Bridging Demonstrations and Decisions: Theory and Algorithms for Provable Imitation Learning

Nived Rajaraman

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-142
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-142.html

July 17, 2025

Bridging Demonstrations and Decisions: Theory and Algorithms for Provable Imitation Learning

By

Nived Rajaraman

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kannan Ramchandran, Co-Chair
Assistant Professor Jiantao Jiao, Co-Chair
Assistant Professor Song Mei

Summer 2025

Bridging Demonstrations and Decisions:
Theory and Algorithms for Provable Imitation Learning

Abstract

Bridging Demonstrations and Decisions:
Theory and Algorithms for Provable Imitation Learning

by

Nived Rajaraman

Doctor of Philosophy in Engineering- Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Kannan Ramchandran, Co-Chair

Assistant Professor Jiantao Jiao, Co-Chair

Classical supervised learning paradigms typically assume that training data samples are independently drawn from a target distribution. However, real-world scenarios frequently violate this assumption, presenting data that are temporally correlated, dynamically evolving, or a result of strategic interaction. Learning in these settings is often significantly more challenging both from a theoretical and a practical point of view for these reasons. Recent advances in reinforcement learning (RL) have shown that it is possible to train agents which can operate and generalize in settings where the number of possible outcomes is huge. However, there are a number of challenges with running RL algorithms: these approaches rely on collecting a large amount of "exploration" data, resulting from interaction with a dynamic environment. This form of active data collection is often prohibitively expensive in practice, making mistakes may be costly, such as in settings involving human interaction, and this form of data collection may be hard to reuse. Mitigating these concerns requires developing new frameworks for RL.

In this dissertation, we develop algorithms and analyses for an alternate learning paradigm that aims to utilize static datasets generated by a demonstrator for training policies. This paradigm broadens the applicability of RL to a variety of decision-making problems where historical datasets already exist or can be collected via domain-specific strategies, and which are infinitely reusable. It also brings the scalability and reliability benefits that modern supervised and unsupervised ML methods enjoy into RL. That said, instantiating this paradigm is challenging as it requires reconciling the static nature of learning from offline datasets (against a fixed distribution of problem instances) with the traditionally interactive nature of RL. A major part of this thesis is geared toward addressing precisely how much of a price one must pay to forgo the power of environment interaction.

Imitation Learning (IL) techniques have found a home in several areas, from policy initialization in game-solving agents like AlphaGo [86], and more recently as a fine-tuning backbone in the form of supervised fine-tuning (SFT) for large language models (LLMs) [16]. The key challenge in all these domains is obtaining sufficiently large, diverse, and high-quality demonstration datasets. While more data typically yields better performance, expert data can be expensive to collect. We see this challenge manifest in several forms: in robotics and control, acquiring teleoperated or human-guided trajectories often requires specialized hardware (e.g. motion-capture rigs or force-feedback devices) and careful calibration, limiting the scale of dataset collection [4, 78]. In autonomous driving, critical "edge-case" scenarios (e.g. collision avoidance in unusual weather) are inherently rare, yet essential for safety; collecting them either in simulation or on-road is time-consuming and costly [22, 23]. On the other hand, for training LLMs: fine-tuning large language models relies on human-annotated data, which is hard to parallelize and incurs substantial annotation time [89].

Thus, it is pertinent to understand how best to utilize the dataset and leverage favorable properties of the environment and demonstrator. In this thesis, we will build an understanding of these questions by studying Imitation Learning from a theoretical point of view. We will formulate a statistical question and analyze the best achievable error that algorithms can achieve in various models of feedback. We will scale up these algorithmic insights and leverage the expressivity and representation power offered by modern function approximators to develop performant, practical algorithms. Along the way, we will develop various insights into the landscape of the IL problem, and build a comprehensive understanding and unification of algorithms that have already been successfully deployed in practice, such as Behavior Cloning [65, 74] and GAIL [44] and provide principled improvements to these approaches.

Overview

This thesis is organized into seven chapters in addition to several appendices, each addressing a different aspect of Imitation Learning under the Markov decision process framework:

Chapter 1. We provide a gentle introduction to IL. After reviewing the standard episodic Markov decision process model (Section 1.1), we define the Imitation Learning problem (Section 1.2) and introduce a statistical framework that will underpin the analysis in the remainder of the thesis (Section 1.3).

Chapter 2. This chapter focuses on Imitation Learning in tabular MDPs. We begin with an introduction to the Behavior Cloning (`BC`) algorithm (Section 2.1), derive worst-case statistical lower bounds on the optimal imitation gap showing optimality of `BC` (Section 2.2), and extend our analysis of this approach to when the expert is stochastic (Section 2.3). We then consider the setting where the transition dynamics are known, and propose a new algorithm known as `Mimic-MD` (Section 2.4).

Chapter 3. We will build up our understanding of the `Mimic-MD` algorithm through the lens of value function estimation. We will introduce and motivate the expert value estimation problem (Section 3.2) and reinterpret Mimic-MD through this lens, which will enable us to prove its statistical optimality (Section 3.3).

Chapter 4. In this chapter we will dive deep into understanding what happens when demonstrations are generated by an agent which carries out the task near optimally. We will revisit some of the hard instances we considered in Section 2.2 in this context (Section 4.1), and analyze the case of mimicking the behavior of the expert in reaching a single target state (Section 4.2), and conclude with some conjectures about the optimal expert setting (Section 4.3).

Chapter 5. Here we will explore Imitation Learning in the setting with active interaction, where the learner may query the expert in an adaptive manner to reduce uncertainty and improve performance. We will establish a provable benefit to interactivity in MDPs where "mistakes" can be recovered from, defined in a formal sense.

Chapter 6. In this chapter, we will build upon the insights we described in Chapters 2 and 3 to develop a practical algorithm. We first provide insights into the suboptimality of methods which are deployed most commonly in practice (Section 6.1), introduce the Replay Estimator (Section 6.2), and then present a full algorithmic recipe (Section 6.3), supported by empirical results (Section 6.4).

Chapter 7. Here, we will extend our study of IL to settings with function approximation, extending beyond tabular MDPs. We will analyze settings with linear representations (Section 7.1), consider known-transition settings (Section 7.2), interpret the main theoretical guarantee (Section 7.3), and discuss open problems (Section 7.4).

Each chapter builds upon the previous ones to present a coherent theory and methodology for effective Imitation Learning across a number of settings and abstractions.

To my family, for their unwavering support and sacrifices.

Contents

List of Figures

List of Tables

Acknowledgments

The journey that culminated in this thesis would not have been possible without the constant encouragement and support of my advisors, mentors, collaborators, family and friends.

First and foremost, I must thank my advisors at Berkeley, Kannan and Jiantao, for the tremendous impact they have had on me, both as a researcher and as a person. I still recall my first interactions with both of them like it was yesterday; perhaps these anecdotes serve better than words to describe my time with them.

Jiantao interviewed me when I applied to the PhD program at Berkeley. After some time discussing my research background and interests, I decided it was a good opportunity to get his opinion about a bold new idea I had been thinking about, one I was convinced nobody could have thought of before. Midway through my pitch, Jiantao paused me, gently pointed out a subtle mistake in what I was describing, and encouraged me to carry on and finish the rest. It soon became clear to me that he had already thought through everything I had said to him. It was the first of many moments when I realized he'd quietly been one step ahead, yet chose to step back and give me the space to express myself. Anyone who has worked with Jiantao knows that he is not only incredibly passionate about his research, but always pushing himself to work on problems that are outside his immediate field of expertise. This is the mark of a great academic - one who questions everything and always strives to improve.

My first interaction with Kannan was during the visit days at Berkeley. I remember talking to a number of previous PhD students about what I might ask a potential advisor, and was prompted to not shy away from the "hard questions". I found these topics a little hard to broach, but mentally prepared myself to do so, as this would be the best opportunity to. But when I approached Kannan, there was a warmth and openness in his demeanor that immediately put me at ease. It wasn't that those questions suddenly became unimportant - they just didn't feel as urgent. I distinctly remember walking away from that conversation not just reassured, but genuinely excited at the thought of learning from someone who seemed so thoughtful and an openness to explore new topics. And over time, I indeed discovered this for myself. He always makes time for his students and is open to hearing new ideas even if they probably are terrible.

I am very grateful to have had Jiantao and Kannan as my advisors. Their mentorship shaped not just my research, but who I am as a person, and I hope to honor that by passing it on in the same spirit.

In July of 2017, I was a confused undergrad figuring out what to do with the rest of my life, and was extremely fortunate to have found several mentors who put me on the path I am currently on: I would like to thank Rahul Vaze, Andrew Thangaraj, and Ravi Krishnaswamy who believed in me at a time when all signs seemed to point the other way.

I am thankful for the community at BLISS and BAIR. Among the faculty, I would like to thank Venkat Anantharam, Gireeja Ranade and Anant Sahai for giving me the terrific opportunity to organize the BLISS seminar at Berkeley. I would also like to extend my

heartfelt thanks to Mike Jordan, Nika Haghtalab and Jacob Steinhardt for the opportunity to work together in the early days of the CLIMB seminar. I am also most thankful to Shirley Salanio and other EECS staff, without whose support I would not be graduating.

This Acknowledgements section would not be complete without a mention of my amazing peers in BLISS. Many fond memories from earlier in my PhD are with members of the BLISS lab: Vipul Gupta, Avishek Ghosh, Koulik Khamaru, Efe Aras, Kwan-Yun Lee, Vignesh Subramanian, Vidya Muthukumar and Ashwin Pananjady. I am equally grateful for the time I have gotten to spend together with the current group: Justin Kang, Mert Cemri, Efe Erginbas, Landon Butler and Syomantak Chaudhuri. Of the many trips, our most recent one to ICLR/AISTATS in Singapore/Thailand is fresh in my memory, and I hope we will get to do that again. While I will miss everyone dearly, I also look forward to never having to work out any more lower bounds for heterogeneous DP again.

It's reasonable to say that almost everything I have learned about research has been through a phenomenal set of co-authors, listed approximately in chronological order below, Rahul Vaze, Devvrit, Ravi Krishnaswamy Prafulla Chandra, Andrew Thangaraj, AT Suresh, Lin Yang, Jiantao Jiao, Kannan Ramchandran, Swanand Kadhe, Yanjun Han, Jingbo Liu, Amirali Aghazadeh, Tony Tu, Gokul Swamy, Matt Peng, Sanjiban Choudhury, Drew Bagnell, Steven Wu, Pranjal Awasthi, Dong Yin, Sridhar Thiagarajan, Nevena Lazic, Botao Hao, Csaba Szepesvari, Aryan Mokhtari, Marco Bondaschi, Ashok Makkuva, Michael Gastpar, Yeshwanth Cherapanamjeri, Sumegha Garg, Ayush Sekhari, Abhishek Shetty, Amrith Setlur, Aviral Kumar, Sergey Levine, Angelos Assos, Yuval Dagan, Mert Cemri, Rishabh Tiwari, Xiaoxuan Liu, Kurt Keutzer, Ion Stoica, Ahmad Beirami and Ziteng Sun. Among this list of amazing people, I would like to particularly call out two of my peers. Swanand Kadhe, who was an incredible mentor for me early in grad school and someone who I learned a lot from, especially the attitude of never giving up. And Mert Cemri, who I learned a lot from about how to navigate research in a new area and the skill of seeking out advice and collaborators.

I am really fortunate to be able to have had an incredible support group throughout my PhD in the form of friends. I will need a second (longer) thesis to be able to do justice to what they all mean to me.

The original Cremposters. To Amit Rajaraman for all the life lessons taught. I've learned firsthand from him that the pun can indeed be mightier than the sword. I'm glad to have synced sines with you, buddy. Thanks for all the music recs and the memes.
And to Gunduboss Sidhanth Mohanty, for being the voice of reason in a sea of treason. I am glad we have those messages saved from Gchat 15 years ago.

I am incredibly grateful to Suraj and Kashish for their parenting, as well as the rest of the meditation community: Swiss Martin for being the enabler of everything eclectic, French Martin for his kindness and chocolate cake, and Vinamra for the long walks. I will miss the entire Bonita gang: Eric and Ajil for bear hugs and Balatro. Sahil and Sunash for religiously going to the gym in place of me. And Shailee and Tanuja for all the frolicking in SF and the dietary fiber. My dear friends from south bay summer: Sunny for introducing me to Grocery

Outlet and for hiking together to the tunes of Howard Shore. Lisa for all the killer banter, for maltesers and also for successfully watching Coco 100% from start to finish. Saurabh for tabla, and for never activating beast mode only unless necessary. Maya and Hari for all the birthday celebrations, best memories and for being people who really care. And Ananya for being the initiator of our SF hangouts and introducing me to Baklavastory.

I am very fortunate for my housemates over the years, and to 1641 Walnut for being the hub for all associated frivolous activities: I am especially thankful for Sid, Fred and Pranav for expanding my horizons, culinary or otherwise. Later came our Summer 2023 Berkeley group, which turned into Dirham Store 1641 Walnut: Amit, Sidhanth, Tim, Bingbin for all the spontaneous decisions, Bean, LOTR, and most importantly for featuring in hundreds and hundreds of clueless pictures, and Yeshwanth for all the world knowledge, wisdom and pineapple fried rice. Other friends in Boston: Rachel for all the amazing jam sessions at MIT and at Simons, Stefan for a four hour rundown of One-Piece and top-notch german banter, and Karna for all the music and all the musings. Friends at the Theory Group at Berkeley: Ansh and Sriram for all cooking and for all the eating (but only after visiting rank 8), Omar for the coffee and hot takes, David for his ambient prog house playlist, Nathan for all the chicken noises and lavatorial salutations, Gurujee Ishaq Aden-Ali for taking me under his wing at the bakwas factory, and honorary member Shivam for all the maximum entropy conversations.

Much of my free time in grad school was spent with friends from college: Bowreddy, for all the weekend trips and for all the kids at home, Gaurav and Trivi for that one Yosemite trip and being ever ready to go with the flow, Hegde and Chaithanya for the OG real July trip (2021), and the rest of the gang back in India: Dhandoo, Dasu, Thakur, Gayu and Avhad. I would also like to thank the Austin gang: Aditi, Devvrit and Nilesh for their amazing hospitality, and all the long walks and jam sessions, in spite of a rather suspicious absence of Brussels sprouts.

I am very fortunate for my extended family in the bay area: Chitra aunty, Kannan uncle, Mekala, Taruna and Kavitha aunty, and the time I got to spend with them the past few years. I am grateful for their kindness and support throughout grad school, and all the nice memories from our time hunkering together through COVID. I will forever cherish my time with Krish. He was a good boy.

Finally, to my family back in India: my mother, Vidya, for her never-ending love and for showing me how to care. My father, Venkat, for laying a safety net beneath every decision I've ever made, large or small. To my grandmother, Santha, for the star we share. And my brother, for keeping his book open for me to take a leaf out of, and occasionally scribble in.

Chapter 1

A gentle introduction to Imitation Learning (IL)

Imitation learning or apprenticeship learning is the study of learning from demonstrations in a sequential decision-making framework in the absence of reward feedback. The Imitation Learning problem differs from the typical setting of reinforcement learning in that the learner no longer has access to reward feedback to learn a good policy. In contrast, the learner is given access to expert demonstrations, with the objective of learning a policy that performs comparably to the expert's with respect to the *unobserved reward function*. This is motivated by the fact that the desired behavior in typical reinforcement learning problems is easy to specify in words, but hard to capture accurately through manually-designed rewards. For instance, in autonomous driving it is easy to state the reward function in English as, "drive safely", but it is challenging to write down a specific reward function capturing this target mathematically [2]. This is especially important in safety-critical applications, where misspecified rewards can lead to unpredictable behavior.

In practice, the reward function is sometimes manually refined [59, 10] until the learner demonstrates satisfactory behavior. In contrast, the Imitation Learning (IL) approach posits to learn directly from expert demonstrations in the absence of *reward feedback*, and without learning an explicit reward model. Imitation learning has shown remarkable success in practice over the last decade - the work of [3] showed that using pilot demonstrations to learn the dynamics and infer rewards can significantly improve performance in autonomous helicopter flight. More recently, the approach of learning from demonstrations has shown to improve the state-of-the-art in numerous areas: autonomous driving [43, 63], robot control [5], game AI [46, 38] and motion capture [56] among others. To build up a framework to understand, compare and build upon these methods, we will first begin by introducing an abstraction for the decision making problems these approaches are designed to solve.

Notation. Throughout this thesis, we will use big O notation, *i.e.*, $\mathcal{O}/\Theta/\Omega$ and their upto-polylogarithmic-factor counterparts $\widetilde{\mathcal{O}}/\widetilde{\Theta}/\widetilde{\Omega}$. In particular, $f(n) = \widetilde{\mathcal{O}}(g(n))$ when

$a = \mathcal{O}(f(n) \cdot \max(1, \mathsf{polylog}(n)))$ (with $\widetilde{\Theta}$ and $\widetilde{\Omega}$ defined analogously). We will also use the notation $f(n) \lesssim g(n)$ when $f(n) = \mathcal{O}(g(n))$, $f(n) \gtrsim g(n)$ when $f(n) = \Omega(g(n))$ and $f(n) \asymp g(n)$ when $f(n) = \Theta(g(n))$. The set of integers $\{1, \dots, n\}$ is denoted as $[n]$. For a (potentially uncountably infinite) set S, the set of all measures over S is given by $\Delta(S)$.

1.1 Markov Decision Processes

In this thesis, we will introduce IL through the framework of Markov Decision Processes (MDPs). The MDP abstraction has been widely adopted in the RL literature for its simplicity and generality. At a high level, an MDP is defined by four components: a set of states that describe every possible configuration of the environment; a set of actions available to the learner; a transition mechanism that specifies, for each state and action, how the environment evolves to a next state; and a reward signal that quantifies the immediate merit of each state–action pair. The cumulative reward collected by an agent is a measure of its performance. We will introduce the notion of non-stationary episodic MDPs below, and introduce more specific/general settings later on in the thesis.

Definition 1.1.1 (Episodic Markov Decision Process)**.** *An* episodic Markov Decision Process $\mathcal{M}$ *is a tuple,*

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho, H)\,,$$

where,

- $\mathcal{S}$ *is the state space,*
- $\mathcal{A}$ *is the action space,*
- *for each* $t = 1, \dots, H$*,*

$$\begin{aligned} P = \{P_t\}_{t=1}^{H}, &\quad \text{where, } P_t \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) \quad \text{and,} \\ r = \{r_t\}_{t=1}^{H}, &\quad \text{where, } r_t \colon \mathcal{S} \times \mathcal{A} \to [0,1], \end{aligned}$$

  are the transition kernel and reward function at time t*,*

- ρ *is an initial distribution over states,*
- $H \in \mathbb{N}$ *is the (finite) horizon.*

Starting from an initial state $s_1 \in \mathcal{S}$ *drawn from* ρ*, an episode proceeds for* $t = 1, 2, \dots, H$ *as follows:*

$$a_t \in \mathcal{A}, \quad r_t = r_t(s_t, a_t), \quad s_{t+1} \sim P_t\left(\cdot \mid s_t, a_t\right). \tag{1.1}$$

The process terminates after H *steps.*

Interaction with the MDP, i.e., the environment, is carried out in episodes. The learner iteratively chooses actions, receives rewards (which lie in $[0, 1]$) and transitions to new states. The objective we will study throughout this thesis is that of maximizing the "expected cumulative reward". In order to be able to define this quantity, we will first have to define the decision rule by means of which a learner acts in the environment. This is known as a policy: a (possibly stochastic) map which specifies how a learner plays actions at states.

Definition 1.1.2 (Policy)**.** *A* policy *in an episodic MDP is a sequence of decision rules*

$$\pi = \{\pi_t\}_{t=1}^{H}, \qquad \pi_t \colon \mathcal{S} \to \Delta(\mathcal{A}), \tag{1.2}$$

where $\pi_t(a \mid s)$ is the probability of choosing action $a \in \mathcal{A}$ when in state $s \in \mathcal{S}$ at time t.
A policy π is said to be deterministic if $\pi_t(\cdot|s)$ is a delta function for all $(s, t) \in \mathcal{S} \times [H]$.

It may be apparent from the definition above, but our focus will be on Markovian policies, where the decision rule does not depend on the entire history of interaction, but only on the current (observable) state which the learner is at. We will introduce the following notation to write down expectations with respect to random trajectories drawn from a policy π more succinctly. Define,

$$\mathbb{E}_{\pi}[\cdot] \equiv \mathbb{E}_{\substack{S_1 \sim \rho, \\ \forall t \in [H],\ A_t \sim \pi_t(\cdot|S_t) \\ S_{t+1} \sim P_t(\cdot|S_t, A_t)}}[\cdot]. \tag{1.3}$$

Namely, this is an expectation computed with respect to the distribution over trajectories $\{(S_1, A_1), \cdots, (S_H, A_H)\}$ induced by rolling out the policy π. With this, we are ready to introduce the learning objective.

Definition 1.1.3 (Reward Maximization Objective)**.** *The reward maximization objective is to find a policy $\pi = \{\pi_t\}_{t=1}^{H}$ that maximizes the expected cumulative reward over the horizon H,*

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} r_t(s_t, a_t)\right]. \tag{1.4}$$

$J(\pi)$ is also known as the "value" of the policy π. Where necessary, we will denote the value of a policy by $J_{\mathcal{M}}(\pi)$ or $J_r(\pi)$ to indicate the underlying MDP $\mathcal{M}$ or reward function r. Furthermore, define an arbitrary optimizer of eq. (1.4) as,

$$\pi^{\star} \in \arg\max_{\pi}\ J(\pi). \tag{1.5}$$

1.2 Imitation Learning

Imitation Learning (IL) can be viewed as a variant of the typical Reinforcement Learning formulation where the interaction model is assumed to be slightly different. Recall the

episodic MDP formulation in Definition 1.1.1, where in a single interaction, upon playing an action at a state, the learner observes a (potentially noisy) reward for picking this action and transitions to a new state. In many practical domains, however, the true reward function is either unavailable, difficult to specify by hand, or too expensive to evaluate online. For example, in autonomous driving it is hard to write down a scalar $r_t(s_t, a_t)$ that simultaneously captures safety, comfort, and legality; in robotic manipulation the "correct" reward may depend on subtle human preferences that are hard to encode; and in healthcare applications, any trial-and-error to recover r_t directly can be dangerous or unethical. In such settings we often instead have access to one or more expert *demonstrations*, i.e., trajectories collected by a human or a pre-trained controller, without explicit reward labels. In this thesis, we will model these demonstrations as being generated by a single unknown policy.

Definition 1.2.1 (Imitation Learning)**.** *In this setting, we will assume the existence of an "expert" policy π^{exp}. Furthermore, when interacting with the MDP, the learner no longer receives any reward feedback from the MDP. Namely, at the state s_t, playing the action a_t only transitions the learner to $s_{t+1} \sim P_t(\cdot \mid s_t, a_t)$ and the learner does not observe $r_t(s_t, a_t)$.*

While the Imitation Learning setting specifies the existence of an expert which generates demonstrations, we will introduce specific models of interaction with the MDP and with π^{exp} later on (notably in Definition 1.2.3, Definition 1.2.4 and Definition 1.2.5).
In this setting, the first question that comes to mind is: what is the best possible value achievable by a learner? Indeed, in the case where the expert policy π^{exp} is fully observed, say, via having infinitely many demonstrations, it is possible to achieve a value of at least $J_r(\pi^{\text{exp}})$ even with no knowledge of the ground truth reward function r. And with a bit more thought, it should become clear that the fact that the fact that the reward function is completely unobserved is a barrier to improving beyond this. It is an exercise left to the reader to show that it is possible to construe a reward function r such that $J_r(\pi^{\text{exp}}) - J_r(\pi)$ is always non-positive.

Theorem 1.2.1. *A class of reward functions $\mathcal{R} \subseteq [0,1]^{\mathcal{S}\times\mathcal{A}\times[H]}$ is said to be symmetric if $r \in \mathcal{R} \iff 1-r \in \mathcal{R}$, where $r' = 1-r \implies r'_t(s,a) = 1-r_t(s,a)$ for all $(s,a,t) \in \mathcal{S}\times\mathcal{A}\times[H]$. For any learner policy π, expert policy π^{exp}, and symmetric reward class $\mathcal{R}$ and, it is possible to find a reward function $r \in \mathcal{R}$ such that $J_r(\pi^{\text{exp}}) - J_r(\pi) \leq 0$.*

A hint toward proving this statement is the observation that $J_r(\pi^{\text{exp}}) - J_r(\pi) = J_{r'}(\pi) - J_{r'}(\pi^{\text{exp}})$ for $r' = 1 - r \in \mathcal{R}$.

It is worth mentioning that while the gold standard for the best achievable value is ideally $J_r(\pi^\star)$, the value of the optimal policy, this is not achievable in the absence of any reward feedback. Thus, with imitation feedback, the learner is forced to readjust its target to be the value of the expert policy $J_r(\pi^{\text{exp}})$. This fact leads us to formulate the reward maximization objective in terms of the "imitation gap", which is the gap in the value achieved by the expert policy and the learner's policy.

Definition 1.2.2. *The imitation gap of the learner's policy π is defined as,*

$$\mathsf{Gap}(\pi) = J(\pi^{\text{exp}}) - J(\widehat{\pi}).$$

While it is apparent that the imitation gap can be made zero when π^{exp} is specified exactly to the learner, in practice the learner only observes the expert through a finite set of demonstrations. This hints at the key statistical question which will be studied through many parts of this thesis - how do we minimize the imitation gap as best as possible when the learner only has a few demonstrations from the expert policy. To begin to formulate this question, we first need to specify how the learner interacts with the expert policy (i.e., how demonstrations are collected) and the MDP conjunctively. There are several interesting models to consider, inspired by practical considerations, and we spend the next few subsections discussing them.

Models of expert interaction

In this section, we introduce several models we consider in how the learner can interact with the expert within the Imitation Learning framework. The simplest model is discussed first.

Definition 1.2.3 (No-interaction IL)**.** *The learner is provided a dataset of N trajectories drawn by independently rolling out the expert policy π^{exp} through the MDP. Each trajectory is of the form $\{(s_t, a_t)\}_{t=1}^{H}$, the states visited and actions played by the expert across time. In the no-interaction setting, the learner is not allowed to roll out policies on the MDP.*

The no-interaction setting models the case where demonstrations of an expert policy in the environment are available, but it is otherwise too challenging, or cost-ineffective for the learner to roll out other policies and interact with the environment. This model is also of course the most restrictive for the same reason. However, it is an interesting model, since a number of practical algorithms, such as Behavior Cloning (BC) [65], fall within this framework.

While the no-interaction setting posits that the learner cannot explore within the environment, may also consider another extreme where the learner is able to interact with the environment infinitely. In the limit, the learner is thereby able to learn the transition structure of the MDP exactly; however it is worth noting that the reward function is still completely unknown, cf. Definition 1.2.1. This is referred to as the known-transition setting, and is defined below.

Definition 1.2.4 (Known-transition IL)**.** *The learner is provided a dataset of N trajectories drawn by independently rolling out the expert policy π^{exp} through the MDP. In the known-transition setting, the learner is assumed to know the MDP transition $\{P_t\}_{t=1}^{H}$ exactly.*

The no-interaction and known-transition IL settings are two extremes of a spectrum where the learner is provided a dataset of N trajectories from the expert policies, and also allowed to interact with the MDP (using any sequence of policies of its choice) for M episodes. These settings respectively correspond to the case of $M = 0$ and $M \to \infty$. Understanding the statistical learning problem for the general case M has been the subject of follow up works to some portions of this thesis, such as [104]. We will avoid discussing these results or introducing

this setting here, and instead introduce a third interaction model which corresponds to a setting where the expert can actively be queried while the learner interacts with the MDP. This "active-interaction" setting models settings like teleoperation, where the expert can intervene to correct the behavior of a policy learning in the environment, in the process minimizing redundant interaction with the expert.

Definition 1.2.5 (Active-interaction IL)**.** *In this setting, the learner is not given access to a dataset of expert demonstrations in advance. However, the learner is allowed to interact with the MDP for N episodes while actively querying the expert policy; at any time t, the learner may query the expert policy π^{exp} to receive an action recommendation $a_t^\star \sim \pi_t^{\text{exp}}(\cdot \mid s_t)$ at the visited state s_t. The learner may choose to follow this recommendation, or play any other action.*

It is worth pointing out that the active setting gives the learner more flexibility than the no-interaction setting: following the expert's recommended action at each time results in the same interaction model as as the no-interaction setting (Definition 1.2.3). However, the learner is instead also allowed to actively design a new exploratory policy on the fly based on expert feedback thus far, which is then used to interact with the MDP. Two popular approaches, DAGGER [79] and AGGRAVATE [77] are IL algorithms which operate in the active-interaction setting.

1.3 A statistical framework for IL

Having established the foundations of IL and modes of interaction with the expert policy, we are ready to formulate the statistical learning problem corresponding to IL below. We use the Probably-Approximately Correct (PAC) learning framework [99].

Definition 1.3.1 (Demonstration dataset and learning rules)**.** *Let D be the random dataset obtained as either (i) N rollouts of π^{exp} in the no-interaction or known-transition IL settings (resp. Definitions 1.2.3 and 1.2.4), or (ii) that collected by the learner's interactions with the MDP in the active-interaction IL setting (Definition 1.2.5). D is referred to as the "demonstration dataset".*

A "learning rule" $\mathsf{Alg}(\cdot)$ is any algorithm which ingests D (and additionally $P = \{P_t\}_{t=1}^H$ and ρ in the known-transition setting) and returns a possibly randomized learner policy $\widehat{\pi}$.

Definition 1.3.2 (PAC guarantees for IL)**.** *Given a collection of IL instances $\mathbb{Z}$, we say that a learning rule $\mathsf{Alg}(\cdot)$ solves the Imitation Learning problem (in the no-interaction / known-transition / active-interaction setting) with confidence $1-\delta$ and imitation gap ε over a class of IL instances $\mathbb{Z}$, if for any instance $\mathcal{Z} \in \mathbb{Z}$, we have*

$$\Pr\left(\mathsf{Gap}(\widehat{\pi}) \geq \varepsilon\right) \leq \delta, \tag{1.6}$$

Where, recall that $\mathsf{Gap}(\widehat{\pi}) = J(\pi^{\exp}) - J(\widehat{\pi})$ *is computed under the reward function and distribution over trajectories induced by rolling out* $\pi^{\exp}$ *and* $\widehat{\pi}$ *on the underlying MDP in the instance* $\mathcal{Z}$*, and the probability* $\Pr(\cdot)$ *is computed over the randomness of the demonstration dataset* D *as well as the internal randomness used by* $\mathsf{Alg}(\cdot)$*.*

We will also define the IL problem with expected error guarantees.

Definition 1.3.3 (Expected error guarantees for IL)**.** *A learning rule* $\mathsf{Alg}(\cdot)$ *is said to solve the Imitation Learning problem (in the no-interaction / known-transition / active-interaction setting) with expected imitation gap* ε *over a class of IL instances* $\mathbb{Z}$*, if for any such instance* $\mathcal{Z} \in \mathbb{Z}$*, we have*

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \leq \varepsilon. \tag{1.7}$$

Note once again, that $\mathsf{Gap}(\widehat{\pi}) = J(\pi^{\exp}) - J(\widehat{\pi})$ *is computed under the reward function and distribution over trajectories induced by rolling out* $\pi^{\exp}$ *and* $\widehat{\pi}$ *on the dynamics in the instance* $\mathcal{Z}$*, and the expectation* $\mathbb{E}[\cdot]$ *is computed over the randomness of the demonstration dataset* D *as well as the internal randomness used by* $\mathsf{Alg}(\cdot)$*.*

While for a large collection of statistical learning problems, PAC error guarantees and expected error guarantees can be translated from one to another at the cost of additional polynomial (and often logarithmic) factors in the inverse of the confidence parameter, δ, via integrating the PAC error over δ ($\Longrightarrow$) or via an application of Markov's inequality to the expected error ($\Longleftarrow$). While these generic reductions often result in loose dependency on the confidence parameter δ, and can be improved by tailored analysis to the respective setting, there is a more important reason because of which we define and treat IL with PAC error guarantees and expected error guarantees separately. This observation is special to the IL learning objective, and one often not shared by other statistical learning settings, such as classification under 0-1 loss or regression settings: the imitation gap $\mathsf{Gap}(\pi)$ is not always a non-negative loss when the underlying MDP is fixed. This subtle observation will have interesting implications later on (cf. Chapter 4).

Across both of the statistical learning settings considered previously, Definition 1.3.2 and Definition 1.3.3, our goal will be to study how small ε can be made as a function of N, the number of episodes of demonstrations from the expert / episodes of active interaction. Of course, across the three interaction models, the learner's has access to different modes of feedback and thereby, the best achievable imitation gap as a function of N is expected to behave quite differently. Below we formalize this definition.

Definition 1.3.4 (Worst-case optimal imitation gap)**.** *Given a collection of IL instances* $\mathbb{Z}$*, the minimax optimal imitation gap (in the no-interaction / known-transition / active-interaction setting) is defined as,*

$$\varepsilon^{\star}(N; \mathbb{Z}) = \inf \left\{ \varepsilon \geq 0 : \exists \mathsf{Alg}(\cdot) \text{ such that } \forall \mathcal{Z} \in \mathbb{Z},\ \mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \leq \varepsilon \right\}.$$

Similarly, the PAC optimal imitation gap is defined as,

$$\varepsilon^\star(N, \delta; \mathbb{Z}) = \inf \{\varepsilon \geq 0 : \exists \mathsf{Alg}(\cdot) \text{ such that } \forall \mathcal{Z} \in \mathbb{Z},\ \Pr(\mathsf{Gap}(\widehat{\pi}) \geq \varepsilon) \leq \delta\}.$$

In both definitions, $\mathsf{Alg}(\cdot)$ is a learning rule which ingests the demonstration dataset D (and additionally $P = \{P_t\}_{t=1}^H$ in the known-transition setting) and returns a possibly randomized learner policy $\widehat{\pi}$.

Remark 1.3.1. Note that for a class of IL instances $\mathbb{Z}$, both $\varepsilon^\star(N; \mathbb{Z})$ and $\varepsilon^\star(N, \delta; \mathbb{Z})$ are functions of N, which is the number of episodes of expert demonstrations / active interactions in the dataset D.

With this, we round out the discussion introducing IL within the MDP formulation, the learning objective, different modes of receiving feedback from an expert, and finally, a statistical framework for this problem. In the next section, we discuss the different modes of interaction with the expert in greater detail and present some results for the corresponding statistical learning problems.

1.4 Related Work

Imitation learning (IL) has a rich history, encompassing a multitude of different algorithmic approaches and observation models. This section broadly discusses prior work in the area, spanning methods which fall into the bucket of classical "inverse-reinforcement" style approaches, reduction-based analyses, and active-query methods.

Classical IL via Inverse Reinforcement and Reduction. Early work framed IL as inverse reinforcement learning, where the learner recovers a reward function that explains expert behavior [60, 2, 72, 95, 110, 102, 41, 35, 73], which is then utilized for policy learning. These algorithms assume the principle that a reward function (learned from observations) may generalize better across across environments. A comprehensive survey of these methods is discussed in [7]. In the training of large language models (LLMs), several recent works have also used reward models learned from offline preference data, also known as verifier models [45]: the LLM is perceived as a policy and optimized by carrying out RL on the learned rewards. Several works also carry out search at test-time [101, 19] against these learn reward models. Several works also learn process reward models which give denser feedback along the generation trajectory, akin to value functions [66, 107, 87, 83].

Reduction-based analyses of IL The reduction paradigm attempts to cast IL as a sequence of supervised learning problems under different notions of loss, and arguing about how the generalization ability of solving the underlying supervised learning task translates to the performance of the resulting policy [30, 9]. In particular, the authors of [78, 77] introduced algorithms (DAGGER, AGGREVATE) and corresponding analyses to bound compounding

errors by iteratively mixing expert and learner data. The authors of [14] analyzed disagreement-regularized reductions for tighter guarantees. However, this line of literature does not focus on whether the obtained guarantees are (statistically) tight, since any particular reduction may be sufficient, but not necessary to solve the IL problem.

Imitation via active expert queries While classical IL algorithms learn from a purely offline dataset, in settings like robotics or game solving, it is possible to learn from an expert via teleoperation or human intervention. In these settings, the learner largely operates autonomously within the environment, but a supervising demonstrator may choose to intervene to course-correct the learner. In the process, the expert is only utilized sparsely, and the learner may be "taught" how to learn to recover when it strays from the desired behavior path. In this vein, several algorithms were proposed which actively query an expert during training. DAgger gathers corrective labels on states encountered by the learner [78], AggreVaTe generalizes this to cost-to-go queries [77], and AggreVaTeD applies the same idea in a stochastic policy gradient framework [91]. Complementary dynamic-regret analyses were developed in [20, 51], while Loki bootstraps IL into on-policy gradient methods [21].

Divergence minimization and Moment Matching based approaches. Beyond pure behavior cloning, value-based approaches self-correct by estimating cost-to-go functions [52] and provably reduce compounding error. A parallel thread of work formulates the algorithmic problem of IL as that of minimizing an f-divergence between the state–action distributions of the expert and the learner [44, 47, 108, 88, 49, 94], also known as moment matching. These algorithms are often based on solving minmax optimization problems (such as via Generative-Adversarial Networks [37]), where the minimizer aims to find a good policy which matches the visitation distributions to that in the training dataset, while the maximizer tries to learn a discriminator which distinguishes between trajectories generated by the learner and the expert. There are also off-policy extensions of this class of approaches, such as ValueDICE and SoftDICE [49, 90].

Function Approximation and Representation Learning. Most tabular-MDP guarantees do not immediately extend to large-scale settings. Apprenticeship IL with linear function approximation was pioneered in [1, 96]. More recently there have been a number of different approaches across different settings, enabling provable sample-efficient IL in high-dimensional spaces such as in [6, 103]. The recent work of [34] is also notable, extending analyses of algorithms for IL to settings where the expert belongs to general function classes. Since this body of work will be closest related to the contents of this thesis, more relevant work will be introduced and discussed in later chapters.

Alternate observation models. Often in IL settings, the actions made by an expert are not explicitly observable, or do not translate to the actions taken by a learner operating in the environment. A good example is that of training autonomous driving agents through video:

here the task is to learn how to carry out the task when given access to video demonstrations of it being carried out, which may have been collected by rolling out a different agent, or even by a human. When only expert state trajectories are available, one must infer latent actions [58, 98, 92, 6], which is often significantly more challenging. Although promising, there are some fundamental questions about statistical identifiability in these settings.

Applications. There is an extremely rich body of work integrating IL approaches into the end-to-end RL pipeline. Early works have used it to design initialization policies from offline data, such as in game solving, with AlphaGo [86] and StarCraft [38], among other games [46]. These techniques have also been used widely in settings where the reward is hard to specify such as in autonomous driving [62, 43] and motion capture [56]. IL has also found wide application in robotics [3, 50, 109, 33, 32] and in the pre-training and fine-tuning of LLMs, such as in the form of supervised finetuning (SFT) [16]. LLMs for reasoning tasks are often finetuned on search traces [36, 61] or responses from larger models, [82] A good survey of early works in the field is [4], and in the context of robotics [18].

Chapter 2

Imitation Learning in Tabular MDPs

In this chapter, we initiate the study of IL, focusing on the setting of tabular MDPs, where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are assumed to be finite. This is arguably the simplest setting of IL, where there are no additional structures imposed on the instances under consideration, or the nature of the expert policy.

