A Brief Survey of Online Imitation Learning

Sagar Shrestha[*]
Department of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331
`shressag@oregonstate.edu`

Abstract

Imitation learning (IL) is an important framework for solving sequential decision making problems. It has been shown that IL problem can be reduced to online learning regret minimization problem. As such, IL can directly make use of the existing adversarial online learning algorithms and also inherits the convergence guarantees of the corresponding online learning methods. However, the special structure of IL problem allows for a more relaxed assumptions on the nature of the loss functions which proves adversarial online learning methods too pessimistic in the IL context. Moreover, important practical issues prevent successful application of the existing online imitation learning methods in real world scenarios. Recently, there have been several attempts at addressing these issues. In this report, we provide a brief account of important developments in the field of imitation learning with discussion on ongoing efforts to resolve the associated issues, which are all tightly coupled with online learning.

1 Introduction

In sequential decision making task, an agent takes a series of actions in its environment in order to achieve a predefined goal. Such problems are usually framed as a Reinforcement Learning (RL) problem. However, the RL objective is a hard optimization problem, and existing methods suffer from various issues, such as high sample complexity and high variance of gradients for direct policy optimization. Imitation Learning (IL) circumvents several issues in RL by utilizing expert policy as a reference to simplify the RL objective. IL enjoys a wide range of applications in almost all sequential decision making tasks, e.g., autonomous driving [1], robot manipulation [2], game playing [3], optimization [4], etc.

It has been shown in [5] that the IL objective can be reduced to adversarial Online Learning (OL) regret minimization problem. This provides an important abstraction on the original problem, and enables the utilization of existing algorithms to solve the IL problem. Moreover, the adopted IL algorithm directly inherits the convergence guarantees derived in the OL literature. [5] showed that using any no-regret OL algorithm ensures that (near)-optimal solution to IL problem can be obtained. As such, [5] proposed to use an instance of the Follow-the-leader (FTL) algorithm, named as Dataset Aggregation (`DAgger`) in the context of IL, with strongly convex and lipschitz loss function.

Following the success of [5], many variants were proposed for the IL problem that sought to improve upon `DAgger`. The key issues in using `DAgger` in straightforward manner stems from the stringent requirement that the expert has to label all states visited by the agent. This is prohibitive in most applications. The reduction from IL to OL also makes an overly stringent assumption that the loss function corresponding to the OL setting is adversarial in nature. However, the special structure of the IL problem means that this loss function is non-adversarial. This allows one to exploit this

[*] Project report for AI 539 Introduction to Online Learning.

special structure to make improvements upon the naive OL setting. In the following sections, we summarize these important developments.

2 Background

In sequential decision making task, an agent sequentially takes an action and observes the state of the environment that its action leads to. This process is often framed in the context of Markov Decision Process (MDP). To make it concrete let $(\mathcal{S}, \mathcal{A}, \mathbb{P}, c)$ represent an MDP, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ the set of actions, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ the transition probabilities, and $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the cost function. The RL objective is then to find a policy such that it maximizes the cumulative cost from the starting to the terminating state. Concretely the following summarizes the RL objective:

$$\min_{\pi \in \Pi} J(\pi) := \mathbb{E}_\pi \left[\sum_{t=1}^{T} c(s_t, a_t) \right].$$

Here, we assume finite horizon setting, i.e., the number of time-steps an agent takes is finite. The above can be equivalently written in terms of the induced state distribution under π, denoted by d_π as follows:

$$\min_{\pi \in \Pi} J(\pi) := \mathbb{E}_{s \sim d_\pi} \mathbb{E}_{a \sim \pi(s)} \left[c(s, a) \right].$$

The above problem is computationally challenging as the sampling distribution is also our optimization variable. Moreover, in practice, the cost function is often sparse, e.g., only the win/loss at the end of the episode in games. As such existing RL algorithms incur high sample complexity [6], and exhibit unstable training [7]. Moreover, it is not easy to formulate or select a cost function in many applications. For example, there is no well-defined cost function that can describe an optimal driving behavior.

