模倣では十分ではない：強化学習による模倣のロバスト化による困難な運転シナリオへの対応

Yiren Lu[1], Justin Fu[1], George Tucker[2], Xinlei Pan[1], Eli Bronstein[1], Rebecca Roelofs[2], Benjamin Sapp[1], Brandyn White[1], Aleksandra Faust[2], Shimon Whiteson[1], Dragomir Anguelov[1], Sergey Levine[2,3]

概要- 模倣学習(IL)は、人間のような行動を生成するために、大規模に収集できる高品質な人間の運転データを利用するシンプルで強力な方法である。しかし、模倣学習のみに基づく政策では、安全性や信頼性の懸念を十分に考慮できないことが多い。本論文では、単純な報酬を用いた強化学習と組み合わせた模倣学習が、模倣のみから学習した方策よりも運転方策の安全性と信頼性を大幅に向上させることができることを示す。特に、100kマイル以上の都市走行データでポリシーを学習し、衝突の可能性のレベル別にグループ化したテストシナリオでその有効性を測定する。我々の分析によると、模倣は実証データで十分にカバーされている低難易度のシナリオでは良好な性能を発揮できるが、我々の提案するアプローチは最も困難なシナリオではロバスト性を大幅に向上させる(失敗が38%以上減少)。我々の知る限り、これは実世界の大量の人間の運転データを利用する自律走行における模倣と強化の複合学習アプローチの最初の応用である。

I. INTRODUCTION

規模に応じた自律走行システムの構築には、多くの困難がある。まず第一に、実走行で発生する数多くの稀で困難なエッジケースを扱うという課題である。このため、利用可能なデータ量に応じて手法の性能を拡張できる模倣学習ベースのアプローチが提案されている[1], [2], [3]。デモデータでよく表現されている状況は、そのような方針で正しく処理される可能性が高いが、データではめったに起こらないような、より珍しい、あるいは危険な状況は、何が危険な、あるいは不適切な対応を構成するかについて明示的に指示されていない模倣方針が、予測不可能な対応をする原因となる可能性がある。問題は複雑な相互作用によって複雑化し、類似シナリオにおける人間の専門家の運転データは乏しく、最適とは言えないかもしれない[4]。

[1] Waymo Research. [2] Google Research, Brain Team [3] UC Berkeley.
Point of contact: maxlu@waymo.com.
Webpage: waymo.com/research/imitation-is-not-enough-robustifying-imitation-with-reinforcement-learning

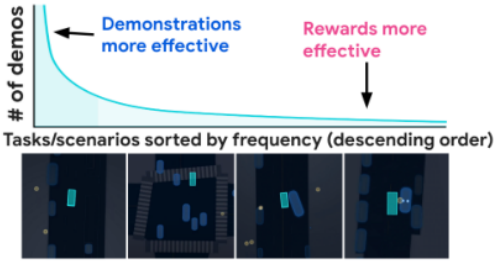

図1:デモンストレーションとリターンのトレードオフ。特定のシナリオのデータ量が減少するにつれて、報酬信号は学習にとってより重要になる。頻度の異なるシナリオを表すいくつかの視覚的な例を示す。

強化学習(RL)は、何が安全または安全でない結果(例えば衝突)を構成するかをポリシーに指示する明示的な報酬関数を活用することで、これを解決する可能性を秘めている。さらに、RL法は閉ループで学習するため、RL政策は観測、行動、結果の間の因果関係を確立することができる。これにより、(1)オープンループIL [5], [6]でよく見られる共変量シフトやスプリアス相関の影響を受けにくく、(2)報酬関数にエンコードされた安全性の考慮を意識するが、デモでは暗黙的なものであるポリシーが得られる。
しかし、例えば[7]、[8]、[9]のようにRLだけに頼ることは、自律走行[10]の未解決の課題である報酬設計に大きく依存するため、問題もある。模倣の忠実度を考慮しなければ、RLで訓練された運転方針は技術的に安全であるが不自然であり、他のエージェントと協調し、運転規約に従うために人間のような運転行動を要求する状況で前進させるのが難しいかもしれない。ILとRLは相補的な強みを提供する：ILはリアリズムを高め、報酬設計の負担を軽減し、RLは特に豊富なデータがない稀で困難なシナリオにおいて、安全性と頑健性を向上させる(図1)。

本論文では、[11]の難易度推定を活用し、安全性と信頼性の懸念を示す可能性が最も高い運転シナリオに焦点を当てる。提案手法BC-SACは、ILとRLを単純な報酬関数で組み合わせ、困難な運転シナリオで学習する。

難易度は、事前に訓練された計画方針で再シミュレーションされたときの衝突またはニアミスの可能性を推定する分類器によって推定される。我々の提案する報酬関数はエージェントの安全性を強制するが、自然な運転行動はILで暗黙的に学習される。学習データは、実世界の人間の運転データ(実世界の都市の運転データ100kマイル以上)のサブセットから得られる[11]。我々は、このアプローチが、人間のような振る舞いを損なうことなく、模倣のみよりも学習されたポリシーの安全性と信頼性を大幅に改善することを実証し、純粋なILとRLベースラインに対して38%と40%の改善を示す。

我々の研究の主な貢献は以下の通りである：(1)実世界の大量の都市人間運転データ(100kマイル以上)と単純な報酬関数を利用した、自律走行におけるILとRLを組み合わせたアプローチの最初の大規模アプリケーションを実施する。(2)難易度別にデータセットをスライスすることで、その性能とベースライン性能を系統的に評価し、ILとRLを組み合わせることで、模倣のみから学習したポリシーよりも安全性と信頼性が向上することを示す(最も難易度の高いバケットでの安全性イベントが38%以上減少)。