From a technical point of view, we aim to characterize the optimal imitation gap $\varepsilon^\star(N, \delta; \mathbb{Z})$ where $\mathbb{Z}$ is the set of IL instances induced by tabular MDPs. To build some intuition, we will first consider the case where the expert policy $\pi^{\exp}$ is deterministic (Definition 1.1.2). Let's introduce some notation first.

Notation. Since we will first study the case where the expert policy is deterministic, define $\Pi_{\mathtt{det}}$ as the set of all deterministic policies, and $\mathbb{Z}_{\text{tabular}}(\mathcal{S}, \mathcal{A}, \mathcal{H})$ is the set of all tabular (i.e., unconstrained) MDPs over the state space $\mathcal{S}$, action space $\mathcal{A}$ and horizon H. We will abbreviate $\mathbb{Z}_{\text{tabular}}(\mathcal{S}, \mathcal{A}, \mathcal{H})$ simply as $\mathbb{Z}_{\text{tabular}}$ where its arguments are clear from context.

Let us take a step back and think about the no-interaction setting. Recall that in the no-interaction setting, the learner is provided a dataset of N demonstrations obtained by rolling out the expert policy $\pi^{\exp}$. In the tabular setting, this dataset is easy to interpret. At time t, at any state s which is visited in D (i.e., in some trajectory, the state at time t is s), then the demonstration dataset exactly reveals the expert policy at this state, namely $\pi_t^{\exp}(\cdot|s)$. On the other hand, at any state which was never visited in D at time t, the learner has no information about which action was played by $\pi^{\exp}$. Thus, a simple learning rule is one which simply mimics the expert policy at those states visited in the demonstration dataset at any time, and plays arbitrarily at the remaining states. It turns out that this simple algorithm is surprisingly powerful, and is a special case of an approach known as Behavior Cloning (`BC`) [65].

2.1 Behavior Cloning

Following the approach pioneered by [11], the authors of [74] show that carrying out supervised learning to learn a policy provides black box guarantees on the performance of the learner, in effect "reducing" the IL problem to supervised learning. This approach is known as behavior cloning.

A more general way of viewing the above approach of mimicking the deterministic expert at those states visited in the dataset, is that of training a classifier. Formally, a deterministic policy π can be treated as a classifier from $\mathcal{S} \times [H] \to \mathcal{A}$. Define the population 0-1 risk of a deterministic policy as the risk under 0-1 loss of the corresponding classifier,

$$\mathcal{L}_{0\text{-}1}(\pi) = \mathbb{E}_{t\sim\text{Unif}([H])}\left[\mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{E}_{A'_t\sim\pi_t(\cdot|S_t)}\left[\mathbb{1}(A'_t \neq A_t)\right]\right]\right]. \tag{2.1}$$

We may also define the empirical 0-1 loss computed on a dataset of demonstrations D by,

$$\mathcal{L}_{0\text{-}1}(\pi; D) = \mathbb{E}_{t\sim\text{Unif}([H])}\left[\mathbb{E}_{S_t\sim f^t_D}\left[\mathbb{E}_{A'_t\sim\pi_t(\cdot|S_t)}\left[\mathbb{1}(A'_t \neq \pi^{\text{exp}}_t(S_t))\right]\right]\right]. \tag{2.2}$$

Here f^t_D is the empirical distribution over states at time t averaged across trajectories in D and $\pi^{\text{exp}}_t(S_t)$ overloads notation to indicate the action played by the deterministic expert at the state S_t at time t. Note that any policy which minimizes the empirical 0-1 risk to 0 in fact mimics the expert at all states observed in D. We define $\Pi^{\texttt{BC}}_{\text{det}}(D)$ as the set of all candidate deterministic policies that minimize the empirical 0-1 risk to zero, namely the set of ERMs.

$$\Pi^{\texttt{BC}}_{\text{det}}(D) \triangleq \left\{\pi \in \Pi_{\texttt{det}} : \mathcal{L}_{0\text{-}1}(\pi; D) = 0\right\} \tag{2.3}$$

Definition 2.1.1 (Tabular BC with a deterministic expert)**.** *In the tabular setting under a deterministic expert policy, BC returns an arbitrary* $\widehat{\pi}^{BC} \in \Pi^{BC}_{\text{det}}(D)$.

Behavior Cloning is an algorithm which has been very well studied in the literature. Indeed, since it is quite close to classification, it is natural to ask whether the imitation gap of the policy learned by BC can be bounded in terms of the performance of the corresponding classifier. The authors of [74] answer this question in the affirmative, and show that if the expert policy is deterministic, any policy $\widehat{\pi}$ that incurs small population 0-1 loss in the sense of eq. (2.1) also incurs small imitation gap.

Theorem 2.1.1 (Theorem 2.1 in [74])**.** *Consider any policy* π *such that* $\mathcal{L}_{0\text{-}1}(\pi) \leq \epsilon$. *Then* $\mathsf{Gap}(\pi) \leq \min\{H, H^2\epsilon\}$.

It is apparent that when $\mathcal{L}_{0\text{-}1}(\pi) = 0$, this must mean that $\pi_t(\cdot) = \pi^{\text{exp}}_t(\cdot)$ almost everywhere. By extension, this implies that $\mathsf{Gap}(\pi) = 0$. On the other hand, the H^2 factor can intuitively be understood as the inability of the learner to get back on track once a mistake is made. This is known as *error compounding*.

Error compounding

To get a better understanding of where this quadratic H factor appears, consider a simplified model where the probability of error of a learner's policy in guessing the deterministic expert's action at each state is ϵ. Indeed, the probability of making the first error at time t is $\epsilon(1-\epsilon)^{t-1}$, and if the learner gets completely lost thereafter the learner fails to collect up to $H-t+1$ units of reward. The imitation gap can therefore be bounded as $H\epsilon + (H-1)\epsilon(1-\epsilon) + \cdots + \epsilon(1-\epsilon)^{H-1} \asymp \min\{H, H^2\epsilon\}$. This dependency is referred to as error compounding, and for a more rigorous analysis we refer the reader to [74]. We also provide a slight generalization of this result in Section 2.3. At a higher level, the issue of error compounding can be interpreted as resulting from a "covariate shift problem": the actual performance of learner depends on the states it encounters under its own rollout distribution; whereas these offline training methods match distributions with respect to the expert's state distribution.

While it remains unclear whether this reduction is optimal, it provides a simple mechanism by which to analyze the performance of BC and understand how the classification error decays as a function of N, the number of demonstrations the learner has access to. In the tabular setting, we next establish a generalization bound for the population 0-1 risk.

Lemma 2.1.2 (Population 0-1 risk of BC)**.** *Consider the no-interaction setting, and assume the expert's policy $\pi^{\exp}$ is deterministic. Consider any policy $\widehat{\pi}^{BC} \in \Pi_{\det}^{BC}(D)$ (defined in eq. (2.3)) which is the set of policies that carry out BC. Then, the expected population* 0-1 *risk of $\widehat{\pi}^{BC}$ (defined in eq. (2.1)) is bounded by,*

$$\mathbb{E}\left[\mathcal{L}_{0\text{-}1}(\widehat{\pi}^{BC})\right] \lesssim \min\left\{1, \frac{|\mathcal{S}|}{N}\right\}.$$

Proof Sketch. The bound on the population 0-1 risk of BC relies on the following observation: at each time t, the learner exactly mimics the expert on the states that were visited in the demonstration dataset at least once. Therefore the contribution to the population 0-1 risk only stems from states that were never visited at time t in any trajectory in D. We identify that for each t, the probability mass contributed by such states has expected value upper bounded by $|\mathcal{S}|/N$. Plugging this back into the definition of the population 0-1 risk completes the proof. □

With this result, invoking [74, Theorem 2.1] immediately results in the upper bound on the expected imitation gap of a learner carrying out BC in Eq. (2.1.3.1). Furthermore, we use a similar approach to establish a high probability bound on the population 0-1 risk of BC.

Theorem 2.1.3 (Upper bounding imitation gap of BC)**.** *Consider any policy $\widehat{\pi}^{BC} \in \Pi_{\det}^{BC}(D)$. Then,*

1. *The expected imitation gap of $\widehat{\pi}^{BC}$ is upper bounded by,*

$$\mathbb{E}\left[\mathsf{Gap}(\widehat{\pi}^{BC})\right] \lesssim \min\left\{H, \frac{|\mathcal{S}|H^2}{N}\right\}. \tag{2.1.3.1}$$

2. *For any $\delta \in (0, \min\{1, H/10\}]$, with probability at least $1-\delta$, the imitation gap of $\widehat{\pi}$ is bounded by,*

$$\mathsf{Gap}(\widehat{\pi}^{\mathtt{BC}}) \lesssim \frac{|\mathcal{S}|H^2}{N} + \frac{\sqrt{|\mathcal{S}|}H^2 \log(H/\delta)}{N}. \tag{2.1.3.2}$$

Proof Sketch. To establish the high probability bound on the population 0-1 risk of BC, we utilize the key observation in the proof of Lemma 2.1.2: for each $t = 1, \cdots, H$, the contribution to the population 0-1 risk in eq. (2.1) stems only from states that were never visited at time t in any trajectory in D. For each t, we show that the mass contributed by such states up to constants does not exceed $\frac{|\mathcal{S}|}{N} + \frac{\sqrt{|\mathcal{S}|}\log(H/\delta)}{N}$ with probability $\geq 1 - \delta/H$. Summing over $t = 1, \cdots, H$ results in an upper bound on the population 0-1 loss that holds with probability $\geq 1-\delta$ (by the union bound). Invoking [74, Theorem 2.1] implies the high probability bound on $\mathsf{Gap}(\widehat{\pi}^{\mathtt{BC}})$. □

Remark 2.1.1. It is worth pointing out that in the deterministic expert setting, Eq. (2.1.3.1) shows that the imitation gap incurred by BC is $\lesssim |\mathcal{S}|H^2/N$ which is interesting in two ways: (i) it is independent of $|\mathcal{A}|$, and (ii) the imitation gap scales inversely in the size of the demonstration dataset N, and not as $1/\sqrt{N}$. Both observations are consequences of the determinism of the expert: at the states where BC is able to guess the expert's action better than a random guess, BC incurs no suboptimality.

Is error compounding inevitable or is it just a consequence of the Behavior Cloning algorithm? In the next section, we will argue that error compounding is *fundamental* to the Imitation Learning problem in the no-interaction or active-interaction settings without any further assumptions. Even if the expert is deterministic, no algorithm can beat the H^2 barrier.

2.2 Statistical lower bounds on the optimal imitation gap

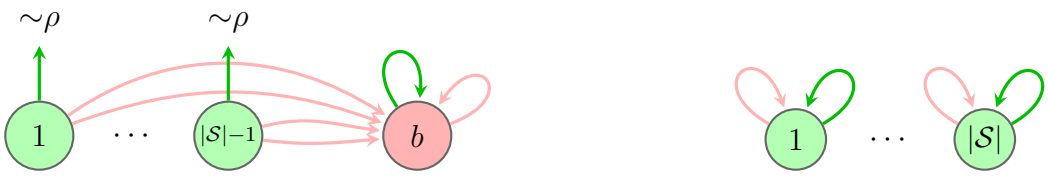
In this section we discuss lower bounds on the optimal imitation gap in the active-interaction setting, which will imply lower bounds for the no-transition setting as well. Our main result is the following lower bound.

Theorem 2.2.1. *For any learning rule $\mathsf{Alg}(\cdot)$ which operates in the active-interaction setting, there exists a tabular MDP $\mathcal{M} \in \mathbb{Z}_{\text{tabular}}$ and a deterministic expert policy π^{exp} such that the expected imitation gap of the policy $\widehat{\pi}$ returned by the learning rule $\mathsf{Alg}(\cdot)$ is lower bounded by,*

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \gtrsim \min\left\{H, |\mathcal{S}|H^2/N\right\}.$$

Where $\mathsf{Gap}(\cdot)$ is computed under the dynamics and rewards induced by the MDP $\mathcal{M}$. This lower bound continues to hold even under the constraint that π^{exp} must be an optimal policy on $\mathcal{M}$.

We construct the worst case MDP templates for the active-interaction setting in Figure 2.1a and defer the formal analysis to the appendix. The key intuition behind this result is to



(a) MDP template when in the no-interaction setting,
Upon playing the expert's action at any state except b, learner is renewed in the initial distribution,
$\rho = \{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-2)\zeta, 0\}$ where $\zeta = \frac{1}{N+1}$

(b) MDP template in the known-transition setting,
Each state is absorbing, initial distribution is given by $\{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-1)\zeta\}$ where $\zeta = \frac{1}{N+1}$

Figure 2.1: MDP templates for lower bounds under different settings: green arrows indicate state transitions under the expert's action, red arrows indicate state transitions under other actions

identify that at states which were never visited during the learner's interactions with the MDP, the learner has no prior knowledge about the expert's policy. Furthermore, at such states the learner also has no knowledge about what state transitions are induced under different actions. With no available information, the learner is essentially forced to play an arbitrary policy on these states. A careful construction of the underlying MDP ultimately forces the learner to incur compounding errors when such states are visited, resulting in the lower bound.

To build upon this intuition a bit further, consider the special case of the no-interaction setting. In Figure 2.1a, at any state of the MDP, except one, every other action, moves the learner to the absorbing state b. Suppose a learner independently plays an action different from the expert at a state with probability ϵ. Upon making a mistake, the learner is transferred to b and collects no reward for the rest of the episode. Thus the imitation gap of the learner is $\geq H\epsilon + (H-1)\epsilon(1-\epsilon) + \cdots + (1-\epsilon)^H \gtrsim \min\{H, H^2\epsilon\}$. By construction of ρ, we identify that any learner must make a mistake with probability $\epsilon \gtrsim |\mathcal{S}|/N$, resulting in the claim. It is interesting to observe that this argument closely resembles the intuition mentioned in the introduction for the $\lesssim \epsilon H^2$ bound on imitation gap that the reduction to supervised learning guarantees. In the following remark, we provide some more reasons as to why to expect that the same lower bound construction carries over to the active-interaction setting.

Remark 2.2.1. The lower bound construction in Figure 2.1a applies even if the learner can actively query the expert while interacting with the MDP. If the expert's queried action is not followed at any state, the learner is transitioned to b with probability 1. Upon doing so, the learner no longer can get any meaningful information about the expert's policy at states for the rest of the episode. Seeing that the "most informative" dataset the learner can collect involves following the expert at each time, it is no different had an demonstration dataset of

N trajectories been provided in advance. This reduces the active case to the no-interaction case for which the existing construction applies.

Remark 2.2.2. In the active-interaction setting, Theorem 2.2.1 in conjunction with Eq. (2.1.3.1) shows that when the expert plays a deterministic policy, BC achieves the optimal expected imitation gap of $\frac{|\mathcal{S}|H^2}{N}$ for Imitation Learning. Furthermore, this shows the optimality of BC even in the no-interaction setting as well. The ability to actively query the expert does not improve the sample complexity of Imitation Learning when the expert is deterministic. An important implication of this result is that DAGGER [74] and other algorithms that actively query an expert, *cannot improve over BC in the worst case.*

Lower bounds in the known-transition setting

Looking at the lower bound construction in the active-interaction setting, we see that the hardness of learning comes from the following property: at states which were never visited in the demonstration dataset, and where the learner has no idea about the transition of the MDP, the best a learner can do is pick an action at random. This is because of the fact that the learner cannot distinguish between actions at all. But in case the learner is able to make observations about the MDP transition (i.e., distinguish between actions), this property breaks down. In this section, we lower bound the worst-case optimal imitation gap incurred in the known-transition setting. By virtue of the above discussion, we arrive at a lower bound which is weaker than the bound in the active-interaction setting, and suggests no error compounding (i.e., no quadratic dependency on H).

Theorem 2.2.2. *In the known-transition setting, for any learning rule $\widehat{\pi} \leftarrow \mathsf{Alg}(D, P, \rho)$, there exists a tabular MDP $\mathcal{M} \in \mathbb{Z}_{\text{tabular}}$ and a deterministic expert policy π^{exp} such that the expected imitation gap of the learner is lower bounded by, $\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \gtrsim \min\{H, |\mathcal{S}|H/N\}$, where $\mathsf{Gap}(\cdot)$ is computed under the dynamics and rewards induced by $\mathcal{M}$.*

The lower bound instance in this construction is provided in Figure 2.1b. In these MDPs, each state is absorbing so a policy only stays at a single state for the whole episode. If the initial state of the MDP was not visited in the dataset, the learner does not see the expert's actions for the rest of the episode which is the only one at each state to offer non-zero reward. Conditioned on being initialized at such a state, the expected imitation gap is $\gtrsim H$. By construction of ρ, we determine that probability of the learner starting in such a state is $\gtrsim |\mathcal{S}|/N$ in expectation over the demonstration dataset D, resulting in the claim.

Remark 2.2.3. The lower bounds on the optimal imitation gap we arrive at in Theorem 2.2.1 for the no-interaction / active-interaction settings and Theorem 2.2.2 for the known-transition setting are universal - they apply for any learning rule $\mathsf{Alg}(\cdot)$. In contrast, the lower bound example in [74] (see Figure 1 and related discussion in their paper) applies only for supervised learning and is not universal. They construct a particular MDP and show that there exists a particular learner policy which (i) plays an action different than the expert with probability ϵ, and (ii) imitation gap $\gtrsim H^2\epsilon$. In fact on this example, the imitation gap of BC is exactly 0

if the learner is provided even *a single expert trajectory*. Thus their example does not provide a lower bound on the imitation gap of all learner algorithms as a function of the size of the dataset, N. In particular, even BC performs well on their example.

2.3 Imitating a stochastic expert

Prompted by the success of BC in the setting where the expert policy is deterministic, it is natural to ask whether it continues to remain a good approach when the expert is a stochastic policy. The version of BC we presented earlier, of training a classifier to minimize the empirical 0-1 risk, must be modified slightly to account for the expert's stochasticity.
Define the population risk under log-loss of a policy as the risk under log-loss of the corresponding classifier,

$$\mathcal{L}_{\log}(\pi) = \mathbb{E}_{t\sim\mathrm{Unif}([H])}\left[\mathbb{E}_{\pi^{\exp}}\left[\log\left(\frac{\pi_t^{\exp}(\cdot \mid S_t)}{\pi_t(\cdot \mid S_t)}\right)\right]\right]. \tag{2.4}$$

We may also define the empirical log-loss computed on a dataset of demonstrations D by,

$$\mathcal{L}_{\log}(\pi; D) = \mathbb{E}_{t\sim\mathrm{Unif}([H])}\left[\mathbb{E}_{S_t\sim f_D^t}\left[\log\left(\frac{\pi_t^{\exp}(\cdot \mid S_t)}{\pi_t(\cdot \mid S_t)}\right)\right]\right]. \tag{2.5}$$

Finally, define $\Pi^{\texttt{BC}}(D)$ as the set of all policies that minimize the empirical risk under log-loss to zero. Namely,

$$\Pi^{\texttt{BC}}(D) \triangleq \left\{\pi \in \Pi : \mathcal{L}_{\log}(\pi; D) = 0\right\} \tag{2.6}$$

Definition 2.3.1 (Tabular BC with a stochastic expert / log-loss BC)**.** *In the tabular setting under a stochastic expert policy, BC returns an arbitrary* $\pi^{BC} \in \widehat{\pi}^{BC}(D)$.

Within the confines of this section, we will abbreviate the version of Behavior Cloning in Definition 2.3.1 simply as "log-loss BC". Note that minimizing the empirical risk under log-loss, precisely translates to the learner playing the empirical expert policy distribution at states observed in the demonstration dataset. This viewpoint will be important later on, in motivating why the log-loss is the "correct" notion of error to consider in the stochastic setting. Furthermore, it is interesting to note that when the expert is deterministic, BC indeed still minimizes the empirical 0-1 risk to 0 and falls back to the version of the algorithm discussed in Definition 2.1.1.

In the deterministic expert setting, the reduction of [74] shows that it suffices for a policy to achieve low supervised learning loss (i.e., population risk under 0-1 loss) to achieve low imitation gap. Motivated by the reduction in [74], a natural question to ask is whether a similar reduction to supervised learning applies when the expert can be stochastic. We will show that when the expert plays a general policy, any learner which minimizes the TV distance to the expert's policy at states drawn by rolling out the expert achieves small imitation gap.

Algorithm 1 BC in the stochastic expert setting (tabular MDPs), i.e., log-loss BC

Input: Demonstration dataset D
for $t = 1, 2, \cdots, H$ **do**
  for $s \in \mathcal{S}$ **do**
    if $s \in \mathcal{S}_t(D)$ **then** ▶ $\mathcal{S}_t(D)$: states visited by trajectories in D at time t
      $\widehat{\pi}_t(\cdot|s) = \pi_t^D(\cdot|s)$ ▶ $\pi_t^D(\cdot|s)$: empirical estimator of $\pi_t^{\exp}(\cdot|s)$ in D
    else
      $\widehat{\pi}_t(\cdot|s) = \mathrm{Unif}(\mathcal{A})$.
    end if
  end for
end for
Return $\widehat{\pi}$

Reduction of IL to prediction under TV distance

Recall that [74, Theorem 2.1] show that if the expert's policy is deterministic, and the time-averaged probability of the learner $\widehat{\pi}$ correctly guessing the expert's action at each state is ϵ, then $\mathsf{Gap}(\widehat{\pi}) \leq \min\{H, \epsilon H^2\}$. In this section we prove a generalization of this result which applies even if the expert plays a stochastic policy. In particular, we consider a supervised learning reduction from Imitation Learning to matching the expert's policy in total variation (TV) distance. To this end, we first introduce the population TV risk,

$$\mathcal{L}_{\mathsf{TV}}(\widehat{\pi}) = \mathbb{E}_{t\sim\mathrm{Unif}([H])}\left[\mathbb{E}_{\pi^{\exp}}\left[D_{\mathrm{TV}}\left(\widehat{\pi}_t(\cdot|S_t), \pi_t^{\exp}(\cdot|S_t)\right)\right]\right]. \tag{2.7}$$

We show that if the learner minimizes the population TV risk to be $\leq \epsilon$ then the expected imitation gap of the learner is $\lesssim \min\{H, H^2\epsilon\}$. The population TV risk of a learner is a generalization of the population 0-1 risk to the case where the expert's policy is stochastic. We formally state the reduction below.

Lemma 2.3.1. *Consider any policy $\widehat{\pi}$ such that $\mathcal{L}_{\mathsf{TV}}(\widehat{\pi}) \leq \epsilon$. Then, $\mathsf{Gap}(\widehat{\pi}) \leq \min\{H, H^2\epsilon\}$.*

Remark 2.3.1. When the expert is deterministic, the definition of $\mathcal{L}_{\mathsf{TV}}(\cdot)$ matches that of $\mathcal{L}_{0\text{-}1}(\cdot)$ (eq. (2.1)) recovering the guarantee in [74, Theorem 2.1]. Thus, Lemma 2.3.1 strictly generalizes the supervised learning reduction for BC, Theorem 2.1.1.

While the reduction approach seems promising at first, there is a catch - the population TV risk in fact converges very slowly to 0. Since it corresponds to the rate at which the empirical action distribution approaches the population distribution in TV distance, the convergence rate is $\asymp \sqrt{|\mathcal{A}|/N}$ even if $|\mathcal{S}| = 1$. In the same setting, the population 0-1 risk which is the counterpart in the deterministic expert setting converges at a much faster $\lesssim 1/N$ rate (Theorem 2.1.2).

This analysis seems to suggest that (no-interaction) IL may be a harder problem to tackle in the setting where the expert is a stochastic policy. However, we will prove a surprising

result in the next section. By circumventing the reduction framework, we will show in Theorem 2.3.2 that the *expected* imitation gap achieved by log-loss BC achieves the same $1/N$ rate of convergence (up to logarithmic factors) even when the expert is stochastic.

Circumventing the reduction approach: log-loss BC

In this section, we consider log-loss BC (Algorithm 1). Via the lower bound we prove in Theorem 2.2.1 the guarantees on expected imitation gap of this policy is statistically optimal up to logarithmic factors. Note that the approach of playing the expert's empirical action distribution at states observed in the demonstration dataset in fact corresponds to minimizing the empirical risk under log-loss.

There is a critical difference between the stochastic expert and deterministic expert settings. In contrast to latter, it is no longer true that BC exactly mimics the expert policy at states which were visited in the demonstration dataset. However, by virtue of playing an empirical estimate of the expert's policy at these states it is plausible the expected imitation gap of the learner is still 0. However, a proof of this claim is not straightforward since the empirical distribution played by the learner at different states is not independent across time as functions of the dataset D.

To circumvent this technical challenge, we constructing a coupling between the dataset drawn from the expert policy, and policy of the learner. Under the coupling it turns out the expected reward gap of the learner will in fact be 0 when the visited states are all observed in the dataset. The remaining task is to bound the probability that at some point in the episode the learner visits a state unobserved in the demonstration dataset. A careful analysis of this probability term shows that it is bounded by $\lesssim |\mathcal{S}|H\log(N)/N$ under the coupling.

Theorem 2.3.2. *In the no-interaction setting, the learner's policy $\widehat{\pi}^{BC}$ returned by log-loss BC with a stochastic expert (Algorithm 1 / Definition 2.3.1) has expected imitation gap upper bounded by,*

$$\mathbb{E}\left[\mathit{Gap}(\widehat{\pi}^{BC})\right] \lesssim \min\left\{H, \frac{|\mathcal{S}|H^2\log(N)}{N}\right\},$$

2.4 Known-transition setting

In this section, we discuss the setting where the learner is not only provided expert demonstrations, but the state transitions of the MDP are known exactly. The learning setting appears frequently in applications like robotic manipulation [109] or game solving [80], where the learner has access to accurate models / simulators representing the dynamics of the system, but where the reward corresponding to the task to be acomplished may be difficult to construct or write down. In this section, we analyze the optimal imitation gap in this setting and surprisingly show that it is possible to break the lower bound in Theorem 2.2.1 and suppress the issue of error compounding. In the known-transition setting, the learning rules

we will consider (in the sense of Definition 1.3.1), $\mathsf{Alg}(\cdot)$, will be measurable functions of the MDP transition $P = \{P_t\}_{t=1}^H$, initial state distribution ρ, and the demonstration dataset D.

It is worth that several prior works such as that of [14] proposed algorithms that claimed to bypass the covariate shift problem, however at the time, we will present the first result to provably do so in the general tabular MDP setting without additional assumptions.

In the known-transition setting, mimicking the expert on states where the expert's policy is known is still a good approach, since there is no contribution to the learner's imitation gap as long as the learner only visits such states in an episode. However compared to the no-interaction setting, with the additional knowledge of P and ρ, the learner can potentially do better on states that are not visited in the demonstrations, and *correct* its mistakes even after it takes a wrong action, to avoid the error compounding problem. The algorithm we propose is known as `Mimic-MD` and introduced next.

Circumventing the error compounding barrier: `Mimic-MD`

`Mimic-MD` can be viewed as a hybrid approach which mimics the expert on some states, and uses a minimum distance (MD) functional [105, 31] to learn a policy on the remaining states. The idea of using minimum distance functionals was considered in [92], proposing to sequentially learn a policy by approximately minimizing a notion of discrepancy between the learner's state distribution and the expert's empirical state distribution. However, we remark that our approach is fundamentally different from matching the state distributions under the expert's and learner's policy: it crucially relies on mimicking the expert actions on some states, and only applying the MD functional approach on the remaining states.
The main guarantee we will establish for this algorithm is the following result.

Theorem 2.4.1. *Consider the learner's policy $\widehat{\pi}$ returned by `Mimic-MD` (Algorithm 2). When the expert policy π^{exp} is deterministic, in the known-transition setting,*

1. *The expected imitation gap of the learner is upper bounded by,*

$$\mathbb{E}\left[\mathsf{Gap}(\widehat{\pi})\right] \lesssim \min\left\{H,\ \sqrt{\frac{|\mathcal{S}|H^2}{N}},\ \frac{|\mathcal{S}|H^{3/2}}{N}\right\}. \tag{2.4.1.1}$$

2. *For any $\delta \in (0, \min\{1, H/5\})$, with probability $\geq 1-\delta$, the imitation gap of the learner satisfies,*

$$\mathsf{Gap}(\widehat{\pi}) \lesssim \frac{|\mathcal{S}|H^{3/2}}{N} \log\left(\frac{|\mathcal{S}|H}{\delta}\right). \tag{2.4.1.2}$$

`Mimic-MD` improves the quadratic dependence on H of the imitation gap incurred by `BC` (Theorem 2.1.3) by at least a $\sqrt{H}$ factor while preserving the dependence of on $|\mathcal{S}|$ and N.

Algorithm 2 `Mimic-MD` on tabular MDPs

1: **Input:** Demonstration dataset D.
2: Let D_1 be $N/2$ trajectories drawn uniformly without replacement from D. Let $D_2 = D \setminus D_1$.
3: For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, define,

$$\mathcal{T}_t^{D_1}(s,a) \triangleq \left\{ \{(s_{t'}, a_{t'})\}_{t'=1}^{H} \;\middle|\; s_t = s, a_t = a,\ \exists \tau \le t : s_\tau \notin \mathcal{S}_\tau(D_1) \right\}.$$

▶ Set of trajectories that visit (s, a) at time t, and at some time $\tau \le t$ visit a state unvisited at time τ in any trajectory in D_1.

4: Define the following optimization problem:

$$\min_{\pi \in \Pi_{\mathrm{det}}^{\mathsf{BC}}(D_1)} \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| \Pr_{\pi} \left[\mathcal{T}_t^{D_1}(s,a) \right] - \frac{\sum_{\mathsf{tr} \in D_2} \mathbb{1}\left(\mathsf{tr} \in \mathcal{T}_t^{D_1}(s,a)\right)}{|D_2|} \right| \tag{OPT-MD}$$

Choose $\widehat{\pi}$ as any optimizer of (OPT-MD).

▶ $\Pi_{\mathrm{det}}^{\mathsf{BC}}(D_1)$ is defined in (Eq. (2.3))

5: **Return** $\widehat{\pi}$.

Proof sketch of Theorem 2.4.1 and interpreting `Mimic-MD`

Eq. (2.4.1.1) shows that `Mimic-MD` (Algorithm 2) breaks the $|\mathcal{S}|H^2/N$ imitation gp compounding barrier which is not possible in the no-/active-interaction setting, as in Theorem 2.2.1. `Mimic-MD` inherits the spirit of mimicking the expert by exactly copying the expert actions in dataset D_1: as a result, the learner only incurs imitation gap upon visiting a state unobserved in D_1 at some point in an episode. Let $\mathcal{E}_{D_1}^{\le t}$ be the event that the learner visits a state at some time $\tau \le t$ which has not been visited in any trajectory in D_1 at time τ. In particular, for any policy $\widehat{\pi}$ which mimics the expert on D_1, we show,

$$\mathsf{Gap}(\widehat{\pi}) \le \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{t=1}^{H} \left| \Pr_{\pi^{\mathrm{exp}}} \left[\mathcal{E}_{D_1}^{\le t}, s_t = s, a_t = a \right] - \Pr_{\widehat{\pi}} \left[\mathcal{E}_{D_1}^{\le t}, s_t = s, a_t = a \right] \right|. \tag{2.8}$$

In the known-transition setting the learner knows the transition functions $\{P_t : 1 \le t \le H\}$ and the initial state distribution ρ, and can exactly compute the probability $\Pr_\pi[\mathcal{E}_{D_1}^{\le t}, s_t = s, a_t = a]$ for any known policy π. However, unfortunately the learner cannot compute $\Pr_{\pi^{\mathrm{exp}}}[\mathcal{E}_{D_1}^{\le t}, s_t = s, a_t = a]$ given only D_1. This is because the expert's policy on states unobserved in D_1 is unknown and the event $\mathcal{E}_{D_1}$ ensures that such states are necessarily visited. Here we use the remaining trajectories in the dataset, D_2 to compute an empirical estimate of $\Pr_{\pi^{\mathrm{exp}}}[\mathcal{E}_{D_1}^{\le t}, s_t = s, a_t = a]$. The form of eq. (2.8) exactly motivates Algorithm 2, which replaces the population term $\Pr_{\pi^{\mathrm{exp}}}[\mathcal{E}_{D_1}^{\le t}, s_t = s, a_t = a]$ by its empirical estimate in the MD functional.

Remark 2.4.1. In the known-transition setting, the maximum likelihood estimate (MLE) for $\pi^{\exp}$ does not achieve the optimal sample complexity. When the expert is deterministic, all policies in $\Pi^{\texttt{BC}}_{\text{det}}(D)$ have equal likelihood given D. This is because the probability of observing a trajectory does not depend on the expert's policy on the states it does not visit. From Eq. (2.1.3.1) and Theorem 2.2.1 the expected imitation gap of the worst policy in $\Pi^{\texttt{BC}}_{\text{det}}(D)$ is $\asymp |\mathcal{S}|H^2/N$. Since the MLE does not give a rule to break ties, this implies that it is not optimal.

Remark 2.4.2. The standard analysis of conventional minimum distance functional / distribution matching approaches rely on convergence of the empirical distribution to the population in the corresponding distance functional. For most non-trivial choices of the distance functional, this convergence rate is slow and is $\gtrsim 1/\sqrt{N}$, given N samples. At a technical level, the state distributions are matched only at states unvisited in the demonstration dataset. In particular, $1/N$ rate of convergence of `Mimic-MD` relies on the fact that the effective mass of the distributions being matched shrinks from 1 to $|\mathcal{S}|H/N$.

Remark 2.4.3. Although data splitting may not be necessary, we conjecture that the conventional minimum distance functional approach, which matches the empirical distribution of either states or state-action pairs does not achieve the rate in Eq. (2.4.1.1) since it does not necessarily exactly mimic the expert on the observed demonstrations. In particular, conventional distribution matching approaches do not take into account the fact that the expert's action is known at every state visited in the dataset. These policies may choose to play a different action at a state, even if the expert's action is observed in the dataset. In contrast, `Mimic-MD` returns a policy that is constrained to mimic the expert at states visited in the demonstration dataset, avoiding this issue.

Remark 2.4.4. The optimization problem OPT-MD solved by `Mimic-MD` is over multivariate degree-H polynomials in $\{\pi_1(\cdot|\cdot), \cdots, \pi_H(\cdot|\cdot)\}$. As such, it is not possible to solve this optimization problem in polynomial (in H) time, however in the next section, we will discuss how to view this algorithm from a different lens which will enable it to be computed efficiently. We will appeal to the fact that the polynomial is sparse having at most N non-zero coefficients. Moreover, our analysis does not require that the optimization problem OPT-MD be solved exactly, which we discuss in Corollary 2.4.1 and remark 2.4.5.

We also provide a guarantee when the learner solves the optimization problem in OPT-MD to an accuracy of ε. The guarantee on imitation gap admit by `Mimic-MD` in Eq. (2.4.1.1) is recovered taking $\varepsilon = 0$.

Corollary 2.4.1. *Consider any policy $\widehat{\pi}$ that minimizes the optimization problem OPT-MD to an additive error of ε. Then, the expected imitation gap of the learner is upper bounded by,*

$$\mathbb{E}\left[\mathit{Gap}(\widehat{\pi})\right] \lesssim \min\left\{H,\ H\sqrt{\frac{|\mathcal{S}|}{N}} + \varepsilon,\ \frac{|\mathcal{S}|H^{3/2}}{N} + \varepsilon\right\}. \tag{2.9}$$

Remark 2.4.5. Corollary 2.4.1 shows that `Mimic-MD` is amenable in the following settings and combinations thereof.

1. As discussed in Remark 2.4.4, optimization problems over multivariate degree H polynomials (as in OPT-MD) in general are not exactly solvable in polynomial time. Corollary 2.4.1 shows that it suffices to solve OPT-MD approximately to result in a policy with small imitation gap.

2. This approach applies in the approximate transition setting, where the transition functions are not known exactly but are known approximately. In particular, suppose the learner's policy $\widehat{\pi}$ solves OPT-MD exactly when the probabilities $\Pr_\pi[\cdot]$ are computed using the approximate transition functions. By the smoothness of $\Pr_\pi[\cdot]$ one can bound the imitation gap of $\widehat{\pi}$ on OPT-MD when the probabilities are instead computed using exact transition functions. Applying Corollary 2.4.1 for this ε controls the imitation gap of the resulting policy.

Remark 2.4.6. In the known-transition setting, in conjunction with eq. (2.4.1.1), this lower bound shows that when the expert is deterministic, then `Mimic-MD` (Algorithm 2) is optimal in the dependence on $|\mathcal{S}|$ and N and is suboptimal by a factor of at most $\sqrt{H}$ in its dependence on the episode length.

Chapter 3

Understanding `Mimic-MD`

In the previous section, we established that with the power of environment interaction, (i.e., in the known-transition setting), it is possible to arrive at an upper bound of $\widetilde{O}(|\mathcal{S}|H^{3/2}/N)$ for the imitation gap of the `Mimic-MD` algorithm. In this chapter, we dig deeper into two questions,

1. Is this dependency on H optimal? Can we really hope to get imitation gap scaling linearly-in-H as Theorem 2.2.2 suggests?

2. The `Mimic-MD` algorithm is an optimization problem over degree-H multivariate polynomials (cf. the training objective (OPT-MD)). Can it be implemented efficiently? Is there a statistical-computational gap in breaking past the quadratic error compounding barrier?

3. How can we connect `Mimic-MD` to practical Imitation Learning algorithms? What insights do these theoretical analyses provide?

We will address all three problems in this chapter by presenting an alternate view of `Mimic-MD` via a two-way reduction to the "value estimation problem" of the unknown expert policy, which will be defined in due time.

We will also show that under the additional assumption that the expert is optimal for the true reward function, there exists an efficient algorithm, which we term as `Mimic-Mixture`, that provably achieves imitation gap $\widetilde{O}(1/N)$ for arbitrary 3-state MDPs with rewards only at the terminal layer. In contrast, we will show that no algorithm can achieve imitation gap scaling as $\widetilde{O}(\sqrt{H}/N)$ with high probability if the expert is not constrained to be optimal. While the optimal expert setting does not help in the no-interaction setting as discussed in Theorem 2.2.1, these results formally establish its benefit in the known transition setting.

3.1 Introduction

As introduced previously, the no-interaction setting precludes the learner from interacting with the environment. While we establish that there is a provable benefit to having information about the MDP transition in the previous chapter, it's worth recalling what the bottleneck in the no-interaction setting is, and why error compounding must be incurred. No algorithm can beat the $\widetilde{\Omega}(|\mathcal{S}|H^2/N)$ lower bound on the imitation gap in this setting. In a nutshell, the idea is that with no information about the MDP transition, the learner cannot recover after making a "mistake", playing an action different from the expert's action at some state. Concretely, at each time t, any learner has an $\Omega(|\mathcal{S}|/N)$ probability of making a mistake due to not having observed the expert action at this time step[1]. The probability that the learner has made a mistake (i.e., deviated from the expert policy) at some point up to time t in an episode can be forced to be $\asymp |\mathcal{S}|t/N$ since the union bound is approximately tight in the lower bound instance, and under this event the learner incurs imitation gap of 1. The expected imitation gap of the learner is therefore, $\sum_{t=1}^{H} |\mathcal{S}|t/N \asymp |\mathcal{S}|H^2/N$.