In order to circumvent these issues, IL offers a suitable alternative whenever an "expert" reference policy is available. For example, in autonomous driving task, it is difficult to define a cost function but easy to obtain driving data from human drivers. With access to expert policy, denoted by $\pi^\star$, one can write the IL objective as finding a policy that performs as well as the expert:

$$\min_{\pi \in \Pi} \; J(\pi) - J(\pi^\star).$$

Performance difference lemma [8] showed that

$$J(\pi) - J(\pi^\star) \leq C_{\pi^\star} \mathbb{E}_{d_\pi} \left[D(\pi || \pi^\star) \right],$$

where $D(\pi||\pi^\star)$ is a divergence measure between two probability distributions, e.g., KL-divergence, and $C_{\pi^\star}$ is a constant. This implies that in order to perform as well as the expert, on simply needs to minimize the following objective:

$$\min_{\pi \in \Pi} \mathbb{E}_{d_\pi} \left[D(\pi || \pi^\star) \right]. \tag{1}$$

Since the expectation is still with respect to the state distribution under the policy that we wish to learn, the problem is still difficult computationally. A naive approach to IL is to take expectation under the state distribution visited by the expert policy, i.e., replacing d_π with $d_{\pi^\star}$. However, it has been shown that the such an approach incurs error quatratic in T. In the following section, we present how Problem 1 can be converted to an OL problem.

3 Reduction from IL to OL

The seminal work of [5] showed that problem (1) can be framed as a regret minimization problem in OL. To understand this, recall that in the framework of OL, the environment reacts with a loss function after every action that the agent takes. At time step n, the action of the agent can be regarded as the the choice of policy π_n and the loss function can be chosen as follows:

$$l_n(\pi) = \mathbb{E}_{d_{\pi_n}} [D(\pi || \pi^\star)].$$

Note that this loss function does not incur the same optimization difficulty because of fixed distribution d_{π_n}. The average regret can then be written as

$$\text{AvgRegret}_N = \frac{1}{N} \sum_{n=1}^{N} l_n(\pi_n) - \min_{\pi \in \Pi} \frac{1}{N} \sum_{n=1}^{N} l_n(\pi) \tag{2}$$

Using any no-regret algorithm, such as FTL, OGD and FTRL, one can show that the $\mathrm{AvgRegret_N} \to 0$ as $N \to \infty$. [5] showed that minimizing the average regret corresponds to approximately minimizing the objective in Problem (1). This can be observed as follows:

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N} l_n(\pi_n) - \min_{\pi\in\Pi}\frac{1}{N}\sum_{n=1}^{N} l_n(\pi) &= \mathrm{AvgRegret}_N \\
\implies \frac{1}{N}\sum_{n=1}^{N} l_n(\pi_n) &= \min_{\pi\in\Pi}\frac{1}{N}\sum_{n=1}^{N} l_n(\pi) + \mathrm{AvgRegret}_N \\
\implies \min_{\pi_1,\ldots,\pi_N} l_n(\pi_n) &\le \min_{\pi\in\Pi}\frac{1}{N}\sum_{n=1}^{N} l_n(\pi) + \mathrm{AvgRegret}_N \\
\implies \min_{\pi_1,\ldots,\pi_N} \mathbb{E}_{d_{\pi_n}}[D(\pi_n||\pi^\star)] &\le \underbrace{\min_{\pi\in\Pi}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{d_{\pi_n}}[D(\pi||\pi^\star)]}_{\epsilon_N} + \mathrm{AvgRegret}_N.
\end{aligned}$$

This implies that there exist a policy π_n within the sequence of policies $\pi_1, \ldots, \pi_N$, whose expected divergence from the expert policy is upper bounded by the sum of $\mathrm{AvgRegret}_N$ and the best policy in the policy class for aggregated loss function. If we use a no-regret algorithm, and our policy class contains the expert policy, then this implies that π_n is a solution to (1). In [5], the authors propose to use Follow-the-leader (FTL) as the no-regret online learning algorithm. The resulting IL algorithm was termed as DAgger (Dataset Aggregation). Success of this algorithm was followed by many works trying to improve upon the algorithm and analysis [9–13] The algorithm has also observed wide adoption in many important applications [1, 2, 14].

