

SS-MAIL: SELF-SUPERVISED MULTI-AGENT IMITATION LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

The current landscape of multi-agent expert imitation is broadly dominated by two families of algorithms - Behavioral Cloning (BC) and Adversarial Imitation Learning (AIL). BC approaches suffer from compounding errors, as they ignore the sequential decision-making nature of the trajectory generation problem. Furthermore, they cannot effectively model multi-modal behaviors. While AIL methods solve the issue of compounding errors and multi-modal policy training, they are plagued with instability in their training dynamics. In this work, we address this issue by introducing a novel self-supervised loss that encourages the discriminator to approximate a richer reward function. We employ our method to train a graph-based multi-agent actor-critic architecture that learns a centralized policy, conditioned on a learned latent interaction graph. We show that our method (SS-MAIL) outperforms prior state-of-the-art methods on real-world prediction tasks, as well as on custom-designed synthetic experiments. We prove that SS-MAIL is part of the family of AIL methods by providing a theoretical connection to cost-regularized apprenticeship learning. Moreover, we leverage the self-supervised formulation to introduce a novel teacher forcing-based curriculum (Trajectory Forcing) that improves sample efficiency by progressively increasing the length of the generated trajectory. The SS-MAIL framework improves multi-agent imitation capabilities by stabilizing the policy training, improving the reward shaping capabilities, as well as providing the ability for modeling multi-modal trajectories.

1 INTRODUCTION

Training an agent to imitate an expert is a promising approach to learning intelligent behavior and can be used in applications such as autonomous driving and robotic manipulation. More specifically, the ability of the agent to robustly learn optimal policies in real-world scenarios is a current challenge facing the field. The most promising approaches for imitation learning are Behavioral Cloning (BC) and Adversarial Imitation Learning (AIL). BC methods have been shown to produce compounding errors (Ross et al., 2011), which makes it unsuitable for complex applications. Adversarial Imitation learning methods, such as GAIL (Ho & Ermon, 2016), iteratively train a discriminator, and use it as a proxy reward function for updating a policy. We show that the proxy reward function learned by GAIL fails to provide dense supervision for policy updates and leads to inefficient and unstable training.

Let’s consider the example task of training two agents to draw the planar letters “ML” on a piece of paper, with the expert trajectories shown in red, and the initial policy shown in blue in Fig. 1(a). Training an RL agent to imitate the expert policy would ideally require a rich reward landscape with a clear gradient starting from the current

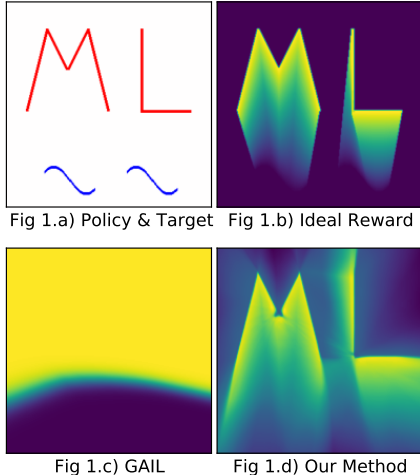


Figure 1: (a) An example task that starts with an initial policy shown in blue, and aims to imitate the expert trajectories shown in red. (b)-(d) show the reward landscapes of the ideal reward function, the learned discriminator in GAIL, and our method (SS-MAIL), respectively.

policies, in the lower half of the image, and terminating at the expert sketches, as depicted in Fig. 1(b). However, the reward function learned by GAIL is almost constant throughout the state space, while abruptly changing at the decision boundary, as depicted in Fig. 1(c). In theory, GAIL would be able to learn the optimal policy with sufficient exploration; however, in practice, the training dynamics resulting from the sparse rewards in the local neighborhood of the current policy lead to sub-optimal policies. We see that apart from a minuscule sliver of the state space, the agent is left in the dark when it comes to a prospective policy update. This leads us to wonder how the agents will ever learn to imitate the expert if they are left to explore in the dark. There have been attempts such as WAIL (Xiao et al., 2019) to address this issue; however, it is also plagued with unstable critic-training, which we will elaborate upon in the upcoming sections. To briefly motivate the rest of the paper, we tease the results of our method (SS-MAIL) in Fig. 1(d), and we will explain in the upcoming sections the details of the framework that helped us get these results.

The problem illustrated in the toy example above would be exacerbated in larger state spaces. We see that despite the recent progress in Adversarial Imitation Learning (AIL), the application of methods such as GAIL in scenarios with larger state spaces may be impeded by the exponential cost of exploring in ever-increasing hyper-spaces. The exploding cost of exploration would make tasks such as self-driving and robotic manipulation infeasible. These shortcomings are addressed in the SS-MAIL method introduced in this work.

2 BACKGROUND: IMITATING EXPERT MULTI-AGENT TRAJECTORIES

In this work, we consider sequential decision making problems, and we model them using the framework of Markov decision processes (MDP). An MDP can be represented as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, T)$ with state-space \mathcal{S} , action-space \mathcal{A} , dynamics $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, reward function $r(s, a)$, initial state distribution ρ_0 , and horizon T .

Imitation learning (Osa et al., 2018) methods aim to learn task policies directly from expert demonstrations. The family of imitation learning algorithms is broadly divided into two classes, behavioral cloning (BC) (Bain & Sammut, 1995) and Adversarial Imitation Learning (AIL) (Fu et al., 2017). BC directly regresses from the expert’s states to its decisions; however, such a straightforward supervised learning approach ignores the sequential nature of the problem and policy errors cascade during execution.