II. 関連研究

自律走行における学習ベースのアプローチ表Iに、プランニングのための様々な学習ベースのアルゴリズムの主要な特性を簡単にまとめる。ILは運転方針の導出に採用された最も古く、最も一般的な学習ベースのアプローチの一つである[1]，[2]，[24]，[25]，[26]，[27]。IL[3]、[28]またはRL[8]で学習された制御可能なモデルにより、ユーザは目標または制御信号(例えば、左、右、直進)の形で高レベルのコマンドを指定し、高レベルのルートプランニングと低レベルの制御を組み合わせることができる。

IL手法の欠点は2つある：(1)オープンループIL(広く使われている行動クローニングアプローチ[12]、[14]、[13]、[29]、[30]など)は共変量シフト[5]に悩まされる(これはクローズドループ学習[15]、[16]で対処できる)、(2)IL手法は衝突回避のような良い運転を構成するものについての明示的な知識を欠いている。RL法は、クローズドループ学習により明示的な報酬信号から学習することを可能にし、車線維持[31]、交差点横断[32]、車線変更[33]などのタスクに適用されている。これらの研究は、特定のシナリオにおけるRLの有効性を示しているが、我々の研究は、実世界のシステムで自律走行を展開することを困難にする、大規模で集約的な性能と、困難でセーフティクリティカルなエッジケースの両方を分析している。

RLやその他の自律走行のための閉ループ法は、通常、シミュレーションを学習に用いる。このような公共環境は数多く存在し、特に、シミュレーションされたエージェントを駆動するもの(例えば、エキスパート追従/ログ再生[34]、[35]、[36]、[37]、知的運転モデル(IDM)[38]、または他のルールベースのシステム[39]、MLベースのエージェント[38]、[40])、シナリオが手続き的に生成されるか(例えば、 [39]、[41]、[40])、実世界の運転シーンから初期化されるか[42]、[36]、[34]、[37]。実験では、他のエージェントがログをたどる実世界のデータに対して、クローズドループで開発・評価を行った。
ILとRLの組み合わせ。DQfD[22]、DDPGfD[43]、DAPG[44]などの手法により、ILはRLが疎な報酬が既知の領域における探索の課題を克服するのに役立つことが示されている。TD3+BC[21]やCQL[20]などのオフラインRLアプローチは、RL目的とIL目的を組み合わせて、Q学習更新を正則化し、分布外値の過大評価を回避する。我々の目標は、ILとRLの新しいアルゴリズムの組み合わせを提案することではなく、むしろ、スケールでの自律走行における課題に対処するために、この一般的なアプローチを活用することである。

自律走行車にとって困難でセーフティクリティカルなシナリオに対応する。[4]は、モデル予測制御と組み合わせたILプランナのアンサンブルを用いて、自律走行におけるロングテールシナリオに対処するポリシーを学習する。安全性を向上させるもう一つのアプローチは、安全性を保証するルールベースのフォールバック層で学習したプランナを補強することである[45]，[25]。我々の研究は、報酬を通じて安全意識をモデル学習プロセスに直接組み込むという点で、これらのアプローチとは異なる。また、本手法は必要に応じてフォールバック層と互換性があるが、これは今後の課題である。警察の頑健性を向上させるもう一つの方法は、研修中に否定的な事例の頻度を増やすことである。[46]は、無人航空機が衝突する様々な方法をカバーする故障データを収集し、負と正のデータを組み合わせることで、よりロバストなポリシーを学習するのに役立つ。[11]は、困難なエッジケースのパフォーマンスを向上させるために、カリキュラムトレーニングの使用を調査している。また、学習中に困難なシナリオにポリシーがさらされる機会を増やす一方で、RLが最も困難なシナリオでどのように大きな改善をもたらすかを示すことで、これらの知見を拡張する。

III. BACKGROUND

A. マルコフ決定過程(MDP)

本研究では、自律走行政策学習問題をマルコフ決定過程として扱う。標準的な形式論に従い、MDPをタプル{S，A，T，R，γ，ρ_0}として定義する。SとAはそれぞれ状態空間と行動空間を表す。

表1：ロボット制御と自律走行に対する様々な学習ベースのアプローチの比較。

	Offline Demo	Closed-loop	Rewards	Example Methods
行動クローニング(BC)	Expert Demos	No	No	Multipath [12], Precog [13], Trajectron++ [14]
敵対的模倣/IRL	Expert Demos	Yes	No	IRL [15], GAIL [16], MGAIL [17]
RL	No	Yes	Yes	DQN [18], SAC [19]
Offline RL	Behavioral Data	No	Yes	CQL [20], TD3+BC [21]
“Imitative” RL	Expert Demos	Yes	Yes	DQfD[22]、DAPG[23]、BC-SAC(当社)

Tは遷移モデルを表す。Rは報酬関数、γは割引係数を表す。ρ_0 は初期状態分布を表す。目的は、報酬の期待割引和$\pi^* = \max_{\pi ET,\pi,\rho 0}\left[\sum_{t=0}^{P\infty}\right.$ γ^t R(s_t, a_t)] を最大化する、SからAへの(確率的)写像である政策πを見つけることである。