To break this quadratic H dependency (under additional assumptions, such as having access to the MDP transition), the above analysis suggests that one needs to beat the union bound in the probability of making a mistake at time t, which implies that one needs to conduct *long-range planning*. The error events of making a mistake at each time t should be made *negatively correlated* with each other, in the sense that we can quickly recover from the mistakes made in the past. This is enabled when the learner has knowledge of the MDP transition function and is the key idea behind the `Mimic-MD` algorithm (Algorithm 2), which has imitation gap upper bounded by $O(|\mathcal{S}|H^{3/2}/N)$.

There is also the question of whether `Mimic-MD` is an efficient algorithm. It is a-priori unclear how to implement the version in Algorithm 2 efficiently, since it is a complicated optimization problem and even the description of the learning objective involves a combinatorially large marginalization over all possible trajectories. In the next section, we will introduce a reduction of the IL problem in the known-transition setting to a different one we refer to as "expert value estimation". This reduction will enable us to address both the questions of efficiency as well as statistical lower bounds in the known-transition setting.

3.2 Expert value estimation

In this section, we begin with a few definitions, and formally introduce an equivalence between Imitation Learning in the known-transition setting and the expert value estimation problem. At a high level, the expert value estimation problem entails that the learner is able to estimate the value of the expert's policy under the unknown ground truth reward with high probability.

[1]The term $|\mathcal{S}|/N$ comes from the *missing mass* in sampling [54], which can also be understood as the V/N regret in binary classification with zero oracle error, where $V = |\mathcal{S}|$ is the VC-dimension when $|\mathcal{A}| = 2$.

Definition 3.2.1 (Expert value estimation)**.** *Given a collection of tuples of MDP instances and expert policies, denoted* $\mathbb{Z} = \{\mathcal{Z}\} = \{(\mathcal{S}, \mathcal{A}, P, r, \rho, H, \pi^{\exp})\}$*, we say that an estimator* $\widetilde{J}_r(\pi^{\exp})$ *is a expert value estimator for* $J_r(\pi^{\exp})$ *with confidence* $1-\delta$ *and error* ϵ *if for any such instance* $\mathcal{Z} = (\mathcal{S}, \mathcal{A}, P, r, \rho, H, \pi^{\exp})$*, we have,*

$$\Pr(|J_r(\pi^{\exp}) - \widetilde{J}_r(\pi^{\exp})| \geq \epsilon) \leq \delta. \tag{3.1}$$

where the estimator $\widetilde{J}$ *is a function of all information available in the known-transition setting, and the candidate reward* r*. Namely,* $\widetilde{J}$ *is a measurable function of* ρ*,* P*,* r *and a demonstration dataset* D *(Definition 1.3.1) of expert trajectories, but not the expert* $\pi^{\exp}$ *directly. The probability in eq. (3.1) is computed over the randomness of* D *and the internal randomness of* $\widetilde{J}$*.*

The related problem of *uniform* expert value estimation is defined as estimating the value of the expert policy uniformly on some class of reward functions, rather than just the ground truth reward, with high probability.

Definition 3.2.2 (Uniform expert value estimation)**.** *Given a collection of IL instances* $\mathbb{Z} = \{\mathcal{Z}\} = \{(\mathcal{S}, \mathcal{A}, P, r, \rho, H, \pi^{\exp})\}$*, we say that an estimator* $\widetilde{J}_r(\pi^{\exp})$ *is a "uniform expert value estimator" for* $J_r(\pi^{\exp})$ *on* $\mathcal{R}_D$ *with confidence* $1-\delta$ *and error* ϵ *if for any such instance* $\mathcal{Z} = (\mathcal{S}, \mathcal{A}, P, r, \rho, \pi^{\exp})$*, we have*

$$\Pr\left(\sup_{r' \in \mathcal{R}_D} |J_{r'}(\pi^{\exp}) - \widetilde{J}_{r'}(\pi^{\exp})| \geq \epsilon\right) \leq \delta,$$

where the estimator $\widetilde{J}$ *is a measurable function of* ρ*,* P*,* r*, and a demonstration dataset* D*, and an input set of reward functions,* $\mathcal{R}_D$*, which contains the true reward* r*, but not the expert* $\pi^{\exp}$ *directly. We add the subscript* D *to emphasize that* $\mathcal{R}_D$ *is allowed to depend on the demonstration dataset, but we may omit the subscript when clear from context.*

The quantity $\sup_{r' \in \mathcal{R}_D} |J_{r'}(\pi^{\exp}) - \widetilde{J}_{r'}(\pi^{\exp})|$ *is referred to as the "uniform expert value estimation error".*

Out first contribution is to propose a general formulation which reduces IL (Definition 1.3.2) to uniform expert value estimation (Definition 3.2.2) with known transitions. Define the learner policy $\widehat{\pi}$ as the solution to the following minimax optimization problem:

$$\widehat{\pi} \leftarrow \arg\min_{\pi} \max_{r \in \mathcal{R}_D} \widetilde{J}_r(\pi^{\exp}) - J_r(\pi), \tag{OPT}$$

where $\mathcal{R}_D$ is the same as that in Definition 3.2.2. The next result shows reductions between IL and expert value estimation when transitions are known.

Theorem 3.2.1 (Reductions between IL and expert value estimation with known transitions)**.** *Consider the following two cases of* $\mathcal{R}_D$*:*

(*i*) *Symmetric class: for any IL instance* $\mathcal{Z} = (\mathcal{S}, \mathcal{A}, P, r, \rho, H, \pi^{\exp})$ *we consider,* $r \in \mathcal{R}_D$*, and in addition,* $(1 - r) \in \mathcal{R}_D$ *is also in the set. A notable special case is when we consider all possible reward functions bounded between zero and one;*

(*ii*) *Optimal expert: for each IL instance,* $\mathcal{Z} = (\mathcal{S}, \mathcal{A}, P, r, \rho, H, \pi^{\exp})$*,* $\pi^{\exp}$ *is an optimal policy for* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho, H)$*.*

Then, under both cases,

(*i*) *If* $\widehat{\pi}$ *admits PAC guarantees for the IL problem (Definition 1.3.2) with confidence* $1 - \delta$ *and error* ϵ*, then* $J_r(\widehat{\pi})$ *solves the expert value estimation (Definition 3.2.1) with confidence* $1 - 2\delta$ *and error* ϵ*;*

(*ii*) *if* $\widetilde{J}_r(\pi^{\exp})$ *solves uniform expert value estimation (Definition 3.2.2) with confidence* $1 - \delta$ *and error* ϵ*, then the minimax algorithm in* (OPT) *solves IL (Definition 1.3.2) with confidence* $1 - \delta$ *and error* 2ϵ*.*

Proof. The proof proceeds in two parts,

(*i*) *IL* $\implies$ *expert value estimation.* Consider the expert value estimator $J_r(\widehat{\pi})$. For the case when the reward function is symmetric, by choosing two MDPs differing in their reward functions, as r (the ground truth reward) and $1 - r$ in eq. (1.6) and union bounding, we have the desired result. For the optimal expert case, we can save one δ factor since we know $|J_r(\pi^{\exp}) - J_r(\widehat{\pi})| = J_r(\pi^{\exp}) - J_r(\widehat{\pi})$.

(*ii*) *Uniform expert value estimation* $\implies$ *IL.* To analyze the imitation gap of the learner in (OPT), observe that,

$$\begin{aligned} J_r(\pi^{\exp}) - J_r(\widehat{\pi}) &\leq \max_{r' \in \mathcal{R}_D} J_{r'}(\pi^{\exp}) - \widetilde{J}_{r'}(\pi^{\exp}) + \max_{r' \in \mathcal{R}_D} \widetilde{J}_{r'}(\pi^{\exp}) - J_{r'}(\widehat{\pi}) \\ &\overset{(i)}{\leq} \max_{r' \in \mathcal{R}_D} |J_{r'}(\pi^{\exp}) - \widetilde{J}_{r'}(\pi^{\exp})| + \max_{r' \in \mathcal{R}_D} \widetilde{J}_{r'}(\pi^{\exp}) - J_{r'}(\pi^{\exp}) \\ &\leq 2\epsilon. \end{aligned}$$

where (i) uses the fact that $\pi^{\exp}$ is a feasible policy to the optimization problem (OPT).

□

Next we instantiate this framework to establish minimax upper and lower bounds for Imitation Learning. We begin with the setting where the expert policy could be arbitrary.

3.3 Mimic-MD through the lens of expert value estimation

For brevity let $\mathcal{R}$ denote the set of all reward functions such that $r_t(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Recall the conclusion of Section 3.2: in the known-transition setting, it suffices to construct a uniform expert value estimator such that $\sup_{r\in\mathcal{R}} |J_r(\pi^{\exp}) - \widetilde{J}_r(\pi^{\exp})|$ is small, to ensure that the IL problem can be solved.

In the known-transition setting, the learner collects N demonstrations from the expert policy. Thus, a natural and unbiased estimator of $J_r(\pi^{\exp})$ is the average reward r collected by the trajectories in D, an empirical estimate. This idea can be generalized. Let f_t^π denote the marginal state-action distribution induced at time t by the policy π. Rolling out the policy $\pi^{\exp}$, we may rewrite,

$$J_r(\pi^{\exp}) = \sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim f_t^{\pi^{\exp}}}[r_t(s,a)]$$

This motivates us to estimate $J_r(\pi^{\exp})$ using a generic estimator of the form,

$$\sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim \widehat{f}_t^{\pi^{\exp}}}[r_t(s,a)],$$

where $\widehat{f}_t^{\pi^{\exp}}$ is some estimator of $f_t^{\pi^{\exp}}$. In this case, the uniform expert value estimation error over reduces to the sum of *total variation* distances $\sum_{t=1}^{H} D_{\mathrm{TV}}\left(f_t^{\pi^{\exp}}, \widehat{f}_t^{\pi^{\exp}}\right)$, where we used the variational representation, $D_{\mathrm{TV}}(P,Q) = \sup_{r\in[0,1]^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}_P[r] - \mathbb{E}_Q[r]$.

As discussed previously, for each t, we obtain N i.i.d. samples from distribution $f_t^{\pi^{\exp}}$ via the demonstration dataset, and therefore it is natural to use the empirical distribution as the estimator $\widehat{f}_t^{\pi^{\exp}}$. It then follows from standard results [42] that with probability at least $1-\delta$, we have [2]

$$\sum_{t=1}^{H} D_{\mathrm{TV}}\left(f_t^{\pi^{\exp}}, \widehat{f}_t^{\pi^{\exp}}\right) \lesssim H\sqrt{\frac{|\mathcal{S}| + \log(H/\delta)}{N}}. \tag{3.2}$$

Although (3.2) seems to suggest that the imitation gap error now grows linear in H, there is a catch: the dependence on N has degraded to $N^{-1/2}$, and this bound becomes even worse than the behavior cloning bound of $\widetilde{O}(|\mathcal{S}|H^2/N)$ when N is large.

We will show that the Mimic-MD algorithm (Algorithm 2) can be interpreted as constructing an *improved* estimator for $f_t^{\pi^{\exp}}$ that achieves smaller TV loss in eq. (3.2) than the empirical estimator. This is surprising at first glance, but there is a good reason for why this is possible. Indeed, the empirical distribution $\widehat{f}_t^{\pi^{\exp}}$ does not utilize any information about the transition kernel, which the learner has access to. Intuitively, given the MDP transition, one

[2] Note that we have assumed the expert policy is deterministic, so the support of $f_t^{\pi^{\exp}}$ is at most $|\mathcal{S}|$.

can potentially simulate many new trajectories and view them as new datapoints in order to improve statistical efficiency. The only case where simulation fails is when the learner encounters a state where we have not visited in the dataset; but the probability of seeing an unseen state within the first t steps is at most $\lesssim |\mathcal{S}|t/N$ by a union bound, and hence the total variation loss in estimating $f_t^{\pi^{\exp}}$ can be improved to be $\lesssim \sqrt{|\mathcal{S}|/N}\sqrt{|\mathcal{S}|t/N} = |\mathcal{S}|\sqrt{t}/N$.[3] Summing up over $t \in [H]$, the overall $\sum_{t=1}^{H} |\mathcal{S}|\sqrt{t}/N \lesssim |\mathcal{S}|H^{3/2}/N$. The following theorem summarizes the performance of Mimic-MD, which is the specific instantiation of (OPT) when we consider all possible rewards and the improved value estimator mentioned above.

It turns out that optimization problem (OPT-MD) can be formulated as a convex optimization, implying that Mimic-MD can be solved efficiently approximately (which suffices to recover the statistical guarantees). In particular, in the space of joint state-action probabilities, $\{f_t^{\pi}(s_t, a_t)\}_{t\in[H], s_t\in\mathcal{S}, a_t\in\mathcal{A}}$, the objective can be represented as a convex program. The learner's policy at each time t can be extracted from this representation using,

$$\pi_t(a|s) = \frac{f_t^{\pi}(s,a)}{\sum_{\tilde{a}\in\mathcal{A}} f_t^{\pi}(s,\tilde{a})}.$$

The convexity of the objective stems from the convexity of the TV distance eq. (3.2) when parameterized by the joint state-action probabilities. Below we present the main result establishing these properties of Mimic-MD, which is essentially a refinement of Theorem 2.4.1.

Theorem 3.3.1. *The optimization problem* (OPT-MD) *in* Mimic-MD *can be formulated as a convex program with* poly$(|\mathcal{S}|, |\mathcal{A}|, H)$ *variables and linear constraints. Moreover, its solution* $\widehat{\pi}$ *achieves expected imitation gap,*

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \lesssim \begin{cases} H\log H/N & |\mathcal{S}| = 2 \\ |\mathcal{S}|H^{3/2}/N & |\mathcal{S}| \geq 3 \end{cases}$$

for all IL instances in $\mathbb{Z}_{\text{tabular}}$*, the set of IL instances over tabular MDPs and with no constraint on the expert policy.*

A formal proof of this result is deferred to Appendix B.1. The proof will follow the same outline as that of Theorem 2.4.1. This refinement shows that when $|\mathcal{S}| = 2$, the expected imitation gap achieved by Mimic-MD is in fact nearly linear in H, which nearly matches the lower bound in Chapter 2 (Theorem 2.2.2) [4].

[3] Precisely, the TV error of estimating a discrete distribution $(p_1, p_2, \ldots, p_k)$ from N i.i.d. samples is upper bounded by $\sum_{i\in[k]}\sqrt{\frac{p_i}{N}} \leq \sqrt{k/N}$ where the worst case is attained when $p_i \equiv 1/k$. In Mimic-MD, we are reducing the effective probability mass from 1 to $\frac{|\mathcal{S}|t}{N}$, so the problem is reduced to upper bounding $\sup_{p_i\geq 0, \sum_{i\in[k]} p_i \leq \frac{|\mathcal{S}|t}{N}} \sum_i \sqrt{\frac{p_i}{N}} = \sqrt{\frac{k}{N}}\sqrt{\frac{|\mathcal{S}|t}{N}}$.

[4] Indeed, whenever we have an unseen state for a distribution with binary values, the total variation distance between the empirical distribution and the real distribution is of order $\widetilde{O}(1/N)$. However, it is not true for distributions with $|\mathcal{S}| \geq 3$ in general: for example, if $p = (1/2, 1/2 - 1/N, 1/N)$, then with constant probability we will not see the third state in the dataset, but the total variation distance between empirical distribution and true distribution scales as $O(1/\sqrt{N})$.

Algorithm 3 Alternate view of Mimic-MD (Algorithm 2)

1: **Input:** Expert dataset D.

2: Let D_1 be $N/2$ trajectories drawn uniformly without replacement from D. Let $D_2 = D \setminus D_1$.

3: Define,

$$\mathcal{T}_t^{D_1}(s,a) \triangleq \{\{(s_{t'}, a_{t'})\}_{t'=1}^{H} | s_t = s, a_t = a,\ \exists \tau \le t : s_\tau \notin \mathcal{S}_\tau(D_1)\}$$

▶ Set of trajectories that visit (s,a) at time t, and at some time $\tau \le t$ visit a state unvisited at time τ in any trajectory in D_1.

4: **(Original view)** Return $\widehat{\pi}$ as any optimizer of the following program:

$$\widehat{\pi} \leftarrow \arg\min_{\pi \in \Pi_{\mathrm{det}}^{\mathsf{BC}}(D_1)} \sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \Pr_{\pi}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{1}{|D_2|} \sum_{\mathsf{tr}\in D_2} \mathbb{1}\left(\mathsf{tr} \in \mathcal{T}_t^{D_1}(s,a)\right) \right|. \quad \text{(OPT-MD)}$$

▶ $\Pi_{\mathrm{det}}^{\mathsf{BC}}(D_1)$ is defined in eq. (2.3)

5: **(Alternate view)** Equivalently, it suffices to output the policy $\widehat{\pi}$ as the solution to the following minmax optimization problem,

$$\widehat{\pi} \leftarrow \arg\min_{\pi} \sup_{r\in\mathcal{R}} \widetilde{J}_r(\pi^{\mathrm{exp}}) - J_r(\pi) \qquad \text{(OPT-MD-value)}$$

where $\mathcal{R}$ is the set of all reward functions and the value estimator $\widetilde{J}$ is defined as,

$$\widetilde{J}_r(\pi^{\mathrm{exp}}) = \sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} r_t(s,a)\widehat{f}_t(s,a),$$

where, for any tuple $(s,a,t) \in \mathcal{S}\times\mathcal{A}\times[H]$,

$$\widehat{f}_t(s,a) = \Pr_{\pi^{\mathrm{exp}}}\left[\neg\mathcal{T}_t^{D_1}(s,a)\right] + \Pr_{\mathrm{Unif}(D_2)}\left[\mathcal{T}_t^{D_1}(s,a)\right]. \qquad (3.3)$$

6: **Return** $\widehat{\pi}$.

Statistical lower bounds in the known-transition setting

While Theorem 3.3.1 resolves the case of $|\mathcal{S}| = 2$ showing a matching upper and lower bound on the imitation gap in , it still leaves open the possibility of improving the H dependency beyond $H^{3/2}$ when $|\mathcal{S}| \ge 3$. In the following result, we show that this is no coincidence. $H^{3/2}/N$ is a statistical barrier on the imitation gap of any learner in the known-transition setting.

Theorem 3.3.2. *Suppose $H \ge 2$ and $N \ge 7$. If $N \ge 6H$, for every learning rule $\mathsf{Alg}(\cdot)$ in the known-transition setting (returning policy $\widehat{\pi}$), there exists an MDP $\mathcal{M}$ on 3 states such*

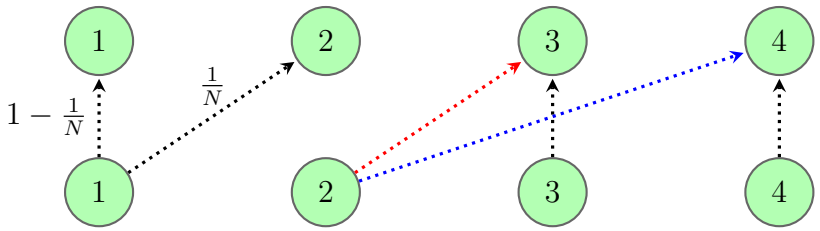


Figure 3.1: The *4-state MDP transition*: The states 1, 3 and 4 have a single action, with the transition probabilities indicated above the arrow. On the other hand, state 2 has 2 actions: with probability 1, the red action transitions the learner to state 3 while the blue action transitions the learner to state 4.

that,

$$\Pr\left(\mathsf{Gap}(\widehat{\pi}) \geq \frac{cH^{3/2}}{N}\right) \geq c',$$

for some constants $c, c' > 0$, and where $\mathsf{Gap}(\cdot)$ is computed under the dynamics induced by $\mathcal{M}$. Here the probability is computed over the randomness of the demonstration dataset D as well as the internal randomness employed by the learning rule $\mathsf{Alg}(\cdot)$.

To best illustrate the insights behind the lower bound construction, we construct a particular lower bound instance, which we name *4-state MDP*, and describe informally the main ideas. In Appendix B.2, we formally prove the lower bound for $|\mathcal{S}| = 3$ as well by essentially combining the state labelled 1 with 3.

The 4-state MDP

Since Theorem 3.2.1 shows that the expert value estimation problem is not harder than IL, it suffices to show that the value estimation error is at least $H^{3/2}/N$. Concretely, the 4-state MDP in Figure 3.1 is time-invariant with the states labelled 1, 2, 3 and 4. All states besides 2 are trivial and without loss of generality have only a single action. The state 2 has exactly 2 actions, one leading deterministically to state 3 and the other to state 4. Furthermore, the reward function of the MDP is all 1 on the state 3 (recall there is only a single action at this state). We assume the initial distribution is $(1 - 1/N, 1/N, 0, 0)$.
Define variables

$$U_i = \begin{cases} 1 & \text{if } \pi_i^{\text{exp}}(\text{red} \mid 2) = 1 \\ 0 & \text{if } \pi_i^{\text{exp}}(\text{blue} \mid 2) = 1 \end{cases}$$

Note that the marginal distribution at state 1 at time t is independent of expert policy and equal to $(1 - 1/N)^t$, and the marginal distribution of state 2 at time t is always $w_t \triangleq (1 - 1/N)^{t-1}/N$. Consider the case that $N \gtrsim H$, in this case the marginal probability

of state 2 for every time step is $w_t \asymp 1/N$. We say that the contribution of time t to final expert value is $v_t = \sum_{i=1}^{t-1} w_i U_i$, and the final *expert value* $V^\star$ is given by

$$V^\star = \sum_{t=1}^{H} v_t, \tag{3.4}$$

and we aim to show the estimation error of $V^\star$ is at least $\asymp H^{3/2}/N$.
The major step towards the final lower bound is to show that the estimation error of v_H, which is the contribution to the final value from the last layer, is at least $\sqrt{H}/N$. This dependence on the time horizon would then accumulate to achieve the $H^{3/2}$ result. Note that $v_H = \sum_{t=1}^{H-1} w_t U_t$, can be viewed as the weighted combination of parameters U_i, but since the marginal probability of state 2 is about $1/N$, in total N trajectories there will be a constant fraction of state 2's across time steps that are not observed in the dataset. If we impose a uniform prior on $U_i \sim \mathsf{Bern}(1/2)$, then the posterior variance of v_H is at least a constant fraction of the prior variance of v_H, which is

$$\mathsf{Var}(v_H) = \sum_{t=1}^{H-1} \frac{w_t^2}{4} \asymp \frac{H}{N^2},$$

which implies that the posterior standard deviation of v_H is at least of order $\sqrt{H}/N$. Then, we can combine this lower bound with (3.4) to show that the overall estimation error of $V^\star$ is at least $\sum_{t=1}^{H} \sqrt{t}/N \asymp H^{3/2}/N$. This result implies that `Mimic-MD` indeed achieves the optimal dependence on the MDP horizon H, growing as $H^{3/2}$.

Chapter 4

Learning an optimal expert

The previous chapter assumes that the learner operates in the known-transition setting, but imposes no assumptions on the nature of the expert. Indeed, the expert policies considered in the 4-state instances invoked in the proof of the lower bound Theorem 3.3.2 are really far from being optimal on the underlying MDP. In practice, however experts often are often not pathological /worst-case policies and may even achieve near optimal performance. Indeed it only makes sense to carry out Imitation Learning when the expert carries out the underlying task well. In this chapter, we study Imitation Learning under the known-transition setting, under the assumption that the expert is an optimal policy on the underlying (unknown) reward function r. This will turn out to be a really hard setting, which we make partial progress toward understanding. A natural first question is whether the lower bound in Theorem 3.3.2 still holds when we impose the additional assumption that the expert policy $\pi^{\exp}$ is the optimal policy for the true reward function r.

4.1 Revisiting the 4-state MDP instances

In the previous chapter, we constructed a special class of MDPs on 4-states under which a statistical lower bound on the imitation gap can be established. Attempting to deploy the same construction, but with the additional assumption that the expert policy is optimal, we encounter the following difficulties:

(i) If we follow the proof and only put rewards on state 3, then the optimal policy would be only choosing the *red* action, hence the posterior uncertainty of v_H would not scale with H since with a single action at state 2 would reveal the whole policy $\pi^{\exp}$;

(ii) If we place rewards on the links from 2 to 3 if $U_i = 1$ and from 2 to 4 if $U_i = 0$, then we can still impose the uniform Bernoulli priors on $\{U_i\}_{i=1}^H$, but the posterior standard deviation of v_H would still be $1/N$ since in this case only state 2 would contribute to the value but its marginal probability is independent of $\pi^{\exp}$ and always $\asymp 1/N$.

These two difficulties beg the question: can we formally prove that for the 4-state MDP instance, and assuming that the expert policy $\pi^{\exp}$ is optimal on the underlying reward function r, can we show that the imitation gap $\mathsf{Gap}(\widehat{\pi}) \lesssim \widetilde{O}(H/N)$ with high probability?

The crucial observation we make here, is that it suffices to find a policy $\widehat{\pi}$ such that its *expected* imitation gap is small to guarantee imitation gap small with constant probability. Indeed, if we can show $\mathbb{E}[\mathsf{Gap}(\widehat{\pi})]$ is small, it immediately implies a concentration bound on $\mathsf{Gap}(\widehat{\pi})$ by an application of Markov's inequality, using the assumption that $\pi^{\exp}$ is optimal on the underlying reward function r. This is not possible when the expert is an arbitrary policy, since $\mathsf{Gap}(\cdot)$ is not a non-negative random variable.

To achieve small expected imitation gap, since r is deterministic, it suffices to find some policy $\widehat{\pi}$ whose expected state-action occupancy measure is close to that of the expert policy. We remark that the unbiased *estimation* of the probability $\Pr_{\pi^{\exp}}(s_H = s^\star)$ is in fact trivial and achieved by the empirical distribution of the state $s^\star$; however, our target of *realization* of this estimated distribution is much more difficult since this requires showing the existence of a policy $\widehat{\pi}$ with small bias. For example, one of the key challenges in proving Theorems 4.1.1 and 4.2.1 is that the empirical distribution of the state $s^\star$ may not be achievable by any policy owing to the possibly limited approximation power of the MDP. The next theorem shows we can solve the 4-state MDP instance with nearly linear dependence on H.

Theorem 4.1.1. *In the known-transition setting, there exists an efficient learning rule (returning policy $\widehat{\pi}$) such that for the family of 4-state MDP instances,*

$$\mathsf{Gap}(\widehat{\pi}) \lesssim \frac{H \log(NH)}{N} \tag{4.1}$$

with probability 0.99 for any ground truth reward r such that $\pi^{\exp}$ is optimal.

We defer the proof of Theorem 4.1.1 to Appendix C.1, but present the explicit policy construction for the single state 3 at layer H here, which conveys the key insights of the algorithm. Let X_t be the number of trajectories in which the expert visits state 2 at time t in the dataset; the X_t's jointly follow a multinomial distribution. Consider the learning rule which returns the policy $\widehat{\pi}$ with,

$$\widehat{\pi}_t(\text{red} \mid 2) = \begin{cases} \frac{\sum_{i=1}^{H-1} X_i U_i}{\sum_{i=1}^{H-1} X_i}, & \text{if } \sum_{i=1}^{H-1} X_i > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{4.2}$$

Note that this policy can be computed since at any time t at which state 2 was not visited in the dataset (i.e. U_t is unknown), $X_t = 0$. Let's work assuming the size of the demonstration dataset $n \sim \text{Poi}(N/2)$, a trick referred to as Poissonization, which enables certain random variables in the analysis to effectively be decoupled. This is permissible since $n \leq N$ with very high probability ($\geq 1 - e^{-3N/16}$ using Poisson tail bounds). Under this assumption,

X_t's are distributed independently as $\mathrm{Poi}\left(\frac{N}{2}\Pr_{\pi^{\exp}}(s_t=2)\right)$. Using the property that for $X\sim\mathrm{Poi}(\mu)$ and independent $Y\sim\mathrm{Poi}(\lambda)$, $\mathbb{E}\left[X/(X+Y)\mid X+Y>0\right]=\mu/(\mu+\lambda)$, we have

$$\mathbb{E}\left[\frac{\sum_{t=1}^{H-1}X_tU_t}{\sum_{t=1}^{H-1}X_t}\;\middle|\;\sum_{t=1}^{H}X_t>0\right]=\frac{\sum_{t=1}^{H-1}\Pr_{\pi^{\exp}}(s_t=2)U_t}{\sum_{t=1}^{H-1}\Pr_{\pi^{\exp}}(s_t=2)},$$

Finally, observe that $\Pr_{\pi^{\exp}}(\sum_{t=1}^{H-1}X_t=0)=\prod_{t=1}^{H-1}\Pr_{\pi^{\exp}}(X_t=0)=e^{-\frac{N}{2}\sum_{t=1}^{H-1}\Pr_{\pi^{\exp}}(s_t=2)}$. Therefore,

$$\left|\Pr_{\pi^{\exp}}(s_H=3)-\mathbb{E}\left[\Pr_{\widehat{\pi}}(s_H=3)\right]\right|\le\Pr_{\pi^{\exp}}\left(\sum_{t=1}^{H-1}X_t=0\right)\sum_{t=1}^{H-1}\Pr_{\pi^{\exp}}(s_t=2)\lesssim\frac{1}{N},$$

since $\sup_x xe^{-tx}=1/(et)$ for any $t>0$.
We remark that (4.2) is carefully constructed such that $\widehat{\pi}_t(\text{red}\mid 2)\in[0,1]$ almost surely to guarantee it is a valid policy, and many natural approaches such as replacing the denominator with the expectation of $\sum_{t=1}^{H-1}X_t$ does not achieve this goal.

4.2 Matching a single state with no error compounding

The proof of Theorem 4.1.1 crucially relies on obtaining a policy whose expected state visitation probability at state 3 of the terminal layer is nearly the same as that of the expert. Can we generalize it to *arbitrary* MDPs and an *arbitrary* target state? The following theorem answers this question affirmatively.

Theorem 4.2.1. *In the known-transition setting, fix any state $s^\star$ at time t of any MDP $\mathcal{M}$. Consider a deterministic expert policy $\pi^{\exp}$, a demonstration dataset D of N trajectories drawn from the expert policy, and any subset $\mathcal{S}_0\subseteq\cup_{t=1}^{H}\mathcal{S}_t$ of states at which the expert actions are known. Let $\Pi^{\mathsf{BC}}(\mathcal{S}_0)\supseteq\Pi^{\mathsf{BC}}(D)$ be the set of policies that mimic the expert action on all states of $\mathcal{S}_0$, there exists a learning rule $\mathsf{Alg}(\cdot)$ (Algorithm 4) returning policies $\widehat{\pi}\in\Pi^{\mathsf{BC}}(\mathcal{S}_0)$ such that*

$$\left|\mathbb{E}[\Pr_{\widehat{\pi}}(s_t=s^\star)]-\Pr_{\pi^{\exp}}(s_t=s^\star)\right|\lesssim\frac{1}{N}.$$

The main message of Theorem 4.2.1 is that, in the known transition setting, there is *no error compounding* for achieving a near-unbiased *realization* of the probability of any single state. Specifically, the upper bound $O(1/N)$ in Theorem 4.2.1 crucially does not depend on H, which is in sharp contrast to the unknown transition setting where the error is $\Theta(H/N)$, as well as the known transition setting but with an absolute error $\Theta(\sqrt{H}/N)$. The construction of the policy $\widehat{\pi}$ relies on a mixture of two deterministic policies inside $\Pi^{\mathsf{BC}}(\mathcal{S}_0)$, where the choice of the mixing coefficient is much more complicated than that in the proof of Theorem 4.1.1 requires a careful inductive procedure detailed later. We also note that the choice of the

subset $\mathcal{S}_0$ is arbitrary, and Theorem 4.2.1 holds even if $\mathcal{S}_0 = \emptyset$; the reason why we introduce $\mathcal{S}_0$ is to show that the near-unbiased realization does not require a costly coordination among all states, and it could always be achieved by properly specifying the actions for a possibly small number of unvisited states.

Algorithm 4 `Mimic-Mixture`

1: **Input:** Demonstration dataset D, states $\mathcal{S}_0$ with known expert action, target state $s^\star$ at time t

2: Compute the following two policies π^{L} and π^{S} based on the known transitions:

$$\pi^{\mathrm{L}} = \arg\max_{\pi \in \Pi^{\mathsf{BC}}(\mathcal{S}_0)} \Pr_{\pi}(s_t = s^\star), \quad \pi^{\mathrm{S}} = \arg\min_{\pi \in \Pi^{\mathsf{BC}}(\mathcal{S}_0)} \Pr_{\pi}(s_t = s^\star). \tag{4.3}$$

3: Draw $n \sim \mathsf{Poi}(N/2)$, and return an arbitrary policy $\widehat{\pi}$ if $n > N$.

4: For every possible trajectory $\mathsf{tr} = (s_1, \cdots, s_H) \in \mathcal{S}^H$, count its number of appearances $X(\mathsf{tr})$ from the first n trajectories in the demonstration dataset.

5: For each $\mathsf{tr} \in \mathcal{S}^H$, compute $\beta^{\mathrm{L}}(\mathsf{tr})$, $\beta^{\mathrm{S}}(\mathsf{tr})$ and $\beta^\star(\mathsf{tr})$ according to Lemma 4.2.2.

6: Subsample each $X(\mathsf{tr})$ independently with probability $\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr})$ to obtain $Y(\mathsf{tr})$.

7: Subsample each $Y(\mathsf{tr})$ independently with probability $(\beta^\star(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))/(\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))$ to obtain $Z(\mathsf{tr})$.

8: Compute the mixing coefficient

$$\widehat{\alpha} = \frac{\sum_{\mathsf{tr} \in \mathcal{S}^H} Z(\mathsf{tr})}{\sum_{\mathsf{tr} \in \mathcal{S}^H} Y(\mathsf{tr})}. \tag{4.4}$$

If the denominator is zero, return any $\widehat{\alpha} \in [0, 1]$.

9: **Return** a randomized policy $\widehat{\pi} = \widehat{\alpha}\pi^{\mathrm{L}} + (1 - \widehat{\alpha})\pi^{\mathrm{S}}$.

The construction of the learner's policy $\widehat{\pi}$ is summarized by `Mimic-Mixture` in Algorithm 4. The idea is to find two extremal policies, i.e. policies π^{L} and π^{S} which maximize and minimize the induced probability of the target state $s^\star$ among all policies in $\Pi^{\mathsf{BC}}(\mathcal{S}_0)$, respectively (cf. eq. (4.3)), and choose the learner's policy $\widehat{\pi}$ as a proper mixture of these extremal policies, i.e. $\widehat{\pi} = \widehat{\alpha}\pi^{\mathrm{L}} + (1 - \widehat{\alpha})\pi^{\mathrm{S}}$. Since $\pi^{\mathrm{L}}, \pi^{\mathrm{S}} \in \Pi^{\mathsf{BC}}(\mathcal{S}_0)$, it is clear that the mixture $\widehat{\pi}$ also belongs to $\Pi^{\mathsf{BC}}(\mathcal{S}_0)$. As the learner's target is to match the expert probability $\Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star)$, the ideal choice of $\widehat{\alpha}$ would be

$$\alpha^\star = \frac{\Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)}{\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)}, \tag{4.5}$$

which by definition of $\pi^{\mathrm{L}}, \pi^{\mathrm{S}}$ always lies in $[0, 1]$. Note that the only unknown quantity in eq. (4.5) is the probability $\Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star)$ induced by the unknown expert policy, we need to replace this probability by a proper estimator. The most natural approach is to use the empirical version of $\Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star)$, which is an unbiased estimator. However, plugging this empirical version into eq. (4.5) may make the final ratio $\alpha^\star$ outside $[0, 1]$, giving an invalid

mixture policy $\widehat{\pi}$; a naïve truncation of $\alpha^\star$ to $[0,1]$ will also incur a too large bias (of the order $\Omega(\sqrt{H}/N)$), for the truncation operation is similar in spirit to the minimum distance projection used in `Mimic-MD`.

To circumvent this difficulty, our idea is to replace all probabilities $\Pr_{\pi^{\exp}}(s_t = s^\star)$, $\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star)$, $\Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)$ in eq. (4.5) by appropriate estimates such that the ratio lies in $[0,1]$ almost surely, even if the latter two probabilities are in fact perfectly known and thus do not require any estimation in principle. To construct these estimators, we consider a Poissonized sampling model as follows: draw an independent Poisson random variable $n \sim \mathsf{Poi}(N/2)$, which does not exceed N with probability at least $1 - e^{-\Omega(N)}$ by the Chernoff bound. For each possible state trajectory $\mathsf{tr} = (s_1, \cdots, s_H) \in \mathcal{S}^H$, define $X(\mathsf{tr})$ to be the total count of this trajectory in the first n trajectories of D:

$$X(\mathsf{tr}) = \sum_{i=1}^{n} \mathbb{1}(\mathsf{tr}_i = \mathsf{tr}).$$

Note that the sample size in the above count is a Poisson random variable $n \sim \mathsf{Poi}(N/2)$, instead of the fixed number N. The advantage of the Poisson sampling is that, the above count $X(\mathsf{tr})$ exactly follows a Poisson distribution $\mathsf{Poi}(N/2 \cdot \Pr_{\pi^{\exp}}(\mathsf{tr}))$, and these counts $\{X(\mathsf{tr})\}$ for different trajectories are mutually independent. We apply the following linear estimators for the probabilities in eq. (4.5):

$$\begin{aligned} \widehat{\Pr}_{\pi^{\exp}}(s_t = s^\star) &\triangleq \frac{2}{N} \sum_{\mathsf{tr} \in \mathcal{S}^H} \beta^\star(\mathsf{tr}) \cdot X(\mathsf{tr}), \quad \widehat{\Pr}_{\pi^{\mathrm{L}}}(s_t = s^\star) \triangleq \frac{2}{N} \sum_{\mathsf{tr} \in \mathcal{S}^H} \beta^{\mathrm{L}}(\mathsf{tr}) \cdot X(\mathsf{tr}), \\ \widehat{\Pr}_{\pi^{\mathrm{S}}}(s_t = s^\star) &\triangleq \frac{2}{N} \sum_{\mathsf{tr} \in \mathcal{S}^H} \beta^{\mathrm{S}}(\mathsf{tr}) \cdot X(\mathsf{tr}), \end{aligned} \tag{4.6}$$

where $\beta^\star(\mathsf{tr}), \beta^{\mathrm{L}}(\mathsf{tr}), \beta^{\mathrm{S}}(\mathsf{tr}) \in [0,1]$ are appropriate coefficients to be specified later. We require the following three properties for these coefficients:

1. *Unbiasedness.* the coefficients should be chosen so that the estimators in eq. (4.6) are unbiased in estimating the corresponding true probabilities. Mathematically, we require that $\sum_{\mathsf{tr} \in \mathcal{S}^H} \beta^\dagger(\mathsf{tr}) \cdot \Pr_{\pi^{\exp}}(\mathsf{tr}) = \Pr_{\pi^\dagger}(s_t = s^\star), \quad \dagger \in \{*, \mathrm{L}, \mathrm{S}\}$.

2. *Order.* for every trajectory $\mathsf{tr} \in \mathcal{S}^H$, it holds that $\beta^{\mathrm{S}}(\mathsf{tr}) \le \beta^\star(\mathsf{tr}) \le \beta^{\mathrm{L}}(\mathsf{tr})$. This requirement ensures that plugging eq. (4.6) into eq. (4.5) always gives a ratio in $[0,1]$.