Adversarial IL methods formulate the imitation learning problem as an adversarial game between the policy and the discriminator. The discriminator measures some divergence between the expert and policy’s state-action distribution, and the policy aims to fool this discriminator. For instance, GaIL (Ho & Ermon, 2016) minimizes the JS divergence between the expert and policy’s state distribution. Their algorithm implements a min-max optimization procedure,

$$\min_{\pi \in \Pi} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} -\lambda_H H(\pi) + \mathbb{E}_{\pi}[\log(D(s, a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s, a))] \quad (1)$$

AIL methods provide a compelling approach to imitation since they do not face the issue of compounding errors. However, such adversarial methods are known to be unstable to train. Our work aims to deal with this instability while retaining the advantages of AIL methods.

Previous work for multi-agent behavior prediction such as NRI (Kipf et al., 2018) and DNRI (Graber & Schwing, 2020) model the expert’s behavior using a VAE (Kingma & Welling, 2013) in which the encoder and the decoder networks are modeled as GNNs (Kipf & Welling, 2016) and the latent code represents the intrinsic interaction graph. Since their network architecture is designed to represent and work on the interaction graph of the agents, they can be used to understand the relationships between the agents qualitatively. These models are trained using multi-step BC and, therefore, cannot model multi-modal behavior. This is a significant drawback since real-world behavior is often multi-modal. Our work builds upon their policy architecture by training them using an actor-critic algorithm in an AIL setting.

3 METHOD

In this section we will introduce SS-MAIL, and its components, namely: **1) SS-AIL:** a self-supervised AIL method to train the Discriminator, **2) MAIL:** an off-policy graph convolutional actor-critic architecture that imitates the expert using supervision from the Discriminator, **3) Trajectory Forcing:** a teacher forcing based curriculum generator for stabilized training dynamics.

3.1 SS-AIL: SELF-SUPERVISED GENERATIVE ADVERSARIAL IMITATION LEARNING

3.1.1 FROM APPRENTICESHIP LEARNING TO SS-AIL

Ho & Ermon (2016) show that GAIL is a form of apprenticeship learning with a specific cost regularization, which optimizes the min-max objective:

$$\psi_{\text{GA}}^*(\rho_{\pi} - \rho_{\pi_E}) = \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} \mathbb{E}_{\pi}[\log(D(s, a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s, a))] \quad (2)$$

In practise, the rewards from the discriminator do not provide rich policy-gradients, which result in sub-optimal policies. We address this issue by introducing a self-supervised loss function for our discriminator that encourages it to provide an enriched reward signal, not only in the local neighborhood of the current trajectory, but globally as well.

Our self-supervised loss is presented in Eq. 3, below. We sample $\alpha \in [-1, 1]$ and take the weighted average of the trajectories resulting from the policy and the expert, specifically $\tau_G \sim \pi$ and $\tau_E \sim \pi_E$. This new trajectory can be mathematically formulated as $\tau_{\alpha} = \alpha\tau_G + (1 - \alpha)\tau_E$. Our intuition behind creating the self-supervised loss is that the state action pairs of τ_{α} should regress smoothly based on their relative distance to τ_G , or in other words $\mathbb{E}_{\tau_{\alpha}}[D_{\theta}(s, a)] = \alpha$. Thus, we convert the Binary Cross Entropy loss to a Mean Squared Error (MSE) loss and use the sampled α values as self-supervised labels.

$$\psi_{\text{SS}}^*(\rho_{\pi} - \rho_{\pi_E}) = \mathbb{E}_{\pi, \pi_E, \alpha} \left[\underbrace{(0 - D(s_G, a_G))^2}_{\text{Generated MSE}} + \underbrace{(1 - D(s_E, a_E))^2}_{\text{Expert MSE}} + \underbrace{(\alpha - D(s_{\alpha}, a_{\alpha}))^2}_{\text{Self-Supervised MSE}} \right] \quad (3)$$

An interesting consequence of this formulation is that the third term in the self-supervised optimization objective, the self-supervised MSE term, serves as a natural form of exploration during the policy optimization stage.

Theorem 3.1. *SS-AIL is an instantiation of cost-regularized apprenticeship learning, i.e. the policy at the saddle point (π, D) of the min-max problem described above is a solution to $\min_{\pi} \max_c (-\psi(c) + E_{\pi}[c(s, a)] - E_{\pi_E}[c(s, a)])$ for some specific ψ .*

Proof. Refer to the appendix. □

Algorithm 1 SS-AIL

Input: Expert Trajectories $\tau_E \sim \pi_E$, Initial Policy π_{ϕ} , Initial Discriminator D_{θ}

Initialize: Policy π_{ϕ} , Discriminator D_{θ}

while Policy Improves **do**

Sample Trajectories $\tau_G \sim \pi_{\phi}$

Sample α and Compute: $\tau_{\alpha} = \alpha\tau_G + (1 - \alpha)\tau_E$

Update D_{θ} using gradient:

$$\mathbb{E}_{\tau_G} \left[\nabla_{\theta} (D_{\theta}(s, a))^2 \right] + \mathbb{E}_{\tau_E} \left[\nabla_{\theta} (1 - D_{\theta}(s, a))^2 \right] + \mathbb{E}_{\tau_{\alpha}, \alpha} \left[\nabla_{\theta} (\alpha - D_{\theta}(s, a))^2 \right]$$

Update ϕ with SAC to increase the following objective: $\mathbb{E}_{\tau_i \sim \{\tau_G, \tau_{\alpha}\}} [D_{\theta}(s, a)]$

end while

3.1.2 ALGORITHM

In the SS-AIL algorithm we start by sampling a trajectory from the current policy. Next, we compute the weighted average between the current trajectory and a sampled expert trajectory. We denote this trajectory as τ_α . According to our Discriminator update rule, derived previously, we train our discriminator to learn a smooth reward function by reconstructing the self-supervised labels, α . Next, we need to update our policy function.