B. 模倣学習(IL)

ILはエキスパートを模倣することで最適な政策を構築する。π_βと表記されるエキスパート(最適政策)が、環境との相互作用により軌道D = {s_0, a_0, – –, s_N, a_N } のデータセットを生成すると仮定する。学習者の目標は、π_βを模倣した方針πを学習することである。実際には、専門家の状態のみを観測するので、逆動力学を用いて専門家の行動を推定する。例えば、行動クローニング(BC)は対数尤度目的$E_{s,a}$ [log π(a¦s)]を介してポリシーを学習する。あるいは、閉ループアプローチには、逆RL(IRL)[15]や敵対的IL(GAIL [16], MGAIL [17])があり、これらは条件付き行動分布を通じて間接的にではなく、政策と専門家の間の占有率や状態行動訪問分布により直接的に一致することを目的としている。原理的には、オープンループ模倣[5]に影響を与える共変量シフトの問題を解決することができる。

C. 強化学習(RL)

RLは、オンライン試行錯誤を繰り返しながら、最適な方針を学習することを目的とする。この研究では、Qlearningのようなオフポリシー、バリューベースのRLアルゴリズムを使用する。これらの方法は、特定の状態と行動から出発したときの将来の期待リターンとして定義される状態-行動価値関数を学習することを目的としている。

$$Q^\pi(s,a) = \mathbb{E}_{\mathcal{T},\pi,\rho_0}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)|s_0=s,a_0=a\right].$$

本研究では、連続制御ポリシーの学習にアクター・クリティック法を用いる。典型的な行為者批判手法は、ベルマン誤差を最小化する批判者Qと価値関数を最大化する行為者πを交互に学習する。我々は、ソフト・アクター・クリティック(SAC)[19]のエントロピー正則化更新を使用する：

$$\min_Q \ \mathbb{E}_{s,a,s'\sim\pi}[(Q(s,a)-\hat{Q}(s,a,s'))^2] \qquad (1)$$

$$\max_\pi \ \mathbb{E}_{s,a\sim\pi}\left[Q(s,a)+\mathcal{H}(\pi(\cdot|s))\right], \qquad (2)$$

where

$$\hat{Q}(s,a,s') = r(s,a)+\gamma\mathbb{E}_{a'\sim\pi}\left[\bar{Q}(s',a')-\log\pi(a'|s')\right] \qquad (3)$$

であり、$\bar{Q}$は勾配を通過しない批評家のコピーであるターゲットネットワークを示す。

IV. RL拡張BCによるドライブの獲得

ILとRLの相補的な強みから恩恵を受けるアプローチを設計したい。模倣は報酬設計を必要とせず、豊富な学習信号源を提供し、RLはデータが乏しい稀で困難なシナリオにおけるILの弱点に対処する。この直感に従い、データが豊富なデモからの学習信号と、データが乏しい報酬信号を利用する目的を定式化する。具体的には、ILとRLの目的を加重混合して利用する：

$$\max_\pi \ \mathbb{E}_{\mathcal{T},\pi,\rho_0}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\right]+\lambda\mathbb{E}_{s,a\sim\mathcal{D}}[\log\pi(a|s)]. \qquad (4)$$

A. 行動クローンソフトアクタークリティック(BC-SAC)

原理的には、様々なRL手法をILと組み合わせて式4を最適化することができるが、効率的な学習のための便利な選択は、アクター・クリティック・アルゴリズムを使用することである。この場合、DAPG [23]やTD3+BC [21]と同様に、Q関数の期待値(すなわち、クリティック)に模倣学習目的を追加するだけで、式4に関してポリシーを最適化することができる。アクターにエントロピー正則化の目的をさらに追加する、広く使われているSACフレームワークを基に、完全なアクター目的を得る：

$$\mathbb{E}_{s,a\sim\pi}[Q(s,a)+\mathcal{H}(\pi(\cdot|s))]+\lambda\mathbb{E}_{s,a\sim\mathcal{D}}[\log\pi(a|s)].$$

批評家更新は式1で概説したSACと同じままである。λを適切に設定することで、この目的は、エキスパートデータがデータ分布D内にあるとき、ポリシーがエキスパートデータを模倣することを促す。図2はこの概念を視覚化したものである。

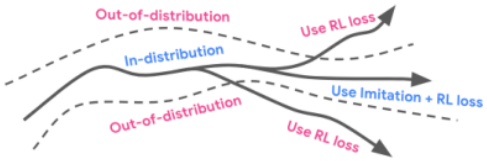

図2:異なる客観的影響力分布内状態の場合、ILとRLの両方の目的が学習シグナルを提供する。分布外の状態では、RL目的が支配的である。

B. 報酬関数

「良い」運転行動を捕捉するための報酬関数を設計することはオープンチャレンジ[10]であるため、単純な報酬関数が安全制約をエンコードするだけでよいのに対し、模倣学習損失に頼って主にポリシーを誘導することでこの問題を回避することができる。このため、報酬信号として衝突距離とオフロード距離の組み合わせを用いる。

$$R_{\text{collision}} = \min(d_{\text{collision}} - d_{\text{c\_offset}}, 0), \tag{5}$$

ここで、$d_{collision}$は自車両と他の車両の最も近いバウンディングボックスとの間の最も近い点のユークリッド距離(メートル)であり、$d_{coffset}$(デフォルト1.0)は車両が近くのオブジェクトから距離を保つように促すために追加されるオフセットである。オフロード報酬は

$$R_{\text{off-road}} = \text{clip}(-d_{\text{o\_offset}} - d_{\text{to-edge}}, -2.0, 0.0), \tag{6}$$

