. *Advances in Neural Information Processing Systems*, Vol. 34, pp. 12340–12350, 2021.
- 3) Wynne A Lee. Neuromotor synergies as a basis for coordinated intentional action. *Journal of motor behavior*, Vol. 16, No. 2, pp. 135–170, 1984.
- 4) Seyed Safavynia, Gelsy Torres-Oviedo, and Lena Ting. Muscle synergies: implications for clinical evaluation and rehabilitation of movement. *Topics in spinal cord injury rehabilitation*, Vol. 17, No. 1, pp. 16–24, 2011.
- 5) Mitsuhiro Hayashibe and Shingo Shimoda. Synergetic learning control paradigm for redundant robot to enhance error-energy index. *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 10, No. 3, pp. 573–584, 2017.
- 6) Reinhard Gentner, Susanne Gorges, David Weise, Kristin aufm Kampe, Mathias Buttmann, and Joseph Classen. Encoding of motor skill in the corticomuscular system of musicians. *Current Biology*, Vol. 20, No. 20, pp. 1869–1874, 2010.
- 7) Jiazheng Chai and Mitsuhiro Hayashibe. Motor synergy development in high-performing deep reinforcement learning algorithms. *IEEE Robotics and Automation Letters*, Vol. 5, No. 2, pp. 1271–1278, 2020.
- 8) Jihui Han, Jiazheng Chai, and Mitsuhiro Hayashibe. Synergy emergence in deep reinforcement learning for full-dimensional arm manipulation. *IEEE Transactions on Medical Robotics and Bionics*, Vol. 3, No. 2, pp. 498–509, 2021.
- 9) Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.
- 10) Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarini, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.