In GAIL, Ho & Ermon (2016) use the on-policy TRPO update rule. However, according to the min-max function derived above, our policy should be optimized on not only the current trajectory, but also τ_α . This serves as an inherent exploration term for our policy training, as we are exploring states that are disparate from the current trajectory. Thus, we use the SAC update policies to accommodate the off-policy updates from τ_α .

We formulate our update rules such that the discriminator converges faster than the actor, similar to the two-timescale approach in the actor-critic setting. Thus, we no longer require balancing the discriminator, as seen in GAIL. Also, the change in our Discriminator values converges to zero, as opposed to WAIL, which constantly updates the fluid surface of the Discriminator output, as there is no grounding of the outputs to any specific value.

3.1.3 SELF-SUPERVISION AS A FORM OF REWARD-SHAPING

Our novel self-supervised loss gives us considerable flexibility in the sampling approaches used to obtain α . As stated previously in our discussion on the discriminator approximating a vector field, we noticed that only taking α values from $[0,1]$ would be insufficient for training as during exploration, the agent would explore states that would naturally correspond to negative α values. As the discriminator was not trained on these states, generalization isn't guaranteed. Thus, to complete the neighborhood around the current policy trajectory, we also consider negative α values. We show an empirical analysis in the next section. The negative α values assist in inducing a complete vector field around the neighborhood of our trajectory, as shown in Fig 3.

We can custom design the sampling approach and the loss functions to suit our need, which will open up greater flexibility and control for the community to experiment with. We elaborate in the Appendix.

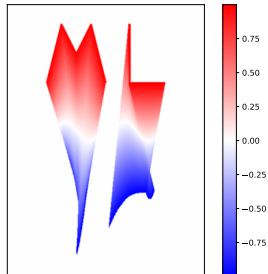


Figure 2: Vector field induced by negative α .

3.2 MAIL: MULTI-AGENT IMITATION LEARNING

In this section we introduce the architecture of the policy network, which is divided into two parts - the graph encoder and the graph soft actor-critic (G-SAC). In our architectures, we use the message passing standard established by Kipf et al. (2018), and model the node-to-edge and edge-to-node message passing operations as follows:

$$\mathbf{h}_{(i,j)}^l = f_e^l([\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{s}_{(i,j)}]); \quad \mathbf{h}_j^{l+1} = f_v^l\left(\left[\sum_{i \in \mathcal{N}_j} \mathbf{h}_{(i,j)}^l, \mathbf{s}_j\right]\right) \quad (4)$$

where the subscripts of (i, j) represent edge related features, while subscripts denoted with single letters, such as i and j represent node related features. h represents the corresponding edge or node embedding, and s represents states. Finally, f represents a neural network function approximator, and \mathcal{N}_j corresponds to indices of the neighboring nodes connected to the node indexed by the subscript, which in this example is j .

3.2.1 GRAPH ENCODER

The goal of the Graph Encoder is to infer the underlying interaction graph of the various agents, conditioned upon the observed history. This latent interaction graph will be used as weights in the graph convolution step in G-SAC. Our Graph Encoder is based on the encoder network Graber & Schwing (2020). We employ a fully connected graph convolutional network to output a distribution

over edge weights. However, unlike in DNRI, we do not include prior networks that train for pre-cognition of the future evolution of the graph based on the observed history. Our model takes in as input the cumulative observable state space, and employs an LSTM to keep track of the history. We model this as follows:

$$\mathbf{h}_{(i,j),\text{enc}}^{t+1} = \text{LSTM}_{\text{enc}} \left(\mathbf{h}_{(i,j)}^t, \mathbf{h}_{(i,j),\text{enc}}^t \right), \quad (5)$$

$$\underbrace{q_\phi \left(\mathbf{z}_{(i,j)}^t \mid \mathbf{x} \right)}_{\text{Graph Encoder Output}} = \text{softmax} \left(f_{\text{enc}} \left(\mathbf{h}_{(i,j),\text{enc}}^t \right) \right) \quad (6)$$

3.2.2 GRAPH SOFT ACTOR-CRITIC (G-SAC)

G-SAC samples a dynamically computed interaction graph from the Graph Encoder and uses the graph as the activations for the respective edges among the nodes of the graphs. It then computes a graph convolutional to obtain the mean and standard deviation values, similar to the output of SAC Haarnoja et al. (2018). We also include another head which functions as a critic. The critic head approximates the Q-function of the policy and is used to train the network. As in SAC, we also keep a target critic, which is updated based on polyak averaging.