ここで、$d_{to-edge}$は最も近い道路端までの車両の距離(メートル)である(負は道路上、正は道路外)。d_{offset}（デフォルト1.0)は、車両が道路端まで距離を保つように促すオフセットである。$R = R_{collision} + R_{off-road}$ となるように、報酬を加法的に結合する。

C. 順方向および逆方向の車両ダイナミクスモデル

ステアリングと加速アクション$a = (a_{steer}, a_{accel})$が与えられたときの車両の次のポーズ$(x, y, \theta)$を計算する運動学的自転車ダイナミクスモデル[47]を用いて車両の状態を更新する。模倣学習のためのエキスパートアクションを得るために、逆ダイナミクスモデルを用いて、データセットのログに記録された軌跡と同じ状態を達成したであろうアクションを解く。これらのエキスパートアクションは、推論された状態$T(s_t, a_t)$と地上真理の次の状態s_{t+1}との間のコーナーの(x, y)位置のMSEを最小化することによって求められる。

D. モデルアーキテクチャ

主な構成要素は、アクターネットワーク$\pi(a|s)$、ダブルQクリティックネットワーク$Q(s, a)$、ターゲットダブルQクリティックネットワーク$Q^{-}(s, a)$である。各ネットワークは個別のTransformer観測を持つが、衝突報酬は[49]に記述されているエンコーダで、全ての車両状態、道路グラフ点、信号機信号、経路目標を含む特徴をエンコードする。アクターネットワークは、平均μと分散σでパラメータ化されたタンスクワッシュ対角ガウス分布を出力する。

E. 困難な例に対する学習

学習ベースの手法の性能は、特にロングテール分布を持つセーフティクリティカルな設定において、学習データ分布に強く依存する[50], [51], [45], [52], [53])。自律走行はこのカテゴリーに入る:ほとんどのシナリオは平凡であるが、かなりの少数のシナリオは重大な安全上の懸念を持っている。より困難な例で学習すると、不偏の学習分布を使用するよりも性能が向上することを実証した[11]に従い、学習分布が手法の性能にどのような影響を与えるかを探索する。

V. EXPERIMENTS

A. 実験セットアップ

データセット。サンフランシスコ(SF)[11]で運行されている車両群から収集された、10秒のセグメントに分割された10万マイル以上の専門家の運転軌跡からなるデータセット(All と表記)を使用する。これらのセグメントをトレーニング用640万個、テスト用10k個に分割する。同じ日に運行されている同じ車両からの軌跡は、訓練とテストの漏れを避けるために、同じパーティションに格納される。15Hzでサンプリングされた軌跡は、自律走行車(AV)の状態と、AVの知覚システムによって測定された環境の状態を記述する特徴を含んでいる。シナリオレベルの分布外推定量を直接構築することは困難であり、困難なシナリオは一般的に頻度が低いため、イベントの希少性を測定するための代理として[11]で説明されている難易度モデルを使用する。実行セグメントが与えられたとき、難易度モデルは、内部のAVプランナで再シミュレーションしたときに、そのセグメントが衝突するかニアミスになるかを予測する。我々は、5.6kの正例と80kの負例からなるデータセットに対して、クロスエントロピー損失を用いて、教師ありの方法で難易度モデルを学習した。Top1、Top10、Top50のサブセットを作成し、それぞれ400万セグメントの時系列的なデータセットから、難易度モデルスコアの上位1%(40k train、1.2k test)、10%(400k train、19k test)、50%(200万 train、66k test)のパーセンタイルを選択する。シミュレーションを行う。第IV-C章で述べたように、車両ダイナミクスは2次元自転車ダイナミクスモデルを用いてモデル化されている。シーン内の他の車両や歩行者の行動は、ログから再生される(ログ再生)。同様に

から[34]、[35]、[36]まで。これは、エージェントが

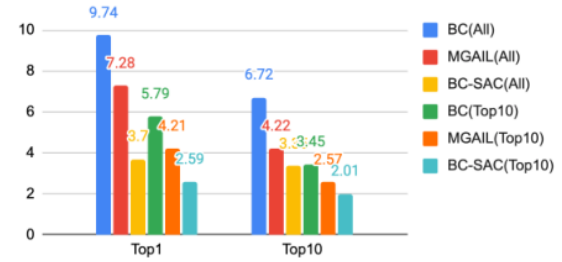

図3:最も困難な評価セットでの失敗率：Top1とTop10(AllとTop10で学習した場合、低い方が良い)。BC-SACは一貫して最も低いエラー率を達成している。

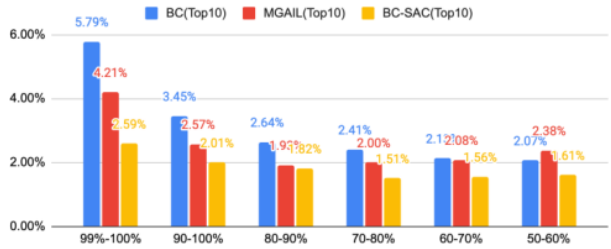

図4:難易度の異なるシナリオ(50%～100%、低いほど良い)におけるBC、MGAIL、BC-SACの失敗率。評価データセットがより困難になるにつれて、どの手法も性能が低下するが、BC-SACは常に最も性能が良く、劣化も最も少ない。

非反応であるため、他のエージェントの行動が人間らしくなることを保証し、模倣的な損失を含めることで、学習した方針がログから大きく逸脱することを抑制し、ログ再生エージェントが非現実的になる原因となる。また、ポーズの発散を緩和するために、10秒の短いセグメントを使用する。

