Towards Offline Imitation Learning: Strictly Batch Settings, Generalization, and Variability in Expertise

Rishabh Agrawal
University of Southern California
rishabha@usc.edu

Abstract

My doctoral research develops a unified framework for offline imitation learning (IL) that tackles three central challenges: achieving sample efficiency in strictly batch settings, ensuring robustness and generalization under dynamics shifts, and learning from demonstrations of varying quality. At the core of this work is a new paradigm for strictly offline IL based on enforcing the Markov Balance Equation (MBE), a fundamental structural property of trajectory data. Using advanced conditional density estimation, I developed two algorithms, CKIL and MBIL, which achieve state-of-the-art performance in high-dimensional continuous-control tasks. Building upon this foundation, I developed the first Distributionally Robust Offline IL framework under a stationarity constraint, enabling robustness to transition-model mismatch without requiring any additional interaction. I am now extending this direction through Robust Behavior Foundation Models (RBFMs), which aim to generalize across dynamics shifts for a wide range of tasks. Finally, I propose a variational approach for learning from crowdsourced demonstrations by inferring and accounting for demonstrator expertise. Together, these contributions yield principled and practical IL algorithms with strong performance and robustness, broadening the applicability of IL to real-world domains such as robotics, healthcare, and autonomous systems.

Introduction and Motivation

Teaching autonomous agents complex skills remains a central goal of AI. IL is particularly compelling because it bypasses the need for manual reward engineering by learning directly from expert demonstrations. However, many IL algorithms require continuous interaction with the environment, which is often impractical, unsafe, or prohibitively expensive. This motivates the **strictly batch offline IL** setting, where the agent must learn from a fixed dataset without any additional interaction. This setting is highly challenging: classical methods such as Behavioral Cloning (BC) suffer from covariate shift (Ross and Bagnell 2010), while modern offline approaches can exhibit termination and reward biases (Sun et al. 2021). Moreover, strictly batch offline IL demands algorithms that efficiently leverage limited data while remaining robust to distributional shifts.

My research addresses this gap by asking the central question: *Can we develop a unified offline IL framework that is sample-efficient, generalizes to new environments, and effectively learns from noisy, mixed-expertise demonstrations?* Answering this requires both principled theoretical foundations and practical algorithmic innovations.

Foundational Work: IL via Markov Balance

The first part of my thesis establishes a novel paradigm for strictly batch IL. The key insight is that the **Markov Balance Equation (MBE)**, a fundamental identity linking a policy with the underlying dynamics, provides a direct and powerful learning signal. I reformulated IL as the problem of finding a policy that best satisfies:

$$P_{\pi_D}(s', a'|s, a) = \pi_D(a'|s')T(s'|s, a),$$

where $T(s'|s, a)$ represents the unknown transition dynamics of the underlying Markov Decision Process (MDP), $\pi_D(a'|s')$ is the demonstrator's policy that we aim to learn, and $P_{\pi_D}(s', a'|s, a)$ is the transition density of the induced Markov chain over state-action pairs. By modeling the unknown transition densities using advanced conditional density estimation, this approach provides a principled pathway to directly learn policies from batch data without interacting with the environment. Key contributions include:

1. **CKIL:** An algorithm based on Conditional Kernel Density Estimators that established both the theoretical foundation and the initial empirical success of the MBE approach (Agrawal et al. 2025b). This work also includes a detailed finite-sample complexity analysis of the transition densities, providing strong theoretical guarantees.
2. **MBIL:** A scalable algorithm leveraging Conditional Normalizing Flows, which overcomes the curse of dimensionality inherent in kernel-based methods and achieves state-of-the-art performance on high-dimensional MuJoCo tasks using a single expert trajectory (Agrawal et al. 2025c).

While MBE-based IL is highly effective under nominal dynamics, real-world systems often exhibit test-time changes (e.g., friction, mass, actuator delay). To address this, I developed **BE-DROIL** (Agrawal et al. 2025a) (Balance Equation-based Distributionally Robust Offline IL), which provides the *first* robust imitation learning framework under a stationarity constraint that operates strictly

offline using demonstrations from a single nominal environment, accounts for worst-case transition shifts via an f-divergence ambiguity set, and enforces a Bellman-flow-consistent triplet occupancy formulation that removes direct dependence on unknown dynamics. Using convex duality, we derived a closed-form importance-weighted objective over nominal data, enabling a practical alternating optimization algorithm, and empirically demonstrated state-of-the-art robustness on MuJoCo environments.