$$\mathbf{h}_{(i,j)}^t = \sum_k z_{ij,k} f_e^k \left([\mathbf{x}_i^t, \mathbf{x}_j^t] \right)$$

$$\mu_j^{t+1} = f_\mu \left(\sum_{i \neq j} \mathbf{h}_{(i,j)}^t \right); \quad \sigma_j^{t+1} = f_\sigma \left(\sum_{i \neq j} \mathbf{h}_{(i,j)}^t \right);$$

$$\underbrace{p \left(\mu_j^{t+1}, \sigma_j^{t+1} \mid \mathbf{x}^t, \mathbf{z} \right)}_{\text{G-SAC Output}} = \mathcal{N} \left(\mu_j^{t+1}, \sigma^2 \mathbf{I} \right)$$

3.3 TRAJECTORY FORCING

Teacher Forcing based curriculums are an important tool for training sequence models (Ranzato et al., 2015). In a sequence generation setting, we expect the learned models to use the outputs of the previous timesteps as inputs. However, training such generative models was shown to be sensitive to weight initialization (Yu et al., 2017), in the absence of pedagogical intervention, such as teacher forcing. Scheduled Sampling (Bengio et al., 2015) is a curriculum based on teacher forcing that attempts to gradually transition from using the ground-truths as inputs, to using the model’s previous outputs. NRI and DNRI use teacher forcing in their approaches; however, such an approach could result in compounding errors during test time, due to the distributional shift. The use of such curriculums is not prevalent in AIL, as the frameworks accommodate only for on-policy updates. There are methods that try to work around this by pre-training using BC (Jena et al., 2020). However, in SS-AIL, the off-policy functionality is built into the formulation. Thus, we can leverage α values close to 1 and use them as a proxy for teacher forcing. Our goal is to take advantage of the exploration functionality of our loss function to gradually reduce the teacher forcing frequency to zero. In other words, our curriculum progressively increases the intervals of pedagogical intervention. We show this in Section 4.5.

We mathematically model the frequency of interventions as $1.5^{-\text{epoch}/\beta}$, where beta is a hyperparameter that assists in generating a progressively increasing sequence. Our intuition behind modeling the length of the generated trajectory as an exponential with respect to the epoch, is that we intend on doubling the size of the generated trajectory every β epochs. A linear model would result in reduced sample efficiency during the later stages of training, as the generalizability of the model would outpace the curriculum. The exponential increase ensures efficient data utilization.

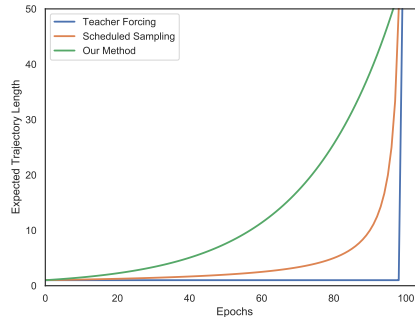


Figure 3: We plot the expected length of the generated trajectory during training in between pedagogical intervention for teacher forcing, scheduled sampling and trajectory forcing. We show that our curriculum provides a gradual increase in expected trajectory lengths for better policy training.

4 RESULTS

In this section, we quantitatively evaluate the strengths of our framework using two custom-built environments and a real-world dataset. These experiments illustrate the advantages of the SS-MAIL framework, namely: i) increased stability of its training dynamics, ii) multi-modal trajectory modeling capabilities, iii) increased sample-efficiency of Trajectory Forcing, iv) enhanced versatility in reward-shaping using self-supervision, and v) robustness to compounding errors on real-world datasets.

4.1 EXPERIMENTAL SETUP

To evaluate the capabilities of our methods against prior SOTA baselines, we evaluate our framework on one custom-built environment (Y-Junction), and one real world prediction task (Noisy Mocap). To minimize the prevalence of confounding variables, the experiment design of Y-Junction was simplified considerably to ensure that any promising improvement in performance is warranted by the inherent attributes of the frameworks, and not any other external confounding factors, such as variable overflow (Dietz et al., 2015). The Y-Junction environment is a trajectory forecasting environment that tests the multi-modal capabilities of AIL and BC based frameworks by simulating a 3-way Y-Junction scenario. The environment has three agents that follow one another on a one-way street until the beginning of a fork or Y-Junction. Then the lane splits into two, thus simulating two potential modes. Our goal is to pick one of the two lanes or modes. We provide more details in the appendix. The Noisy Mocap prediction task involves training a multi-agent policy on the CMU Motion Capture Dataset (CMU) for subject #35, and observing the zero-shot generalization in the presence of noisy inputs.

4.2 STABILITY IN TRAINING DYNAMICS

Robustness of the training algorithm to initialization, stochasticity in the environment, and training data is very important to assure predictability in the training dynamics. The Y Junction experiment, upon which we evaluate the robustness, provides a simple multi-agent testing scenario to illustrate the differences between various AIL approaches. To have a uniform evaluating strategy, we use the same architecture for all methods, and solely swap-out the corresponding AIL loss functions. We elaborate further in the Appendix.