ベースライン。本手法をオープンループ(BC [1])およびクローズドループ(MGAIL [17])の模倣手法と比較する。後者は、閉ループ学習とシミュレータダイナミクスの微分可能性を利用する。完全性を期すため、RLのみのアプローチを表すSACベースラインも含める。

メトリクス。我々は2つのメトリクスを用いてエージェントを評価する：

1) 失敗率：どのタイムステップでも、少なくとも1つの衝突またはオフロードイベントがあるランセグメントの割合。エゴ・ビークルのバウンディング・ボックスが他のオブジェクトのバウンディング・ボックスと交差する場合、衝突は真である。マップに従って自車両のバウンディングボックスが走行可能な表面から逸脱している場合、オフロードは真である。
2) ルート進行率：専門家によるデモンストレーションと比較した、ポリシーによるルート上の移動距離の比率。自車両の状態を経路上に投影し、経路の始点から全長を計算する。

B. Results

ベースライン手法(BC，MGAIL，SAC)と我々の手法(BC-SAC)を学習データセットのいくつかのサブセット(All，Top10，Top1)で学習し、評価セットのサブセット(Top1，Top10，Top50，All)に対して評価を表IIに示す。すべての構成は3つのランダムシードで評価され、平均値と標準偏差が報告される。以前、[11]は、Top10でMGAILをトレーニングすると、Allでトレーニングしても同様の性能が得られることを示した。同様に、Top10で学習した場合、全ての手法が最も良い性能を示すことが分かる。これは、模倣学習法が大量のデータに依存して暗黙のうちに運転嗜好を推測していることを反映している。一方、BC-SACはTop1で学習した場合、ロバストな性能を発揮する。Top10で学習した場合、全ての手法が最も良い性能を発揮することを考慮し、以下のサブセクションではその設定に注目する。BC-SACと模倣手法(BC、MGAIL)の困難なシナリオでの比較。図4は、難易度別に評価データセットスライスでBC-SACとBCおよびMGAILを比較したものである。BC-SACは全体的に良好な性能を達成し、特にBCとMGAILの両方の性能が大幅に低下する、より困難なスライスにおいて顕著である。さらに、BC-SACは、性能の難易度が異なるシナリオ(σ=0.37)対BC(σ=1.29)およびMGAIL(σ=0.78)において、最も分散が小さい。BC-SACとRLのみのトレーニング(SAC)の比較。すべての構成において、BC-SACは安全性指標においてSACを上回った(表II)。

は、大量のデモからの学習信号を利用する。SACは、より多くの境界アクション値で、デモから大きく逸脱したアクションを生成し、不自然な(より多くの)運転行動と不快な(急加速)運転行動をもたらす(図5)。BC-SACはBCロスにより、ログと同様のアクション分布を生成する。

報酬シェーピングとRL / IL重み。報酬関数の形式とRLと模倣成分の重みが最終的な性能にどのように影響するかを答えるために、一連のアブレーション研究を実施する。Top10データの10%をサンプリングして構築したより小さなデータセットを使用し、(1)我々の完全報酬対離散バイナリ報酬(図6右)、(2)オフロードと衝突報酬項の重み(図6左)、(3)オフロードと衝突オフセットのパラメータ(図7)、(4)目的語のRL項とIL項の重み(図8)を比較する。

Method	Training	Top1 (%)	Top10 (%)	Top50 (%)	All (%)	Route Progress Ratio, All(%)
BC	All	9.74±0.49	6.72±0.47	5.14±0.39	4.35±0.27	99.00±0.39
MGAIL	All	7.28±0.98	4.22±0.77	3.40±0.97	**2.48±0.29**	**99.55±1.91**
SAC	All	5.29±0.66	4.64±1.08	4.12±0.74	6.66±0.44	77.82±8.21
BC-SAC	All	**3.72±0.62**	**2.88±0.23**	**2.64±0.21**	3.35±0.31	95.26±8.64
BC	Top10	5.79±0.82	3.45±0.72	2.71±0.57	3.64±0.31	**98.06±0.18**
MGAIL	Top10	4.21±0.95	2.57±0.52	2.20±0.52	**2.45±0.35**	96.57±1.19
SAC	Top10	4.33±0.47	4.11±0.63	3.66±0.47	5.60±0.86	71.05±2.47
BC-SAC	Top10	**2.59±0.31**	**2.01±0.29**	**1.76±0.20**	2.81±0.26	87.63±0.58
BC	Top1	7.66±1.13	7.84±0.92	6.63±0.78	6.85±0.65	**94.10±1.00**
MGAIL	Top1	4.24±0.95	3.16±0.43	2.74±0.46	3.79±0.46	93.10±11.72
SAC	Top1	4.15±0.31	3.87±0.12	3.46±0.16	5.98±1.03	75.63±2.19
BC-SAC	Top1	**3.61±0.87**	**2.96±1.11**	**2.69±0.87**	**3.38±0.48**	75.00±17.21