3. *Feasibility.* this requirement is a bit subtle. We require that all coefficients $\beta^\star(\mathsf{tr})$, $\beta^{\mathrm{L}}(\mathsf{tr})$, and $\beta^{\mathrm{S}}(\mathsf{tr})$ only depend on public information (known transition probabilities, initial distribution, expert actions at states in $\mathcal{S}_0$, policies $\pi^{\mathrm{L}}, \pi^{\mathrm{S}}$, and $s^\star$) and the private information associated with tr (expert actions at states visited in trajectory tr). Importantly, these coefficients cannot depend on expert actions not in $\mathcal{S}_0 \cup \mathsf{tr}$, as those actions may not be observable to the learner, leaving the coefficients not always well-defined. In contrast, dependence on the expert actions at states in tr is feasible, for

these actions are observed if $X(\mathsf{tr}) > 0$, and the coefficients could be arbitrarily chosen with $\beta^{\dagger}(\mathsf{tr}) \cdot X(\mathsf{tr}) \equiv 0$ if $X(\mathsf{tr}) = 0$, for $\dagger \in \{*, \mathrm{L}, \mathrm{S}\}$.

The following lemma shows that we can indeed construct coeffs. $\{\beta^{\star}(\mathsf{tr})\}, \{\beta^{\mathrm{L}}(\mathsf{tr})\}, \{\beta^{\mathrm{S}}(\mathsf{tr})\}$ satisfying the above three requirements, and they can be used to construct a policy such that Theorem 4.2.1 holds.

Lemma 4.2.2. *There exist coefficients $\beta^{\star}(\mathsf{tr}), \beta^{\mathrm{L}}(\mathsf{tr}), \beta^{\mathrm{S}}(\mathsf{tr}) \in [0,1]$ such that all of the unbiasedness, order, and feasibility properties hold. Furthermore, given any such coefficients, one can efficiently construct a policy $\widehat{\pi}$ such that Theorem 4.2.1 holds.*

The proof of Lemma 4.2.2 is via a careful inductive argument and is deferred to Appendix C.3. Armed with the result of Theorem 4.2.1, we can prove an improved imitation gap bound in 3-state MDPs when rewards are only present in the last layer.

Corollary 4.2.1. *Suppose $|\mathcal{S}| = 3$ and $r_t \equiv 0$ for all $t = 1, 2, \ldots, H-1$, $r_H \in [0,1]$, π^{exp} is optimal for r, and the transitions are known. Then, there exists an efficient algorithm based on `Mimic-MD` and `Mimic-Mixture` such that the imitation gap is upper bounded by $\widetilde{O}(1/N)$ with probability 0.99.*

4.3 Conjectures for the optimal expert setting

Corollary 4.2.1 shows that in case the reward is only on the terminal state for MDPs on 3 states, the optimal imitation gap scales as $\widetilde{\Theta}(1/N)$ when the expert is an optimal policy. In case the expert is not optimal, the rate degrades to $\widetilde{O}(\sqrt{H}/N)$ using the same analysis as in Theorem 3.3.1. In this section, we conjecture an optimal algorithm for Imitation Learning with an optimal expert policy, which is based on using Inverse Reinforcement Learning (IRL) to instantiate the reward family $\mathcal{R}_D$ in (OPT). In particular, we instantiate $\mathcal{R}_D$ as,

$$\mathcal{R}_{\mathrm{opt}}(D) = \{r : \exists \pi \in \Pi^{\mathtt{BC}}_{\mathrm{det}}(D) \text{such that } \pi \text{ is optimal on } r\}, \tag{4.7}$$

as the set of rewards which induce an optimal policy consistent with what is observed of the expert policy. The natural value estimator, $\widehat{J}_r(\pi^{\mathrm{exp}})$ is to choose the value of the optimal policy on a given reward r.

In this section we will conjecture an approach (Algorithm 5) for the optimal expert setting which achieves imitation gap bounded by $\widetilde{O}(H/N)$. While proving this result in its full generality turns out to be quite challenging, we will establish it on a significant generalization of the hard 4-state MDP instances considered in the previous chapter. Our proof relies on proving certain structural properties of the optimal expert policy on these MDP instances. Showing these properties hold for arbitrary MDPs would resolve a major open problem.

The remainder of this section is dedicated to providing evidence that this algorithm in fact might achieve the conjectured statistical guarantee of $\widetilde{O}(H/N)$. In particular, we consider a family of 4-state MDPs which generalize the one considered in the lower bound in Theorem 3.3.2 and show that Algorithm 5 in fact achieves this guarantee.

Algorithm 5 Conjectured Optimal Algorithm

1: **Input:** Expert dataset D,
2: Define $\Pi^{\texttt{BC}}_{\text{det}}(D)$ as in Eq. (2.6) ▶ Policies which mimics expert on states visited in D
3: Define $\mathcal{R}_{\text{opt}}(D)$ as in eq. (4.7) ▶ Set of rewards such that there exists an optimal policy on this reward belonging to $\Pi^{\texttt{BC}}_{\text{det}}(D)$. This corresponds to Inverse Reinforcement Learning.
4: Define the learner's policy as the solution to the minimax optimization problem,

$$\widehat{\pi} = \min_{\pi \in \Pi^{\texttt{BC}}_{\text{det}}(D)} \max_{r \in \mathcal{R}_{\text{opt}}(D)} \left(\max_{\pi'} J_r(\pi') \right) - J_r(\pi) \tag{4.8}$$

▶ This corresponds to using the value estimation framework in (OPT) with $\widetilde{J}_r(\pi^{\text{exp}}) = \max_{\pi'} J_r(\pi')$.

A generalized family of 4-state MDPs

Consider a family of MDPs on 4 states structured as in fig. 4.1. This is an extension of the 4-state MDP in fig. 3.1 with the probability $1/N$ of transitioning from state 1 to 2 changed to an arbitrary time-varying $p_t > 0$, as well as with the two actions at state 2, a_+ and a_- inducing a general next state distribution supported on states 3 and 4. Likewise, the singular action at states 3 and 4 induces an arbitrary distribution supported on states 3 and 4. The reward function for this MDP is completely arbitrary.

Theorem 4.3.1. *On the family of MDPs depicted in fig. 4.1, Algorithm 5 achieves expected imitation gap upper bounded by* $\widetilde{O}(H/N)$.

The key idea behind analyzing the conjectured optimal algorithm is to show how to construct a reference policy such that its imitation gap on all feasible rewards in $\mathcal{R}_{\text{opt}}(D)$ as compared against any policy in $\Pi^{\texttt{BC}}_{\text{det}}(D)$ is bounded. First we bound the learner's imitation gap in the following lemma.

Lemma 4.3.2. *The imitation gap of the conjectured optimal algorithm in Algorithm 5 is bounded by,*

$$\mathit{Gap}(\widehat{\pi}) \leq \max_{r \in \mathcal{R}_{opt}(D)} \left(\max_{\pi' \in \Pi^{\mathit{BC}}_{\text{det}}(D)} J_r(\pi') \right) - J_r(\pi_{ref}) \tag{4.9}$$

for any reference policy $\pi_{ref} \in \Pi^{\mathit{BC}}_{\text{det}}(D)$.

Proof. Observe that since the ground truth reward $r \in \mathcal{R}_{\text{opt}}(D)$,

$$\begin{aligned}
\mathsf{Gap}(\widehat{\pi}) &\leq \max_{r \in \mathcal{R}_{\text{opt}}(D)} J_r(\pi^{\text{exp}}) - J_r(\widehat{\pi}) \\
&\leq \max_{r \in \mathcal{R}_{\text{opt}}(D)} \left(\max_{\pi' \in \Pi^{\mathsf{BC}}_{\text{det}}(D)} J_r(\pi') \right) - J_r(\widehat{\pi}) \\
&\leq \max_{r \in \mathcal{R}_{\text{opt}}(D)} \left(\max_{\pi' \in \Pi^{\mathsf{BC}}_{\text{det}}(D)} J_r(\pi') \right) - J_r(\pi_{\text{ref}})
\end{aligned}$$

where the second inequality uses the definition of $\mathcal{R}_{\text{opt}}(D)$, and the last inequality uses the fact that $\widehat{\pi}$ is the minimizer of the objective in eq. (4.8). □

In order to use this inequality, we show how to construct a reference policy such that on any reward $r \in \mathcal{R}_{\text{opt}}(D)$, the value achieved by this policy, in expectation, is at most $\widetilde{O}(H/N)$ suboptimal compared to the optimal policy on that reward function. The key idea in constructing this reference policy is to notice that whenever the next-state distribution at a state can vary significantly across actions, observing the action at that state provides a lot of information to the learner about value functions across various actions. However, picking the wrong action at these states might also induce suboptimality. On the other hand, when the next-state distribution at a state does not vary significantly across actions, the opposite happens - playing the wrong action at this state does not significantly hurt the learner, but observing the action at this state is also not very informative. Therefore, the strategy to construct the reference policy will be to combine information across various actions at the same level of "informativeness" to balance the risk of picking wrong actions. For the family of instances we consider, the informativeness of state 2 at any time is evaluated using the metric $|P_t(3|2, a_+) - P_t(3|2, a_-)|$, which is the TV distance between the next-state distributions induced by a_+ and a_- at state 2 at time t.

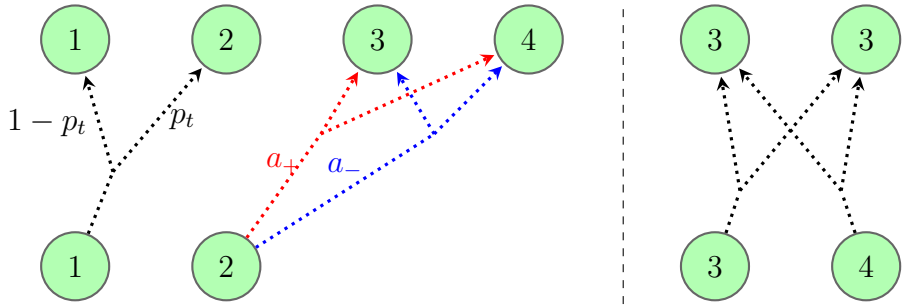


Figure 4.1: The generalized *4-state MDPs*: The states 1, 3 and 4 have a single action. On the other hand, state 2 has 2 actions, $\{a_-, a_+\}$ with next state distribution supported on states 3 and 4. Likewise, states 3 and 4 with next state distribution also supported on states 3 and 4.

Chapter 5

Imitation Learning with active interaction

In this section, we will revisit the topic of IL with active interaction, which we introduced in Chapter 1 and prove a statistical lower bound for in Section 2.2. There, we showed that in the absence of any assumptions there is no statistical benefit to active interaction, compared to algorithms which learn from a static dataset of demonstrations. It begs the question as to why approaches such as DAGGER [74] and AGGREVATE [77] which actively query the expert often perform better than `BC` in practice. To explain this gap our previous results imply that additional assumptions must be imposed.

To motivate this assumption, we turn to the statistical lower bound we prove in Section 2.2 in the active-interaction setting. The key idea in the lower bound is to include an absorbing "bad" state in the MDP which is never visited in the demonstration dataset and offers no reward. Any policy which visits this state is doomed to incur a large reward gap - in the absence of full information, the learner is forced to visit this state often. The lower bound instance is pathological in the sense that even if the expert itself visits the bad state, it is never able to "recover" and return to the remaining states. Indeed in practical situations such as driving a car, experts often can recover and collect a high reward even if a mistake is made locally. The authors of [74] introduce an assumption to this effect, which we refer to as μ-recoverability.

Definition 5.0.1 (μ-recoverability)**.** *An IL instance is said to satisfy μ-recoverability if for each $t \in [H]$ and $s \in \mathcal{S}$, $\mathbb{E}_{a \sim \pi_t^{\exp}(\cdot|s)} \left[Q_t^{\pi^{\exp}}(s,a) \right] - Q_t^{\pi^{\exp}}(s,a) \leq \mu$ for all actions $a \in \mathcal{A}$. Informally, if the expert plays an "incorrect" action at any state s at a single time t and goes back to choosing the correct actions afterwards, the expected reward collected is less by at most μ.*

The μ-recoverability assumption captures the ability of an expert to recover and collect a high reward at a state even upon locally deviating from its action distribution at states. The reduction in [78, Theorem 2] shows that under μ-recoverability, a learner policy $\widehat{\pi}$ which minimizes the 0-1 loss with respect to the expert's policy under the learner's own state distribution. We will define the loss under a generic sequence of state distributions

$f = (f_1, \cdots, f_H)$,

$$\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f) \triangleq \frac{1}{H} \sum_{t=1}^{H} \mathbb{E}_{s \sim f_t} \left[\mathbb{E}_{a \sim \widehat{\pi}_t(\cdot|s)} \left[\mathbb{1}(a \neq \pi_t^{\text{exp}}(s)) \right] \right]. \tag{5.1}$$

And the reduction argues that if $\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f^{\widehat{\pi}}) \leq \epsilon$, then $\mathsf{Gap}(\widehat{\pi}) \leq \mu H \epsilon$. However, it is a-priori unclear how small ϵ can be made as a function of the number of the size of demonstration dataset / number of MDP interactions, N in the no-interaction / active-interaction settings. This is a drawback of the reduction approach followed by [74, 78] since it cannot distinguish between the power of learners in different interaction models. Indeed one challenge for a learner to minimize the 0-1 under its own state distribution is that the learner's policy changes over the course of optimization.

In this chapter, we will propose a learner in the active-interaction setting with expected 0-1 loss under the learner's own state distribution bounded by $|\mathcal{S}|/N$. This "completes" the reduction in a sense, and establishes imitation gap bounds for the active setting as an explicit function of the number of states $|\mathcal{S}|$, interactions N and horizon H.

Theorem 5.0.1. *In the active-interaction setting there exists a learning rule such that the resulting policy $\widehat{\pi}$ satisfies, $\mathbb{E}[\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f^{\widehat{\pi}})] \lesssim |\mathcal{S}|/N$. Furthermore, under μ-recoverability, $\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \lesssim \mu|\mathcal{S}|H/N$.*

The proof of this result utilizes the no-regret reduction of [78]. Indeed, observe that it suffices for the learner to find a sequence of policies $\widehat{\pi}^1, \cdots, \widehat{\pi}^T$ such that the *online-learning regret*, defined as,

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; f^{\widehat{\pi}^i}) - \min_{\pi} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\pi; f^{\widehat{\pi}^i}) \lesssim \frac{|\mathcal{S}|}{N}. \tag{5.2}$$

is sufficiently small. Note that in eq. (5.2), the oracle loss $\min_\pi \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\pi; f^{\widehat{\pi}^i})$ is in fact 0, achieved by $\pi = \pi^{\text{exp}}$, and so, the mixture policy over the learner's policies, $\frac{1}{N} \sum_{i=1}^{N} \widehat{\pi}^i$ satisfies,

$$\mathcal{L}_{0\text{-}1}\left(\widehat{\pi}; f^{\widehat{\pi}}\right) \lesssim \frac{|\mathcal{S}|}{N}.$$

Note that while the regret being minimized in eq. (5.2) involves losses not observable without full knowledge of π^{exp}, it is possible to compute an unbiased estimate of $\mathcal{L}_{0\text{-}1}\big(\widehat{\pi}^i; f^{\widehat{\pi}^i}\big)$ by rolling out *just a single trajectory*. This is only enabled in the active-interaction model, and is not possible given just a fixed dataset of demonstrations.

Toward this end, suppose for each i, the learner rolls out a single trajectory according to $\widehat{\pi}^i$ and denote the resulting empirical state visitation distribution $\widehat{f^i} = (\widehat{f_1^i}, \cdots, \widehat{f_H^i})$ where $\widehat{f_t^i}$ is the empirical distribution at time t. Observe that,

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; \widehat{f^i})$$

is an unbiased estimate of $\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; f^{\widehat{\pi}^i})$ if $\widehat{\pi}^i$ is a measurable function the first $i-1$ rolled out trajectories (according to $\widehat{\pi}^1, \cdots, \widehat{\pi}^{i-1}$), and concentrates around this via martingale concentration arguments. Thus, it will suffice for the learner to find a sequence of policies $\widehat{\pi}^1, \cdots, \widehat{\pi}^T$ which minimize the *empirical online-learning regret*: $\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; \widehat{f}^i) - \min_\pi \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{0\text{-}1}(\pi; \widehat{f}^i)$ to be $\lesssim \frac{|\mathcal{S}|}{N}$. As we discuss in more detail later in the full proof of this result, it is possible to construct a sequence of policies $\widehat{\pi}^1, \cdots, \widehat{\pi}^N$ using entropy-regularized mirror descent [85] which minimizes the empirical online-learning regret to be $\lesssim |\mathcal{S}|/N$. The resulting policy $\widehat{\pi} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}^i$ minimizes the expected 0-1 loss under its own state distribution to be $\lesssim |\mathcal{S}|/N$ in expectation. The guarantee on the expected imitation gap of this policy directly follows from [78, Theorem 2] under μ-recoverability.

This guarantee on the imitation gap is optimal for any learner in the active-interaction setting assuming μ-recoverability. Optimality essentially follows from the lower bound we proved in the active interaction setting in Theorem 2.2.1, where if $N \geq |\mathcal{S}|H$, the expected imitation gap incurred is $\gtrsim \frac{|\mathcal{S}|H^2}{N}$. By scaling each reward by a factor of μ/H, the same family of IL instances now satisfies μ-recoverability and results in the lower bound for active learners.

Theorem 5.0.2 (Corollary of Theorem 2.2.1)**.** *Suppose $N \geq |\mathcal{S}|H$. For every learning rule in the active-interaction setting, there exists an IL instance such that the resulting policy $\widehat{\pi}$ incurs expected imitation gap $\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \gtrsim \min\{\mu, \mu|\mathcal{S}|H/N\}$ under some worst-case tabular IL instance.*

Now, under the same μ-recoverability assumption, we study learners in the no-interaction setting. We prove a lower bound that in the worst case, showing that even though there exist actions that allow a learner to recover at pathological states, error compounding is unavoidable for such learners.

Theorem 5.0.3. *Suppose $|\mathcal{S}| \geq 3$ and $|\mathcal{A}| \geq H$. For any learning rule in the no-interaction setting, there exists an IL instance which satisfies μ-recoverability for $\mu \geq 1$, and such that the resulting policy $\widehat{\pi}$ incurs expected imitation gap lower bounded by, $\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \gtrsim \min\{H, |\mathcal{S}|H^2/N\}$.*

This is the first result to establish a clear *separation* in the statistical minimax rate of the imitation gap incurred by learners in the no-interaction setting such as `BC`, and learners which can interact with the MDP, such as DAGGER [78] and AGGRAVATE [77]. The instances we construct in this lower bound are a modification of those considered in Theorem 2.2.1. There, the MDPs considered have a "bad" state in the MDP never visited by the expert. We modify the instance to add a single "recovery" action at the bad state; the instance now satisfies μ-recoverability for any $\mu \geq 1$. If the number of actions are large $|\mathcal{A}| \geq H$, any no-interaction learner still fails to identify the recovery action with constant probability. In essence this reduces the instance to the lower bound considered in Theorem 2.2.1 and any no-interaction learner is forced to incur expected imitation gap $\gtrsim \min\{H, |\mathcal{S}|H^2/N\}$.

Chapter 6

Toward a practical algorithm

Many practically performant algorithms in the IL literature fall under the empirical moment matching framework (e.g. GAIL [44], MaxEnt IRL [110]); see Table 3 of [93] for more examples. Reward moment matching corresponds to finding a policy which best matches the state-action visitation measure of π^{exp}, in the sense of minimizing an Integral Probability Metric (IPM) [57]. Formally, the learning rule takes the form,

$$\min_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi}[f] - \mathbb{E}_{\pi^{\text{exp}}}[f], \tag{6.1}$$

f is referred to as the "moment" here, and is simply a discriminator which tries to distinguish between rollouts under π and π^{exp}.
This formulation is void of any statistical considerations, indeed it invokes the expectation $\mathbb{E}_{\pi^{\text{exp}}}[f]$ which cannot be computed by the learner. This includes the algorithm we discuss in Chapter 3, namely (OPT), which reduces the IL problem in the known-transition setting to the uniform expert value estimation problem (Definition 3.2.2). Recall, which is defined as returning any minimizer of

$$\min_{\pi} \max_{r \in \mathcal{R}_D} \widetilde{J}_r(\pi^{\text{exp}}) - J_r(\pi).$$

This is but a finite sample implementation of eq. (6.1), where f is the cumulative reward of trajectory, and we replace $\mathbb{E}_{\pi^{\text{exp}}}[f]$ by an estimator $\widetilde{J}_r(\pi^{\text{exp}})$. The simplest approach toward writing down a D-measurable optimization problem is to replace $\mathbb{E}[f]$ by an empirical estimate $\mathbb{E}_D[f]$, where $\mathbb{E}_D[\cdot]$ indicates an expectation computed over a random trajectory drawn from the demonstration dataset D.

Definition 6.0.1. *The empirical moment matching learner $\pi^{\texttt{MM}}$ attempts to best match the empirical state-visitation measure under a set of discriminators $\mathcal{F}$. Namely,*

$$\pi^{\texttt{MM}} \in \arg\min_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi}\left[\frac{\sum_{t=1}^{H} f_t(s_t, a_t)}{H}\right] - \mathbb{E}_{D}\left[\frac{\sum_{t=1}^{H} f_t(s_t, a_t)}{H}\right]. \tag{6.2}$$

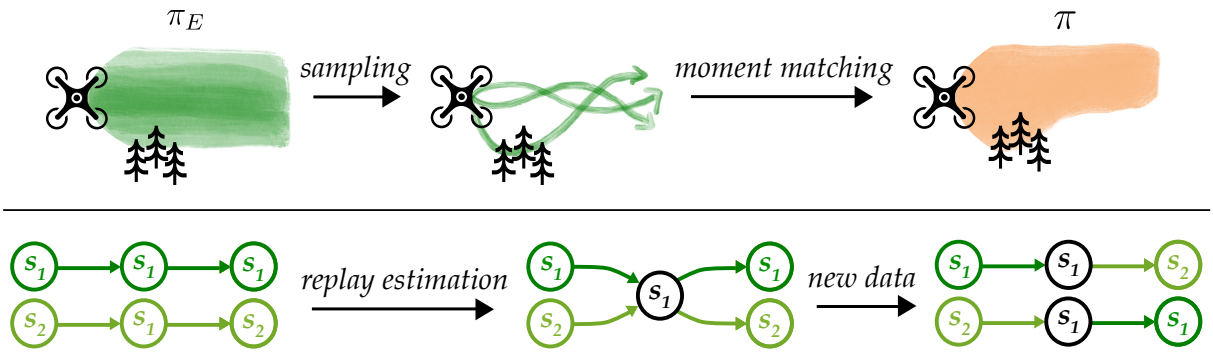


Figure 6.1: *Top*: Attempting to exactly match a finite-sample approximation of expert moments can cause a learner to reproduce chance occurrences (e.g. the relatively unlikely flight through the trees). This can lead to policies that perform poorly at test time (e.g. because the learner flies through the trees relatively often). *Bottom*: Replay estimation reduces the empirical variance in expert demonstrations by repeatedly executing observed expert actions in a stochastic simulator. By generating new trajectories (e.g. $s_1 \to s_1 \to s_2$ on the right) that are consistent with expert actions, one can augment the original demonstration set and compute expert moments more accurately.

When a small set of demonstrations are used within the trained objective, the learner may choose to take incorrect actions in order to match the noisy moments estimated from the dataset, leading to policies that perform poorly at test-time.

One solution to this problem is to query the expert to generate more demonstrations in an online / interactive fashion, as discussed in Chapter 2 [76]. However, when we are unable to do so, we still have to grapple with the practical question of "*how can we smooth out a noisy empirical estimate of moments f?*"

In this chapter, we propose a practical algorithm, *Replay Estimation* (`RE`) which builds upon the ideas in eq. (OPT) and Algorithm 2 to result in a performant algorithm for IL when given access to a simulator. In its most basic form, `RE` consists of repeatedly executing observed expert actions within a stochastic simulator, terminating rollouts whenever one ventures out of the support of the expert demonstrations. Effectively, this approach stitches together parts of different trajectories to generate a smoothed estimate of expert moments. By using the simulator where we know the expert's actions, we can generate more diverse training data that is nevertheless consistent with the expert demonstrations.

6.1 Suboptimality of `MM` and `BC`

In this section, we will do a deeper dive into the suboptimality of `BC` and the empirical moment matching (`MM`) introduced previously. We begin with a vignette to illustrate some

key issues in greater detail.

Suboptimality of MM Consider the MDP in Figure 6.2b, where the expert always takes the green action. Doing so puts them in s_1 or s_2 with equal probability. Given that the expert is deterministic and there are few states, BC could easily recover the expert's policy by learning to simply output the observed green action on both states, even when there are very few demonstrations.

Now, what would happen if we tried to match moments of the expert's state-action visitation distribution for this problem? It is rather unlikely that we see *exactly* equal probabilities for both states in the observed data. If by chance we see s_2 more than we see s_1, the learner might realize that the only way to match the observed state distribution (a prerequisite for matching the observed state-action distribution) is to occasionally take the red action at s_2. In general, this could cause the learner to spend an unnecessary amount of time in s_2 which may be undesirable (e.g. if s_2 corresponds to the tree-filled area in Figure 6.1 (top)). The core issue we hope to illustrate in this example is that by treating the empirical estimate of the expert's behavior as perfectly accurate, distribution matching can force the learner to take incorrect actions to minimize training error, leading to test-time performance degradation. This can lead to slow statistical rates $\propto H/\sqrt{N}$.

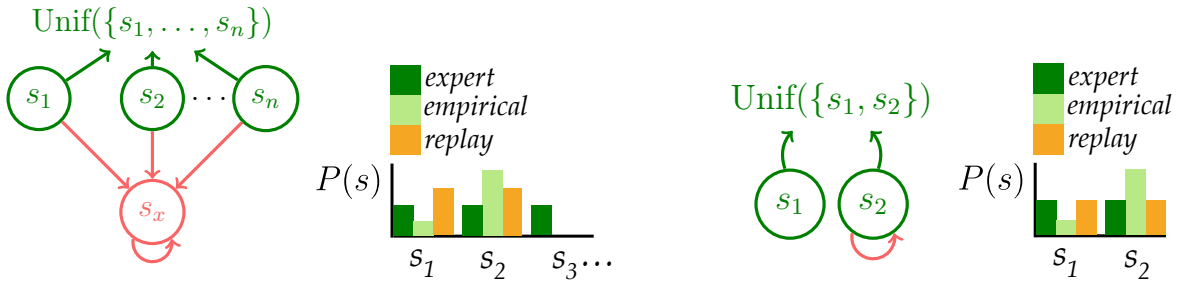
Theorem 6.1.1. *If $H \geq 4$, there is a tabular IL instance on an MDP with 2 states and actions on which with constant probability, the empirical moment matching learner (Definition 6.0.1) incurs, $\mathsf{Gap}(\pi^{\texttt{MM}}) \gtrsim H/\sqrt{N_{\text{exp}}}$.*

The proof of this result is deferred to appendix E.2. The proof of this lower bound exploits the fact that the data generation process in the dataset is inherently random. Consider a slight modification of the MDP instance shown in fig. 6.2b, where the reward function is 0 for $t = 1$. For $t \geq 2$, the transition function is absorbing at both states; the reward function equals 1 at the state s_1 for any action and is 0 everywhere else. Then, the expert state distribution at time 2 and every time thereon is in uniform across the two states, $\{1/2, 1/2\}$. However, in the dataset D, the learner sees a noisy realization of this distribution in the dataset of the form $\{1/2 - \delta, 1/2 + \delta\}$ for $|\delta| \approx \pm 1/\sqrt{N_{\text{exp}}}$. Because of this noise, the empirical moment matching learner may be encouraged to *deviate from the expert's observed behavior* and pick the red action at s_2 as this results in a better match to the empirical *state* visitation measures at every point in the rest of the episode - a prerequisite to matching the empirical state-action visitation measure. The learner is willing to pick an action different from what the expert played in order to better match the *inherently noisy* empirical state-action visitation distribution.

Remark 6.1.1. Theorems 2.2.1 and 6.1.1 are separate lower bound IL instances against the performance of BC and empirical moment matching. On the uniform mixture of the two MDPs (i.e. deciding the underlying MDP based on the outcome of a fair coin), with constant probability, *both* $\mathsf{Gap}(\pi^{\texttt{BC}}) \gtrsim |\mathcal{S}|H^2/N_{\text{exp}}$ and $\mathsf{Gap}(\pi^{\texttt{MM}}) \gtrsim H/\sqrt{N_{\text{exp}}}$. On this

mixture instance, training both BC and empirical moment matching and choosing the better of the two is also statistically suboptimal.

Suboptimality of BC. Because it does not account for the covariate shift that results from policy action choices, behavioral cloning can lead to a quadratic compounding of errors and poor test time performance [76]. Consider, for example, the MDP in Figure 6.2a.



(a) The expert always takes the green action, which places it in a uniform distribution over $s_1, \ldots, s_n$. At states where we have demonstrations (e.g. s_1, s_2), both BC and MM will take the same, correct action. However, at states where we have no demonstrations (e.g. s_3), MM will correctly take the green action to get back to states with demonstration support, while BC might not.

(b) An MDP where the expert always takes the green action that puts them in the uniform distribution over s_1 and s_2. Because of full expert support, BC will learn to always take this action at both states. However, if the empirical state distribution is more tilted towards s_2, MM will take the incorrect red action.

Figure 6.2: A deeper dive into the suboptimality of BC and MM

Let us assume that the expert always takes the green action, dropping them in a state in the top row with uniform probability. In a small demonstration set, we might not see expert actions at some states in the top row. At all such states, BC will have no idea of what to do. In contrast, MM will take the green action as doing so might send the learner back to a state with positive demonstration support. Thus for this problem, MM will recover the optimal policy while BC will not. As we saw in Chapter 2, this leads to errors $\propto H^2/N$ in the worst case.

6.2 The Replay Estimator

The previous two examples show us that there exist simple MDPs for which BC or MM will not recover the expert's policy. This begs the question: is it possible to do *better than both worlds* and recover the optimal policy on both problems with a single algorithm? We answered this question in the tabular setting in Theorem 2.4.1 via Mimic-MD (Algorithm 2) which achieves better performance than BC. We will establish a result showing that it outperforms MM later in

this section. This improvement is possible because BC does not use any dynamics information and MM does not leverage the knowledge of where expert actions are known. However, it is unclear how to extend this approach beyond the tabular setting as the algorithm relies on a large measure of states being visited in the demonstrations.

It turns out it is indeed possible to do so, via the technique of *replay estimation* (RE). In its simplest form, RE builds on top of the insights of Mimic-MD (Algorithm 2) and involves exploring in the environment by playing a cached expert's action whenever possible and re-starting the rollout if one ventures out of the support of the demonstration dataset. Then, one appends these rollouts to the demonstration set, treating them as additional training data – while biased, they are consistent with observed expert behavior. Intuitively, repeated simulation has a *smoothing* effect on the training data as doing so marginalizes out the statistical error that comes from the stochasticity of the dynamics. We can see this point more explicitly by considering the above two MDP examples: in Figure 6.2b, repeatedly playing the green action and appending these rollouts to the demonstration dataset would bring us much closer to a uniform distribution over s_1 and s_2. Similarly, in Figure 6.2a, replay estimation would bring us toward a uniform distribution over the states $\{s_1, \cdots, s_n\}$ in the expert demonstrations.

We could then plug in this improved distribution estimate into the MM procedure eq. (6.1). Notice how doing so would cause MM to be highly likely to recover the optimal policy on both MDPs. For example, in Figure 6.2b, replay estimation would make the learner much less likely to play the red action in s_2. This fact is sufficient to establish *statistical optimality* in the tabular setting and with linear function approximation, with an error rate $\propto \min(H^{3/2}/N, H/\sqrt{N})$. In short, replay estimation is a practical technique for reducing some of the finite-sample variance in expert demonstrations that enables MM to perform optimally in the finite sample regime. We now provide some intuition on how to generalize this approach to beyond the tabular setting.

Leaving the Tabular Setting. Mimic-MD was introduced in the tabular setting; this characteristic property of the setting makes it easy to answer the question of "*on what states do we know the expert's action?*", since one can enumerate over all states efficiently. To enable us to answer this question in more general settings (with infinite state spaces), we introduce the notion of a *membership oracle* $\texttt{MEM} : \mathcal{S} \to \{0, 1\}$. Explicitly, $\texttt{MEM}(s) = 1$ for states where we know the expert's action well (e.g. states where we have lots of similar demonstrations) and $\texttt{MEM}(s) = 0$ otherwise. We can then compute expert moments by splitting on the output of the membership oracle:

$$\mathbb{E}_{\pi^{\text{exp}}}[f(s,a)] = \underbrace{\mathbb{E}_{\pi^{\text{exp}}}\left[f(s,a)\mathbb{1}(\texttt{MEM}(s)=1)\right]}_{(i)} + \underbrace{\mathbb{E}_{\pi^{\text{exp}}}\left[f(s,a)\mathbb{1}(\texttt{MEM}(s)=0)\right]}_{(ii)} \tag{6.3}$$

Note that the indicators in (i) and (ii) are complements of each other, rendering the above sum a valid estimate of the expert moment. As we know the expert action well wherever

$\texttt{MEM}(s) = 1$, simulated rollouts of the BC policy approximates (i) well; on the other hand we resort to a naive empirical estimate to approximate (ii), as we do not know enough about the expert's action at these states to accurately generate additional demonstrations via BC rollouts. In general, we relax MEM to a *soft membership oracle* [106], in order to handle uncertainty in how well we know the expert's action at a given state. We proceed by first analyzing the statistical properties of applying MM to this bipartite estimator before discussing practical constructions of performant membership oracles.

Algorithm 6 Replay Estimation (RE)

1: **Input:** Expert demonstrations D, policy class Π, moment class $\mathcal{F} = \bigoplus_{t=1}^{H} \mathcal{F}_t$, simulator SIM, TRAIN which returns a membership oracle given a dataset
2: Partition the dataset D into D_1 and D_2
3: Using TRAIN, learn a membership oracle MEM on D_1
4: Train $\pi^{\texttt{BC}}$ using behavior cloning on D_1
5: Roll out $\pi^{\texttt{BC}}$ in SIM N_{replay} times to construct a new dataset, D_{replay}
6: Define prefix weights $\mathcal{P}(s_{1\ldots t-1}) = \prod_{t'=1}^{t-1} \texttt{MEM}(s_{t'}, t')$
7: Define,

$$\widehat{E}(f) = \mathbb{E}_{D_{\text{replay}}}\left[\frac{1}{H}\sum\nolimits_{t=1}^{H} f_t(s_t, a_t)\left(\mathcal{P}(s_{1\ldots t})\right)\right] + \mathbb{E}_{D_2}\left[\frac{1}{H}\sum\nolimits_{t=1}^{H} f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\ldots t})\right)\right].$$

8: **Return:** $\pi^{\texttt{RE}}$, a solution to the moment-matching problem:

$$\arg\min_{\pi\in\Pi} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi}\left[\frac{1}{H}\sum\nolimits_{t=1}^{H} f_t(s_t, a_t)\right] - \widehat{E}(f) \tag{6.4}$$

To handle this challenge, we introduce the notion of a *soft membership oracle* [106], $\texttt{MEM} : \mathcal{S} \times [H] \to [0, 1]$ which captures the learner's inherent uncertainty in the expert's actions at a state at each point in an episode. The soft membership oracle assigns high weight to a state if BC is likely to closely agree with the expert policy and gives a lower weight to states where BC is likely to be inaccurate. By this definition, if the membership oracle is consistently large at all the states visited in a trajectory, we can be confident that *a trajectory generated by BC is as though it was a rollout from the expert policy.* Formally, for any function g and time $t = 1, \cdots, H$, we have the decomposition,

$$\mathbb{E}_{\pi^{\text{exp}}}[g(s_t, a_t)] = \underbrace{\mathbb{E}_{\pi^{\text{exp}}}\left[g(s_t, a_t)\mathcal{P}(s_{1\ldots t})\right]}_{(i)} + \underbrace{\mathbb{E}_{\pi^{\text{exp}}}\left[g(s_t, a_t)\left(1 - \mathcal{P}(s_{1\ldots t})\right)\right]}_{(ii)} \tag{6.5}$$

where $\mathcal{P}(s_{1\ldots t})$ is defined as the *prefix weight* $\prod_{t'=1}^{t} \texttt{MEM}(s_{t'}, t')$. We need to use prefix weights instead of the single-sample weights sketched in the previous section to account for the probability of BC getting to the current state in the same manner the expert would have.

Because of the high accuracy of BC on segments with high prefix weights, in eq. (6.5), (i) can be approximated by replacing the expectation over π^{exp} by that over $\pi^{\texttt{BC}}$, i.e. replay estimation. On the other hand, since the prefix weight is low on the remaining trajectories in (ii), we know that BC is inaccurate, so we resort to using a simple empirical estimate to estimate this term.

While we leave the particular choice of the soft membership oracle flexible, intuitively, states at which BC closely agrees with the expert policy should be given high weight while where those where BC is inaccurate should be weighted lower. In Section 6.3, we discuss several practical approaches to designing such a soft membership oracle. We first prove a generic policy performance guarantee for the outputs of our algorithm as a function of the choice of MEM.

Theorem 6.2.1. *Consider the policy $\pi^{\texttt{RE}}$ returned by Algorithm 6. Assume that $\pi^{\text{exp}} \in \Pi$ and the ground truth reward function $r_t \in \mathcal{F}_t$, which is assumed to be symmetric ($f_t \in \mathcal{F}_t \iff -f_t \in \mathcal{F}_t$) and bounded (For all $f_t \in \mathcal{F}_t$, $\|f_t\|_\infty \leq 1$). Choose $|D_1|, |D_2| = \Theta(N)$ and suppose $N_{\text{replay}} \to \infty$. With probability $\geq 1 - 3\delta$,*

$$\mathit{Gap}(\pi^{\texttt{RE}}) \lesssim \mathcal{L}_1 + \mathcal{L}_2 + \frac{\log\left(F_{\max} H/\delta\right)}{N} \tag{6.6}$$

where $F_{\max} \triangleq \max_{t\in[H]} |\mathcal{F}_t|$, and,

$$\mathcal{L}_1 \triangleq H^2 \, \mathbb{E}_{\pi^{\text{exp}}}\left[\frac{\sum_{t=1}^{H} \mathit{MEM}(s_t, t) D_{\text{TV}}\left(\pi_t^{\text{exp}}(\cdot|s_t), \pi_t^{\texttt{BC}}(\cdot|s_t)\right)}{H}\right], \tag{6.7}$$

$$\mathcal{L}_2 \triangleq H^{3/2}\sqrt{\frac{\log\left(F_{\max} H/\delta\right)}{N}\frac{\sum_{t=1}^{H} \mathbb{E}_{\pi^{\text{exp}}}\left[1 - \mathit{MEM}(s_t, t)\right]}{H}}.$$

We discuss a proof of this result in Appendix E.3 and include bounds when N_{replay} is finite.

Remark 6.2.1. Note that Theorem 6.2.1 can be extended to infinite function families using the standard technique of replacing $|\mathcal{F}_t|$ by, $\mathcal{N}(\mathcal{F}_t, \xi, \|\cdot\|_\infty)$ the ξ log-covering number (metric entropy) of $\mathcal{F}_t$ in the L_∞ norm, for $\xi = \frac{1}{NH}$. Likewise, we may appropriately replace $F_{\max}$ by $\mathcal{N}_{\max} \triangleq \max_{t\in[H]} \mathcal{N}(\mathcal{F}_t, \frac{1}{NH}, \|\cdot\|_\infty)$. For ease of exposition here, we stick to the case where $\mathcal{F}_t$ is finite.