In Fig. 4, we observe that the training loss of SS-AIL reduces to zero, implying that the algorithm is able to successfully imitate the expert during training. SS-AIL quickly converges to the expert policy for all initializations; however, GAIL and WAIL are unable to converge despite good policy initializations. We see that GAIL and WAIL start-out with better performances, due to their policy initializations; however, are unable to consistently improve and fluctuate considerably. As the Y-Junction experiment is unbounded in its state-space, there is no boundary that constrains diverging policies, thus resulting in the observed exploding losses. In contrast, we see that the performance of the policy trained with SS-AIL consistently improves, and converges to zero. These results can be attributed to the richer family of reward functions being approximated by the Discriminator. The Binary Cross Entropy loss of GAIL (Ho & Ermon, 2016) results in sparse reward signals for the policy training, thus resulting in unstable training. Furthermore, the inability of the WAIL Discriminator to converge upon a designated output surface results in a non-stationary value-function, thereby destabilizing the training of the critic. This instability becomes more pronounced as the policy and expert approach each other in the space of trajectories, as small variations in the surface of the discriminator output would lead to prominent shifts in the value-function.

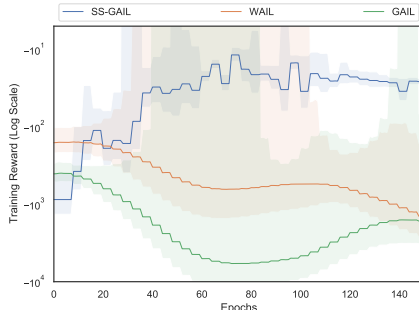


Figure 4: Training error over time in the Y-Junction environment shows that SS-AIL successfully imitates the expert. The low standard deviation of error for SS-AIL despite unfavorable initialization demonstrates the increased robustness of its training dynamics.

4.3 MULTI-MODAL CAPABILITIES

Differentiating between the various modes in the expert-trajectory dataset is essential for training multi-modal policies. The gradient of the policy training loss should be in the direction of a specific mode, instead of the weighted average over all the modes. The Y Junction example provides a simple multi-agent testing scenario that demonstrates how BC based methods learn sub-optimal policies in the presence of multiple modes. To evaluate SS-AIL and DNRI on their multi-modal modeling capabilities, we ensure both algorithms have similar policy architectures and differ solely in their BC and AIL based policy training steps.

We plot the trajectories in Fig 5, and illustrate issue of modal averaging, observed in BC methods. We observe that SS-AIL converges to the different expert modes, implying that the Discriminator successfully approximates a reward function that has a discernible preference among prospective modalities. We do not observe this property in BC based DNRI algorithm, as shown in Fig. 5. The self-supervised loss function has an inherent positive feedback loop that progressively increases the gradients of the rewards for imitating the closest expert modality. This property is a direct consequence of the self-supervised loss, as the slope of the reward is inversely proportional to the distance between the policy and expert trajectories.

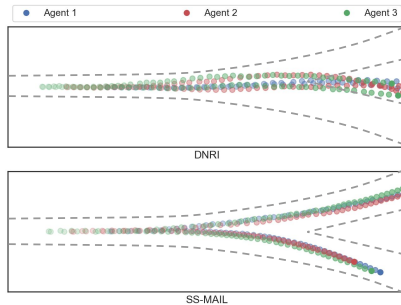


Figure 5: A visualization of the multi-modal trajectories learned by SS-AIL and DNRI on the Y-Junction task. We observe that DNRI averages over the different modes, while SS-AIL successfully differentiates between them. The inability to distinguish between multi-modal expert trajectories can prove disastrous in continuous state settings, as this may lead to visiting states that differ considerably from the expert state-distribution.

4.4 SAMPLE EFFICIENCY AND CURRICULUMS

Teacher Forcing alleviates the issues of unfavorable initializations and low sample efficiency during the initial stages of training; however, it results in compounding errors during testing. The Y Junction experiment is a simple multi-agent scenario that demonstrates how Trajectory Forcing alleviates unfavorable initializations, increases sample efficiency during the initial stages of training, and eliminates the issue of compounding errors. To evaluate the effectiveness of the Trajectory Forcing curriculum, we train SS-AIL with different values of β to show an ablation over β , which ranges from no Teacher Forcing (0%), seen in AIL approaches, to complete Teacher Forcing (100%), seen in BC approaches. In Figure 6, we focus on two aspects of the plot, the mean and standard deviation. A smaller mean loss, averaged over multiple random seeds, implies robustness in policy training to different model initializations. Moreover, a low standard deviation implies superior generalization during testing, and thereby lower compounding errors. we see that without teacher forcing the mean and standard deviation of the testing loss is considerably high, which is common for AIL methods. Teacher Forcing reduces the mean loss considerably; however, the standard deviation is high due to the compounding errors. We observe that for a β value of around 15%, the trajectory forcing curriculum considerably reduces the mean loss as well as the standard deviation during testing. This demonstrates the robustness to weight initializations, improvement in sample efficiency, and reduction in compounding errors.

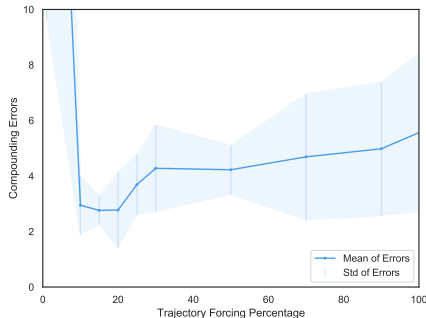


Figure 6: Running an ablation, over β , for the mean and standard deviation of testing error in the Y-Junction environment illustrates the existence of β values that result in both low mean and standard deviation. This highlights Trajectory Forcing’s ability to improve the robustness to unfavorable weight initializations, increase the sample efficiency, as well as eliminate the issue of compounding errors.