表II:異なる訓練/評価サブセットにおけるBC-SACとベースラインの失敗率(低い方が良い)と進歩率(高い方が良い)。

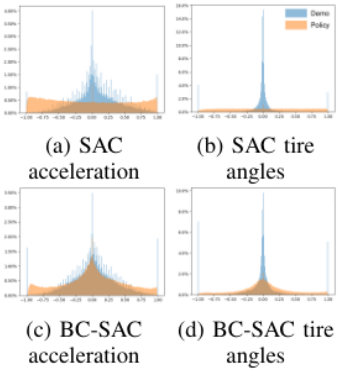

図5:限界行動分布。SAC/BC-SAC(オレンジ)対ログ(青)。

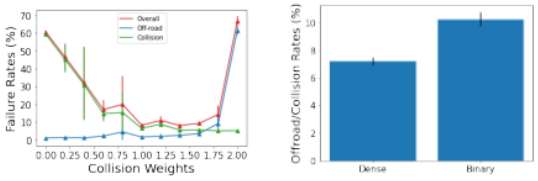

図6:左:オフロード/衝突重み。オフロードウェイトとコリジョンウェイトの合計は2.0である。x軸はコリジョンウェイトである。オフロードとコリジョンの重みをバランスよく選択することで、最高のパフォーマンスが得られる。右: 密な報酬と二値報酬。バイナリ報酬は、安全事象が発生した場合に-1、そうでない場合に0と定義される。報酬が濃いと、安全事象が少なくなる。

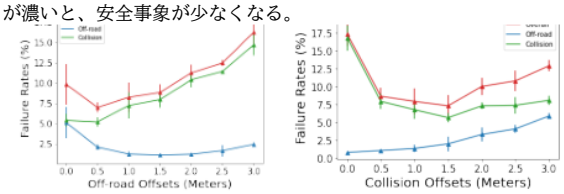

図7:オフロードオフセットd_{offset}と衝突オフセットd_{offset}のアブレーション。オフセットが少量あれば、全体的なパフォーマンスが向上する。

この結果から、提案する形状報酬は、報酬パラメータを適切に選択することで、より単純な疎報酬よりも全体的な性能を向上させ、模倣項とRL項のバランスが最良の性能につながることが示された。

進捗と安全性のバランス。我々の研究はセーフティクリティカルなシナリオに焦点を当てているが、図8右では、安全性メトリクスの大きな回帰なしに、少量の進捗報酬を導入することで、有意に多くの進捗につながることを示している。しかし、大きな進歩報酬はパフォーマンスの低下につながる。

詳細な故障解析。表IIIは、Top1とTop10のバケットからサンプリングされた80のシナリオのセットにおける故障モードの詳細な解析を示している。

Method	CLIP	COLL	OFF	RED	LAN	DIV
BC-SAC	8	7	2	1	7	15
MGAIL	16	8	8	2	0	6

表III:小規模サンプルセット(N=80)において、BC-SACとMGAILが発生させた故障頻度分類(タイプ別)。BC-SACは一般に直接衝突やオフロードの事象が少ないが、他の物体に衝突する頻度が高い。

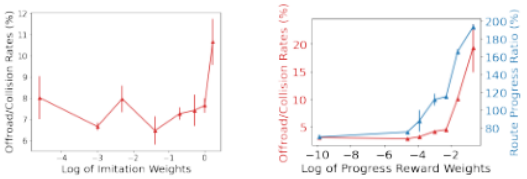

図8:左の模倣重み(対数スケール)と失敗率。右進捗報酬重み(対数スケール) vs 政策評価性能:安全イベント率と経路進捗率。

失敗を6つのバケットに大別する。CLIP(クリッピング):移動中に車両が側面の物体と衝突したときに発生する小さな衝突。OFF(オフロード): エージェントが道路を走行する際に失敗する。LAN(不良車線): エージェントが別の車線に侵入し、間違った車線か、悪い合流のどちらかになり、衝突する。COLL(衝突):計画エージェントが故障し、他の車両に走行する衝突。RED(赤色光):衝突の原因となる赤色光の違反。最後に、DIV(log divergence): 対数からの発散により、模擬エージェントが計画エージェントと衝突する衝突。

全体として、MGAILはクリッピング衝突やオフロードイベントが多い傾向がある。図9は、RLがILより改善される2つのケースを示している。MGAILは模倣手法であるため、衝突に対する明示的なペナルティがなく、現実的でない動作中の小さな衝突に対して敏感でないため、我々の手法はこれらのケースで改善されると仮定する。一方、BC-SACが遭遇する衝突は、衝突がAVプランナーのアクションの直接の結果ではなく、プランナーが他の車両に衝突するような方法でログから乖離するケースである傾向がある。BC-SACもトラフィックルールに従うことで明示的に報酬を得ることができないため(模倣によってこの動作を受け継いでいるが)、それによる失敗も少量見られる。

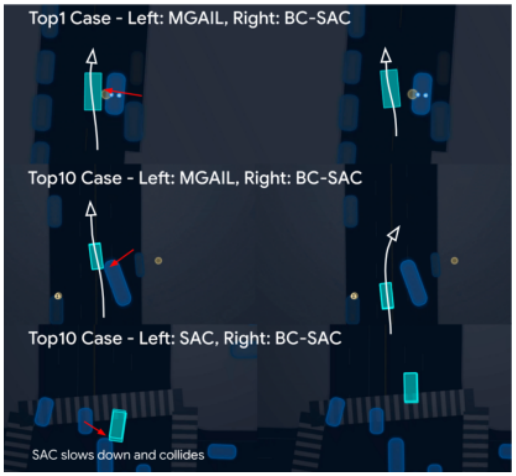

図9:BC-SACが模倣(MGAIL)とRLのみ(SAC)よりも改善するいくつかのシナリオの視覚化。シアンの車が制御されている。例1:MGAILは歩行者が二重駐車車から出るのと衝突し、BC-SACは十分なクリアランスを残す。例2:MGAILは十分なクリアランスを提供せず、進入車両と衝突する。例3:SACは交差点で減速し、後方衝突を引き起こす。BC-SACは、交差点を通る適切な速度プロファイルを衝突なく維持する。