The term $\mathcal{L}_1$ measures how accurate BC is on states from expert trajectories where $\texttt{MEM}(s_t, t)$ is large. Intuitively, if we set $\texttt{MEM}(s_t, t) = 1$ on states where BC is accurate and $\texttt{MEM}(s_t, t) = 0$ elsewhere, we would expect this term to be small. $\mathcal{L}_2$ can be thought of a measure of BC's coverage: it tells us how much of the expert's visitation distribution we believe BC to be inaccurate on. If BC has good coverage (i.e. $1 - \texttt{MEM}(s_t, t)$ is small on expert trajectories), we expect this term to be small.

Prima facie, one might think that because $\mathcal{L}_1$ resembles the imitation gap of BC and $\mathcal{L}_2$ resembles that of MM, RE can only perform as well as the best of BC ($\propto H^2/N$) and MM ($\propto H/\sqrt{N}$) on a given instance. However, with a careful choice of MEM, one can achieve "better than both worlds" statistical rates. In particular, since RE is a generalization of Mimic-MD of [71], in the tabular setting, an appropriately initialized version of RE achieves the optimal imitation gap of

$$\min\left\{\frac{|\mathcal{S}|H^{3/2}}{N}, H\sqrt{\frac{|\mathcal{S}|}{N}}\right\}\log\left(\frac{|\mathcal{S}|H}{\delta}\right)$$

and strictly improves over both BC and MM.

6.3 Practical Algorithm

When considering the implementation of RE (Alg. 6) in practice, two main questions arise:

1. How does one construct a membership oracle in practice, especially when action spaces may be continuous?
2. How does one design good solvers for the moment matching problem in eq. (6.4)?

We now provide potential answers to both of these questions.

Membership Oracle. Recall from the interpretation of Theorem 6.2.1 that the membership oracle MEM intuitively should capture how uncertain BC is about the expert's action at a state. For continuous action spaces, ideally one would assign the membership oracle at that state based on an appropriate notion of distance between the action played by BC and that played by the expert. For example, for a sigmoid function σ and constants μ, β,

$$\texttt{MEM}_{\texttt{EXP}}(s,t) = \sigma\left(\frac{\mu - \left\|\pi^{\texttt{BC}}(s) - \pi^{\exp}(s)\right\|_2}{\beta}\right), \tag{6.8}$$

However, in the non-interactive setting where the demonstrator cannot be queried at states, we can only approximate this quantity. The first approximation we propose is inspired by Random Network Distillation (RND) [17], used by [100] to estimate the support of the expert policy. We instead propose to use RND as a measure of the epistemic uncertainty of BC about expert actions. That is,

$$\texttt{MEM}_{\texttt{RND}}(s,t) = \sigma\left(\frac{\mu - \left\|\pi^{\texttt{BC}}(s) - \widehat{\pi^{\texttt{BC}}}(s)\right\|_2}{\beta}\right), \tag{6.9}$$

where $\widehat{\pi^{\texttt{BC}}}$ is a network trained to imitate the output of the classifier $\pi^{\texttt{BC}}$ on the states observed in the demonstration dataset. To train $\widehat{\pi^{\texttt{BC}}}$, we evaluate $\pi^{\texttt{BC}}$ on states observed in the demonstration dataset, and plug this new dataset into the standard BC pipeline.

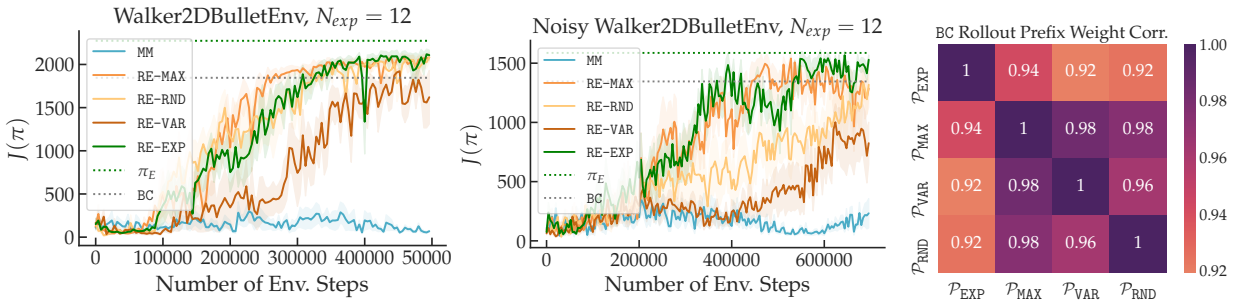


Figure 6.3: **Left:** All variants of RE are able to nearly match expert performance while MM struggles to make any progress. **Center:** We add i.i.d. noise to the environment to make the control problem more challenging. RE is still able to match expert performance, unlike MM. **Right:** We compute correlations between the idealized prefix weights of $\texttt{MEM}_{\texttt{EXP}}$ and the other oracles and see $\texttt{MEM}_{\texttt{MAX}}$ correlate most.

We can also utilize other uncertainty measures like the disagreement of an ensemble to measure epistemic uncertainty, which has previously shown success on various simulated sequential decision making tasks [64]. Past work by [15] proposes to regularize the standard BC (classification) error, by the variance of an ensemble of independently trained BC learners at the states visited by the learner's policy. This encourages the learner to mimic BC on the states where the action predicted by all the policies in the ensemble are similar, and avoid states where they are different (i.e. the variance at these states is high). In contrast, we define,

$$\texttt{MEM}_{\texttt{VAR}}(s,t) = \sigma\left(\frac{\mu - \text{Var}(\pi^{\texttt{BC}(1)}(s), \ldots, \pi^{\texttt{BC}(k)}(s))}{\beta}\right), \tag{6.10}$$

where $\{\pi^{\texttt{BC}(1)}, \cdots, \pi^{\texttt{BC}(k)}\}$ are BC policies trained with different initializations, which produces sufficient diversity when using deep networks as function approximators. Lastly, we can also use the maximum difference across the ensemble as a measure of uncertainty:

$$\texttt{MEM}_{\texttt{MAX}}(s,t) = \sigma\left(\frac{\mu - \max_{i,j\in[k]} \left\|\pi_i^{\texttt{BC}}(s) - \pi_j^{\texttt{BC}}(s)\right\|_2}{\beta}\right), \tag{6.11}$$

as suggested by the work of [48]. We compare these four choices below. For computing prefix weights, we use the average of distances up till the current timestep. This modification serves to improve the numerical stability of our method.

Empirical Moment Matching. We implement approximate Nash equilibrium computation of eq. (6.4) by running a no-regret learner against a best-response counterpart [93]. Our approach is related to the GAIL algorithm of [44] which we improve in 4 ways: *(i)* we use a general Integral Probability Metric [57] instead of the Jensen-Shannon Divergence used

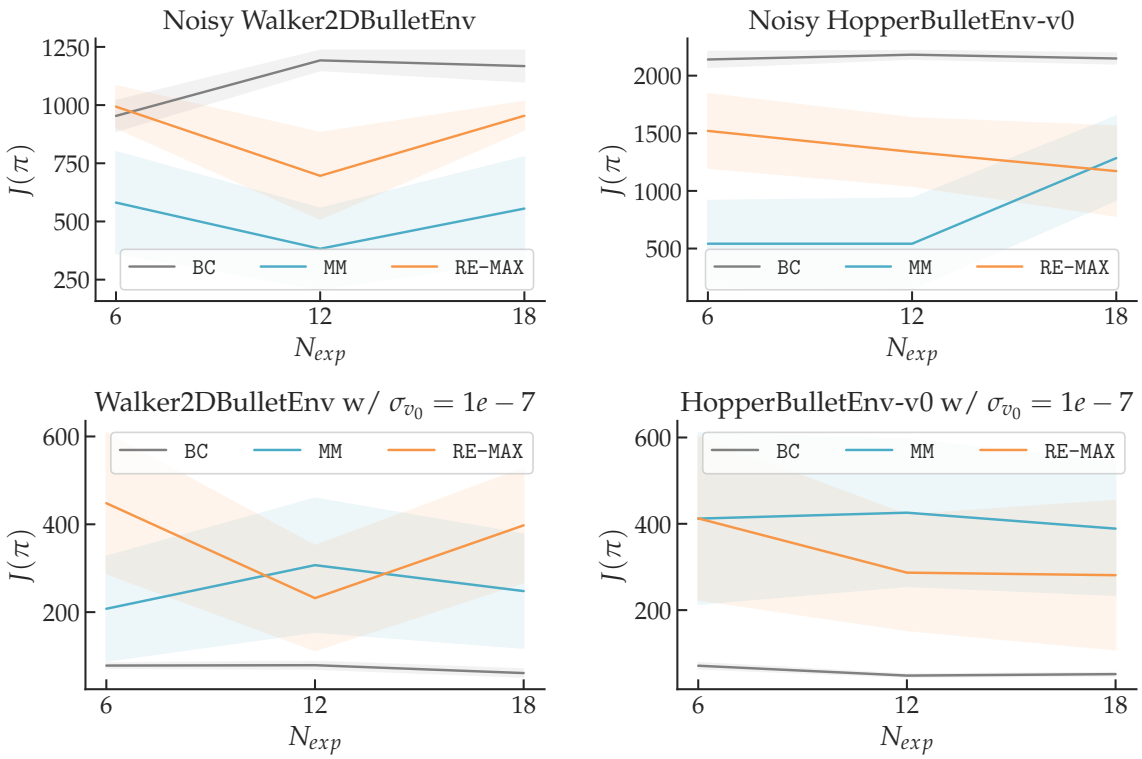


Figure 6.4: We see RE with $\texttt{MEM}_{\texttt{MAX}}$ improve the performance of MM on the Noisy Walker2DBulletEnv and HopperBulletEnv tasks. We see RE (and MM) out-perform BC on the initial-state-perturbed Walker2DBulletEnv and HopperBulletEnv tasks .

in the original paper which improves the representation power of the the discriminator, *(ii)* we add in gradient penalties to the discriminator, which improves convergence rates [39], *(iii)* we solve the entropy-regularized forward problem via Soft-Actor Critic [40] as the policy optimizer, for improved sample efficiency, and *(iv)* we use optimistic mirror descent instead of gradient descent as our optimization algorithm for both players, giving us faster convergence to Nash equilibria, both in theory [97] and in practice [29]. Together, these changes lead to an implementation which *significantly out-performs the original*, giving us a strong baseline to compare against. We include an ablation to confirm this fact in Section 6.4. We emphasize that the RE technique can be used to improve *any* online moment matching algorithm.

6.4 Experimental Results

We now quantify the empirical benefits of RE on several continuous control tasks from the the PyBullet suite [24]. All the task we consider have long horizons ($H \approx 1000$) and we use relatively few demonstrations. ($N \leq 20$). We set N_{replay} as 100 BC rollouts (Line 4 of

Algorithm 6). We test all four membership oracles from the previous section ($\texttt{MEM}_{\texttt{EXP}}$ as an idealized target, $\texttt{MEM}_{\texttt{RND}}$, $\texttt{MEM}_{\texttt{VAR}}$, and $\texttt{MEM}_{\texttt{MAX}}$ as practical solutions). In Figure 6.3 (left), we see that with only twelve trajectories, RE is able to reliably match expert performance for all oracles considered, while MM is not. The environment considered in this experiment is nearly deterministic, indicating that RE can help even when the environment is not stochastic. We hypothesize that the randomness in the initial state is sufficient for replay estimation to generate a significant improved estimate of the state-action visitation measure. This improvement is especially interesting considering both of the hard examples we studied for MM and BC in Section 6.1 were heavily stochastic.

Performance under perturbations. In Figure 6.3 (center), we add in i.i.d. noise to the environment dynamics at each timestep, making it stochastic. This makes the problem significantly more challenging than the standard version of the Walker task. RE is still able to match expert performance, with $\texttt{MEM}_{\texttt{MAX}}$ working notably well. The correlation plot in Figure 6.3 (right) shows us $\texttt{MEM}_{\texttt{MAX}}$ appears to be best correlated with the idealized prefix weights, $\texttt{MEM}_{\texttt{EXP}}$ under the state distribution induced by BC. Because of its superior performance, we use $\texttt{MEM}_{\texttt{MAX}}$ for the rest of our experiments. In the left half of Figure 6.4, we see RE improve the performance of MM. In the right half, we see RE out-perform BC in responding to an *extremely tiny* amount of noise added to the initial velocity of the agent (similar to the experiments of [73]).

Prefix weight distributions. In fig. 6.5, we plot the distributions of the prefix weights generated by each membership oracle on simulated BC rollouts on WalkerBulletEnv. Note that $\texttt{MEM}_{\texttt{VAR}}$ is significantly overconfident in prefix weights compared to $\texttt{MEM}_{\texttt{EXP}}$, as indicated by the heavier right-tail. On the other hand, $\texttt{MEM}_{\texttt{RND}}$ and $\texttt{MEM}_{\texttt{MAX}}$ are less overconfident and better overlap with the idealized prefix weights induced by $\texttt{MEM}_{\texttt{EXP}}$. This aligns with the correlation plot between the various membership oracles in Figure 6.3. Moreover, in terms of policy performance, this further justifies the superior behavior of $\texttt{MEM}_{\texttt{MAX}}$ compared to $\texttt{MEM}_{\texttt{VAR}}$.

Ablations

In Figure 6.6, we consider how each of the changes we described earlier in this chapter, lead to improved performance of our RE baseline. The first, using a Wasserstein distance, leads to lower expected return but is required for solving the full moment-matching problem – see [93] for more details. Switching from PPO to the more sample-efficient SAC [40] leads to fast learning. Adding in gradient penalties for discriminator stability [93, 39] also improves final performance and learning speed. The last change we employ, using Optimistic Mirror Descent [29] in both the discriminator and RL algorithm also (slightly) improves performance. To our knowledge, we are the first to utilize this technique in the Imitation Learning literature and reccomend it as best practice for future moment-matching algorithms. We refer interested readers to the work of [97] for theoretical details of why OMD enables superior performance.

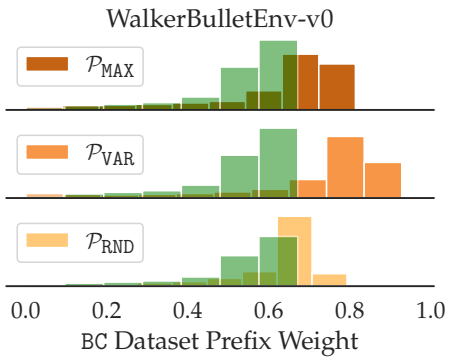


Figure 6.5: Histogram of prefix weights generated by rolling out trajectories from BC. The green superimposed histogram represents prefix weights generated by $\texttt{MEM}_{\texttt{EXP}}$

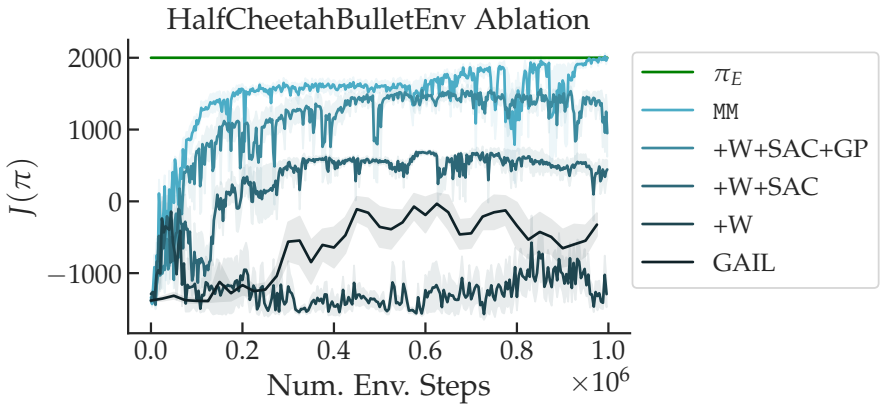


Figure 6.6: We ablate the four key changes we made to off-the-shelf GAIL to improve performance / theoretical guarantees. We see that each improved performance, with MM significantly out-performing options with fewer changes. Our improvements upon MM with the Replay Estimation technique are therefore improving upon an already strong baseline.

Experimental Setup

We begin with the hyperparameters for our Standard Bullet and Noisy Bullet experiments.

Expert

We use the Stable Baselines 3 [68] implementation of PPO [81] or SAC [40] to train experts for each environment. For the most part, we use already tuned hyperparameters from [67] in the implementation. The modifications we used are are shown in table 6.1.

Parameter	Value
buffer size	300000
batch size	256
γ	0.98
τ	0.02
Training Freq.	64
Gradient Steps	64
Learning Rate	Lin. Sched. 7.3e-4
policy architecture	256 x 2
state-dependent exploration	true
training timesteps	1e6

Table 6.1: Expert hyperparameters for Walker Bullet Task and Hopper Bullet Task

Noisy Experts

In addition to the default Bullet Tasks, we test performance of algorithms on noisy environments. Namely, we generate noisy expert data by re-training expert policies with Gaussian noise added to the actions of the expert during the exploration phase while training. We then re-generate expert data by sampling from the expert policies trained on noisy data to analyze the performance of our method under stochasticity. Table 6.2 lists the standard deviation of the (i.i.d.) noise we applied to the actions in the different environments.

env.	Noise Distribution.
hopper	$\mathcal{N}(0, 0.1)$
walker	$\mathcal{N}(0, 0.5)$

Table 6.2: Noise we applied to all policies in each environment.

Baselines

We average over 5 runs and use a common architecture of 256 x 2 with ReLU activations for both our method and the `MM` baseline we compare against. For each datapoint, the cumulative reward is averaged over 10 trajectories. For all tasks, we train on $\{6, 12, 18\}$ expert trajectories with a maximum of 400k iterations of the optimization procedure. Table 6.3 shows the hyperparameters we used for `MM`. Empirically, smaller learning rates, large batch sizes, and gradient penalties were critical for the stable convergence of our method.

Parameter	Value
Batch Size	2048*
Learning Rate	Linear Schedule of 8e-3*
f Update Freq.	5000
f gradient target	0.4
f gradient penalty weight	10

Table 6.3: Learner hyperparameters for MM. * indicates the parameter was different for the Hopper Initial State shift experiments (4096 for batch size and Linear Schedule of 8e-4, respectively.).

We note that MM requires careful tuning of f UPDATE FREQ. for strong performance. We searched over step sizes of $\{1250, 2500, 5000\}$ and selected the one which achieved the most stable updates. In practice, we recommend evaluating a trained policy on a validation set to set this parameter. We also used similar parameters for training SAC, also from the Stable Baselines 3 [68] implementation, as we did for training the expert policy. Table 6.4 shows the choice of hyperparmeters we used for training SAC. We directly added in the optimistic mirror descent optimizers [29] for both the critic and actor objectives of SAC. Table 6.5 shows the learning hyperparameters for any BC policies used for generating simulated data for the membership oracles. Table 6.6 shows the number of training steps per task we used for both the baseline and our method.

Parameter	Value
γ	0.98
τ	0.02
Training Freq.	64
Gradient Steps	64
Learning Rate	Linear Schedule of 7.3e-4
policy architecture	256 x 2

Table 6.4: Leaning hyperparameters for the SAC component of MM

Algorithm hyperparameters

In this section, we use **bold text** to highlight sensitive hyperparameters. We use the same network architecture choices as the MM baseline. For all environments, we generated 100 trajectories of simulated behavior cloning data to use with our method.

Parameter	Value
entropy weight	0
l2 weight	0
training timesteps	1e5

Table 6.5: Learner hyperparameters for Behavioral Cloning

env.	training steps
walker (no noise)	400000
walker (with noise)	400000
hopper (no noise)	400000
hopper (with noise)	400000

Table 6.6: Number of training steps for the different tasks

For all tasks, we rolled out **100 trajectories** from a `BC` trained network to use with our membership oracle. Table 6.7 shows how we partitioned our dataset between the `BC` training set and the expert membership oracle dataset. We also use the full dataset for moment matching, not just D_2, as we found this lead to slightly better performance.

Expert Size	D_1	D_2
6 trajs	4	2
12 trajs	10	2
18 trajs	16	2

Table 6.7: Partition of D into D_1 and D_2 based on the number of expert trajectories provided. For the Noisy Walker experiments, we used $5, 10, 14$ trajectories for D_1 instead of the above.

Membership Oracle hyperparameters

For both $\texttt{MEM}_{\texttt{VAR}}$ and $\texttt{MEM}_{\texttt{MAX}}$, we use 5 `BC` networks in the ensemble. We followed the exact same parameters in Table 6.5 to train each `BC` imitator. Table 6.8 shows the choice of μ and β values we used for each membership oracle.

env	parameter	$\texttt{MEM}_{\texttt{EXP}}$	$\texttt{MEM}_{\texttt{RND}}$	$\texttt{MEM}_{\texttt{VAR}}$	$\texttt{MEM}_{\texttt{MAX}}$
walker	β	0.1	0.1	0.01	0.1
walker	μ	0.33	0.22	0.015	0.35
hopper	β	0.8	0.25	0.08	0.1
hopper	μ	0.68	0.4	0.05	0.25

Table 6.8: Membership oracle hyperparameters across different environments

env	parameter	$\texttt{MEM}_{\texttt{MAX}}$
walker	β	0.01
walker	μ	0.0001
hopper	β	0.01
hopper	μ	0.0001

Table 6.9: Membership oracle hyperparameters across different initial state shift environments.

Initial State Shift Experiments

We use demonstrations generated by an expert trained on the standard Bullet tasks but subject the learner (both at train and test time) to a initial velocity perturbation of a zero-mean Gaussian with variance ($\sigma = 1e-7$). We refer interested readers to our code for our precise method of injecting noise as we believe it might be of interest for future experiments. In all demonstrations, the expert starts from rest. Despite this relatively small shift, we see BC performance drop significantly, as is characteristic of real-world problems where it significantly under-performs on-policy IL methods. All results are averaged over five seeds and for all environments, we train BC for 1e5 steps (as well as for the query policies for RE). For RE, we train 5 policies and use the $\texttt{MEM}_{\texttt{MAX}}$ approximate membership oracle. We use the above parameters for MM for our base moment-matcher.

Chapter 7

IL with parametric function approximation

In practical settings, RL algorithms are deployed in state and action spaces which are often continuous or unbounded. Carrying out IL efficiently requires imposing additional assumptions. In this chapter, we go beyond the tabular setting and study IL in the presence of function approximation. We will begin with the linear-expert setting (Definition 7.1.1) where $\mathcal{S}$ and $\mathcal{A}$ may be unbounded, but the learner is provided a set of feature representations of state-actions, and the expert policy is constrained to be realizable by a unknown linear (in the feature representations) classifier. We will then extend these ideas to the setting of general function approximation, extending beyond the linear setting.

7.1 Linear function approximation

In this section we study IL with linear function approximation. We first formally introduce the linear-expert setting and show that it generalizes several settings which are interesting and practically relevant.

Definition 7.1.1 (Linear-expert setting)**.** *In this setting, for each (s, a, t) tuple, the learner is provided a feature representation $\phi_t(s, a) \in \mathbb{R}^d$. For each $t \in [H]$ there exists an unknown vector $\theta_t^\star \in \mathbb{R}^d$ such that $\forall s \in \mathcal{S}$, $\pi_t^{\exp}(s) = \arg\max_{a \in \mathcal{A}} \langle \theta^\star, \phi_t(s, a) \rangle$.*

Remark 7.1.1. The linear-expert setting (Definition 7.1.1) generalizes the linear-$Q^\star$ setting with an optimal expert. Under this assumption, the optimal expert policy plays actions according to $\pi_t^\star(s) = \arg\max_{a \in \mathcal{A}} Q_t^\star(s, a) = \arg\max_{a \in \mathcal{A}} \langle \theta_t^\star, \phi_t(s, a) \rangle$ for an unknown $\theta_t^\star \in \mathbb{R}^d$. Thus the expert policy is realizable by a linear multi-class classifier. Since the tabular setting is a special case of the linear-$Q^\star$ setting with $d = |\mathcal{S}||\mathcal{A}|$, with features for each t chosen as the standard basis vectors in $\mathbb{R}^d$, the linear-expert setting with $d = |\mathcal{S}||\mathcal{A}|$ generalizes the tabular setting with an optimal expert.

In the tabular setting, we showed in Theorem 2.1.3 that the expected imitation gap of BC is $O(|\mathcal{S}|H^2/N)$ which is also optimal in the no-interaction setting. We first introduce a version of BC for the linear-expert setting.

Definition 7.1.2 (BC in the linear-expert setting)**.** *For $t = 1, \cdots, H$, denote $(D)_t$ as a collection of N state-action pairs visited at time t across trajectories in D. In the linear-expert setting BC trains a policy $\widehat{\pi}$ as follows: for each $t = 1, \cdots, H$, the learner trains a linear multi-class classifier $\widehat{h}_t : \mathcal{S} \to \mathcal{A}$ on the dataset $(D)_t$ using the algorithm of [25] and plays the policy $\widehat{\pi}_t(s) = \widehat{h}_t(s)$.*

We next establish an upper bound on the imitation gap incurred by BC in the linear-expert setting.

Theorem 7.1.1. *In the linear-expert setting, with probability $1-\delta$, the imitation gap of the policy $\widehat{\pi}$ output by BC in the linear-expert setting (Definition 7.1.2) is upper bounded with probability $1-\delta$ by $\mathsf{Gap}(\widehat{\pi}) \lesssim \frac{H^2(d+\log(1/\delta))\log(N)}{N}$.*

Linear function approximation with parameter sharing

Definition 7.1.3 (Linear-expert setting with parameter sharing)**.** *This is a special case of the linear-expert setting where $\theta^\star_t$ is the same across $t \in [H]$.*

Remark 7.1.2. Linear-expert setting with dimension d is a special case of the linear-expert setting with parameter sharing, with dimension dH. Define $\theta^\star = (\theta^\star_1, \cdots, \theta^\star_H) \in \mathbb{R}^{dH}$ and $\phi'_t(s,a) = (0^d, \cdots, \phi_t(s,a), \cdots, 0^d) \in \mathbb{R}^{dH}$ where $\phi_t(s,a)$ is embedded in the coordinates $td+1$ to $(t+1)d$. Then, $\pi^{\exp}_t(s) = \arg\max_{a\in\mathcal{A}} \langle \theta^\star_t, \phi_t(s,a)\rangle = \arg\max_{a\in\mathcal{A}} \langle \theta^\star, \phi'_t(s,a)\rangle$, satisfying Definition 7.1.1.

Our main contribution is to show that in the linear-expert setting with parameter sharing, where the expert plays actions according to the *same linear classifier* at each time in an episode, the imitation gap incurred by BC is $\widetilde{\mathcal{O}}(\frac{dH}{N})$, breaking the quadratic dependence on H suffered more generally (cf. Theorem 2.2.1). With such a parameter sharing assumption on the expert, intuitively, each trajectory now provides H training examples to learn a single classifier capturing the learner's policy. This setting is motivated by the fact that often in practice, BC is are implemented to learn a single classifier across the episode [23].
The proof of this result is derived by a supervised learning reduction of IL to *sequence multi-class linear classification* where we learn linear classifiers from $\mathcal{S}^H \to \mathcal{A}^H$. The supervised learning reduction of [74] posits to learn separate classifiers from $\mathcal{S} \to \mathcal{A}$ or $\mathcal{S} \times [H] \to \mathcal{A}$: this fails to account for the shared parameter $\theta^\star$ across time. While in both cases the resulting policy is an ERM classifier, the imitation gap grows quadratically in H using the supervised learning reduction. In contrast, using the multi-class classification algorithm of [25], we also provide an algorithm $\widehat{\pi}$ with imitation gap *growing linearly in H*. To begin with, define Θ as

the set of linear multi-class classifiers for sequences, of the form,

$$\mathcal{S}^H \ni (s_1, \cdots, s_H) \mapsto \underset{a_1,\cdots,a_H \in \mathcal{A}}{\arg\max} \left\langle \theta, \sum_{t=1}^{H} \phi_t(s_t, a_t) \right\rangle \in \mathcal{A}^H. \tag{7.1}$$

for $\theta \in \mathbb{R}^d$. Note that under the linear-expert assumption with parameter sharing, the expert's policy can be identified as a classifier in the family described above. At each state s, the expert plays the action according to $\arg\max_{a\in\mathcal{A}} \langle \theta^\star, \phi_t(s,a)\rangle$ at time t. Summing over any sequence of states $s_1, \cdots, s_H$, the expert's policy therefore satisfies $(\pi_1^{\exp}(s_1), \cdots, \pi_H^{\exp}(s_H)) = \arg\max_{a_1,\cdots,a_H} \langle \theta^\star, \sum_{t=1}^H \phi_t(s_t, a_t)\rangle$. This can be interpreted as a multi-class linear classifier from $\mathcal{S}^H \to \mathcal{A}^H$, in contrast to the usual implementation of `BC` which learns a sequence of classifiers from $\mathcal{S} \to \mathcal{A}$. In particular, for each input sequence of states $(s_1, s_2, \cdots, s_H)$ the expert "classifier" outputs the label, which is a sequence of actions $(\pi_1^{\exp}(s_1), \pi_2^{\exp}(s_2), \cdots, \pi_H^{\exp}(s_H))$.

Note that classifiers of the form eq. (7.1) indeed correspond to meaningful (Markovian) policies which drawn actions a_t from a policy which is a function of only the current state s_t. Indeed the map in eq. (7.1) is separable as $\sum_{t=1}^H \arg\max_{a_t\in\mathcal{A}} \langle \theta, \phi_t(s_t, a_t)\rangle$ where we carry out the optimization for each variable $a_1, \cdots, a_H$ separately. By contradiction, the action played by the classifier at any state s_t at time t must be $\arg\max_{a_t\in\mathcal{A}} \langle \theta, \phi_t(s_t, a_t)\rangle$ which is Markovian. More importantly, such classifiers can be learned from the demonstration dataset which essentially contains N i.i.d. examples of input-label pairs drawn from a "ground-truth" classifier (i.e., $\pi^{\exp}$).

Based on these insights, we prove a bound on the imitation gap of the policy induced by `BC` via the reduction of Theorem 2.1.1 [76]. The intuition is that in any trajectory where the learner's actions exactly match the expert's actions, no suboptimality is incurred. In contrast, in any trajectory where the learner plays an action different from the expert at some time, the reward gap incurred is H.

Lemma 7.1.2. *Consider any linear multi-class classifier $\widehat{\theta} : \mathcal{S}^H \to \mathcal{A}^H$ (in eq. (7.1)) with expected 0-1 loss, $\mathrm{E}_{\pi^{\exp}}[\mathbb{1}(\widehat{\theta}(s_1, \cdots, s_H) \neq (a_1, \cdots, a_H))] \leq \gamma$. Then, the policy $\widehat{\pi}$ corresponding to the linear classifier $\widehat{\theta}$, satisfies $\mathsf{Gap}(\widehat{\pi}) \leq H\gamma$.*

[25] provide a compression based algorithm for linear multi-class classification in the realizable setting. Indeed, invoking [25, Theorem 5], it is possible to learn a linear classifier $\widehat{\theta} \in \Theta$ such that the expected 0-1 loss of the classifier is upper bounded by $\frac{(d+\log(1/\delta))\log(N)}{N}$ given N expert trajectories. In conjunction with Lemma 7.1.2 this results in an upper bound on the imitation gap of the resulting policy.

Theorem 7.1.3. *Consider a learner which uses `BC` (Definition 7.1.2) instantiated with the compression based linear classification subroutine of [25]. Under the linear-expert assumption with parameter sharing (Definition 7.1.3), with probability $\geq 1-\delta$, the imitation gap of the learned policy $\widehat{\pi}$ satisfies,*

$$\mathsf{Gap}(\widehat{\pi}) \lesssim \frac{H(d + \log(1/\delta))\log(N)}{N}.$$

Remark 7.1.3. The linear dependence on the horizon can be interpreted in a different way: with parameter sharing, the learner can aggregate information across time steps in an episode to learn a single linear classifier with improved guarantees. The amount of training data the learner has access to is effectively larger by a factor of H since each trajectory provides H samples of data for learning a single classifier common across time. The reduction analysis of [74] shows an imitation gap gap of $H^2\epsilon$ for learners with expected 0-1 loss under the expert state distribution upper bounded by ϵ. If data can be aggregated across time to learn a single parameter, the expected 0-1 loss can be brought down by a factor of H, showing that it is possible to achieve imitation gap scaling linearly on the length of the horizon.

The results in this section can also be extended to more general forms of function approximation, such as bounded Natarajan dimension classes, as considered in [28]. We will skip past this discussion and move on to studying the known-transition setting.

7.2 Function approximation with known transitions

In this section, we will study IL in the known-transition setting with general forms of function approximation. Our discussion in Chapter 2 proposes the `Mimic-MD` approach which in effect can be thought of as "simulating artificial trajectories" to improve the estimation power of the learner. This connection is reminded to the reader below through the lens of *uniform expert value estimation* introduced in Chapter 3: the problem of estimating what the value of the unknown expert policy $\pi^{\exp}$ is simultaneously under all reward functions belonging to some class. Given a uniform expert value estimator $\widetilde{J}_r(\pi^{\exp})$, which with probability $1-\delta$ (over the demonstration dataset and external randomness) for all reward functions r, satisfies $|J_r(\pi^{\exp}) - \widetilde{J}_r(\pi^{\exp})| \leq \epsilon$, then the policy $\widehat{\pi}$ output by the following optimization problem (a restatement of eq. (OPT)),

$$\widehat{\pi} \leftarrow \arg\min_{\pi} \max_{r} \widetilde{J}_r(\pi) - J_r(\pi) \tag{OPT}$$

incurs imitation gap $\mathsf{Gap}(\widehat{\pi}) \leq 2\epsilon$ with the same probability $1-\delta$. [69] also show that this objective is a convex program and can be approximately solved in an efficient manner in the tabular setting. In this context, to execute the approach of simulating artificial trajectories, observe that a learner can construct a good estimate of the expert's value under some reward function r by decomposing it as the sum of two parts:

$$J_r^1(\pi^{\exp}) = \mathbb{E}_{\pi^{\exp}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\mathbb{1}(\mathcal{E})\right], \text{ and } J_r^2(\pi^{\exp}) = \mathbb{E}_{\pi^{\exp}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\mathbb{1}(\mathcal{E}^c)\right]. \tag{7.2}$$

where $\mathcal{E}$ is the event that the all the states $(s_1, \cdots, s_H)$ visited in the trajectory are observed in the demonstration dataset. The first term, J_r^1, can be estimated to an arbitrary level of accuracy for any reward function r by rolling out many artificial trajectories using $\pi^{\exp}$, known at all states observed in the dataset. The remaining term, J_r^2 can be tackled using a simple empirical estimate, as explained below. The event $\mathcal{E}^c$ guarantees that states in a

trajectory are visited are observed in the dataset D (where the expert's policy is known). Therefore, by holding out some trajectories in the dataset, the learner may carry out an empirical estimate of $J_r^2(\pi^{\exp})$ using these trajectories. The error in uniform value estimation precisely stems from the error incurred by the empirical estimate, which is shown to be $O(|\mathcal{S}|H^{3/2}/N)$ in [70], translating to the imitation gap of the policy $\widehat{\pi}$ in (OPT).

It is a natural question to ask whether this approach of simulating artificial trajectories can be applied when state and action spaces may be unbounded, such as with function approximation. To effectively use such an approach, the learner should be able to infer the expert's action at a large fraction of states in spite of *observing the expert's actions only on a measure-0 subset of states*. This will be the critical discussion of this section, but prior to jumping in, we will set up the formulation we consider.

Definition 7.2.1 (IL with function-approximation)**.** *In this setting, for each $t \in [H]$, there is a parameter class $\Theta_t \subseteq \mathbb{B}_2^d$, the unit L_2 ball in d dimensions, and an associated function class $\{f_{\theta_t} : \theta_t \in \Theta_t\}$. For each $t \in [H]$ there exists an unknown $\theta_t^E \in \Theta_t$ such that $\forall s \in \mathcal{S}$,*

$$\pi_t^{\exp}(s) = \arg\max_{a \in \mathcal{A}} f_{\theta_t^E}(s, a). \tag{7.3}$$

Lipschitz parameterization

We impose a condition on the nature of the parameterization of the function classes. The "Lipschitz parameterization" condition can be interpreted as saying that a small change to the underlying classifier does not all of a sudden change the label of a large mass of points. In particular, points which are classified with a large enough "margin" continue to stay in the same class even if the underlying classifier/parameter is perturbed.

Definition 7.2.2 (Lipschitz parameterization)**.** *A function class $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$ where $g_\theta(\cdot) : \mathcal{X} \to \mathbb{R}$ is said to satisfy L-Lipschitz parameterization if, $\|g_\theta(\cdot) - g_{\theta'}(\cdot)\|_\infty \leq L\|\theta - \theta'\|_2$. In other words, for each $x \in \mathcal{X}$, $g_\theta(x)$ is an L-Lipschitz function in θ, in the L_2 norm.*

Assumption 7.2.1. *For each t, the class $\{f_{\theta_t} : \theta_t \in \Theta_t\}$ is L-Lipschitz in its parameterization, $\theta_t \in \Theta_t$.*

Recall that `BC` essentially trains an offline classification algorithm on the demonstration dataset to predict actions. Moreover, the reduction of [76] (i.e., Theorem 2.1.1) bounds the imitation gap of the resulting policy in terms of the 0-1 generalization error of the learned classifier. In order to extend `RE` (which uses the transition of the MDP) to deal with parametric function approximation, and show error guarantees which surpass that achieved by `BC`, we assume that the learner has access to a slightly stronger offline classification oracle, which, given access to a dataset of classification examples, *returns an approximate version of the underlying ground truth parameter*. More formally,

Assumption 7.2.2 (Offline classification oracle)**.** *We assume that the learner has access to a multi-class classification oracle, which given n examples of the form, (s^i, a^i) where $s^i \overset{i.i.d.}{\sim} \mathcal{D}$ and $a^i = \arg\max_{a \in \mathcal{A}} f_{\theta^\star}(s^i, a)$, returns a $\widehat{\theta} \in \Theta$ such that, with probability $\geq 1 - \delta$, $\|\widehat{\theta} - \theta^\star\|_2 \leq \mathcal{E}_{\Theta,n,\delta}$.*

Remark 7.2.1. We will assume that a classification oracle satisfying Assumption 7.2.2 (trained on a slice of the demonstration dataset at each time t) is used by `RE` to train the `BC` policy in Line 3 of Algorithm 6. We denote the resulting offline classifiers as $\widehat{\theta}^{\mathsf{BC}} = (\widehat{\theta}^{\mathsf{BC}}_1, \cdots, \widehat{\theta}^{\mathsf{BC}}_H)$.