3.1 Challenges and Existing Methods

Although the work of [5] converts the hard optimization problem (1) into a familiar online learning objective, there are many practical and theoretical issues with the proposed reduction. In this section we discuss these issues and existing works that address these issues.

3.2 Limited Expert Demonstration Learning

One of the major practical issues with the `DAgger` algorithm is that it requires the expert policy $\pi^\star$ to label all the states visited by the policy π_n for all $n \in [N]$. This becomes prohibitive for many application domains. Consider, for example, the driving task, where it essentially means that the human driver needs to label the states visited by the current learnt policy. This is clearly not feasible. In order to address this issue, there have been several efforts that tries to make use of limited expert query for imitation learning [10, 11, 13].

[13] proposed `SafeDAgger` which sought to minimize the number of expert queries in the `DAgger` algorithm. Specifically, it introduces a safety policy that takes into account both the state and primary policy's actions to predict the likelihood of primary policy to deviate from the expert trajectory. This allows the method to only query the expert whenever there is sufficient deviation from the expert trajectory.

Another line of work considers expert intervention learning [10, 11], where the system tries to learn the policy from intervention data of the expert. Specifically, in [10], it is assumed that the agent executes its policy in the environment and expert only intervenes whenever the agent's action on any state is beyond certain threshold away from the expert's policy. For this, the authors define two spaces: a space of good state-action (s, a) pairs, denoted by $\mathcal{G}$, and a space of bad (s, a), denoted by $\overline{\mathcal{G}}$. It is assumed that the expert only intervenes whenever the agent is in $\overline{\mathcal{G}}$. Let d_{π_θ} denote the state distribution visited by π_θ, and $d^I_{\pi_\theta}$ denote the state distribution visited under the expert policy during intervention. Then the following objective is minimized:

$$\underset{\theta}{\text{minimize}}\ \ l(\theta) = \underbrace{\mathbb{E}_{s\sim d_{\pi_\theta}}[1_{(s,\pi_\theta(s))\notin\mathcal{G}}]}_{\text{stay in good enough region}} + \lambda\underbrace{\mathbb{E}_{s\sim d^I_{\pi_\theta}}[1_{\pi_\theta(s)\neq\pi^\star(s)}]}_{\text{learn intervention actions}}.$$

Essentially, the agent is encouraged to stay in good state-action region and learn from expert while in bad region. The loss function $l(\theta)$ is similarly converted into an online loss function $l_n(\theta)$ by fixing the distribution at time step n to be $d_{\pi_{\theta_n}}$, and online learning algorithm can be used. [10] and [11] showed great empirical success of the approach in driving scenario, with huge improvement over `DAgger` in terms of the number of expert queries.

3.3 Exploiting the Non-adversarial Nature of the Loss Function

Another important issue with the reduction of IL to OL is that the adopted OL methods consider the loss function as adversarial in nature, and optimize for the worst case. However this is hardly true for IL. The online loss function in IL is determined by the expert policy and the state distribution under the selected policy. Since the expert policy is usually fixed for all time steps, and the selected policy is under the control of the agent, the loss function is far from adversarial. As such, there have been works that attempt to exploit the structure of the online loss function of IL in order to derive better approach to solving (2) [9, 15, 16].

In [15], the authors propose to learn a predictive model of the future loss function. The main idea is to estimate the future cost function $l_n(\cdot)$ so that the agent could select the policy $\pi_n \approx \min_{\pi\in\Pi}\sum_{i=1}^{n} l_i(\pi)$. This leads to accelerated regret minimization. To that end [15] analyse the first order optimality condition of the loss function for convex Π, and build a predictive model for the gradient of the loss. The authors show that the resulting algorithm that leverages such prediction offers a provable acceleration on the covergence of the existing online imitation learning algorithms.