These foundational works were conducted with my advisor Prof. Jain and collaborators Prof. Nayyar and Prof. Dahlin. My primary contributions included formulating the MBE objective, designing and implementing the algorithms, performing finite-sample complexity analysis, and leading the experimental evaluations. Collaborators guided the work in terms of technical correctness and overall validation.

Current and Future Research Thrusts

Generalization via Robust Behavior Foundation Models

While an optimal policy in an MDP is tied to a specific reward function, changes in the task typically require retraining. Behavior Foundation Models (BFMs) have been proposed to enable zero- or few-shot adaptation by learning the occupation measure of all possible policies (Pirotta et al. 2024). However, a critical limitation of BFMs, and of IL more broadly, is their poor generalization under test-time dynamics shifts. In my ongoing work, I am developing Robust Behavior Foundation Models (RBFMs) that preserve the adaptive capabilities of BFMs while being robust to worst-case dynamics changes. This work builds upon my BE-DROIL framework, but its direct application is non-trivial because BFMs learn occupation measures across an entire class of policies, and the worst-case transition distribution may differ for each individual policy. Hence, addressing this challenge requires a principled robustification strategy that jointly considers learning occupancy measures for all policies and defining the corresponding uncertainty set.

Learning from Crowdsourced Data with Varying Expertise

Real-world demonstration data is often crowdsourced and imperfect, containing a mix of expert and novice trajectories that, if imitated directly, can yield suboptimal policies (Wang et al. 2023). The variability in demonstrator skill and consistency further complicates offline IL, which must rely entirely on the provided dataset. To address this challenge, I propose a variational framework (Kingma and Welling 2013) that infers a latent expertise variable for each demonstrator without requiring expertise labels. The learned policy conditions on this inferred expertise, adopting more stochastic behavior for low-quality demonstrations and more deterministic behavior for high-quality ones. This approach leverages the full dataset while selectively emphasizing reliable trajectories, enabling robust and high-performance policy learning from mixed-quality demonstrations.

Anticipated Thesis Contributions

1. A principled framework for strictly batch offline IL based on the Markov Balance Equation, offering stronger learning signals and improved performance in high-dimensional continuous environments.
2. A distributionally robust strictly offline IL formulation that enables reliable generalization under dynamics shifts for a given task.
3. Robust Behavior Foundation Models (BFMs) that train policies to generalize across environmental dynamics variations, with the promise of generalizing seamlessly across tasks with minimal fine-tuning.
4. A practical variational-inference approach for learning from noisy, unlabeled, and crowdsourced demonstrations, allowing the extraction of high-quality policies.

Conclusion

This thesis introduces a new paradigm for strictly offline IL centered on the Markov Balance Equation and complements it with approaches that address key challenges in robustness, generalization, and learning from imperfect data. Foundational work validates the balance-based framework, while ongoing research develops robust, general-purpose IL agents for real-world deployment. Overall, this research pushes the boundaries of offline IL, producing algorithms that are both high-performing and dependable in practical settings.

References

Agrawal, R.; Alvi, Y.; Jain, R.; and Nayyar, A. 2025a. Balance Equation-based Distributionally Robust Offline Imitation Learning. *arXiv preprint arXiv:2511.07942*.

Agrawal, R.; Dahlin, N.; Jain, R.; and Nayyar, A. 2025b. Conditional Kernel Imitation Learning for Continuous State Environments. In *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, Proceedings of Machine Learning Research. PMLR.

Agrawal, R.; Dahlin, N.; Jain, R.; and Nayyar, A. 2025c. Markov Balance Satisfaction Improves Performance in Strictly Batch Offline Imitation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15311–15319.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Pirotta, M.; Tirinzoni, A.; Touati, A.; Lazaric, A.; and Ollivier, Y. 2024. Fast Imitation via Behavior Foundation Models. In *The Twelfth International Conference on Learning Representations*.

Ross, S.; and Bagnell, D. 2010. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 661–668. JMLR Workshop and Conference Proceedings.

Sun, M.; Mahajan, A.; Hofmann, K.; and Whiteson, S. 2021. Softdice for imitation learning: Rethinking off-policy distribution matching. *arXiv preprint arXiv:2106.03155*.

Wang, Y.; Dong, M.; Zhao, Y.; Du, B.; and Xu, C. 2023. Imitation learning from purified demonstrations. *arXiv preprint arXiv:2310.07143*.