

Imitation learning framework based on Fully Probabilistic Design

Siavash Fakhimi Derakhshan, Ladislav Jirsa, Tatiana Valentine Guy

Adaptive System Department, Institute of Information Theory and Automation, Czech Academy of Sciences

fakhimi@utia.cas.cz, jirsa@utia.cas.cz, guy@utia.cas.cz

Keywords

Fully probabilistic design, imitation learning, Kullback-Liebler divergence, learning from demonstration, optimal policy

1 Introduction

Imitation learning is an important branch of machine learning and control theory that aims to construct decision-making policies from observed demonstrations. Instead of learning through repeated trial-and-error interaction with the environment, the learner tries to extract useful behavioral patterns from expert trajectories. This paradigm is especially attractive in applications where online exploration is costly, risky, or impractical, such as robotics, autonomous systems, industrial control, and human-centered decision support. In such settings, demonstrations provide a valuable source of information about how a task can be performed successfully.

However, traditional imitation learning methods face several limitations. Behavior Cloning, one of the most common approaches, directly maps states to actions using supervised learning. Although simple and computationally convenient, it often suffers from compounding errors when the learner encounters states that are insufficiently represented in the training data. Inverse Reinforcement Learning addresses this issue by inferring a latent reward function that explains expert behavior, but this comes at the cost of increased computational complexity and reliance on strong assumptions regarding optimality and reward structure. More recent adversarial methods improve flexibility and empirical performance, yet they typically require complex optimization procedures and substantial data.

This study introduces an alternative framework for imitation learning by building on the theory of Fully Probabilistic Design (FPD). Instead of formulating the control objective in terms of a reward function, FPD describes desired behavior through an ideal probability distribution over future states and actions. The learner then searches for a stochastic policy whose induced closed-loop behavior is as close as possible to this ideal model. This perspective provides a principled probabilistic foundation for decision-making under uncertainty and opens a new direction for imitation learning in environments where demonstrations are available but explicit rewards are not.

The importance of this idea lies in the fact that it does not seek to reproduce demonstrations mechanically. Rather, it evaluates demonstrated behavior according to how well it aligns with the learner's own probabilistic objective. As a result, demonstrations become a source of structured evidence rather than fixed templates to be copied. This makes the framework particularly suitable when demonstrations come from different experts, contain exploratory actions, or are only partially optimal.

2 Main Contribution

The main contribution of this study is the development of a novel imitation learning method that integrates demonstration-based learning with the Fully Probabilistic Design framework. The method learns a stochastic policy directly from trajectories while explicitly incorporating the learner's desired probabilistic behavior. This is a meaningful contribution because it moves beyond the two dominant paradigms in imitation learning: direct action replication and reward inference. Instead of asking what action the expert chose in a given state, or what hidden reward the expert might have optimized, the proposed method asks which demonstrated behaviors are most compatible with the learner's ideal closed-loop objective.

A central strength of the method is its ability to work with heterogeneous and imperfect demonstrations. The dataset may include trajectories generated by multiple experts, each with different styles, preferences, or levels of quality. In many real-world scenarios, such variability is unavoidable. Standard imitation learning methods often struggle in this setting because they implicitly assume that all demonstrations are equally informative and equally aligned with the target objective. The proposed framework overcomes this difficulty by assigning importance weights to demonstrated trajectory fragments according to their compatibility with the ideal probability distribution. Fragments that better support the desired behavior receive more influence, while less useful or poorly aligned fragments are naturally down-weighted.

Another major contribution is the algorithmic mechanism through which the policy is learned. At each decision step, the learner observes the current state and searches the demonstration dataset for subsequences beginning in that state. For each candidate action, it extracts relevant future trajectory fragments and uses them to estimate the local closed-loop behavior that would follow from that choice. This estimated behavior is then compared with the ideal probabilistic model through the Kullback–Leibler divergence. Candidate actions that produce smaller divergence are given a higher probability in the learned policy. This construction yields a stochastic decision rule that is both data-driven and explicitly goal-oriented.

This research, therefore, contributes not only a conceptual framework but also a practical computational procedure. The learner does not need to solve a full reinforcement learning problem, infer a reward function, or assume that demonstrations are perfectly optimal. Instead, it uses available data in a selective and principled way. This makes the method computationally attractive and theoretically elegant. It also offers a bridge between probabilistic control and imitation learning by showing that demonstration data can be filtered through an explicit probabilistic objective rather than treated as unquestionable ground truth.

A further contribution is the reinterpretation of imitation learning itself. In this framework, imitation is not merely copying observed actions; it is the process of extracting useful behavioral structure from demonstrations under a user-defined objective. This makes the learner more autonomous and robust. The policy that emerges is informed by demonstrations, but it is ultimately shaped by the learner’s own model of desirable behavior. Such a viewpoint is especially valuable in uncertain environments where demonstrations may be informative yet incomplete, inconsistent, or only locally relevant.

3 Conclusion

In conclusion, *Policy Learning via Fully Probabilistic Design* makes an important contribution to imitation learning by introducing a probabilistic alternative to reward-based and purely supervised approaches. This study shows that it is possible to learn stochastic control policies directly from demonstrations while avoiding explicit reward engineering and while remaining robust to imperfect or heterogeneous data. By embedding imitation learning into the FPD framework, the authors provide a method that is theoretically grounded, practically meaningful, and well-suited to uncertain decision-making settings.

The key value of the proposed approach is its ability to combine two important ideas: learning from demonstrations and optimizing toward a desired probabilistic behavior. Demonstrations are not copied blindly; instead, they are evaluated and weighted according to their relevance to the learner’s ideal objective. This allows the method to exploit useful information even when the available data are diverse, noisy, or partially suboptimal. At the same time, the use of stochastic policies makes the framework naturally compatible with uncertainty and complex control tasks.

Overall, the proposed work presents a clear and original perspective on imitation learning. Its main significance lies in showing that demonstration-based policy learning can be formulated as probabilistic matching between actual and desired closed-loop behavior. This insight broadens the methodological toolbox of imitation learning and suggests promising opportunities for future work in adaptive control, robotics, autonomous systems, and data-driven sequential decision-making.

Acknowledgement

The authors would like to acknowledge the partial support of the COST Action CA21169 and the IAMI Project funded by the European Union under Grant Agreement No 101168272.