3.4 Closing the Gap between Theory and Practice

The reduction from IL to OL allows IL methods to inherit convergence analysis from the adopted OL methods. For example, `DAgger` inherits the convergence guarantees of the FTL algorithm for strongly convex and Lipschitz loss functions. However, empirically it has been observed that the convergence rate is much faster; e.g., `DAgger` in [5] learned to mimic a model predictive control in just three rounds. One of the reasons for this mismatch is because of the non-adversarial loss functions in IL, whereas the OL algorithms were analyzed under the worst-case adversarial loss functions [17]. As such, there have been attempts at closing this gap by deriving custom analysis for IL algorithms [18]. Mainly, analysis in [18] concluded that increasing the capacity of the policy class improves the convergence rate for convex and smooth loss functions.

Additionally, in practice, using neural network based policies has also observed empirical success [16]. However, this makes the loss function non-convex, and the existing analysis does not cover this case. In order to address these issues, recent works have focused on analysis that does not define convexity in terms of the parameters of the policy but on the statistics of the policy distribution [12]. Although, these issues have not been fully addressed, there are ongoing efforts to closing the theory-practice gap.

4 Discussion and Conclusion

Reduction of IL to OL provides a powerful framework for solving imitation learning problems, which has been effective in practice. The online imitation learning framework has enjoyed a decade of success in various applications. However, the resulting abstract OL problem loses many intrinsic structure of the IL problem, such as non-adversarial nature of the loss function. Further, off-the-shelf OL algorithms used to solved the reduced problem face many issues in practice. One of the most important issues is that of limited expert query. Many methods have been proposed to address these issues, some of which have been discussed in the report. Nonetheless, solving these important issues appears to be open research problems with many ongoing research from different fields, such as robotics and autonomous driving.

References

[1] L. Le Mero, D. Yi, M. Dianati, and A. Mouzakitis, “A survey on imitation learning techniques for end-to-end autonomous vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[2] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun, “Survey of imitation learning for robotic manipulation,” *International Journal of Intelligent Robotics and Applications*, vol. 3, pp. 362–369, 2019.

[3] C. Thurau, C. Bauckhage, and G. Sagerer, “Imitation learning at all levels of game-ai,” in *Proceedings of the international conference on computer games, artificial intelligence, design and education*, vol. 5, 2004.

[4] M. Gasse, D. Chételat, N. Ferroni, L. Charlin, and A. Lodi, “Exact combinatorial optimization with graph convolutional neural networks,” *Advances in neural information processing systems*, vol. 32, 2019.

[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] S. M. Kakade, *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

[7] J. Tsitsiklis and B. Van Roy, “Analysis of temporal-diffference learning with function approximation,” *Advances in neural information processing systems*, vol. 9, 1996.

[8] S. Kakade and J. Langford, “Approximately opti- mal approximate reinforcement learning,” in *in proc. ICML*, 2002.

[9] C.-A. Cheng, X. Yan, E. Theodorou, and B. Boots, “Model-based imitation learning with accelerated convergence,” *arXiv preprint arXiv:1806.04642*, 2018.

[10] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa, “Learning from interventions: Human-robot interaction as both explicit and implicit feedback,” in *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.

[11] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, “Hg-dagger: Interactive imitation learning with human experts,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.

[12] J. W. Lavington, S. Vaswani, and M. Schmidt, “Improved policy optimization for online imitation learning,” in *Conference on Lifelong Learning Agents*. PMLR, 2022, pp. 1146–1173.

[13] J. Zhang and K. Cho, “Query-efficient imitation learning for end-to-end autonomous driving,” *arXiv preprint arXiv:1605.06450*, 2016.

[14] H. He, H. Daume III, and J. M. Eisner, “Learning to search in branch and bound algorithms,” *Advances in neural information processing systems*, vol. 27, 2014.

[15] C.-A. Cheng, X. Yan, E. Theodorou, and B. Boots, “Accelerating imitation learning with predictive models,” in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3187–3196.

[16] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, and B. Boots, “Agile autonomous driving using end-to-end deep imitation learning,” *arXiv preprint arXiv:1709.07174*, 2017.

[17] C.-A. Cheng and B. Boots, “Convergence of value aggregation for imitation learning,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1801–1809.

[18] X. Yan, B. Boots, and C.-A. Cheng, “Explaining fast improvement in online imitation learning,” in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 1874–1884.