



POLITECNICO
MILANO 1863



Reformulating Imitation Learning as Return Distribution Matching

F. Lazzati, A. M. Metelli

The 2nd Workshop on Advances in Financial AI

Outline

Background

Contribution 1 - New Imitation Learning Formulation

Contribution 2 - Algorithmic Solution

Contribution 3 - Theoretical Analysis

Conclusion

Outline

Background

Contribution 1 - New Imitation Learning Formulation

Contribution 2 - Algorithmic Solution

Contribution 3 - Theoretical Analysis

Conclusion

Background

- AI nowadays $\rightarrow$ sequential decision-making problems

Background

- AI nowadays $\rightarrow$ sequential decision-making problems
- Objective $\rightarrow$ reward r^* (e.g., money)

Background

- AI nowadays → sequential decision-making problems
- Objective → reward r^* (e.g., money)
- Examples → chatbots (LLMs), autonomous robots (VLAs)

Background

- AI nowadays → sequential decision-making problems
- Objective → reward r^* (e.g., money)
- Examples → chatbots (LLMs), autonomous robots (VLAs)
- Training technique → **Imitation Learning** (Pomerleau 1988):
 1. collect human demonstrations
 2. train the agent to imitate them

Background

- AI nowadays → sequential decision-making problems
- Objective → reward r^* (e.g., money)
- Examples → chatbots (LLMs), autonomous robots (VLAs)
- Training technique → **Imitation Learning** (Pomerleau 1988):
 1. collect human demonstrations
 2. train the agent to imitate them
- **However**, existing algorithms are *risk-neutral* → unsuitable for financial applications

Outline

Background

Contribution 1 - New Imitation Learning Formulation

Contribution 2 - Algorithmic Solution

Contribution 3 - Theoretical Analysis

Conclusion

Contribution 1 - New Imitation Learning Formulation

- Given human behavioral data $\mathcal{D} = \{(s_i, a_i)\}_i \sim \pi^E$
- Collected from some policy $\pi^E : \mathcal{S} \rightarrow \mathcal{A}$
- **Find $\hat{\pi}$ minimizing $\mathcal{W}(\eta^{\hat{\pi}}, \eta^{\pi^E})$**
- Where η^{π} =distribution over rewards r^* induced by π , $\mathcal{W}$ =Wasserstein distance

Contribution 1 - New Imitation Learning Formulation

- Given human behavioral data $\mathcal{D} = \{(s_i, a_i)\}_i \sim \pi^E$
- Collected from some policy $\pi^E : \mathcal{S} \rightarrow \mathcal{A}$
- **Find $\hat{\pi}$ minimizing $\mathcal{W}(\eta^{\hat{\pi}}, \eta^{\pi^E})$**
- Where η^{π} =distribution over rewards r^* induced by π , $\mathcal{W}$ =Wasserstein distance

Key Insights:

- In Imitation Learning, we minimize $|\mathbb{E}[\eta^{\hat{\pi}}] - \mathbb{E}[\eta^{\pi^E}]|$
- Allows us to train foundation models with the same risk preferences as the human demonstrator

Outline

Background

Contribution 1 - New Imitation Learning Formulation

Contribution 2 - Algorithmic Solution

Contribution 3 - Theoretical Analysis

Conclusion

Contribution 2 - Algorithmic Solution

Algorithm 1: RS-BC (Risk-Sensitive Behavior Cloning)

Input: Dataset $\mathcal{D} = \{(s_i, a_i)\}_i$, reward r^* , policy parameters θ

// Enrich the dataset with reward information

1 $\mathcal{D}' \leftarrow \{(s_i, g_i, a_i)\}_i$, where $g_i = \sum_{j < i} r^*(s_j, a_j)$

// Train $\hat{\pi}_\theta$ to predict a_i given (s_i, g_i)

2 $\theta^* \leftarrow \arg \min_{\theta} \sum_i \mathcal{L}(\hat{\pi}_\theta(s_i, g_i), a_i)$

3 **Return** θ^*

Outline

Background

Contribution 1 - New Imitation Learning Formulation

Contribution 2 - Algorithmic Solution

Contribution 3 - Theoretical Analysis

Conclusion

Contribution 3 - Theoretical Analysis

Theorem (Informal)

For any human demonstrating policy π^E and reward r^* , the RS-BC algorithm requires just a polynomial number of samples to output a policy $\hat{\pi}$ whose induced distribution over rewards $\eta^{\hat{\pi}}$ is close to the distribution η^{π^E} induced by the demonstrator policy in Wasserstein distance with high probability.

Outline

Background

Contribution 1 - New Imitation Learning Formulation

Contribution 2 - Algorithmic Solution

Contribution 3 - Theoretical Analysis

Conclusion

Conclusion

In this work:

- Risk-sensitive formulation of imitation learning
- Algorithm for training risk-aware foundation models
- Theoretical guarantees of sample efficiency

Conclusion

In this work:

- Risk-sensitive formulation of imitation learning
- Algorithm for training risk-aware foundation models
- Theoretical guarantees of sample efficiency

Future directions:

- Extensive empirical validation of RS-BC
- Fine-tuning the output of RS-BC with reinforcement learning

Thank you for your attention!

Questions?

References |



Pomerleau, Dean A. (1988). "ALVINN: An Autonomous Land Vehicle in a Neural Network". In: *Advances in Neural Information Processing Systems 1 (NeurIPS)*.