VI. CONCLUSIONS

模倣学習とRLを組み合わせ(BC-SAC)、単純な安全報酬と組み合わせ、実世界の運転の大規模データセットで学習する、困難な運転シナリオにおけるロバストな自律走行のための手法を提示した。全体として、本手法は困難なシナリオにおける安全性と信頼性を大幅に改善し、その結果、ILのみ、RLのみのベースラインと比較して、最も困難なシナリオの安全性事象が38%以上減少した。我々の広範な実験では、訓練データセット、報酬シェーピング、IL / RLの目的語の役割を調べた。BC-SACは暗黙の人間のような運転行動を模倣から継承し、RLは分布外安全シナリオを扱うためのフェイルセーフである。ILのみの設定と同様に、最も困難なシナリオの上位10%でトレーニングを行うことで、ILとRLを組み合わせた設定において最もロバストな性能が得られる。この研究は主に安全関連の報酬の最適化に焦点を当てたが、自然な拡張は、進捗状況、交通ルールの遵守、乗客の快適性など、他の要因を目的に組み込むことである。報酬関数以外に、このアプローチは、エゴ・ビークル側の分布外行動に対する他のエージェントの予期せぬ行動を考慮しておらず、それでもなお、IL目的とRL目的の間のトレードオフを発見的に選択する必要がある。

将来有望な研究の方向性は、訓練と評価のための反応性シムエージェントを可能にし、おそらく分布シフトを緩和する方法論と組み合わせて、明示的な制約として安全性を強制するアプローチを拡張することであろう。

REFERENCES

[1] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[2] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[3] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.

[4] W. Zhou, Z. Cao, N. Deng, X. Liu, K. Jiang, and D. Yang, "Long-tail prediction uncertainty aware trajectory planning for self-driving vehicles," *arXiv preprint arXiv:2207.00788*, 2022.

[5] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[6] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[7] X. Pan, Y. You, Z. Wang, and C. Lu, "Virtual to real reinforcement learning for autonomous driving," *arXiv preprint arXiv:1704.03952*, 2017.

[8] X. Liang, T. Wang, L. Yang, and E. Xing, "Cirl: Controllable imitative reinforcement learning for vision-based self-driving," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 584–599.

[9] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 15 222–15 232.

[10] W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone, "Reward (mis) design for autonomous driving," *arXiv preprint arXiv:2104.13906*, 2021.

[11] E. Bronstein, S. Srinivasan, S. Paul, A. Sinha, M. O'Kelly, P. Nikdel, and S. Whiteson, "Embedding synthetic off-policy experience for autonomous driving via zero-shot curricula," in *6th Annual Conference on Robot Learning*, 2022.

[12] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[13] N. Rhinehart, R. McAllister, K. M. Kitani, and S. Levine, "PRECOG: prediction conditioned on goals in visual multi-agent settings," *CoRR*, vol. abs/1905.01296, 2019.

[14] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control," *arXiv preprint arXiv:2001.03093*, 2020.

[15] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of 17th International Conference on Machine Learning, 2000*, 2000, pp. 663–670.

[16] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[17] N. Baram, O. Anschel, and S. Mannor, "Model-based adversarial imitation learning," *arXiv preprint arXiv:1612.02179*, 2016.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *CoRR*, vol. abs/1801.01290, 2018.

[20] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[21] S. Fujimoto and S. S. Gu, "A minimalist approach to offline reinforcement learning," *Advances in neural information processing systems*, vol. 34, pp. 20 132–20 145, 2021.

[22] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband *et al.*, "Deep q-learning from demonstrations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[23] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Robotics: Science and Systems*, 2018.

[24] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 222–15 232.

[25] M. Vitelli, Y. Chang, Y. Ye, A. Ferreira, M. Wołczyk, B. Osiński, M. Niendorf, H. Grimmett, Q. Huang, and A. Jain, "Safetynet: Safe planning for real-world self-driving vehicles using machine-learned policies," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 897–904.

[26] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," *arXiv preprint arXiv:2207.05844*, 2022.

[27] V. Lioutas, A. Scibior, and F. Wood, "Titrated: Learned human driving behavior without infractions via amortized inference," *Transactions on Machine Learning Research*, 2022.

[28] N. Rhinehart, R. McAllister, and S. Levine, "Deep imitative models for flexible inference, planning, and control," *arXiv preprint arXiv:1810.06544*, 2018.

[29] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," *arXiv preprint arXiv:2007.13732*, 2020.

[30] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, D. Weiss, B. Sapp, Z. Chen, and J. Shlens, "Scene transformer: A unified multi-task model for behavior prediction and planning," *CoRR*, vol. abs/2106.08417, 2021.

[31] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8248–8254.

[32] D. Isele, R. Rahimi, A. Cosgun, K. Subramanian, and K. Fujimura, "Navigating occluded intersections with autonomous vehicles using deep reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2034–2039.

[33] P. Wang, C.-Y. Chan, and A. de La Fortelle, "A reinforcement learning based approach for automated lane change maneuvers," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1379–1384.

[34] E. Vinitsky, N. Lichtlé, X. Yang, B. Amos, and J. Foerster, "Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world," *arXiv preprint arXiv:2206.09889*, 2022.

[35] P. Kothari, C. Perone, L. Bergamini, A. Alahi, and P. Ondruska, "Drivergym: Democratising reinforcement learning for autonomous driving," *arXiv preprint arXiv:2111.06889*, 2021.