4.5 ZERO-SHOT GENERALIZATION ON NOISY REAL-WORLD DATA

Compounding errors occur when the model is unable to correct for deviations from the training distribution of states, and fails to compensate for the observed deviations. This can be attributed to poor generalization, and is commonly caused by noise. The Noisy Mocap experiment is a real-world multi-agent prediction dataset that demonstrates how negligible amounts of Gaussian noise, with standard deviation of 0.05, can accumulate and permanently derail the generation of trajectories. To evaluate the issue of compounding errors in the presence of noise during test time, we evaluate SS-AIL and DNRI using similar policy architectures.

In this experiment, we observe that DNRI progressively deviates further from the expert trajectories as the trajectory length increases. We also observe a positive correlation between the standard deviation and the trajectory length. However, in the case of MAIL we do not observe a similar linear increase in the compounding errors. Moreover, the slope decreases with the increase in trajectory length. This observed deviation can be attributed to the loss functions used to train the respective algorithms. BC approaches are trained to precisely match the expert. Despite having a lower validation loss, generalizing zero-shot to noisy environments is not guaranteed, as seen in Fig 7. SS-AIL, on the other hand, learns a policy that maximizes the cumulative discriminator return. Thus, the trained policy is akin to a vector field surrounding the expert state-distribution, as it learns to optimize for reaching rewarding states, thereby resulting in robust recovery even in noisy environments.

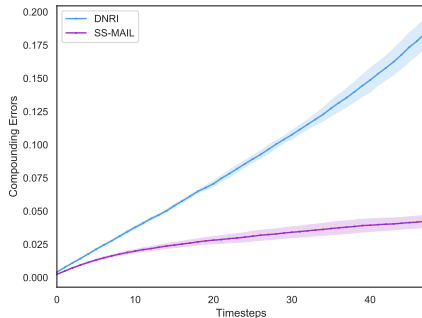


Figure 7: We plot the Compounding Errors over time for the Noisy Mocap environment shows that SS-AIL successfully compensates for noisy inputs during test time, zero-shot. The low standard deviation and the decrease in slope for SS-AIL illustrate its generalization capabilities.

4.6 REWARD SHAPING USING SELF-SUPERVISION

Reward-Shaping provides engineers and scientists with a way to meticulously control the reward function, resulting in more efficient policy training and reduction in training time. In our method, the self-supervised sampling allows for precise control over the reward function the discriminator is approximating. In this experiment, we investigate one such customization of the reward function that results in enhanced training dynamics. The Noisy Mocap experiment provides us with a real-world dataset to investigate the ramifications of different sampling procedures on the training dynamics of SS-AIL. To evaluate the effect of the self-supervised sampling, we keep the overall SS-AIL algorithm constant and solely change the sampling distributions of α .

We see that sampling only positive values of α results in training dynamics that take longer to converge. We observe that with the introduction of negative α values, the training dynamics substantially improves. This demonstrates the importance of reward shaping using self-supervision, which opens up the possibilities for researchers to design more intricate reward function approximators. The improved performance observed upon the introduction of negative α values can be attributed to the resulting completion of the reward function in the local neighborhood of the current trajectory. As the current trajectory corresponds to an α value of zero, then the absence of negative α values during self-supervised sampling would result in an incomplete reward function in the local neighborhood of the policy’s state-distribution.

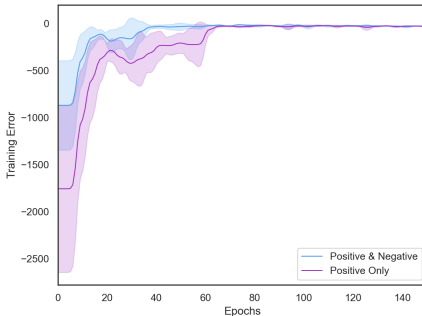


Figure 8: We plot the training loss over epochs for the Noisy Mocap environment and show that sampling negative α values improves the final loss and speeds up training. This can be attributed to the richer reward gradient in the local neighborhood of the current trajectory.

The inclusion of negative α values is one such example of reward-shaping for enhanced performance and stable training dynamics. Based on the application, engineers and scientists can design custom rewards for the enhancing the policy training, based on the application constraints.

5 RELATED WORK

In this section we provide a brief overview of the various research contributions that share similarities to the approaches described in our work. We start by discussing the recent work on graph-structured data, and then delve into multi-agent trajectory forecasting, and finally, imitation learning.

Graph structured data does not possess a fixed structure, and thus can not be used with traditional neural networks. To address these issues, there have been approaches such as (Li et al., 2015), (Scarselli et al., 2008), and (Kipf & Welling, 2016) that introduce the notion of message passing to accommodate for the irregular, yet structured nature of graphical data. In this paper, we use Graph Convolutional Networks (Kipf & Welling, 2016) for our internal message passing; however, there are other methods proposed, such as Graph Attention Networks (Veličković et al., 2017), and Graph Recurrent Networks (GRN) (Hajiramezanali et al., 2019) as prospective forms of message passing. GRN also provides the notion of recurrence, which can be used to model time-series data. The Graph Convolutional RL (Jiang et al., 2018) work is a relevant bridge between graph structured data and the sequential decision making setting, which employs message passing to deal with multi-agent reinforcement learning. Finally, we move on to the multi-agent trajectory prediction task. This setting provides is an interesting application of message passing, as the nature of the interactions among the agents may or may not be known a priori. Some recent work are: Evolve Graph (Li et al., 2020), Social Attention (Vemula et al., 2018), SocialGAN (Gupta et al., 2018), and STGAT (Huang et al., 2019).