A careful reader might note that Assumption 7.2.2 asks for a slightly stronger requirement than just finding a classifier with small generalization error (which need not be close to the ground truth $\theta^\star$). The generalization error, i.e. $\mathbb{P}\mathbb{r}_{s \sim \mathcal{D}}\left[\arg\max_{a \in \mathcal{A}} f_{\theta^\star}(s, a) \neq \arg\max_{a \in \mathcal{A}} f_{\widehat{\theta}}(s, a)\right]$ in the notation of Assumption 7.2.2, was previously studied in [27] for multi-class classification. The authors show that up to log-factors in the number of classes (i.e. the number of actions), the *Natarajan dimension* characterizes the generalization error of the best learner, which scales as $\Theta(1/n)$ given n classification examples. Under certain assumptions on the input distribution and the function family (e.g. for linear families, which we study in Section 7.3), we later show that the generalization error guarantee can be extended to approximately learning the parameter as well (up to problem dependent constants).

We are now ready to define the membership oracle under which we study `RE` below,

$$\texttt{MEM}(s, t) = \begin{cases} +1 & \text{if } \exists a \in \mathcal{A} \text{ such that, } \forall a' \in \mathcal{A},\ f_{\widehat{\theta}^{\mathsf{BC}}_t}(s, a) - f_{\widehat{\theta}^{\mathsf{BC}}_t}(s, a') \geq 2L\mathcal{E}_{\Theta_t,N,\delta/H} \\ 0 & \text{otherwise.} \end{cases} \tag{7.4}$$

The intuitive interpretation of `MEM` is that, state which are classified by `BC` as some action with a significant margin are assigned as $+1$, and the remaining states are assigned as 0 by the membership oracle.

Finally, we impose an assumption on the IL instances we study. We assume that the classification problems solved by `BC` at each $t \in [H]$ satisfy a margin condition.

Assumption 7.2.3 (Weak margin condition)**.** *For $t \in [H]$ and $\theta \in \Theta_t$, define $a^\theta_s = \arg\max_{a \in \mathcal{A}} f_\theta(s, a)$ as the classifier output on the state s. The weak margin condition with parameter $\mu > 0$ assumes that for each t, there is no classifier $\theta \in \Theta_t$ such that for a large mass of states, $f_\theta(s_t, a^\theta_{s_t}) - \max_{a \neq a^\theta_{s_t}} f_\theta(s_t, a)$, i.e. the "margin" from the nearest classification boundary, is small. Formally, the weak-margin condition with parameter μ states that for any $\eta \leq 1/\mu$,*

$$\forall \theta \in \Theta_t, \quad \Pr_{\pi^{\exp}}\left(f_\theta(s_t, a^\theta_{s_t}) - \max_{a \neq a^\theta_{s_t}} f_\theta(s_t, a) \geq \eta\right) \geq e^{-\mu\eta}. \tag{7.5}$$

The weak margin condition only assumes that there is at least an exponentially small (in η) mass of states with margin at least η. A smaller μ indicates a larger mass away from any decision boundary.

Remark 7.2.2. Note that the weak margin condition is the multi-class extension of the Tsybakov margin condition of [53, 8] defined for the binary case. In particular, in eq. (7.5), we may replace the RHS by $1-\mu\eta$, or $1-(\mu\eta)^\alpha$ for $\alpha>0$ to get different analogs of the margin condition and the main guarantee, Theorem 7.2.1, as we discuss in Appendix F.2.

Remark 7.2.3. It suffices to assume that for each t, eq. (7.5) is only true for the singular choice $\theta=\widehat{\theta}_t^{\mathtt{BC}}\in\Theta_t$, for our main guarantee (Theorem 7.2.1) to hold.

Under the previously discussed assumptions, we provide a strong guarantee for `RE` which uses the classification oracle in Assumption 7.2.2 to define `BC`, and the membership oracle as defined in eq. (7.4).

Theorem 7.2.1. *For IL with parametric function approximation, under Assumptions 7.2.1 to 7.2.3, appropriately instatiating* ***RE*** *ensures that with probability* $\geq 1-4\delta$,

$$\mathsf{Gap}(\pi^{\mathtt{RE}})\lesssim H^{3/2}\sqrt{\frac{\mu L\log\left(F_{\max}H/\delta\right)}{N}\frac{\sum_{t=1}^{H}\mathcal{E}_{\Theta_t,N,\delta/H}}{H}}+\frac{\log\left(F_{\max}H/\delta\right)}{N}. \tag{7.6}$$

where $F_{\max}=\max_{t\in[H]}|\mathcal{F}_t|$ *is as defined in Theorem 6.2.1.*
Note that we impose the same assumptions on the policy and discriminator classes employed by ***RE*** *as in Theorem 6.2.1. Namely,* (i) $\pi^{\exp}\in\Pi$, *and for each* $t\in[H]$, (ii) *the ground truth reward function* $r_t\in\mathcal{F}_t$, (iii) $\mathcal{F}_t$ *is symmetric, i.e.* $f_t\in\mathcal{F}_t\iff -f_t\in\mathcal{F}_t$, *and* (iv) $\mathcal{F}_t$ *is* 1*-bounded, i.e. for all* $f_t\in\mathcal{F}_t$, $\|f_t\|_\infty\leq 1$.

As discussed later in Remark 6.2.1, this result can be extended to infinite discriminator families by replacing $|\mathcal{F}_t|$ by the appropriate log-covering number of $\mathcal{F}_t$ in L_∞ norm.
The intuition behind the result is as follows. By Assumption 7.2.2, the learner is able to approximately learn $\widehat{\theta}_t^{\mathtt{BC}}\approx\theta_t^\star$ at each time t. Since the discriminator functions are Lipschitz (Assumption 7.2.1) and there are not too many states classified with small margin (Assumption 7.2.3), this means that the states classified with large margin by $\widehat{\theta}_t^{\mathtt{BC}}$ are correctly classified by θ_t^E, while the states which are close to a decision boundary (induced by `BC`) may be misclassified by `BC`. Therefore, we may set the membership oracle as $+1$ on states classified by `BC` with a large margin, and 0 on states classified with small margin. In particular, the membership oracle considered in eq. (7.4) ensures that on states at which $\mathtt{MEM}(s,t)>0$, $\pi_t^{\exp}(\cdot|s)=\pi_t^{\mathtt{BC}}(\cdot|s)$. Likewise, the states at which $\mathtt{MEM}(s,t)=0$ correspond to the states which are classified by `BC` with a small margin (i.e. are close to a decision boundary), the probability mass of which is bounded by the weak margin condition, Assumption 7.2.3. All in all, in the language of Theorem 6.2.1, these results ensure that $\mathcal{L}_1=0$ and $\mathcal{L}_2$ is significantly smaller than $H^{3/2}\sqrt{\log(F_{\max}H/\delta)/N}$ which essentially results in the proof of Theorem 7.2.1.

7.3 Interpretations of Theorem 7.2.1

In order to interpret Theorem 7.2.1, we draw the connection back with `BC` and `MM`. As discussed earlier, [76] prove the best known general statistical guarantee for `BC` in terms of

the generalization error of the underlying classifiers. In particular, with probability $\geq 1-\delta$,

$$\mathsf{Gap}(\pi^{\mathtt{BC}}) \lesssim \overline{\mathsf{Gap}}(\pi^{\mathtt{BC}}) \triangleq H^2 \frac{\sum_{t=1}^{H} \mathcal{E}^{\text{class}}_{\Theta_t,N,\delta/H}}{H} \tag{7.7}$$

Here, $\mathcal{E}^{\text{class}}_{\Theta,n,\delta}$ denotes the best achievable 0-1 generalization error for multi-class classification: in the notation of Assumption 7.2.2, there exists a learner $\widehat{\theta}$, such that on any classification instance $\mathbb{E}_{s\sim\mathcal{D}}[\max_{a\in\mathcal{A}} f_{\widehat{\theta}}(s,a) \neq \max_{a\in\mathcal{A}} f_{\theta^\star}(s,a)] \leq \mathcal{E}^{\text{class}}_{\Theta,n,\delta}$ with probability $\geq 1-\delta$. On the other hand, the best general statistical guarantee for `MM` is,

$$\mathsf{Gap}(\pi^{\mathtt{MM}}) \lesssim \overline{\mathsf{Gap}}(\pi^{\mathtt{MM}}) \triangleq H\sqrt{\frac{\log\left(F_{\max}H/\delta\right)}{N}} \tag{7.8}$$

(we can extend to infinite classes using the covering number argument in Remark 6.2.1).

Now, whenever the statistical error for parameter estimation matches with the best statistical error for generalization in offline classification, namely, $\mathcal{E}_{\Theta,n,\delta} \asymp \mathcal{E}^{\text{class}}_{\Theta,n,\delta}$ up to problem dependent constants, the guarantee in Theorem 7.2.1 can be reinterpreted as,

$$\mathsf{Gap}(\pi^{\mathtt{RE}}) \lesssim \overline{\mathsf{Gap}}(\pi^{\mathtt{MM}})\sqrt{\frac{\overline{\mathsf{Gap}}(\pi^{\mathtt{BC}})}{H}}. \tag{7.9}$$

The interpretation of this result is that, whenever `BC` admits a non-trivial gap on the imitation gap, namely $\mathsf{Gap}(\pi^{\mathtt{BC}}) \ll H$, the performance gap in eq. (7.9) is $\ll \mathsf{Gap}(\pi^{\mathtt{MM}})$. Furthermore, from [27], for multi-class classification,

$$\mathcal{E}^{\text{class}}_{\Theta,n,\delta} \lesssim \frac{(\mathfrak{n}_\Theta + \log(1/\delta))\log(n)\log|\mathcal{A}|}{n}$$

where $\mathfrak{n}_\Theta$ denotes the Natarajan dimension of the function class $\{f_\theta : \theta \in \Theta\}$ and $\mathcal{A}$ denotes the set of labels (which here, are the set of actions). Therefore, from eq. (7.9), we get the guarantee,

$$\mathsf{Gap}(\pi^{\mathtt{RE}}) \leq \widetilde{O}\left(\frac{H^{3/2}}{N}\left(\log(F_{\max}) \times \frac{\sum_{t=1}^{H}\mathfrak{n}_{\Theta_t}}{H}\right)^{1/2}\right), \tag{7.10}$$

whenever the underlying classification problem allows $\mathcal{E}^{\text{class}}_{\Theta,n,\delta} \asymp \mathcal{E}_{\Theta,n,\delta}$ up to problem dependent constants. Note that the polylogarthmic factors in eq. (7.10) are in $|\mathcal{A}|$, N and $1/\delta$. Under these conditions, we essentially recover a performance guarantee for `RE` which scales as $H^{3/2}/N$. This improves on the quadratic H dependence incurred by `BC`, and is optimal in the worst case, even in the tabular setting with just 3 states, as shown in [69]. This guarantee also suggests a natural measure of complexity for IL - the average Natarajan dimension, $\frac{\sum_{t=1}^{H}\mathfrak{n}_{\Theta_t}}{H}$ multiplied by the maximum log-covering number of the discriminator (or reward) class, $\log(F_{\max})$.

From this discussion, an important problem stands out: *For what kind of classification problems is* $\mathcal{E}_{\Theta,n,\delta} \asymp \mathcal{E}^{class}_{\Theta,n,\delta}$*?* While we do not answer this question in its full generality, focusing on the special case of linear classification and provide a set of sufficient conditions under which $\mathcal{E}_{\Theta,n,\delta} \asymp \mathcal{E}^{\text{class}}_{\Theta,n,\delta}$ up to problem dependent constants. In conjunction with eq. (7.10), this results in a novel guarantee for IL with linear function approximation, under significantly weaker conditions compared to prior work.

Revisiting the linear-expert setting: known transitions

In this section, we provide an upper bound on the imitation gap of `RE` in the presence of linear function approximation, which we defined previously in Definition 7.1.1 as the linear-expert setting. This setting will turn out to be a special case of the case of IL under parametric function approximation with Lipschitzness. In this section, we will make the additional assumption that the the reward function admits a linear parameterization.

Definition 7.3.1 (Policy induced by a linear classifier)**.** *Consider a set of vectors* $\theta = \{\theta_1, \cdots, \theta_H\}$ *where each* $\theta_t \in \mathbb{R}^d$*. A policy* π^θ *is said to be induced by the set of linear classifiers defined by* θ *if for all* $s \in \mathcal{S}$ *and* $t \in [H]$*,*

$$\pi^\theta_t(s) = \arg\max_{a \in \mathcal{A}} \langle \theta_t, \phi_t(s,a) \rangle. \tag{7.11}$$

By this definition, $\pi^{\text{exp}} = \pi^{\theta^E}$*.*

Definition 7.3.2 (Linear reward setting)**.** *Define* $\mathcal{R}_{\text{lin,t}}$ *as the family of linear reward functions (defined at the single time-step t) which takes the form of an unknown linear function of a set of the features,*

$$\mathcal{R}_{\text{lin,t}} = \left\{ \{ r_t(s,a) = \langle \omega, \phi_t(s,a) \rangle : s \in \mathcal{S}, a \in \mathcal{A} \} : \omega \in \mathbb{R}^d, \|\omega\|_2 \le 1 \right\}. \tag{7.12}$$

For the rewards to be 1*-bounded, we assume the features satisfy* $\|\phi_t(s,a)\|_2 \le 1$*. Define* $\mathcal{R}_{\text{lin}} = \otimes_{t=1}^H \mathcal{R}_{\text{lin,t}}$*. The linear reward setting assumes the true reward function of the MDP,* $r \in \mathcal{R}_{\text{lin}}$*.*

Remark 7.3.1. Note that our guarantees in Theorem 7.3.1 hold even if the set of features in the definition of $\mathcal{R}_{\text{lin,t}}$ in Definition 7.3.2 differ from those used to define the expert classifier Assumption 7.2.2. Regardless, we assume that both sets of features are known to the learner.

In the case of parametric function approximation with Lipschitzness, note that we assume both the weak margin condition (Assumption 7.2.3), as well as the existence of a linear classification oracle (Assumption 7.2.2). Below, in the linear expert case, we show a sufficient condition which implies both of these conditions. In particular, define the positive hemisphere with pole at θ, i.e. $\{x : \mathbb{B}^d_2 : \langle \theta, x \rangle \ge 0\}$ as $\mathbb{H}^d_\theta$. We abbreviate $\mathbb{H}^d_{\theta^E_t}$ as $\mathbb{H}^d_t$.

Assumption 7.3.1 (Bounded density assumption)**.** *For each time $t \in [H]$, state $s \in \mathcal{S}$, action $a \in \mathcal{A}$ and $\theta \in \Theta_t$, define $\overline{\phi}_t(s,a) = \phi_t(s,a^\theta_s) - \phi_t(s,a)$ where $a^\theta_s = \arg\max_{a' \in \mathcal{A}} \langle \theta, \phi_t(s,a') \rangle$. Consider the measure $\Pr_{\pi^{\exp}}\left(\exists a \neq a^\theta_{s_t} : \overline{\phi}_t(s_t,a) \in \cdot\right)$. Let $\overline{d}^E_t$ represent the Radon-Nikodym derivative of this measure against the uniform measure on $\mathbb{H}^{d-1}_t$. The bounded density assumption states that for each $t \in [H]$ there are constants $c_{\min} > 0$ and $c_{\max} < \infty$ such that for all $x \in \mathbb{H}^d_t$,*

$$c_{\min} \leq \overline{d}^E_t(x) \leq c_{\max}. \tag{7.13}$$

We now state the main result we prove for IL in the linear setting.

Theorem 7.3.1. *Assuming the linear-expert and linear reward setting (Definitions 7.1.1 and 7.3.2) and under the bounded density assumption (Assumption 7.3.1), appropriately instantiating* `RE` *ensures that with probability $\geq 1-\delta$,*

$$\mathsf{Gap}(\pi^{\texttt{RE}}) \lesssim \sqrt{\frac{c_{\max}}{c_{\min}}} \frac{H^{3/2} d^{5/4} \log^{\frac{3}{2}}(NdH/\delta)}{N}.$$

7.4 Discussion and open problems

In this chapter we proved a general meta theorem (Theorem 6.2.1) bounding the imitation gap of the policies produced by the algorithm in terms of the parameter estimation error in offline classification. This results in a guarantee for IL with the optimal dependence on the horizon, H, and number of expert rollouts available, N, under the assumption that the parameter estimation guarantee matches the generalization guarantee for the underlying offline classification problem. Under these conditions, the analysis also suggests a natural measure of complexity for IL depending on the average Natarajan dimension and log-covering number (metric entropy) of the discriminator class. It is a significant open question to extend the analysis of `RE` to depend on less stringent classification oracles, and only require constructing learners with bounded generalization error, as required by `BC`. There are reasons this may not generically be possible for classes with bounded VC dimension, but it is plausible that such guarantees can be extended under stronger assumptions. We motivate this briefly below.

Drawbacks of the offline classification oracle (Assumption 7.2.2). Note that the offline classification oracle we consider in Assumption 7.2.2 requires a learner with bounded parameter estimation error, compared to one with bounded generalization error which is typically studied in practice [27]. While under certain conditions, both measures of error have the same asymptotic scaling in n and the Natarajan dimension $\mathfrak{n}_\Theta$, in general they can be different. For instance, consider the case of linear (binary) classification under very poor coverage: denoting the true classifier as $\theta^\star \in \mathbb{R}^d$, and with the input distribution as $s \sim \mathrm{Unif}(\{e_1, -e_1\})$ where $e_1 = (1, 0, \cdots, 0) \in \mathbb{R}^d$. Then, given $n = \Omega(\log(1/\delta))$ samples,

there exists a learner which with probability $1 - \delta$ learns a classifier with 0 generalization error. However, regardless of the number of samples n, no learner can guarantee to learn $\theta^\star$, since the remaining $d - 1$ coordinates of $\theta^\star$ aside from $\theta^\star_1$ do not affect the labels of the inputs, which are what are observed by a learner. Thus, under poor coverage, the parameter $\theta^\star$ cannot be learned consistently even though there exists a trivial classifier with bounded generalization error.

Designing practical algorithms which do not require parameter tuning. The membership oracle resulting from the analysis in theory (eq. (7.4)) as well as those implemented in practice (eqs. (6.8) to (6.11)) require tuning either the margin threshold $2L\mathcal{E}_{\Theta_t,N,\delta/H}$, or the scale parameters μ and β. An interesting next direction would be to develop an algorithm which uses data dependent scales (e.g. [40]) to reduce the effort required to fine-tuning these parameters.

Bibliography

[1] Pieter Abbeel and Andrew Y Ng. "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 1.

[2] Pieter Abbeel and Andrew Y. Ng. "Apprenticeship Learning via Inverse Reinforcement Learning". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 1. ISBN: 1581138385. DOI: 10.1145/1015330.1015430. URL: https://doi.org/10.1145/1015330.1015430.

[3] Pieter Abbeel et al. "An Application of Reinforcement Learning to Aerobatic Helicopter Flight". In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, 2007, pp. 1–8. URL: http://papers.nips.cc/paper/3151-an-application-of-reinforcement-learning-to-aerobatic-helicopter-flight.pdf.

[4] Brenna D Argall et al. "A survey of robot learning from demonstration". In: *Robotics and autonomous systems* 57.5 (2009), pp. 469–483.

[5] Brenna D. Argall et al. "A survey of robot learning from demonstration". In: *Robotics and Autonomous Systems* 57.5 (2009), pp. 469–483. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2008.10.024. URL: http://www.sciencedirect.com/science/article/pii/S0921889008001772.

[6] Sanjeev Arora et al. "Provable Representation Learning for Imitation Learning via Bi-level Optimization". In: *arXiv preprint arXiv:2002.10544* (2020).

[7] Saurabh Arora and Prashant Doshi. "A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress". In: *arXiv preprint arXiv:1806.06877* (2018).

[8] Jean-Yves Audibert and Alexandre B. Tsybakov. "Fast learning rates for plug-in classifiers". In: *The Annals of Statistics* 35.2 (2007), pp. 608–633. DOI: 10.1214/009053606000001217. URL: https://doi.org/10.1214/009053606000001217.

[9] Ashwin Balakrishna et al. "On-policy robot imitation learning from a converging supervisor". In: *Conference on Robot Learning*. PMLR. 2020, pp. 24–41.

[10] Christopher Berner et al. "Dota 2 with large scale deep reinforcement learning". In: *arXiv preprint arXiv:1912.06680* (2019).

[11] Alina Beygelzimer et al. "Error limiting reductions between classification tasks". In: *ICML*. 2005, pp. 49–56. URL: https://doi.org/10.1145/1102351.1102358.

[12] Patrick Billingsley. "Statistical Methods in Markov Chains". In: *Ann. Math. Statist.* 32.1 (Mar. 1961), pp. 12–40. DOI: 10.1214/aoms/1177705136. URL: https://doi.org/10.1214/aoms/1177705136.

[13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. "Concentration Inequalities - A Nonasymptotic Theory of Independence". In: *Concentration Inequalities*. 2013.

[14] Kiante Brantley, Wen Sun, and Mikael Henaff. "Disagreement-Regularized Imitation Learning". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rkgbYyHtwB.

[15] Kianté Brantley, Wen Sun, and Mikael Henaff. "Disagreement-regularized imitation learning". In: *International Conference on Learning Representations*. 2019.

[16] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[17] Yuri Burda et al. "Exploration by random network distillation". In: *arXiv preprint arXiv:1810.12894* (2018).

[18] Carlos Celemin et al. "Interactive imitation learning in robotics: A survey". In: *Foundations and Trends® in Robotics* 10.1-2 (2022), pp. 1–197.

[19] Lingjiao Chen et al. "Are more llm calls all you need? towards scaling laws of compound inference systems". In: *arXiv preprint arXiv:2403.02419* (2024).

[20] Ching-An Cheng and Byron Boots. *Convergence of Value Aggregation for Imitation Learning*. 2018. arXiv: 1801.07292 [cs.LG].

[21] Ching-An Cheng et al. *Fast Policy Learning through Imitation and Reinforcement*. 2018. arXiv: 1805.10413 [cs.LG].

[22] Felipe Codevilla et al. "End-to-end driving via conditional imitation learning". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018.

[23] Felipe Codevilla et al. "Exploring the limitations of behavior cloning for autonomous driving". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9329–9338.

[24] Erwin Coumans and Yunfei Bai. *PyBullet, a Python module for physics simulation for games, robotics and machine learning*. http://pybullet.org. 2016.

[25] Amit Daniely and S. Shalev-Shwartz. "Optimal learners for multiclass problems". In: *COLT*. 2014.

[26] Amit Daniely and Shai Shalev-Shwartz. "Optimal Learners for Multiclass Problems". In: *CoRR* abs/1405.2420 (2014). arXiv: 1405.2420. URL: http://arxiv.org/abs/1405.2420.

[27] Amit Daniely et al. "Multiclass learnability and the ERM principle". In: *CoRR* abs/1308.2893 (2013). arXiv: 1308.2893. URL: http://arxiv.org/abs/1308.2893.

[28] Amit Daniely et al. "Multiclass learnability and the erm principle". In: *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings. 2011, pp. 207–232.

[29] Constantinos Daskalakis et al. "Training gans with optimism". In: *arXiv preprint arXiv:1711.00141* (2017).

[30] Hal Daumé III. "Unsupervised search-based structured prediction". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 209–216.

[31] David L. Donoho and Richard C. Liu. "The "Automatic" Robustness of Minimum Distance Functionals". In: *Ann. Statist.* 16.2 (June 1988), pp. 552–586. DOI: 10.1214/aos/1176350820. URL: https://doi.org/10.1214/aos/1176350820.

[32] Yan Duan et al. "One-shot imitation learning". In: *Advances in neural information processing systems* 30 (2017).

[33] Chelsea Finn et al. "One-shot visual imitation learning via meta-learning". In: *Conference on robot learning*. PMLR. 2017, pp. 357–368.

[34] Dylan J Foster, Adam Block, and Dipendra Misra. "Is behavior cloning all you need? understanding horizon in imitation learning". In: *arXiv preprint arXiv:2407.15007* (2024).

[35] Justin Fu, Katie Luo, and Sergey Levine. "Learning Robust Rewards with Adverserial Inverse Reinforcement Learning". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=rkHywl-A-.

[36] Kanishk Gandhi et al. "Stream of Search (SoS): Learning to Search in Language". In: *arXiv preprint arXiv:2404.03683* (2024).

[37] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[38] "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (2019), pp. 350–354. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1724-z. URL: https://doi.org/10.1038/s41586-019-1724-z.

[39] Ishaan Gulrajani et al. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG].

[40] Tuomas Haarnoja et al. *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*. 2018. arXiv: 1801.01290 [cs.LG].

[41] Dylan Hadfield-Menell et al. "Cooperative Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NeurIPS) 2016*. Vol. 29. 2016, pp. 3919–3927.

[42] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. "Minimax Estimation of Discrete Distributions Under ℓ_1 Loss". In: *IEEE Transactions on Information Theory* 61.11 (2015), pp. 6343–6354.

[43] Todd Hester et al. "Deep Q-learning From Demonstrations". In: *AAAI*. 2018.

[44] Jonathan Ho and Stefano Ermon. "Generative adversarial imitation learning". In: *Advances in neural information processing systems*. 2016, pp. 4565–4573.

[45] Arian Hosseini et al. "V-star: Training verifiers for self-taught reasoners". In: *arXiv preprint arXiv:2402.06457* (2024).

[46] Borja Ibarz et al. "Reward learning from human preferences and demonstrations in Atari". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 8011–8023. URL: http://papers.nips.cc/paper/8025-reward-learning-from-human-preferences-and-demonstrations-in-atari.pdf.

[47] Liyiming Ke et al. "Imitation Learning as f-Divergence Minimization". In: *arXiv preprint arXiv:1905.12888* (2019).

[48] Rahul Kidambi et al. *MOReL : Model-Based Offline Reinforcement Learning*. 2021. arXiv: 2005.05951 [cs.LG].

[49] Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. "Imitation learning via off-policy distribution matching". In: *arXiv preprint arXiv:1912.05032* (2019).

[50] Michael Laskey et al. "Dart: Noise injection for robust imitation learning". In: *Conference on robot learning*. PMLR. 2017, pp. 143–156.

[51] Jonathan N Lee et al. "A dynamic regret analysis and adaptive regularization algorithm for on-policy robot imitation learning". In: *International Workshop on the Algorithmic Foundations of Robotics*. Springer. 2018, pp. 212–227.

[52] Yuping Luo, Huazhe Xu, and Tengyu Ma. "Learning Self-Correctable Policies and Value Functions from Demonstrations with Negative Sampling". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rke-f6NKvS.

[53] Enno Mammen and Alexandre B. Tsybakov. "Smooth Discrimination Analysis". In: *The Annals of Statistics* 27.6 (1999), pp. 1808–1829. ISSN: 00905364. URL: http://www.jstor.org/stable/2673938 (visited on 05/18/2022).

[54] David Mcallester et al. "Concentration Inequalities for the Missing Mass and for Histogram Rule Error". In: *Journal of Machine Learning Research*. 2003, pp. 895–911.

[55] David A. McAllester and Robert E. Schapire. "On the Convergence Rate of Good-Turing Estimators". In: *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*. COLT '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 1–6. ISBN: 155860703X.

[56] Josh Merel et al. "Learning human behaviors from motion capture by adversarial imitation". In: *ArXiv* abs/1707.02201 (2017).

[57] Alfred Müller. "Integral probability metrics and their generating classes of functions". In: *Advances in Applied Probability* 29.2 (1997), pp. 429–443.

[58] Ashvin Nair et al. "Combining self-supervised learning and imitation for vision-based rope manipulation". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 2146–2153.

[59] Andrew Y Ng, Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping". In: *Icml*. Vol. 99. 1999, pp. 278–287.

[60] Andrew Y. Ng and Stuart J. Russell. "Algorithms for Inverse Reinforcement Learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 663–670. ISBN: 1558607072.

[61] Allen Nie et al. "EVOLvE: Evaluating and Optimizing LLMs For Exploration". In: *arXiv preprint arXiv:2410.06238* (2024).

[62] Yunpeng Pan et al. "Agile autonomous driving using end-to-end deep imitation learning". In: *arXiv preprint arXiv:1709.07174* (2017).

[63] Yunpeng Pan et al. "Imitation learning for agile autonomous driving". In: *The International Journal of Robotics Research* 39.2-3 (2020), pp. 286–302. DOI: 10.1177/0278364919880273. eprint: https://doi.org/10.1177/0278364919880273. URL: https://doi.org/10.1177/0278364919880273.

[64] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. *Self-Supervised Exploration via Disagreement*. 2019. arXiv: 1906.04161 [cs.LG].

[65] Dean A. Pomerleau. "ALVINN: An Autonomous Land Vehicle in a Neural Network". In: *Advances in Neural Information Processing Systems 1*. Ed. by D. S. Touretzky. Morgan-Kaufmann, 1989, pp. 305–313. URL: http://papers.nips.cc/paper/95-alvinn-an-autonomous-land-vehicle-in-a-neural-network.pdf.

[66] Rafael Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *arXiv preprint arXiv:2305.18290* (2023).

[67] Antonin Raffin. *RL Baselines3 Zoo*. https://github.com/DLR-RM/rl-baselines3-zoo. 2020.

[68] Antonin Raffin et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8. URL: http://jmlr.org/papers/v22/20-1364.html.

[69] Nived Rajaraman et al. *Provably Breaking the Quadratic Error Compounding Barrier in Imitation Learning, Optimally*. 2021. arXiv: 2102.12948 [cs.LG].

[70] Nived Rajaraman et al. "Toward the Fundamental Limits of Imitation Learning". In: *Advances in Neural Information Processing Systems*. 2020.

[71] Nived Rajaraman et al. "Toward the fundamental limits of imitation learning". In: *arXiv preprint arXiv:2009.05990* (2020).

[72] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. "Maximum margin planning". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 729–736.

[73] Siddharth Reddy, Anca D Dragan, and Sergey Levine. "Sqil: Imitation learning via reinforcement learning with sparse rewards". In: *arXiv preprint arXiv:1905.11108* (2019).

[74] Stephane Ross and Drew Bagnell. "Efficient Reductions for Imitation Learning". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, May 2010, pp. 661–668. URL: http://proceedings.mlr.press/v9/ross10a.html.

[75] Stephane Ross and Drew Bagnell. "Efficient Reductions for Imitation Learning". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, May 2010, pp. 661–668. URL: https://proceedings.mlr.press/v9/ross10a.html.

[76] Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning*. 2011. arXiv: 1011.0686 [cs.LG].

[77] Stéphane Ross and J. Andrew Bagnell. "Reinforcement and Imitation Learning via Interactive No-Regret Learning". In: *ArXiv* abs/1406.5979 (2014).

[78] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning." In: *AISTATS*. Ed. by Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík. Vol. 15. JMLR Proceedings. JMLR.org, 2011, pp. 627–635. URL: http://dblp.uni-trier.de/db/journals/jmlr/jmlrp15.html#RossGB11.

[79] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning." In: *AISTATS*. Ed. by Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík. Vol. 15. JMLR Proceedings. JMLR.org, 2011, pp. 627–635. URL: http://dblp.uni-trier.de/db/journals/jmlr/jmlrp15.html#RossGB11.

[80] Tim Salimans and Richard Chen. "Learning Montezuma's Revenge from a Single Demonstration". In: *arXiv preprint arXiv:1812.03381* (2018).

[81] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.

[82] Bilgehan Sel et al. "Algorithm of thoughts: Enhancing exploration of ideas in large language models". In: *arXiv preprint arXiv:2308.10379* (2023).

[83] Amrith Setlur et al. "RL on Incorrect Synthetic Data Scales the Efficiency of LLM Math Reasoning by Eight-Fold". In: *arXiv preprint arXiv:2406.14532* (2024).

[84] Shai Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194.

[85] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014. ISBN: 1107057132.

[86] David Silver et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[87] Avi Singh et al. "Beyond human data: Scaling self-training for problem-solving with language models". In: *arXiv preprint arXiv:2312.06585* (2023).

[88] Jiaming Song et al. "Multi-agent generative adversarial imitation learning". In: *Advances in neural information processing systems* 31 (2018).

[89] Nisan Stiennon et al. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems*. 2020.

[90] Mingfei Sun et al. "Softdice for imitation learning: Rethinking off-policy distribution matching". In: *arXiv preprint arXiv:2106.03155* (2021).

[91] Wen Sun et al. "Deeply aggrevated: Differentiable imitation learning for sequential prediction". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 3309–3318.

[92] Wen Sun et al. "Provably Efficient Imitation Learning from Observation Alone". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, June 2019, pp. 6036–6045. URL: http://proceedings.mlr.press/v97/sun19b.html.

[93] Gokul Swamy et al. "Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap". In: 2021. URL: https://arxiv.org/abs/2103.03236.

[94] Gokul Swamy et al. "Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 10022–10032. URL: https://proceedings.mlr.press/v139/swamy21a.html.

[95] Umar Syed, Michael Bowling, and Robert E Schapire. "Apprenticeship learning using linear programming". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1032–1039.

[96] Umar Syed and Robert E Schapire. "A Game-Theoretic Approach to Apprenticeship Learning". In: *Advances in Neural Information Processing Systems 20*. Ed. by J. C. Platt et al. Curran Associates, Inc., 2008, pp. 1449–1456. URL: http://papers.nips.cc/paper/3293-a-game-theoretic-approach-to-apprenticeship-learning.pdf.

[97] Vasilis Syrgkanis et al. "Fast convergence of regularized learning in games". In: *arXiv preprint arXiv:1507.00407* (2015).

[98] Faraz Torabi, Garrett Warnell, and Peter Stone. "Behavioral cloning from observation". In: *arXiv preprint arXiv:1805.01954* (2018).

[99] Leslie G Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[100] Ruohan Wang et al. "Random expert distillation: Imitation learning via expert policy support estimation". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6536–6544.

[101] Sean Welleck et al. "From decoding to meta-generation: Inference-time algorithms for large language models". In: *arXiv preprint arXiv:2406.16838* (2024).

[102] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. "Maximum Entropy Deep Inverse Reinforcement Learning". In: *arXiv preprint arXiv:1507.04888* (2015).

[103] Tian Xu, Ziniu Li, and Yang Yu. "Error Bounds of Imitating Policies and Environments". In: *arXiv preprint arXiv:2010.11876* (2020).

[104] Tian Xu et al. "Provably efficient adversarial imitation learning with unknown transitions". In: *Uncertainty in Artificial Intelligence*. PMLR. 2023, pp. 2367–2378.

[105] Yannis G. Yatracos. "Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy". In: *Ann. Statist.* 13.2 (June 1985), pp. 768–774. DOI: 10.1214/aos/1176349553. URL: https://doi.org/10.1214/aos/1176349553.

[106] L.A. Zadeh. "Fuzzy sets". In: *Information and Control* 8.3 (1965), pp. 338–353. ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(65)90241-X. URL: https://www.sciencedirect.com/science/article/pii/S001999586590241X.

[107] Eric Zelikman et al. "Star: Bootstrapping reasoning with reasoning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15476–15488.

[108] Yufeng Zhang et al. "Generative adversarial imitation learning with neural networks: Global optimality and convergence rate". In: *arXiv preprint arXiv:2003.03709* (2020).

[109] Yuke Zhu et al. *Reinforcement and Imitation Learning for Diverse Visuomotor Skills*. 2018. URL: https://openreview.net/forum?id=HJWGdbbCW.

[110] Brian D Ziebart et al. "Maximum entropy inverse reinforcement learning." In: *Aaai*. Vol. 8. Chicago, IL, USA. 2008, pp. 1433–1438.

Appendix A

Proof of main results in Chapter 2

We provide proofs for the theorems introduced in Chapter 2 within this appendix. We push the proofs of some of the results invoked in the first few sections later on to Appendices A.3 to A.7.

A.1 No-interaction setting with a determinstic expert

In this section, we discuss the no-interaction setting where the learner is provided access to a dataset D of N trajectories generated by rolling out the expert's policy π^{exp}, and is otherwise not allowed to interact with the MDP. Our goal is to provide guarantees on the expected imitation gap of a policy that carries out BC when the expert's policy is deterministic. As stated previously, we realize this guarantee by first bounding the population 0-1 risk of BC (Theorem 2.1.2) and then invoking the black box reduction guarantee from [74].

Analysis of expected imitation gap of BC

We first discuss the proof of Lemma 2.1.2 and Eq. (2.1.3.1), which bounds the expected imitation gap of a policy carrying out BC, assuming the expert's policy is deterministic. Recall that the population 0-1 loss is defined as,

$$\mathcal{L}_{0\text{-}1}(\widehat{\pi}) = \frac{1}{H}\sum_{t=1}^{H} \mathbb{E}_{s_t \sim f^t_{\pi^{\text{exp}}}}\left[\mathbb{E}_{a\sim\widehat{\pi}_t(\cdot|s_t)}\left[\mathbb{1}(a \neq \pi^{\text{exp}}_t(s))\right]\right]. \tag{A.1}$$

where $f^t_{\pi^{\text{exp}}}$ is the state distribution induced at time t rolling out the expert's policy π^{exp}. We consider a learner $\widehat{\pi}$ that carries out BC given the demonstration dataset D in advance. In particular, the learner's policy $\widehat{\pi}$ is a member of $\Pi^{\texttt{BC}}_{\text{det}}(D)$ since it exactly mimics the expert on the states that were visited at each time in some trajectory in the demonstration dataset.

Thus the contribution to the population 0-1 risk comes from the remaining states $s \in \mathcal{S}_t(D)$,

$$\mathcal{L}_{0\text{-}1}(\widehat{\pi}) \leq \frac{1}{H}\sum_{t=1}^{H} \mathbb{E}_{s_t \sim f^t_{\pi^{\text{exp}}}}\Big[\mathbb{1}(s_t \notin \mathcal{S}_t(D))\Big], \tag{A.2}$$

$$= \frac{1}{H}\sum_{t=1}^{H}\sum_{s\in\mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_t = s]\mathbb{1}(s \notin \mathcal{S}_t(D)). \tag{A.3}$$

Taking expectation on both sides gives,

$$\mathbb{E}\left[\mathcal{L}_{0\text{-}1}(\widehat{\pi})\right] \leq \frac{1}{H}\sum_{t=1}^{H}\sum_{s\in\mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_t = s]\Pr(s \notin \mathcal{S}_t(D)). \tag{A.4}$$

In Lemma A.1.1 we show that this expression is bounded by $\lesssim |\mathcal{S}|/N$, which completes the proof of the population 0-1 risk bound of BC in Theorem 2.1.2.

Lemma A.1.1. $\mathbb{E}\left[\sum_{t=1}^{H}\sum_{s\in\mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_t = s]\ \Pr[s \notin \mathcal{S}_t(D)]\right] \leq \frac{4}{9}\frac{|\mathcal{S}|H}{|D|}$.

As stated previously, we subsequently invoke the supervised learning reduction in [74, Theorem 2.1] to upper bound the expected imitation gap of a learner carrying out BC in Theorem 2.1.3.1. The previous discussion is also amenable for establishing a high probability bound on the expected imitation gap of BC. Indeed, consider the upper bound on the population 0-1 risk of BC in eq. (A.3), which is a function of the demonstration dataset D. It captures the probability mass under the expert's state distribution contributed by states unobserved in the demonstration dataset.

High probability bounds on BC

It turns out that the contribution to the upper bound on population 0-1 risk of BC in eq. (A.3) is captured by the notion of "missing mass" of the time-averaged state distribution under the expert's policy. The high probability result for BC (Theorem 2.1.3.2) follows shortly by invoking existing concentration bounds for missing mass.