[36] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[37] V. Lioutas, J. W. Lavington, J. Sefas, M. Niedoba, Y. Liu, B. Zwartsenberg, S. Dabiri, F. Wood, and A. Scibior, "Critic sequential monte carlo," *arXiv preprint arXiv:2205.15460*, 2022.

[38] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.

[39] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[40] K. Ramamohanarao, H. Xie, L. Kulik, S. Karunasekera, E. Tanin, R. Zhang, and E. B. Khunayn, "Smarts: Scalable microscopic adaptive road traffic simulator," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 2, pp. 1–22, 2016.

[41] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[42] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.

[43] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," *arXiv preprint arXiv:1707.08817*, 2017.

[44] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv preprint arXiv:1709.10087*, 2017.

[45] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[46] D. Gandhi, L. Pinto, and A. Gupta, "Learning to fly by crashing," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3948–3955.

[47] R. Rajamani, *Vehicle dynamics and control*. Springer Science & Business Media, 2011.

[48] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[49] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.

[50] J. Frank, S. Mannor, and D. Precup, "Reinforcement learning in the presence of rare events," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 336–343.

[51] N. Kalra and S. M. Paddock, *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* RAND Corporation, 2016.

[52] S. Paul, K. Chatzilygeroudis, K. Ciosek, J.-B. Mouret, M. Osborne, and S. Whiteson, "Alternating optimisation and quadrature for robust control," in *AAAI Conference on Artificial Intelligence*, 2018.

[53] S. Paul, M. A. Osborne, and S. Whiteson, "Fingerprint policy optimisation for robust reinforcement learning," in *International Conference on Machine Learning*, 2019.

[54] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning *et al.*, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," in *International conference on machine learning*. PMLR, 2018, pp. 1407–1416.

[55] E. Bronstein, M. Palatucci, D. Notz, B. White, A. Kuefler, Y. Lu, S. Paul, P. Nikdel, P. Mougin, H. Chen *et al.*, "Hierarchical model-based imitation learning for planning in autonomous driving," *arXiv preprint arXiv:2210.09539*, 2022.

APPENDIX

A. IL + RL Distributed Actor-Learner Training Architecture

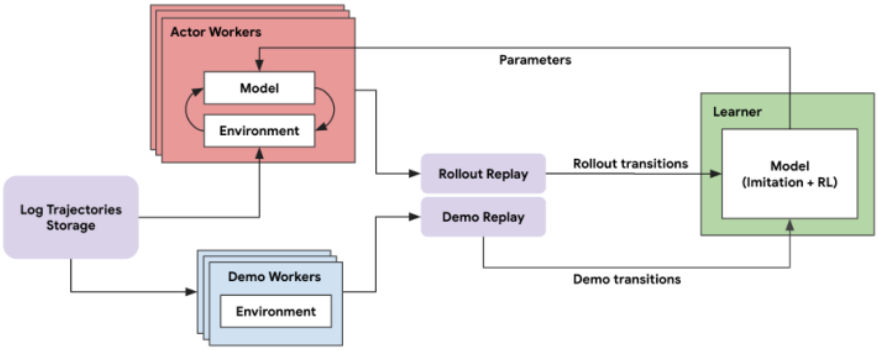

図10:IL+RL分散アクター学習器の学習アーキテクチャ。我々は、分散IMPALAアーキテクチャ[54]を、追加のデモロールアウトワーカーと、アクターワーカーと同じフォーマットのロールアウト遷移を生成するデモ再生バッファで拡張する。学習者ワーカーはロールアウト再生バッファとデモ再生バッファの両方からサンプリングし、オフポリシーで学習更新を実行する。

B. モデルアーキテクチャとハイパーパラメータ設定に関する追加情報

TD3とSAC[48]、[19]に似たデュアル・アクター・クリティック・アーキテクチャを使用する:主なコンポーネントであるアクター・ネットワーク$\pi(a|s)$、ダブルQクリティック・ネットワーク$Q(s, a)$、ターゲット・ダブルQクリティック・ネットワーク$Q^{-}(s, a)$のそれぞれは、[55]で説明されている別のトランスフォーマー観測エンコーダを持ち、エンコーダの埋め込みは(256, 256)完全連結ヘッドに供給される。アクターネットワークは、平均μと分散σでパラメータ化されたtanh-squared対角ガウス分布を出力する。

アクター学習率は1e-4、批判者学習率は1e-4、模倣学習率は5e-5、バッチサイズは64、報酬割引率は0.92である。リプレイのサンプル対挿入比率は8であり、これは学習者がアイテムの全生涯にわたってリプレイバッファ内の各アイテムをサンプリングする平均回数である。実際には、ILとRLの両方の目的語を組み合わせた勾配ステップを実行する代わりに、異なる更新頻度でILとRLの間で学習ステップを交互に行う。8回のRL更新ごとに、IL損失で1回更新する。ハイパーパラメータはグリッドサーチを行うことで求められる。

SACについては、ILステップを実行しないことを除き、BC-SACと同じネットワーク設計とハイパーパラメータを使用する。

BCについては、2次元行動空間(ステア、加速度)を31×7=217アクションに離散化し、同じダイナミクスモデルを基礎とする。BC-SACのアクターネットワークと同様のネットワーク設計を用い、離散行動の確率を表すソフトマックス予測ヘッドを用いる。学習率1e-4、バッチサイズ256のクロスエントロピー損失を学習に使用する。

For MGAIL, we follow the network design and hyper-parameters setting presented in [55].