Adversarial imitation learning methods train a discriminator network to estimate the divergence between the expert and the policy’s state distribution, and use this discriminator for policy improvement. Ghasemipour et al. (2020) provide a unified perspective on this family of algorithms and show that they can be used to minimize any f -divergence between the expert and policy state distributions. Fu et al. (2017) extend this line of work, and present an inverse reinforcement learning algorithm based on this adversarial setup. Ni et al. (2020) move away from this adversarial setup, and present an IRL algorithm that directly optimizes the reward function to minimize any f -divergence between the expert and policy state distributions. Another line of work focuses on utilizing meta-learning methods to enable few-shot imitation or reward inference (Yu et al., 2019; Xu et al., 2019).

6 CONCLUSION

We introduced SS-MAIL to effectively model multi-agent experts and improve the stability of their training dynamics compared to previous AIL methods. Our method comprises of:

1. **SS-AIL**: a self-supervised cost-regularized apprenticeship learning framework that has considerably more stable training dynamics compared to previous AIL methods, while also providing the flexibility for reward-design.
2. **MAIL**: A graph convolutional multi-agent actor-critic framework, which aids in multi-modal trajectory generation and alleviates compounding errors.
3. **Trajectory Forcing**: a teacher forcing based curriculum that takes advantage of our SS-AIL formulation to stabilize the training dynamics and alleviates the issue of domain shift between training and testing.

The improved stability of the training dynamics and the increased sample efficiency of SS-MAIL allows us to train policies that are robust to weight initializations, which would enable efficient training of multi-agent interactions. Moreover, the ability to handle compounding errors would ensure enhanced generalization during test time, in the presence of noisy inputs. Finally, learning from expert imitation provides a useful framework to train control policies in the absence of the ground truth reward function, and SS-MAIL can be applied to such multi-agent control setting, such as autonomous driving and robotic interactions to learn proficient policies that achieve human-level performance.

7 REPRODUCIBILITY

This section describes our efforts towards ensuring reproducibility of our experiments. First, we provide anonymized code for reproducing all of our experimental results. The code contains documentation for how to generate and plot each experimental result in this paper. In addition, we also provide details about our experimental setup, training hyperparameters, and datasets used in the appendix, which should be sufficient for an independent implementation. All of our experiments are done on at least five seeds to ensure reproducibility. We use author-provided implementations of DNRI and SAC to ensure consistency. Our experiments were done using a single Quadro RTX 6000 GPU, and therefore, are reasonably easy to reproduce within a small computational budget.

REFERENCES

- Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*, 2015.
- Graphics Lab Motion Capture Database CMU. Cmu graphics lab motion capture database.
- Will Dietz, Peng Li, John Regehr, and Vikram Adve. Understanding integer overflow in c/c++. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 25(1):1–29, 2015.
- Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.
- Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pp. 1259–1277. PMLR, 2020.
- Colin Graber and Alexander G Schwing. Dynamic neural relational inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8513–8522, 2020.
- Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264, 2018.
- Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.
- Ehsan Hajiramezani, Arman Hasanzadeh, Nick Duffield, Krishna R Narayanan, Mingyuan Zhou, and Xiaoning Qian. Variational graph recurrent neural networks. *arXiv preprint arXiv:1908.09710*, 2019.
- Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.
- Yingfan Huang, HuiKun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6272–6281, 2019.
- Rohit Jena, Changliu Liu, and Katia Sycara. Augmenting gail with bc for sample efficient imitation learning. *arXiv preprint arXiv:2001.07798*, 2020.
- Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. *arXiv preprint arXiv:1810.09202*, 2018.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

- Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pp. 2688–2697. PMLR, 2018.
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *arXiv preprint arXiv:2003.13924*, 2020.
- Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.
- Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Benjamin Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. *arXiv preprint arXiv:2011.04709*, 2020.
- Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*, 2018.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.
- Stéphane Ross, Geoffrey Gordon, and Drew 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*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.
- Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pp. 4601–4607. IEEE, 2018.
- Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.
- Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, and Chelsea Finn. Learning a prior over intent via meta-inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 6952–6962. PMLR, 2019.
- Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *arXiv preprint arXiv:1909.09314*, 2019.

A APPENDIX

In this section, we provide the derivation and the proof for the Theorem 3.1, to establish the connection between Apprenticeship learning and SS-AIL.

A.1 PROOF

Ho and Ermon show that the cost-regularized apprenticeship learning problem

$$\text{AL}_{\psi}(\pi_E) = \min_{\pi} \max_c -\psi(c) + E_{\pi}[c(s, a)] - E_{\pi_E}[c(s, a)] \quad (7)$$

the above is equivalent to min-max problem

$$\min_{\pi} \psi^*(\rho_{\pi}, \rho_E) \quad (8)$$

where ψ^* is the convex conjugate of the cost regularizer ψ .