Definition A.1.1 (Missing mass)**.** *Consider some distribution ν on $\mathcal{X}$, and let $X^N \overset{i.i.d.}{\sim} \nu$ be a dataset of N samples drawn i.i.d. from ν. Let $\mathfrak{n}_x(X^N) = \sum_{i=1}^{N}\mathbb{1}(X_i = x)$ be the number of times the symbol x was observed in X^N. Then, the missing mass $\mathfrak{m}_0(\nu, X^N) = \sum_{x\in\mathcal{X}}\nu(x)\mathbb{1}(\mathfrak{n}_x(X^N) = 0)$ is the probability mass contributed by symbols never observed in X^N.*

It turns out that the missing mass of an arbitrary discrete distribution admits sub-Gaussian concentration. Invoking [54, Lemma 11] establishes the following concentration guarantee for missing mass. A proof of the result is provided in Appendix A.3.

Theorem A.1.2. *Consider an arbitrary distribution ν on $\mathcal{X}$, and let $X^N \overset{i.i.d.}{\sim} \nu$ be a dataset of N samples drawn i.i.d. from ν. Consider any $\delta \in (0, 1/10]$. Then,*

$$\Pr\left(\mathfrak{m}_0(\nu, X^N) - \mathbb{E}[\mathfrak{m}_0(\nu, X^N)] \geq \frac{3\sqrt{|\mathcal{X}|\log(1/\delta)}}{N}\right) \leq \delta. \tag{A.5}$$

Consider the upper bound to the population 0-1 loss in eq. (A.4). Observe that for each fixed $\tau \in [H]$, $\sum_{s \in \mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_\tau = s]\mathbb{1}(s \notin \mathcal{S}_\tau(D))$ is the missing mass of $f^\tau_{\pi^{\text{exp}}}$, given N samples from the distribution. Recall that $f^\tau_{\pi^{\text{exp}}}$ is the distribution over states at time τ rolling out π^{exp}. Thus we can invoke the concentration bound from Theorem A.1.2 to prove that the upper bound on 0-1 loss in eq. (A.4) concentrates. We formally state this result in Lemma A.1.3.

Lemma A.1.3. *For any δ such that $\delta \in (0, \min\{1, H/10\}]$, with probability $\geq 1 - \delta$ over the randomness of the demonstration dataset D,*

$$\frac{1}{H}\sum_{\tau=1}^{H}\sum_{s \in \mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_\tau = s]\mathbb{1}(s \notin \mathcal{S}_\tau(D)) \leq \frac{4|\mathcal{S}|}{9N} + \frac{3\sqrt{|\mathcal{S}|}\log(H/\delta)}{N}. \tag{A.6}$$

Plugging this result into eq. (A.4) provides an upper bound on the population 0-1 risk of `BC`. Subsequently invoking [74, Theorem 2.1], we arrive at the high probability bound on $\mathsf{Gap}(\widehat{\pi})$ for `BC` in Eq. (2.1.3.2).

A.2 No-interaction setting with a stochastic expert

In this section we continue to discuss the no-interaction setting, but drop the assumption that the expert plays a deterministic policy. We assume the expert plays a general stochastic policy.

Analyzing expected imitation gap of `Mimic-MD`

In this section we discuss the proof of Theorem 2.3.2 which bounds the expected imitation gap of `Mimic-MD`. Recall that the objective is to upper bound $\mathbb{E}[\mathsf{Gap}(\widehat{\pi})]$ when the learner carries out `Mimic-MD`. The outline of the proof is to construct two policies π^{first} and $\pi^{\text{orc}-\text{first}}$ that are functions of the dataset D.

The policy π^{first} is easy to describe: order the demonstration dataset arbitrarily, and at a state, play the action in the first trajectory in D that visits it, if it exists. If no such trajectories exist, the policy plays $\text{Unif}(\mathcal{A})$. In particular, we show that the value of π^{first} and `Mimic-MD` are the same, taking expectation over the demonstration dataset D (Lemma A.2.1).

On the other hand, we consider an oracle policy $\pi^{\text{orc}-\text{first}}$ which is very similar. Indeed, $\pi^{\text{orc}-\text{first}}$ first orders the demonstration dataset in the same manner as π^{first}. At any state, it too plays the action in the first trajectory in D that visits it, if it exists. However, if such a trajectory does not exist, $\pi^{\text{orc}-\text{first}}$ simply samples an action from the expert's action distribution and plays it at this state. This explains the namesake of the policy, since it requires oracle access to the expert's policy. By virtue of choosing actions this way, we show that the value of $\pi^{\text{orc}-\text{first}}$ in expectation equals $J(\pi^{\text{exp}})$ (Lemma A.2.3).

At an intuitive level the elements of the proof seem to be surfacing: $\pi^{\text{orc}-\text{first}}$ matches π^{exp} in value, but is not available to the learner. However, it shares a lot of similarity to π^{first}, which in expectation matches $\widehat{\pi}$ in value, the policy we wish to analyze. Informally,

$$\widehat{\pi} \iff \pi^{\text{first}} \approx \pi^{\text{orc}-\text{first}} \iff \pi^{\text{exp}}. \tag{A.7}$$

Thus to establish the bound, we carry out an analysis of $J(\pi^{\text{orc}-\text{first}}) - J(\pi^{\text{first}})$. Indeed we show that since the two policies are largely the same, the learner is suboptimal only on the trajectories where at some point a state is visited where the policies do not match. The final element of the proof is to show that this event in fact occurs with low probability given an demonstration dataset of sufficiently large size.
Before delving into the formal definitions of π^{first} and $\pi^{\text{orc}-\text{first}}$ and other elements of the proof, we introduce a modicum of relevant notation.

Notation. Let the trajectories in the demonstration dataset D be ordered arbitrarily as $\{\mathsf{tr}_1, \cdots, \mathsf{tr}_N\}$. In addition, we denote each trajectory tr_n explicitly as $\{(s_1^n, a_1^n), \cdots, (s_H^n, a_H^n)\}$. For each state $s \in \mathcal{S}$ we define,

$$N_{t,s} = \{n \in [N] : s_t^n = s\}, \tag{A.8}$$

as the (totally) ordered set of indices of trajectories in D which visit the state s at time t.
The policy $\widehat{\pi}$ returned by `Mimic-MD` samples an action from the empirical estimate of the expert's policy at each state wherever available. On the remaining states, the learner plays the distribution $\text{Unif}(\mathcal{A})$.
Given the ordered dataset D, we define the policy $\pi^{\text{first}}(D)$ as,

$$\pi_t^{\text{first}}(\cdot|s) = \begin{cases} \delta_{a_t^n} & \text{if } |N_{t,s}| \geq 1, \text{ where } n = \min(N_{t,s}), \\ \text{Unif}(\mathcal{A}) & \text{otherwise.} \end{cases} \tag{A.9}$$

That is, $\pi^{\text{first}}(D)$ plays the action in the first trajectory that visits the state s at time t.

In order to analyze the expected imitation gap of $\widehat{\pi}(D)$, we first show that $\widehat{\pi}(D)$ and $\pi^{\text{first}}(D)$ have the same value in expectation, and instead study the policy $\pi^{\text{first}}(D)$.

Lemma A.2.1. $\mathbb{E}[J(\widehat{\pi}(D))] = \mathbb{E}[J(\pi^{\text{first}}(D))]$.

With this result, we can write the expected imitation gap of the learner $\widehat{\pi}$ as,

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] = J(\pi^{\text{exp}}) - \mathbb{E}[J(\pi^{\text{first}}(D))]. \tag{A.10}$$

We next move on to the discussion of $\pi^{\text{orc}-\text{first}}$ which is an oracle version of π^{first}. Informally, at any state $\pi^{\text{orc}-\text{first}}$ plays the action from the first trajectory that visits it in D, if available. However on the remaining states instead of playing $\text{Unif}(\mathcal{A})$, $\pi^{\text{orc}-\text{first}}$ samples an action from the expert's action distribution and plays it at this state. Thus, $\pi^{\text{orc}-\text{first}}$ is coupled with the demonstration dataset D.

Prior to discussing $\pi^{\text{orc}-\text{first}}$ in greater depth, we first introduce some preliminaries. In particular, we adopt an alternate view of the process generating the demonstration dataset D which will play a central role in formally defining $\pi^{\text{orc}-\text{first}}$. We mention that this approach is inspired by the alternate view of Markov processes in [12].

To this end, we first define an "expert table" which is a fixed infinite collection of actions at each state and time which the expert draws upon while generating the trajectories in D.

Definition A.2.1 (Expert table)**.** *The expert table,* $\mathbf{T}^\star$ *is a collection of random variables* $\mathbf{T}^\star_{t,s}(i)$ *indexed by* $t \in [H]$*,* $s \in \mathcal{S}$ *and* $i = 1, 2, \cdots$*. Fixing* $s \in \mathcal{S}$ *and* $t \in [H]$*, for* $i = 1, 2, \cdots$*, each* $\mathbf{T}^\star_{t,s}(i)$ *is drawn independently* $\sim \pi^{\text{exp}}_t(\cdot|s)$*.*

In a sense, the expert table fixes the randomness in the expert's non-determinstic policy. As promised, we next present the alternate view of generating the demonstration dataset D, where the expert sequentially samples actions from the expert table at visited states.

Lemma A.2.2 (Alternate view of generating D)**.** *Generate a dataset* D *of* N *trajectories as follows: For the* n^{th} *trajectory* tr_n*, the state* s^n_1 *is drawn independently from* ρ*. The action* a^n_1 *is assigned as the first action from* $\mathbf{T}^\star_{1,s^n_1}(\cdot)$ *that was not chosen in a previous trajectory. Then the MDP independently samples the state* $s^n_2 \sim P_1(\cdot|s^n_1, a^n_1)$*. In general, at time* t *the action* a^n_t *is drawn as the first action in* $\mathbf{T}^\star_{t,s^n_t}(\cdot)$ *that was not chosen at time* t *in any previous trajectory* $n' < n$*. The subsequent state* s^n_{t+1} *is drawn independently* $\sim P_{t+1}(\cdot|s^n_t, a^n_t)$*.*
The probability of generating a dataset $D = \{\mathsf{tr}_1, \cdots, \mathsf{tr}_N\}$ *by this procedure is* $= \prod_{n=1}^N \Pr_{\pi^{\text{exp}}}[\mathsf{tr}_n]$*. This is the same as if the trajectories were generated by independently rolling out* π^{exp} *for* N *episodes.*

Proof. Starting from the initial state $s^n_1 \sim \rho$, the probability of $\mathsf{tr}_n = \{(s_1, a_1), \cdots, (s_H, a_H)\}$ is,

$$\Pr\Big(\mathsf{tr}_n = \{(s_1, a_1), \cdots, (s_H, a_H)\}\Big) = \rho(s_1)\left(\prod\nolimits_{t=1}^{H-1} \pi^{\text{exp}}_t(a_t|s_t)P_t(s_{t+1}|s_t, a_t)\right)\pi^{\text{exp}}_H(a_H|s_H).$$

This relies on the fact that each action in $\mathbf{T}^\star_{t,s}(\cdot)$ is sampled independently from $\pi^{\text{exp}}_t(\cdot|s)$. Carrying out the same calculation for the n trajectories jointly (which we avoid to keep notation simple) results in the claim. The important element remains the same: each action in $\mathbf{T}^\star_{t,s^n_t}(\cdot)$ is sampled independently from $\pi^{\text{exp}}_t(\cdot|s^n_t)$. □

Note that the process in Lemma A.2.2 generates a dataset having the same distribution as if the trajectories were generated by independently rolling out π^{exp} for N episodes. Without loss of generality we may therefore assume that the expert generates D this way. We adopt this alternate view to enable the coupling between the expert's and learner's policies.

Remark A.2.1. We emphasize that the infinite table $\mathbf{T}^\star$ is not known to the learner and is only used by the expert to generate the dataset D. However, by virtue of observing the trajectories in D the learner is revealed some part of $\mathbf{T}^\star$. In particular at the state s and time t, the first $|N_{t,s}|$ actions in $\mathbf{T}^\star_{t,s}$ are revealed to the learner.

Recall that $\pi^{\text{first}}_t(\cdot|s)$ defined in eq. (A.9) deterministically plays the action in the first trajectory in D that visits a state s at time t, if available, and otherwise plays the uniform distribution $\text{Unif}(\mathcal{A})$.

Using the alternate view of generating D in Lemma A.2.2, this policy can be equivalently defined as one which plays the action at the first position in the table $\mathbf{T}^\star$ if observed, and

otherwise plays the uniform distribution.

$$\pi_t^{\text{first}}(\cdot|s) = \begin{cases} \delta_{\mathbf{T}^\star_{t,s}(1)} & \text{if } |N_{t,s}| > 0, \\ \text{Unif}(\mathcal{A}), & \text{otherwise.} \end{cases} \tag{A.11}$$

We now define the oracle policy $\pi^{\text{orc}-\text{first}}$, which plays the first action at each time $t \in [H]$ at each state $s \in \mathcal{S}$. That is,

$$\pi_t^{\text{orc}-\text{first}}(\cdot|s) = \delta_{\mathbf{T}^\star_{t,s}(1)}. \tag{A.12}$$

With this definition, we first identify that the expected value of $\pi^{\text{orc}-\text{first}}$ equals $J(\pi^{\text{exp}})$.

Lemma A.2.3. $J(\pi^{\text{exp}}) = \mathbb{E}\left[J(\pi^{\text{orc}-\text{first}})\right]$.

Plugging this into eq. (A.10), we see that,

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] = \mathbb{E}\left[J(\pi^{\text{orc}-\text{first}}) - J(\pi^{\text{first}})\right]. \tag{A.13}$$

Observe that $\pi^{\text{orc}-\text{first}}$ and π^{first} are in fact identical on all the states that were visited at least once in the demonstration dataset (i.e. having $|N_{t,s}| > 0$). Therefore, as long as the state s visited at each time t in an episode has $|N_{t,s}| > 0$, both policies collect the same cumulative reward.

Lemma A.2.4. *Fix the expert table $\mathbf{T}^\star$ and the demonstration dataset D. Define $\mathcal{E}^c$ as the "good" event that the trajectory under consideration only visits a state s_t at each time $t \in [H]$ such that $|N_{t,s_t}| > 0$, i.e. states that have been observed in the demonstration dataset D at time t. Then,*

$$\mathbb{E}_{\pi^{\text{first}}}\left[\left(\sum_{t=1}^H r_t(s_t, a_t)\right) \mathbb{1}\left(\mathcal{E}^c\right)\right] = \mathbb{E}_{\pi^{\text{orc}-\text{first}}}\left[\left(\sum_{t=1}^H r_t(s_t, a_t)\right) \mathbb{1}\left(\mathcal{E}^c\right)\right]. \tag{A.14}$$

Proof. Both policies are identical on the states such that $|N_{t,s}| > 0$. The event $\mathcal{E}^c$ guarantees that only such states are visited in a trajectory. Therefore both expectations are equal. □

With these preliminaries, we have most of the ingredients to prove the bound on the expected imitation gap of `Mimic-MD`. To this end, from eq. (A.13) we see that,

$$\mathsf{Gap}(\widehat{\pi}) = \mathbb{E}_{\pi^{\text{orc}-\text{first}}}\left[\sum_{t=1}^H r_t(s_t, a_t)\right] - \mathbb{E}_{\pi^{\text{first}}}\left[\sum_{t=1}^H r_t(s_t, a_t)\right]. \tag{A.15}$$

Subsequently invoking Lemma A.2.4, we see that

$$\begin{aligned} \mathsf{Gap}(\widehat{\pi}) &= \mathbb{E}_{\pi^{\text{orc}-\text{first}}}\left[\left(\sum_{t=1}^H r_t(s_t, a_t)\right) \mathbb{1}\left(\mathcal{E}\right)\right] - \mathbb{E}_{\pi^{\text{first}}}\left[\left(\sum_{t=1}^H r_t(s_t, a_t)\right) \mathbb{1}\left(\mathcal{E}\right)\right], \\ &\leq \mathbb{E}_{\pi^{\text{orc}-\text{first}}}\left[\left(\sum_{t=1}^H r_t(s_t, a_t)\right) \mathbb{1}\left(\mathcal{E}\right)\right], \qquad \text{(A.16)} \\ &\leq H \Pr_{\pi^{\text{orc}-\text{first}}}[\mathcal{E}], \qquad \text{(A.17)} \end{aligned}$$

where in the last inequality we use the fact that pointwise $0 \leq r_t \leq 1$ for all $t \in [H]$. Taking expectation gives the inequality,

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \leq H\mathbb{E}[\Pr_{\pi^{\text{orc}-\text{first}}}[\mathcal{E}]]. \tag{A.18}$$

In Lemma A.2.5 we show that $\mathbb{E}[\Pr_{\pi^{\text{orc}-\text{first}}}[\mathcal{E}]]$ is upper bounded by $|\mathcal{S}|H\ln(N)/N$, which completes the proof.

Lemma A.2.5. $\mathbb{E}\left[\Pr_{\pi^{\text{orc}-\text{first}}}[\mathcal{E}]\right] \leq \frac{|\mathcal{S}|H\ln(N)}{N}$.

Although the oracle policy $\pi^{\text{orc}-\text{first}}$ and the dataset D are coupled, the key intuition behind showing that the event $\mathcal{E}$ occurs with low probability is that: it is not possible that, in expectation $\pi^{\text{orc}-\text{first}}$ visits some state s with high probability, but the same state s visited in the dataset D with low probability. This is by virtue of the fact that in expectation $\pi^{\text{orc}-\text{first}}$ matches π^{exp} which is the policy that generates D.

Known-transition setting under deterministic expert policy

In this section, we describe the proof of Eq. (2.4.1.1) which upper bounds the expected imitation gap of `Mimic-MD` (Algorithm 2).
Recall that `Mimic-MD`, true to its name, mimics the expert on the states observed in half the dataset D_1. By virtue of the learner mimicking the expert on states visited in D_1, we show that the learner is suboptimal only upon visiting a state unobserved in D_1 at some point in an episode.

Lemma A.2.6. *Define $\mathcal{E}_{D_1}^{\leq t} = \{\exists \tau < t : \; s_t \notin \mathcal{S}_t(D_1)\}$ as the event that the policy under consideration visits some state at time t that no trajectory in D_1 has visited at time t. Fixing the expert datset D, for any policy $\widehat{\pi}^{\texttt{BC}} \in \Pi_{\text{det}}^{\texttt{BC}}(D_1)$,*

$$\mathsf{Gap}(\widehat{\pi}^{\texttt{BC}}) = \sum_{t=1}^{H} \left\{ \mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{1}(\mathcal{E}_{D_1}^{\leq t}) r_t(s_t, a_t)\right] - \mathbb{E}_{\widehat{\pi}(D)}\left[\mathbb{1}(\mathcal{E}_{D_1}^{\leq t}) r_t(s_t, a_t)\right] \right\}. \tag{A.19}$$

Simplifying this result further using the fact that the reward function is bounded in $[0, 1]$ results in eq. (2.8), recall which we used as a basis for motivating the design of `Mimic-MD` in Section 2.4. In particular, any policy $\widehat{\pi}^{\texttt{BC}}$ that exactly mimics the expert on states observed in D_1 has imitation gap bounded by,

$$\mathsf{Gap}(\widehat{\pi}^{\texttt{BC}}) \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{t=1}^{H} \left| \Pr_{\pi^{\text{exp}}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right] - \Pr_{\widehat{\pi}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right] \right|.$$

The minimum distance functional considered in `Mimic-MD` simply replaces the population term $\Pr_{\pi^{\text{exp}}}[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a]$ by its empirical estimate computed using the dataset D_2. We follow the standard analysis of minimum distance function estimators using the triangle inequality, which in effect reduces the analysis to a question of convergence of the empirical estimate of $\Pr_{\pi^{\text{exp}}}[\mathcal{E}_{D_1}^{\leq t}, s_t = \cdot, a_t = \cdot]$ to the population in ℓ_1 distance.

Before stating the formal lemma, recall that

$$\mathcal{T}_t^{D_1}(s,a) \triangleq \left\{ \{(s_1,a_1),\cdots,(s_H,a_H)\} \middle| s_t = s, a_t = a,\ \exists \tau < t : s_\tau \notin \mathcal{S}_\tau(D_1) \right\}. \tag{A.20}$$

is defined as the set of trajectories that (i) visits the state s at time t, (ii) plays the action a at this time, and (iii) at some time $\tau \leq t$ visits a state unobserved in D_1.

Lemma A.2.7. *Consider any policy $\widehat{\pi}^{\varepsilon}$ which solves the optimization problem in (OPT-MD) to an additive error of ε. Fixing the demonstration dataset D,*

$$\mathsf{Gap}(\widehat{\pi}^{\varepsilon}) \leq 2 \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{t=1}^{H} \left| \Pr_{\pi^{\exp}} \left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a \right] - \frac{\sum_{\mathsf{tr}\in D_2} \mathbb{1}(\mathsf{tr} \in \mathcal{T}_t^{D_1}(s,a))}{|D_2|} \right| + \varepsilon.$$

We emphasize here that $\frac{\sum_{\mathsf{tr}\in D_2} \mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}$ is the empirical estimate of $\Pr_{\widehat{\pi}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right]$ computed using the trajectories in the dataset D_2.

Remark A.2.2. Taking $\varepsilon = 0$ in Lemma A.2.7 captures the case where $\widehat{\pi}^{\varepsilon}$ is the policy returned by `Mimic-MD`.

The last remaining ingredient in proving the guarantee on the expected imitation gap of `Mimic-MD` in Eq. (2.4.1.1) is to bound the convergence rate of the expectation of the RHS of Lemma A.2.7. We carry out this analysis roughly in two parts:

(i) fixing the dataset D_1, for each $t \in [H]$ we bound the convergence rate of the empirical distribution estimate (computed using D_2) of $\Pr_{\widehat{\pi}}[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a]$ to the population in ℓ_1 distance, and

(ii) we show that the resulting bound (which is a function of D_1) has small expectation and converges to 0 quickly.

This establishes the following bound on the expected imitation gap incurred by `Mimic-MD`.

Lemma A.2.8.

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{t=1}^{H} \mathbb{E}\left[\left| \Pr_{\pi^{\exp}} \left[\mathcal{T}_t^{D_1}(s,a) \right] - \frac{\sum_{\mathsf{tr}\in D_2} \mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|} \right| \right] \leq \min\left\{ \sqrt{\frac{8|\mathcal{S}|H^2}{N}},\ \frac{8}{3}\frac{|\mathcal{S}|H^{\frac{3}{2}}}{N} \right\}. \tag{A.21}$$

In conjunction with Lemma A.2.7 this completes the proof of Eq. (2.4.1.1) (by plugging in $\varepsilon = 0$ and noting Remark A.2.2) and also Corollary 2.4.1.
To show the high probability guarantee on `Mimic-MD` in Eq. (2.4.1.2), the key approach is similar. However, we instead

(i) fix D_1 and use sub-Gaussian concentration [13] to establish high probability deviation bounds on the empirical estimate of $\Pr_{\widehat{\pi}}[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a]$, and

(*ii*) use missing mass concentration (Theorem A.1.2) to show that the resulting deviations (which are a function of D_1) concentrate.

Lemma A.2.9. *Fix $\delta \in (0, \min\{1, H/5\})$. Then, with probability $\geq 1-\delta$,*

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right| \\ \lesssim \frac{|\mathcal{S}|H^{3/2}}{N}\left(1+\frac{3\log(2|\mathcal{S}|H/\delta)}{\sqrt{|\mathcal{S}|}}\right)^{1/2}\sqrt{\log(2|\mathcal{S}|H/\delta)}. \tag{A.22}$$

The high probability guarantee for `Mimic-MD` follows suit by invoking Lemma A.2.7 with $\varepsilon = 0$ and Lemma A.2.9.

Proof of lower bounds

In this section we discuss lower bounds on the expected imitation gap of any algorithm in the no-interaction, active and known-transition settings.

Active and no-interaction settings

In this section we discuss the proof of the lower bound in Theorem 2.2.1 which applies in the no-interaction and active settings. We emphasize that the active setting is strictly a generalization of the no-interaction setting: they are no different if the learner queries and plays the expert's action at each time while interacting with the MDP.
Formally, in the active setting, we assume the learner sequentially rolls out policies $\pi_1, \cdots \pi_N$ to generate trajectories $\mathsf{tr}_1, \cdots, \mathsf{tr}_N$. The learner is aware of the expert's action at each state visited in each trajectory tr_n, however may or may not choose to play this action while rolling out π_n. We assume that the policy π_n is learned causally, and can depend on all the previous information collected by the learner: the trajectories $\mathsf{tr}_1, \cdots, \mathsf{tr}_{n-1}$, as well as the expert's policy at each state visited in these trajectories.

Notation. We use $D = \mathsf{tr}_1, \cdots, \mathsf{tr}_n$ to denote the trajectories collected by the learner by rolling out $\pi_1, \cdots, \pi_N$. In addition the learner exactly knows the expert's policy $\pi_t^{\exp}(\cdot|s)$ at all states $s \in \mathcal{S}_t(D)$. We also define $A = \{\pi_t^{\exp}(\cdot|s) : t \in [H], s \in \mathcal{S}_t(D)\}$ as the expert's policy at states visited in D, which is also known to the learner by virtue of actively querying the expert.

The expert policy is deterministic in the lower bound instances we construct. Therefore, we define $\Pi_{\mathrm{det}}^{\mathtt{BC}}(D, A)$ (similar to $\Pi_{\mathrm{det}}^{\mathtt{BC}}(D)$ in eq. (2.3)) as the family of deterministic policies which mimics the expert on the states visited in D. Namely,

$$\Pi_{\mathrm{det}}^{\mathtt{BC}}(D, A) \triangleq \left\{\pi \in \Pi_{\mathrm{det}} : \forall t \in [H], s \in \mathcal{S}_t(D),\ \pi_t(s) = \pi_t^A(s)\right\}, \tag{A.23}$$

where $\delta_{\pi_t^A(s)}$ is the policy observed by the learner upon actively querying the expert in a trajectory that visits s at time t. Informally, $\Pi^{\mathtt{BC}}_{\mathrm{det}}(D, A)$ is the family of expert policies which are "compatible" with the dataset (D, A) collected by the learner.
Define $\mathbb{M}_{\mathcal{S},\mathcal{A},H}$ as the family of MDPs over state space $\mathcal{S}$, action space $\mathcal{A}$ and epsiode length H. In order to prove the lower bound on the worst-case expected imitation gap of any learner $\widehat{\pi}(D, A)$, it suffices to lower bound the Bayes expected imitation gap. Namely, it suffices to find a joint distribution $\mathcal{P}$ over MDPs and expert policies supported on $\mathbb{M}_{\mathcal{S},\mathcal{A},H} \times \Pi_{\mathrm{det-exp}}$ such that,

$$\mathbb{E}_{(\mathcal{M},\pi^{\mathrm{exp}})\sim\mathcal{P}}\Big[J_{\mathcal{M}}(\pi^{\mathrm{exp}}) - \mathbb{E}\left[J_{\mathcal{M}}(\widehat{\pi})\right]\Big] \gtrsim \min\left\{H, \frac{|\mathcal{S}|H^2}{N}\right\}. \tag{A.24}$$

Construction of $\mathcal{P}$. First we choose the expert's policy uniformly from Π_{det}. That is, for each $t \in [H]$ and $s \in \mathcal{S}$, $\pi_t^{\mathrm{exp}}(s) \sim \mathrm{Unif}(\mathcal{A})$. Conditioned on π^{exp}, the distribution over MDPs induced by $\mathcal{P}$ is deterministic and given by the MDP $\mathcal{M}[\pi^{\mathrm{exp}}]$ in fig. A.1. $\mathcal{M}[\pi^{\mathrm{exp}}]$ is defined with respect to a fixed initial distribution over states $\rho = \{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-2)\zeta, 0\}$ where $\zeta = \frac{1}{N+1}$. In addition, there is a special state $b \in \mathcal{S}$ which we refer to as the "bad state". At each state $s \in \mathcal{S} \setminus \{b\}$, choosing the expert's action renews the state in the initial distribution ρ and dispenses a reward of 1, while any other choice of action deterministically transitions to the bad state and offers no reward. In addition, the bad state is absorbing and dispenses no reward irrespective of the choice of action. That is,

$$P_t(\cdot|s,a) = \begin{cases} \rho, & s \in \mathcal{S} \setminus \{b\},\ a = \pi_t^{\mathrm{exp}}(s) \\ \delta_b, & \text{otherwise,} \end{cases} \tag{A.25}$$

and the reward function of the MDP is given by,

$$r_t(s,a) = \begin{cases} 1, & s \in \mathcal{S} \setminus \{b\},\ a = \pi_t^{\mathrm{exp}}(s) \\ 0, & \text{otherwise.} \end{cases} \tag{A.26}$$

We first state a simple consequence of the construction of the MDP instances and $\mathcal{P}$.

Lemma A.2.10. *Consider any policy $\pi^{\mathrm{exp}} \in \Pi_{\mathrm{det}}$. Then, the value of π^{exp} on the MDP $\mathcal{M}[\pi^{\mathrm{exp}}]$ is H.*

Proof. Playing the expert's action at any state in $\mathcal{S} \setminus \{b\}$ is the only way to accrue non-zero reward, and in fact accrues a reward of 1. In addition, note that the expert never visits the bad state b by virtue of the distribution ρ placing no mass on b. Therefore, the value of π^{exp} on the MDP $\mathcal{M}[\pi^{\mathrm{exp}}]$ is H. □

The intuition behind the lower bound construction is as follows. Although the learner can actively query the expert, at the states unvisited in the dataset D, the learner has no idea about the expert's policy or the transitions induced under different actions. Intuitively it is clear that the learner cannot guess the expert's action with probability $\geq 1/2$ at such states,

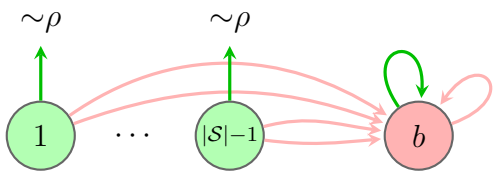


Figure A.1: MDP template when $N_{\text{sim}} = 0$: Upon playing the expert's (green) action at any state except b, learner is renewed in the initial distribution $\rho = \{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-2)\zeta, 0\}$ where $\zeta = \frac{1}{N+1}$. Any other choice of action (red) deterministically transitions the state to b.

a statement which we prove by leveraging the Bayesian construction. In turn, the learner is forced to visit the bad state b at the next point in the episode, and then on collects no reward.

Therefore, to bound the expected reward collected by a learner, it suffices to bound the probability that a learner visits a state unvisited in the demonstration dataset. The remainder of the proof is in showing that in this MDP construction, in expectation any learner visits such states with probability $\epsilon \gtrsim |\mathcal{S}|/N$ at each point in an episode. Moreover, conditioned on the dataset D, these events occur independently across time. Thus informally, the expected imitation gap of a learner is lower bounded by,

$$H\epsilon + (H-1)\epsilon(1-\epsilon) + \cdots + (1-\epsilon)^H \gtrsim \min\{H, H^2\epsilon\}. \tag{A.27}$$

where $\epsilon = |\mathcal{S}|/N$.

We return to a more formal exposition of the proof of the lower bound. Recall that our objective is to lower bound the Bayes expected imitation gap of $\widehat{\pi}$. Invoking Lemma A.2.10, the objective is to lower bound

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\Big[H - \mathbb{E}\Big[J_{\mathcal{M}}(\widehat{\pi}(D,A))\Big]\Big]. \tag{A.28}$$

To this end, we first try to understand the conditional distribution of the expert's policy given the dataset (D, A) collected by the learner. Recall that the dataset D contains trajectories generated by rolling out a sequence of policies $\pi_1, \cdots, \pi_n$, and A captures the expert's policy at states visited in D.

Lemma A.2.11. *Conditioned on the dataset (D, A) collected by the learner, the expert's deterministic policy π^{exp} is distributed $\sim \text{Unif}(\Pi^{BC}_{\text{det}}(D, A))$. In other words, at each state visited in the demonstration dataset, the expert's choice of action is fixed as the one returned when the expert was actively queried at this state. At the remaining states, the expert's choice of action is sampled uniformly from $\mathcal{A}$.*

Definition A.2.2. *Define $\mathcal{P}(D, A)$ as the joint distribution of $\mathcal{Z} = (\mathcal{M}, \pi^{\text{exp}})$ conditioned on the dataset (D, A) collected by the learner. In particular, $\pi^{\text{exp}} \sim \text{Unif}(\Pi^{BC}_{\text{det}}(D, A))$ and $\mathcal{M} = \mathcal{M}[\pi^{\text{exp}}]$.*

From Lemma A.2.11 and the definition of $\mathcal{P}(D, A)$ in Definition A.2.2, applying Fubini's theorem gives,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\Big[H - \mathbb{E}\left[J_{\mathcal{M}}(\widehat{\pi})\right]\Big] = \mathbb{E}\left[\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\left[H - J_{\mathcal{M}}(\widehat{\pi}(D,A))\right]\right]. \tag{A.29}$$

Next we relate this to the first time the learner visits a state unobserved in D.

Lemma A.2.12. *Define the stopping time τ as the first time t that the learner encounters a state $s_t \neq b$ that has not been visited in D at time t. That is,*

$$\tau = \begin{cases} \inf\{t : s_t \notin \mathcal{S}_t(D) \cup \{b\}\} & \exists t : s_t \notin \mathcal{S}_t(D) \cup \{b\} \\ H & \textit{otherwise.} \end{cases} \tag{A.30}$$

Then, conditioned on the dataset (D, A) collected by the learner,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[J(\pi^{\text{exp}}) - \mathbb{E}\left[J(\widehat{\pi})\right]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\left[\mathbb{E}_{\widehat{\pi}(D,A)}\left[H - \tau\right]\right] \tag{A.31}$$

Plugging the result of Lemma A.2.12 into eq. (A.29), we have that,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\Big[J(\pi^{\text{exp}}) - \mathbb{E}\left[J(\widehat{\pi})\right]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)\mathbb{E}\left[\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\left[\mathbb{E}_{\widehat{\pi}}\left[H - \tau\right]\right]\right], \tag{A.32}$$

$$\overset{(i)}{\geq} \left(1 - \frac{1}{|\mathcal{A}|}\right)\frac{H}{2}\mathbb{E}\left[\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right], \tag{A.33}$$

$$= \left(1 - \frac{1}{|\mathcal{A}|}\right)\frac{H}{2}\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right], \tag{A.34}$$

where (i) uses Markov's inequality and the last equation uses Fubini's theorem.
The last remaining element of he proof is to indeed bound the probability that the learner visits a state unobserved in the dataset before time $\lfloor H/2 \rfloor$. In Lemma A.2.13 we prove that for any learner $\widehat{\pi}$, $\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right]$ is lower bounded by $\gtrsim \min\{1, |\mathcal{S}|H/N\}$. Therefore,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\Big[J(\pi^{\text{exp}}) - \mathbb{E}\left[J(\widehat{\pi})\right]\Big] \gtrsim \left(1 - \frac{1}{|\mathcal{A}|}\right)\frac{H}{2}\min\left\{1, \frac{|\mathcal{S}|H}{N}\right\}. \tag{A.35}$$

Since $\left(1 - \frac{1}{|\mathcal{A}|}\right)$ is a constant for $|\mathcal{A}| \geq 2$ the statement of Theorem 2.2.1 follows.

Lemma A.2.13. *For any learner policy $\widehat{\pi}$,*

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right] \geq 1 - \left(1 - \frac{|\mathcal{S}| - 2}{e(N+1)}\right)^{\lfloor H/2 \rfloor} \gtrsim \min\left\{1, \frac{|\mathcal{S}|H}{N}\right\}. \tag{A.36}$$

Figure A.2: MDP template when $N_{\text{sim}} \to \infty$, Each state is absorbing, initial distribution is given by $\{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-1)\zeta\}$ where $\zeta = \frac{1}{N+1}$

Known-transition setting

As in the proof of Theorem 2.2.1, in order to prove the lower bound on the expected imitation gap of any learner $\widehat{\pi}(D, A)$, it suffices lower bound the Bayes expected imitation gap. Namely, it suffices to find a joint distribution $\mathcal{P}$ over MDPs and expert policies supported on $\mathbb{M}_{\mathcal{S},\mathcal{A},H} \times \Pi_{\text{det}-\text{exp}}$ such that,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}}\Big[J(\pi^{\text{exp}}) - \mathbb{E}\left[J(\widehat{\pi}(D,P))\right]\Big] \gtrsim \min\left\{H, \frac{|\mathcal{S}|H}{N}\right\}. \tag{A.37}$$

Construction of $\mathcal{P}$ As in the proof of Theorem 2.2.1, we first sample the expert's policy uniformly from Π_{det}. That is, for each $t \in [H]$ and $s \in \mathcal{S}$, the action $\pi_t^{\text{exp}}(s)$ is drawn uniformly from $\mathcal{A}$. Conditioned on π^{exp}, the distribution over MDPs induced by $\mathcal{P}$ is deterministic and given by the construction $\mathcal{M}[\pi^{\text{exp}}]$ in fig. A.1. $\mathcal{M}[\pi^{\text{exp}}]$ is defined with initial distribution over states $\rho = \{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-1)\zeta\}$ where $\zeta = \frac{1}{N+1}$. Each state $s \in \mathcal{S}$ is absorbing in $\mathcal{M}[\pi^{\text{exp}}]$. Formally, for each $s \in \mathcal{S}$ the transition function of $\mathcal{M}[\pi^{\text{exp}}]$ is,

$$P_t(\cdot|s,a) = \delta_s. \tag{A.38}$$

At any state s, choosing the expert's action $\pi_t^{\text{exp}}(s)$ returns a reward of 1, while any other choice of action offers 0 reward.

$$r_t(s,a) = \begin{cases} 1, & a = \pi_t^{\text{exp}}(s) \\ 0, & \text{otherwise.} \end{cases} \tag{A.39}$$

Note that all the MDPs $\mathcal{M}[\pi^{\text{exp}}]$ for $\pi^{\text{exp}} \in \Pi_{\text{det}}$ share a common set of transition functions and initial state distribution. Therefore, fixing P and ρ, we define $\mathcal{P}'$ to be the joint distribution over expert policies and reward functions induced by $\mathcal{P}$. Then the objective is to lower bound the Bayes expected imitation gap,

$$\mathbb{E}_{(\pi^{\text{exp}},r)\sim\mathcal{P}'}\Big[J_r(\pi^{\text{exp}}) - \mathbb{E}\left[J_r(\widehat{\pi}(D,P))\right]\Big]. \tag{A.40}$$

In this construction, it is yet again the case that the expert's policy π^{exp} collects maximum reward H on $\mathcal{M}[\pi^{\text{exp}}]$.

Lemma A.2.14. *Consider any policy $\pi^{\text{exp}} \in \Pi_{\text{det}}$. Then, the value of π^{exp} on the MDP $\mathcal{M}[\pi^{\text{exp}}]$ is H.*

Proof. At each state visited π^{exp} plays the only action which accrues a reward of 1. By accumulating a local reward of 1 at each step, π^{exp} has value equal to H on the MDP $\mathcal{M}[\pi^{\text{exp}}]$. □

With this explanation, invoking Lemma A.2.14 shows that out objective is to now lower bound,

$$\mathbb{E}_{(\pi^{\text{exp}},r)\sim\mathcal{P}'}\left[H - \mathbb{E}\left[J_r(\widehat{\pi}(D,P))\right]\right]. \tag{A.41}$$

Similar to Lemma A.2.11, we can compute the conditional distribution of the expert's policy (which marginally follows the uniform prior) given the demonstration dataset D.