For SS-AIL,

$$\psi_{\text{SS}}^*(\rho_{\pi} - \rho_{\pi_E}) = \mathbb{E}_{\pi, \pi_E, \alpha} \left[\underbrace{(0 - D(s_G, a_G))^2}_{\text{Generated MSE}} + \underbrace{(1 - D(s_E, a_E))^2}_{\text{Expert MSE}} + \underbrace{(\alpha - D(s_{\alpha}, a_{\alpha}))^2}_{\text{Self-Supervised MSE}} \right] \quad (9)$$

The SS-AIL algorithm aims to solve the min-max optimization problem $\max_{\pi} \min_D \mathbb{E}_{\pi, \pi_E, \alpha} [(0 - D(s_G, a_G))^2 + (1 - D(s_E, a_E))^2 + (\alpha - D(s_{\alpha}, a_{\alpha}))^2]$ using the gradient descent-ascent algorithm. The inner discriminator optimization is a supervised learning problem with the self-supervised loss. The outer policy optimization is an off-policy policy improvement problem, and we use the soft actor-critic algorithm.

Therefore, SS-AIL is an instantiation of cost-regularized apprenticeship learning with the cost regularizer ψ_{SS} .

A.2 FURTHER ELABORATION ON SELF-SUPERVISION FOR REWARD SHAPING

We start by elaborating on the self-supervised sampling used to create Fig 1. We observed that sampling $\alpha \in [-1, 1]$ was not sufficient for reproducing a drop in the reward function after the expert trajectories. Thus, we increased our sampling range to $[-1, 1.5]$ and set the reward function to 0 for values of $\alpha > 1$. This helped our discriminator learn a reward function that does not monotonically increase as we get farther from the current policy.

We can always split our sampling procedure to get different results. We can make use of the non-linearity of the neural-network architectures to even approximate piece-wise distributions. Another extension could be to have steeper slopes farther away from the current policy and smoother slopes closer to the policy state-distribution.

B IMPLEMENTATION DETAILS

We begin by delineating all the experimental details that are common to each of the experiments. Our code is built on Pytorch version 1.2, and we use Adam as our optimizer of choice. We follow the teacher forcing training regiment provided in DNRI (Graber & Schwing, 2020) to train our baseline reference models for comparison.

We save the best models after every epoch, and use the models with the best validation loss for testing. Similar to Graber & Schwing (2020), we normalize our inputs between -1 to 1 using the maximum and the minimum state norms. For all the experimental evaluations, we average the results over 5 seeds.

B.1 CODE

We have provided a driver code in the supplementary material, and the specific hyperparameters in separate folders, one for each figure.

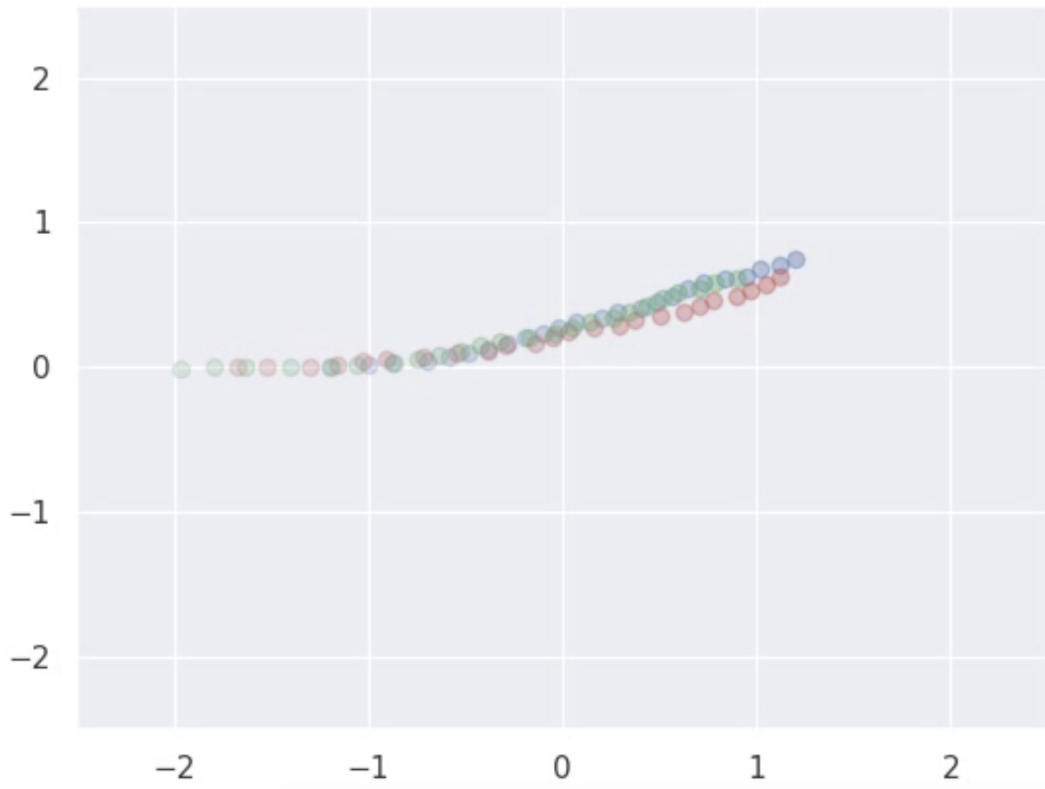


Figure 9: Sample snapshot of the Y-Junction Experiment

B.2 Y-JUNCTION EXPERIMENTAL SETUP

We expand upon the simulated synthetic experiments of (Graber & Schwing, 2020), and use the same number

B.3 NOISY MOCAP SETUP

For noisy mocap, we add gaussian noise of 0.05 standar deviation.