Lemma A.2.15. *Conditioned on D, the distribution of the expert policy π^{exp} is uniform over the family of deterministic policies $\Pi^{BC}_{\text{det}}(D)$ (as defined in eq. (2.3)).*

For brevity of notation, we define this conditional distribution of the expert policy given the dataset D by $\mathcal{P}'(D)$.

Definition A.2.3. *Define $\mathcal{P}'(D)$ as the joint distribution of (π^{exp}, r) conditioned on the demonstration dataset D. In particular, $\pi^{\text{exp}} \sim \text{Unif}(\Pi^{BC}_{\text{det}}(D))$ and $r = r[\pi^{\text{exp}}]$.*

From Lemma A.2.15 and Definition A.2.3 and applying Fubini's theorem,

$$\mathbb{E}_{(\pi^{\text{exp}},r)\sim\mathcal{P}'}\left[\mathbb{E}\left[H - J_r(\widehat{\pi}(D,P))\right]\right] = \mathbb{E}\left[\mathbb{E}_{(\pi^{\text{exp}},r)\sim\mathcal{P}'(D)}\left[H - J_r(\widehat{\pi}(D,P))\right]\right]. \tag{A.42}$$

Fixing the demonstration dataset D, we subsequently show that the imitation gap of the learner is $\Omega(H)$ if initialized in a state unobserved in the demonstration dataset D. The key intuition is to identify that here the learner's knowledge of the transition function plays no role as each state in the MDP is absorbing. Therefore, once again at states unvisited in the demonstration dataset, the learner cannot guess the expert's action with high probability at states, leading to errors that grow linearly in H.

Lemma A.2.16. *For any learner's policy $\widehat{\pi}$ conditioned on the demonstration dataset D,*

$$\mathbb{E}_{(\pi^{\text{exp}},r)\sim\mathcal{P}'(D)}\left[H - J_r(\widehat{\pi}(D,P))\right] \geq H\left(1 - \frac{1}{|\mathcal{A}|}\right)\Big(1 - \rho(\mathcal{S}_1(D))\Big). \tag{A.43}$$

Therefore, from Lemma A.2.16 and eq. (A.42),

$$\mathbb{E}_{(\pi^{\text{exp}},r)\sim\mathcal{P}'}\left[\mathbb{E}\left[H - J_r(\widehat{\pi}(D,P))\right]\right] \geq H\left(1 - \frac{1}{|\mathcal{A}|}\right)\mathbb{E}\Big[1 - \rho(\mathcal{S}_1(D))\Big]. \tag{A.44}$$

The last ingredient left to show is that the probability mass on states unobserved in the demonstration dataset, $1-\rho(\mathcal{S}_1(D))$, is not too small in expectation. Here we realize that this boils down to calculating the expected missing mass of the distribution ρ given N samples drawn independently. By construction of ρ, we show that this is $\gtrsim |\mathcal{S}|/N$ in expectation.

Lemma A.2.17. $\mathbb{E}[1 - \rho(\mathcal{S}_1(D))] \geq \frac{|\mathcal{S}|-1}{e(N+1)}$.

Plugging Lemma A.2.17 back into eq. (A.44) certifies a lower bound on the Bayes expected imitation gap of any learner $\widehat{\pi}$. This implies the existence of an MDP on which the learner's expected imitation gap is $\gtrsim |\mathcal{S}|H/N$.

A.3 Missing proofs in the analysis of BC

Proof of Lemma A.1.1

Since the expert dataset D is composed of trajectories generated by i.i.d. rollouts of π^{exp}, we have that $\Pr[s \notin \mathcal{S}_\tau(D)] = (1 - \Pr_{\pi^{\text{exp}}}[s_\tau = s])^{|D|}$. Therefore,

$$\sum_{t=1}^{H} \sum_{s \in \mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_t = s] \ \Pr[s \notin \mathcal{S}_t(D)] \leq \sum_{\tau=1}^{H} \sum_{s \in \mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_\tau = s] \Big(1 - \Pr_{\pi^{\text{exp}}}[s_\tau = s]\Big)^{|D|}. \tag{A.45}$$

Noting that $\max_{x \in [0,1]} x(1-x)^N = \frac{1}{N+1}\left(1 - \frac{1}{N+1}\right)^N \leq \frac{4}{9N}$, from eq. (A.45),

$$\sum_{\tau=1}^{H} \sum_{s \in \mathcal{S}} \Pr_{\pi^{\text{exp}}}[s_\tau = s] \Big(1 - \Pr_{\pi^{\text{exp}}}[s_\tau = s]\Big)^{|D|} \leq \sum_{\tau=1}^{H} \sum_{s \in \mathcal{S}} \frac{4}{9|D|} \leq \frac{4}{9} \frac{|\mathcal{S}|H}{|D|}. \tag{A.46}$$

Proof of Theorem A.1.2

To prove this theorem, we invoke a result of [54] on the concentration of missing mass.

Theorem A.3.1 (Concentration of missing mass [54]). *Consider an arbitrary distribution ν on $\mathcal{X}$, and let $X^N \overset{i.i.d.}{\sim} \nu$ be a dataset of N samples drawn i.i.d. from ν. Let $\beta \geq 0$ and $\sigma \geq 0$ be constants such that $\sum_{x \in \mathcal{X}} (\nu(x))^2 e^{-(N-\beta)\nu(x)} \leq \sigma^2$. For any $0 \leq \varepsilon \leq \beta\sigma^2$, we have the following,*

$$\Pr\left(\mathfrak{m}_0(\nu, X^N) - \mathbb{E}[\mathfrak{m}_0(\nu, X^N)] \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \tag{A.47}$$

We prove Theorem A.1.2 by an appropriate choice of parameters β, σ^2 and ϵ (as functions of the confidence parameter δ). In particular, choose $\beta = N - \frac{N}{\sqrt{\log(1/\delta)}} \geq \frac{N}{3}$. For this choice of β,

$$\sum_{x \in \mathcal{X}} (\nu(x))^2 e^{-(N-\beta)\nu(x)} = \sum_{x \in \mathcal{X}} (\nu(x))^2 e^{-\frac{N}{\sqrt{\log(1/\delta)}}\nu(x)}, \tag{A.48}$$

$$\leq |\mathcal{X}| \sup_{\nu \in [0,1]} \nu^2 e^{-\frac{N}{\sqrt{\log(1/\delta)}}\nu}, \tag{A.49}$$

$$\overset{(i)}{=} |\mathcal{X}| \left(4e^{-2} \frac{\log(1/\delta)}{N^2}\right). \tag{A.50}$$

where (i) involves computing the supremum explicitly by differentiation. Therefore, for $\beta = N - \frac{N}{\sqrt{\log(H/\delta)}}$, a feasible choice of σ^2 in Theorem A.3.1 that upper bounds $\sum_{x \in \mathcal{X}} (\nu(x))^2 e^{-(N-\beta)\nu(x)}$ is $\frac{3|\mathcal{X}|\log(1/\delta)}{N^2}$. Choose $\varepsilon = \frac{3\sqrt{|\mathcal{X}|\log(1/\delta)}}{N}$ (note that this choice satisfies $\varepsilon \leq \beta\sigma^2$ since $\beta \geq N/3$

and $\sigma^2 = \frac{9|\mathcal{X}|\log(1/\delta)}{N^2}$). Invoking Theorem A.3.1 with this choice of β, σ^2 and ϵ,

$$\Pr\left(\mathfrak{m}_0(\nu, X^N) - \mathbb{E}[\mathfrak{m}_0(\nu, X^N)] \geq \frac{3\sqrt{|\mathcal{X}|}\log(1/\delta)}{N}\right) \leq \exp\left(-\frac{\left(3\sqrt{|\mathcal{X}|}N^{-1}\log(1/\delta)\right)^2}{9|\mathcal{X}|N^{-2}\log(1/\delta)}\right) = \delta. \tag{A.51}$$

This proves Theorem A.1.2.

Proof of Lemma A.1.3

We decompose $\sum_{\tau=1}^{H}\sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}[s_\tau = s]\mathbb{1}(s \notin \mathcal{S}_\tau(D))$ as $\sum_\tau Z_\tau$ where $Z_\tau = \sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}[s_\tau = s]\mathbb{1}(s \notin \mathcal{S}_\tau(D))$. Observe that for each fixed τ, Z_τ is in fact the missing mass of the distribution over states at time τ rolling out $\pi^{\exp}$, given N samples from the distribution. Applying the missing mass concentration inequality from Theorem A.1.2, with probability $\geq 1 - \delta/H$,

$$Z_\tau - \mathbb{E}[Z_\tau] \leq \frac{3\sqrt{|\mathcal{S}|}\log(H/\delta)}{N}. \tag{A.52}$$

Therefore, by union bounding, with probability $\geq 1 - \delta$,

$$\sum_{\tau=1}^{H} Z_\tau \leq \sum_{\tau=1}^{H}\mathbb{E}[Z_\tau] + H\cdot\frac{3\sqrt{|\mathcal{S}|}\log(H/\delta)}{N}. \tag{A.53}$$

Using $\sum_{\tau=1}^{H} Z_\tau = \sum_{\tau=1}^{H}\sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}[s_\tau = s]\mathbb{1}(s \notin \mathcal{S}_\tau(D))$ and applying Lemma A.1.1 to claim that $\sum_{\tau=1}^{H}\mathbb{E}[Z_\tau] \leq 4|\mathcal{S}|H/9N$ completes the proof.

A.4 Reduction of IL $\to$ TV matching

In this section we will prove Lemma 2.3.1. For each $\tau \in [H]$, define the policy $\widetilde{\pi}^\tau = \{\pi_1^{\exp}, \cdots, \pi_\tau^{\exp}, \widehat{\pi}_{\tau+1}, \cdots, \widehat{\pi}_H\}$ with $\widetilde{\pi}^0 = \widehat{\pi}$. The policy $\widetilde{\pi}^\tau$ plays the expert's policy till time τ and the learner's policy for the remainder of the episode. Then,

$$\mathsf{Gap}(\widehat{\pi}) = \sum_{\tau=1}^{H} J(\widetilde{\pi}^\tau) - J(\widetilde{\pi}^{\tau-1}). \tag{A.54}$$

For any fixed $\tau \in [H]$, observe that $\widetilde{\pi}^\tau$ and $\widetilde{\pi}^{\tau-1}$ roll out the same policy till time $\tau - 1$. Therefore the expected reward collected until time $\tau - 1$ for both policies is the same. By linearity of expectation,

$$J(\widetilde{\pi}^\tau) - J(\widetilde{\pi}^{\tau-1}) = \sum_{t=\tau}^{H}\mathbb{E}_{\widetilde{\pi}^\tau}\left[r_t(s_t, a_t)\right] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t, a_t)\right]. \tag{A.55}$$

Now fix some $t \geq \tau$ and consider $\mathbb{E}_{\widetilde{\pi}^\tau}\left[r_t(s_t,a_t)\right] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t,a_t)\right]$. First observe that,

$$\mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t,a_t)\right] = \mathbb{E}_{\substack{s_\tau \sim f^\tau_{\pi^{\exp}} \\ a_\tau \sim \widehat{\pi}_\tau(\cdot|s_\tau)}}\left[\ \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t,a_t)|s_\tau,a_\tau\right]\ \right], \tag{A.56}$$

$$= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^\tau_{\pi^{\exp}}(s)\ \widehat{\pi}_\tau(a|s)\ \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t,a_t)|s_\tau=s,a_\tau=a\right], \tag{A.57}$$

$$= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^\tau_{\pi^{\exp}}(s)\ \widehat{\pi}_\tau(a|s)\ \mathbb{E}_{\widehat{\pi}}\left[r_t(s_t,a_t)|s_\tau=s,a_\tau=a\right]. \tag{A.58}$$

where in the last equation we use the fact that $\widetilde{\pi}^{\tau-1}$ rolls out $\widehat{\pi}$ time τ onwards, and the fact that we condition on the state visited and action played at time τ. Moreover, we also use the fact that $r_t(s_t,a_t)$ only depends on (s_t,a_t) which appears at time $t \geq \tau$. Noting that $\widetilde{\pi}^\tau = (\pi_1^{\exp},\cdots,\pi_\tau^{\exp},\widehat{\pi}_{\tau+1},\cdots,\widehat{\pi}_H)$, a similar decomposition gives,

$$\mathbb{E}_{\widetilde{\pi}^{\tau}}\left[r_t(s_t,a_t)\right] = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^\tau_{\pi^{\exp}}(s)\ \pi_\tau^{\exp}(a|s)\ \mathbb{E}_{\widetilde{\pi}^{\tau}}\left[r_t(s_t,a_t)|s_\tau=s,a_\tau=a\right], \tag{A.59}$$

$$= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^\tau_{\pi^{\exp}}(s)\ \pi_\tau^{\exp}(a|s)\ \mathbb{E}_{\widehat{\pi}}\left[r_t(s_t,a_t)|s_\tau=s,a_\tau=a\right], \tag{A.60}$$

where in the last equation we similarly use the fact that $\widetilde{\pi}^\tau$ rolls out $\widehat{\pi}$ time $\tau+1$ onwards, and the fact that we condition on the action played at time τ. Subtracting eq. (A.58) from eq. (A.60),

$$\mathbb{E}_{\widetilde{\pi}^{\tau}}\left[r_t(s_t,a_t)\right] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t,a_t)\right]$$
$$\leq \sum_{s\in\mathcal{S}} f^\tau_{\pi^{\exp}}(s) \sum_{a\in\mathcal{A}} \mathbb{E}_{\widehat{\pi}}\left[r_t(s_t,a_t)|s_\tau=s,a_\tau=a\right]\left(\pi_\tau^{\exp}(a|s) - \widehat{\pi}_\tau(a|s)\right). \tag{A.61}$$

Observe that $\mathbb{E}_{\widehat{\pi}}\left[r_t(s_t,a_t)|s_\tau=s,a_\tau=a\right]$ is a function of (s,a) and is bounded in $[0,1]$ (since pointwise $0 \leq r_t \leq 1$). Therefore,

$$\mathbb{E}_{\widetilde{\pi}^{\tau}}\left[r_t(s_t,a_t)\right] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[r_t(s_t,a_t)\right] \leq \sum_{s\in\mathcal{S}} f^\tau_{\pi^{\exp}}(s) \sup_{g:\mathcal{A}\to[0,1]} \sum_{a\in\mathcal{A}} g(a)\left(\pi_\tau^{\exp}(a|s) - \widehat{\pi}_\tau(a|s)\right), \tag{A.62}$$

$$\overset{(i)}{=} \sum_{s\in\mathcal{S}} f^\tau_{\pi^{\exp}}(s) D_{\mathrm{TV}}\left(\pi_\tau^{\exp}(a|s), \widehat{\pi}_\tau(a|s)\right) \tag{A.63}$$

$$= \mathbb{E}_{s\sim f^\tau_{\pi^{\exp}}}\left[D_{\mathrm{TV}}\left(\pi_\tau^{\exp}(a|s), \widehat{\pi}_\tau(a|s)\right)\right]. \tag{A.64}$$

where (i) uses the dual representation of TV distance. Summing over $t \geq \tau$ and $\tau \in [H]$ and invoking eqs. (A.54) and (A.55) we get,

$$\mathsf{Gap}(\widehat{\pi}) \leq H \sum_{\tau=1}^{H} \mathbb{E}_{s\sim f^\tau_{\pi^{\exp}}}\left[D_{\mathrm{TV}}\left(\pi_\tau^{\exp}(a|s), \widehat{\pi}_\tau(a|s)\right)\right]. \tag{A.65}$$

Using the definition of $\mathcal{L}_{0\text{-}1}(\cdot)$ (eq. (2.7)) completes the proof.

A.5 Missing proofs in the analysis of log-loss BC

We will produce the proofs excluded in the analysis of BC in the stochastic expert setting, namely, Theorem 2.3.2.

Proof of Lemma A.2.1

Recall that we assume that the trajectories in the expert dataset are ordered arbitrarily as $\{\mathsf{tr}_1, \cdots, \mathsf{tr}_N\}$ where $\mathsf{tr}_n = \{(s_1^n, a_1^n), \cdots, (s_H^n, a_H^n)\}$. $N_{t,s} = \{n \in [N] : s_t^n = s\}$ as defined in eq. (A.8) is the set of indices of trajectories in D that visit the state s at time t. In order to prove this result, suppose the learner's policy $\widehat{\pi}$
With this, we define the randomized stochastic policy $X^{\text{unif}}(D)$ as,

$$X_t^{\text{unif}}(\cdot|s) = \begin{cases} \delta_{a_t^{n(t,s)}} & \text{if } |N_{t,s}| \geq 1, \\ \text{Unif}(\mathcal{A}) & \text{otherwise.} \end{cases} \tag{A.66}$$

where each $n(t,s)$ is a random variable independently sampled from $\text{Unif}(N_{t,s})$ whenever $N_{t,s} \neq \emptyset$. Note that fixing D and $n(t,s)$ for all t, s such that $N_{t,s} \neq \emptyset$, the random variable X^{unif} is a fixed stochastic policy.
The policy $X^{\text{unif}}(D)$ in a sense corresponds to just extracting the randomness in the actions chosen at visited states in the policy $\widehat{\pi}(D)$ returned by `Mimic-MD`.
In particular, it is a short proof to see that the random variables $J(X^{\text{unif}}(D))$ and $J(\widehat{\pi}(D))$ have the same expectation.

Lemma A.5.1. $\mathbb{E}[J(\widehat{\pi}(D))] = \mathbb{E}[J(X^{\text{unif}}(D))]$.

Proof. Consider some trajectory $\mathsf{tr} = \{(s_1, a_1), \cdots, (s_H, a_H)\}$. Fixing the expert dataset D,

$$\mathbb{E}\Big[\Pr_{X^{\text{unif}}(D)}[\mathsf{tr}]\Big|D\Big] = \mathbb{E}\left[\rho(s_1)\left(\prod_{t=1}^{H-1} X_t^{\text{unif}}(a_t|s_t)P_t(s_{t+1}|s_t,a_t)\right)X_t^{\text{unif}}(a_H|s_H)\right]. \tag{A.67}$$

From eq. (A.66) and Algorithm 1, observe that $X_t^{\text{unif}(\cdot|s)} = \widehat{\pi}(\cdot|s) = \text{Unif}(\mathcal{A})$ at states $s : N_{t,s} = \emptyset$ (i.e. which were not visited in the expert dataset). Moreover, on the remaining states $X_t^{\text{unif}}(a_t|s_t)$ is independently sampled from the empirical distribution over states at time t. In particular, this means that $\mathbb{E}[X_t^{\text{unif}}(a_t|s_t)] = \widehat{\pi}_t(a_t|s_t)$. Plugging this in gives,

$$\mathbb{E}[\Pr_{X^{\text{unif}}(D)}[\mathsf{tr}]] = \Pr_{\widehat{\pi}(D)}[\mathsf{tr}]. \tag{A.68}$$

Multiplying both sides by $\sum_{t=1}^{H} r_t(s_t, a_t)$, summing over all trajectories tr and taking expectation with respect to the expert dataset D completes the proof. □

First we provide an auxiliary result that is critical to showing that the policies $J(X^{\text{unif}}(D))$ and $\pi^{\text{first}}(D)$ have the same value in expectation.

To this end, first define $D_{\leq\tau,<\tau} = \{((s_1^n, a_1^n), \cdots, (s_{\tau-1}^n, a_{\tau-1}^n), s_\tau^n) : n \in [N]\}$ to be the truncation of the expert dataset D till time τ, excluding the actions played at this time. $D_{\leq\tau,\leq\tau}$ and other similar notations are defined analogously.

Lemma A.5.2. *Condition on $D_{\leq\tau,<\tau}$ which represents the truncation of trajectories in the expert dataset D till the state visited at time τ. At any state s that is visited at least once in D at time τ (namely with $|N_{\tau,s}| > 0$), the actions $\{a_\tau^n : n \in N_{\tau,s}\}$ played at trajectories that visit the state s at time τ are drawn independently and identically $\sim \pi_\tau^{\exp}(\cdot|s)$.*

Proof. Recall that we condition on $D_{\leq\tau,<\tau}$ which captures trajectories in the expert dataset truncated till the state visited at time τ. Since each trajectory $\mathsf{tr}_n \in [N]$ is rolled out independently, the action a_τ^n in each trajectory tr_n is drawn independently from $\pi_\tau^{\exp}(\cdot|s_\tau^n)$. More importantly, conditioned on $D_{\leq\tau,<\tau}$ the states s_τ^n visited in different trajectories is determined. This implies that $N_{\tau,s}$ for $s \in \mathcal{S}$ is a measurable function of $D_{\leq\tau,<\tau}$.
These two statements together imply that states $s \in \mathcal{S}$ having $N_{\tau,s} > 0$ (which is a measurable function of $D_{\leq\tau,<\tau}$) are such that all the actions $\{a_\tau^n : n \in N_{\tau,s}\}$ are independent. □

Proof of Lemma A.2.1. In order to prove this result, we use an inductive argument. The induction hypothesis is that the expected value of $X^{\text{unif}}(D)$ and $\pi^{\text{first}}(D)$ are the same, conditioned on the expert dataset till time t and the actions from the empirical distribution sampled by $X^{\text{unif}}(D)$ at different states till time t. We formalize this hypothesis in equations after first proving the base case. To recognize the fact that we prove the statement starting from $t = H$, we define $\mathcal{H}_H$ as the base case, and inductively prove $\mathcal{H}_{t-1}$ assuming the hypothesis $\mathcal{H}_t$.
First observe that,

$$\mathbb{E}\Big[J(X^{\text{unif}}(D))\Big|D_{\leq H,<H}, \Big\{n(t,s) \;\Big|\; t \leq H,\ s : N_{t,s} > 0\Big\}\Big] = \mathbb{E}\Big[J(\pi^{\text{first}}(D))\Big|D_{\leq H,<H}\Big]. \quad \text{(A.69)}$$

This is because conditioned on $D_{\leq H,<H}$, the only randomness is in the actions that are played in the different trajectories at time H. By Lemma A.5.2 these are distributed i.i.d. $\sim \pi_t^{\exp}(\cdot|s)$. Taking expectation with respect to $\{n_{H,s}|s : N_{t,s} > 0\}$, results in proof of the base case for $t = H$,

$$\mathcal{H}_H : \mathbb{E}\Big[J(X^{\text{unif}}(D))\Big|D_{\leq H,<H}, \Big\{n(t,s) \;\Big|\; t < H,\ s : N_{t,s} > 0\Big\}\Big] = \mathbb{E}\Big[J(\pi^{\text{first}}(D))\Big|D_{\leq H,<H}\Big].$$

In general consider the hypothesis $\mathcal{H}_\tau$,

$$\mathcal{H}_\tau : \mathbb{E}\Big[J(X^{\text{unif}}(D))\Big|D_{\leq\tau,<\tau}, \Big\{n(t,s) \;\Big|\; t < \tau,\ s : N_{t,s} > 0\Big\}\Big] = \mathbb{E}\Big[J(\pi^{\text{first}}(D))\Big|D_{\leq\tau,<\tau}\Big].$$

Taking expectation with respect to $\{s_\tau^n : n \in [N]\}$, where conditionally $s_\tau^n \sim P_\tau(\cdot|s_{\tau-1}^n, a_{\tau-1}^n)$,

$$\mathbb{E}\Big[J(X^{\text{unif}}(D))\Big|D_{<\tau,<\tau}, \Big\{n(t,s) \;\Big|\; t < \tau,\ s : N_{t,s} > 0\Big\}\Big] = \mathbb{E}\Big[J(\pi^{\text{first}}(D))\Big|D_{<\tau,<\tau}\Big]. \quad \text{(A.70)}$$

Next we take expectation with respect to the actions $\{a_\tau^n : n \in [N]\}$ where each a_τ^n is drawn independently from $\pi_t^{\exp}(\cdot|s_\tau^n)$. This results in,

$$\mathbb{E}\Big[J(X^{\text{unif}}(D))\Big|D_{<\tau,<\tau-1}, \Big\{n(t,s) \;\Big|\; t<\tau,\ s: N_{t,s}>0\Big\}\Big] = \mathbb{E}\Big[J(\pi^{\text{first}}(D))\Big|D_{<\tau,<\tau-1}\Big]. \tag{A.71}$$

Note that on both sides we condition on $D_{<\tau,<\tau-1}$ which is the set of partial trajectories in the expert dataset till time $\tau-1$ (excluding the action at this time). In particular, this conditioning determines the set of states visited at time $\tau-1$ in the expert dataset. Consider any state $s \in \mathcal{S}$:

(*i*) If s was not observed in the dataset D at time $\tau-1$, then with probability 1 over the randomness of X^{unif}, both the policies X^{unif} and π^{first} play the policy $\text{Unif}(\mathcal{A})$;

(*ii*) On the other hand, if s was observed in the dataset D in some trajectory at time $\tau-1$, then X^{unif} samples from an empirical distribution over actions played at the state s in the dataset at time $\tau-1$, which by Lemma A.5.2 are drawn independently from $\pi_{\tau-1}^{\exp}(\cdot|s)$. On the other hand, the action played by π^{first} is also drawn independently from $\pi_{\tau-1}^{\exp}(\cdot|s)$. This shows that the expectation on the LHS does not depend on the choice of $n(\tau-1,s)$ for any state $s \in \mathcal{S}$.

Thus in both cases, the expectation of the random variable on the RHS does not depend on $\{n(\tau-1,s)|s\in\mathcal{S}\}$. Therefore, we can drop the conditioning on this random variable to give,

$$\mathbb{E}\Big[J(X^{\text{unif}}(D))\Big|D_{<\tau,<\tau-1}, \Big\{n(t,s) \;\Big|\; t<\tau-1,\ s: N_{t,s}>0\Big\}\Big] = \mathbb{E}\Big[J(\pi^{\text{first}}(D))\Big|D_{<\tau,<\tau-1}\Big] \tag{A.72}$$

This proves the induction hypothesis $\mathcal{H}_{\tau-1}$ and consequently the hypothesis $\mathcal{H}_1$. Taking expectation on both sides of $\mathcal{H}_1$ with respect to $s_1^n \overset{\text{i.i.d.}}{\sim} \rho$ proves the claim. □

Proof of Lemma A.2.3

Fixing the table $\mathbf{T}^\star$, the probability of observing the trajectory $\mathsf{tr} = \{(s_1,a_1),\cdots,(s_H,a_H)\}$ under the deterministic policy $\pi^{\text{orc}-\text{first}}$ is,

$$\Pr_{\pi^{\text{orc}-\text{first}}}(\mathsf{tr}) = \rho(s_1)\left(\prod_{t=1}^{H-1} \mathbb{1}\left(a_t = \mathbf{T}^\star_{t,s_t}(1)\right) P_t(s_{t+1}|s_t,a_t)\right) \mathbb{1}\left(a_H = \mathbf{T}^\star_{H,s_H}(1)\right). \tag{A.73}$$

Since the actions $\mathbf{T}^\star_{t,s_t}(1)$ are independently drawn from $\pi_t^{\exp}(\cdot|s_t)$, taking expectation, we see that

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc}-\text{first}}}(\mathsf{tr})\right] = \rho(s_1)\left(\prod_{t=1}^{H-1} \pi_t^{\exp}(a_t|s_t)P_t(s_{t+1}|s_t,a_t)\right)\pi_H^{\exp}(a_H|s_H) = \Pr_{\pi^{\exp}}(\mathsf{tr}). \tag{A.74}$$

Multiplying both sides by $\sum_{t=1}^H r_t(s_t,a_t)$ and summing over all trajectories completes the proof.

Proof of Lemma A.2.5

Recall that the "failure" $\mathcal{E}$ is defined as the event that at some time $t \in [H]$, a state s_t is visited such that $|N_{t,s_t}| = 0$, i.e. that was not visited in the expert dataset. By union bounding,

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[\mathcal{E}]\right] \leq \sum_{t=1}^{H}\sum_{s\in\mathcal{S}} \mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[\mathcal{E}_{s,t}]\right], \tag{A.75}$$

where $\mathcal{E}_{s,t}$ is the event that a failure occurs at the state s at time t, i.e. the state s is visited at time t and $|N_{t,s}| = 0$. $\mathcal{E}_{s,t}$ is the intersection of two events. Therefore we have the upper bound,

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[\mathcal{E}_{s,t}]\right] \leq \min\left\{\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[s_t = s]\right], \mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[|N_{s,t}| = 0]\right]\right\}. \tag{A.76}$$

Observe that these two terms in the minimum are easy to compute. Firstly, using eq. (A.74), we have that,

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[s_t = s]\right] = \Pr_{\pi^{\text{exp}}}[s_t = s]. \tag{A.77}$$

On the other hand,

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[|N_{s,t}| = 0]\right] = \mathbb{E}[\mathbb{1}(|N_{s,t}| = 0)] = (1 - \Pr_{\pi^{\text{exp}}}[s_t = s])^N \tag{A.78}$$

where the last equation uses Lemma A.2.2. Putting together eqs. (A.77) and (A.78) with eq. (A.76),

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[\mathcal{E}_{s,t}]\right] \leq \min\left\{\Pr_{\pi^{\text{exp}}}[s_t = s],\ \left(1 - \Pr_{\pi^{\text{exp}}}[s_t = s]\right)^N\right\}. \tag{A.79}$$

In Lemma A.5.3 we show that the RHS is upper bounded by $\log(N)/N$. Therefore,

$$\mathbb{E}\left[\Pr_{\pi^{\text{orc-first}}}[\mathcal{E}_{s,t}]\right] \leq \frac{\log N}{N}. \tag{A.80}$$

Plugging back into eq. (A.75) completes the proof.

Lemma A.5.3. *For any $x \in [0,1]$ and $N > 1$, $\min\{x, (1-x)^N\} \leq \frac{\log N}{N}$.*

Proof. x is an increasing function, while $(1-x)^N$ is decreasing. For $x = \frac{\log N}{N}$,

$$(1-x)^N = \left(1 - \frac{\log N}{N}\right)^N \leq e^{-\log N} \leq N^{-1} \tag{A.81}$$

Therefore for $x \geq \frac{\log(N)}{N}$, $\min\{x, (1-x)^N\} \leq \frac{1}{N}$. Therefore $\min\{x, (1-x)^N\} \leq \frac{\log N}{N}$. □

A.6 Missing proofs in the analysis of Mimic-MD

In this section we will provide proofs of the results excluded in the proof of Theorem 2.4.1.

Proof of Lemma A.2.6

Observe that the complement $\left(\mathcal{E}_{D_1}^{\leq t}\right)^c$ is the event that the policy under consideration until (and including) time $t-1$, only visits states that were visited in at least one trajectory in the expert dataset.
First observe that, fixing the expert dataset D,

$$\mathsf{Gap}(\widehat{\pi}) \tag{A.82}$$

$$= \mathbb{E}_{\pi^{\exp}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\right] - \mathbb{E}_{\widehat{\pi}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\right] \tag{A.83}$$

$$= \sum_{t=1}^{H} \mathbb{E}_{\pi^{\exp}}\left[\left(\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) + \mathbb{1}\left(\mathcal{E}_{D_1}^{\leq t}\right)\right) r_t(s_t,a_t)\right] - \mathbb{E}_{\widehat{\pi}}\left[\left(\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) + \mathbb{1}\left(\mathcal{E}_{D_1}^{\leq t}\right)\right) r_t(s_t,a_t)\right]. \tag{A.84}$$

Indeed, to prove the statement it suffices to prove that,

$$\sum_{t=1}^{H} \mathbb{E}_{\pi^{\exp}}\left[\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) r_t(s_t,a_t)\right] = \sum_{t=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) r_t(s_t,a_t)\right]. \tag{A.85}$$

Recall that the learner $\widehat{\pi}$ mimics the expert at all the states observed in the dataset D_1, i.e. having $|N_{t,s}| > 0$. Observe that when the event $\left(\mathcal{E}_{D_1}^{\leq t}\right)^c$ occurs, all the states visited in a trajectory have $|N_{t,s}| > 0$. Thus, both expectations are carried out with respect to the same policy and are hence equal. More precisely, for any $t \in [H]$,

$$\mathbb{E}_{\widehat{\pi}}\left[\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) r_t(s_t,a_t)\right] = \mathbb{E}_{\substack{s_1\sim\rho,\ \tau\leq t,\\ a_\tau\sim\widehat{\pi}_\tau(\cdot|s_\tau)\\ s_{\tau+1}\sim P(\cdot|s_\tau,a_\tau)}}\left[\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) r_t(s_t,a_t)\right] \tag{A.86}$$

$$\overset{(i)}{=} \mathbb{E}_{\substack{s_1\sim\rho,\ \tau\leq t,\\ a_\tau\sim\pi^\star_\tau(\cdot|s_\tau)\\ s_{\tau+1}\sim P(\cdot|s_\tau,a_\tau)}}\left[\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) r_t(s_t,a_t)\right] \tag{A.87}$$

$$= \mathbb{E}_{\pi^{\exp}}\left[\mathbb{1}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) r_t(s_t,a_t)\right] \tag{A.88}$$

where (i) uses the fact that when $s_\tau \in \mathcal{S}_t(D_1)$ (as implied by $\left(\mathcal{E}_{D_1}^{\leq t}\right)^c$ for each $\tau \leq t$), then, $\pi^{\exp}_\tau(\cdot|s_\tau) = \widehat{\pi}_\tau(\cdot|s_\tau)$. Moreover.

Proof of Lemma A.2.7

First observe that we can write the reward $r_t(s_t, a_t)$ accrued in some trajectory at time t equals $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r_t(s,a)\mathbb{1}((s_t,a_t)=(s,a))$. Therefore, from Lemma A.2.6,

$$\begin{aligned}
\mathsf{Gap}(\widehat{\pi}^{\varepsilon}) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H} r_t(s,a)\left(\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t=s, a_t=a\right] - \Pr_{\widehat{\pi}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t=s, a_t=a\right]\right) \\
&\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left|\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t=s, a_t=a\right] - \Pr_{\widehat{\pi}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t=s, a_t=a\right]\right| \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \Pr_{\widehat{\pi}}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right| \qquad \text{(A.89)}
\end{aligned}$$

where the inequality follows from the assumption that $0 \leq r_t(s,a) \leq 1$ and the last equation follows from the definition $\mathcal{T}_t^{D_1}(s,a) = \{\{(s_\tau,a_\tau)\}_{\tau=1}^{H} | s_t=s, a_t=a,\ \exists\tau\in[H] : s_\tau \notin \mathcal{S}_\tau(D_1)\}$ is the set of trajectories that visit (s,a) at time t and at some point t' in the episode visit a state not visited in any trajectory at time t' in D_1. Using the definition of the learner's policy $\widehat{\pi}$ in the optimization problem (OPT-MD) and applying the triangle inequality,

$$\begin{aligned}
\mathsf{Gap}(\widehat{\pi}^{\varepsilon}) \leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H} &\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right| \\
&+ \left|\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|} - \Pr_{\widehat{\pi}}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right|. \qquad \text{(A.90)}
\end{aligned}$$

Observe that the expert's policy $\pi^{\exp}$ is a feasible policy to the optimization problem (OPT-MD). Since $\widehat{\pi}$ solves (OPT-MD) up to an additive error of ε, we have the upper bound,

$$\mathsf{Gap}(\widehat{\pi}^{\varepsilon}) \leq 2\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right| + \varepsilon. \qquad \text{(A.91)}$$

Proof of Lemma A.2.8

Recall that we carry out sample splitting in Algorithm 2 to give datasets D_1 and D_2. We first fix the trajectories in D_1 and compute the expectation with respect to the dataset D_2. Sample splitting implies that, conditioned on D_1, the trajectories in D_2 are still generated by independently rolling out $\pi^{\exp}$. By Jensen's inequality, we can upper bound by the quadratic

deviation,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\mathbb{E}\left[\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]-\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right|\right]$$
$$\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left(\mathbb{E}\left[\left(\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]-\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right)^2\right]\right)^{1/2} \tag{A.92}$$

Observe that each trajectory $\mathsf{tr}\in D_2$ is generated by independently rolling out $\pi^{\exp}$. Therefore, $\frac{1}{|D_2|}\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))$ is an unbiased estimate of $\Pr_{\pi^{\exp}}[\mathcal{T}_t^{D_1}(s,a)]$. Therefore the expectation term in eq. (A.92) is nothing but the variance: letting tr_1 be an arbitrary trajectory in D_2,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\mathbb{E}\left[\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]-\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right|\right]$$
$$\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left(\frac{1}{|D_2|}\mathrm{Var}\left[\mathbb{1}(\mathsf{tr}_1\in\mathcal{T}_t^{D_1}(s,a)\right]\right)^{1/2} \tag{A.93}$$
$$\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\left(\frac{1}{|D_2|}\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right)^{1/2} \tag{A.94}$$

where the last inequality uses the fact that the variance of an indicator function is at most its mean, and that each $\mathsf{tr}\in D_2$ is independently drawn by rolling out $\pi^{\exp}$. Now, taking expectation with respect to the dataset D_1, and by another application of Jensen's inequality,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\mathbb{E}\left[\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]-\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right|\right]$$
$$\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\frac{1}{|D_2|^{1/2}}\left(\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right]\right)^{1/2} \tag{A.95}$$
$$=\sum_{s\in\mathcal{S}}\sum_{t=1}^{H}\frac{1}{|D_2|^{1/2}}\left(\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t},s_t=s,a_t=\pi_t^{\exp}(s_t)\right]\right]\right)^{1/2}, \tag{A.96}$$

where in the last equation, we use the definition of $\mathcal{T}_t^{D_1}(\cdot,\cdot)$. By an application of the Cauchy Schwarz inequality,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=1}^{H}\mathbb{E}\left[\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,a)\right]-\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,a))}{|D_2|}\right|\right]$$

$$\leq\sum_{t=1}^{H}\frac{|\mathcal{S}|^{1/2}}{|D_2|^{1/2}}\left(\sum_{s\in\mathcal{S}}\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t},s_t=s,a_t=\pi_t^{\exp}(s)\right]\right]\right)^{1/2} \tag{A.97}$$

$$\leq\sum_{t=1}^{H}\frac{|\mathcal{S}|^{1/2}}{|D_2|^{1/2}}\left(\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t}\right]\right]\right)^{1/2}. \tag{A.98}$$

Therefore, to prove the result it suffices to bound $\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t}\right]\right]$, which we carry out in Lemma A.6.1. Here we show that it is upper bounded by $\lesssim 1\wedge|\mathcal{S}|H/|D_1|$. Subsequently using $|D_1|=|D_2|=N/2$ completes the proof.

Lemma A.6.1. *For any $t\in[H]$, the probability of failure under the expert's policy is upper bounded by,*

$$\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t}\right]\right]\leq\frac{4}{9}\frac{|\mathcal{S}|H}{|D_1|} \tag{A.99}$$

Proof. Conditioned on D_1, we decompose based on the first failure time (i.e. the first time the event $\mathcal{E}_{D_1}^{\leq t}$ is satisfied),

$$\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}^{\leq t}\middle|D_1\right]=\Pr_{\pi^{\exp}}\left[\exists\tau\leq t:s_\tau\notin\mathcal{S}_\tau(D_1)\middle|D_1\right], \tag{A.100}$$

$$=\sum_{\tau=1}^{t}\Pr_{\pi^{\exp}}\left[\forall\tau'<\tau,\ s_{\tau'}\in\mathcal{S}_{\tau'}(D_1),s_\tau\notin\mathcal{S}_\tau(D_1)\middle|D_1\right] \tag{A.101}$$

$$\leq\sum_{\tau=1}^{t}\Pr_{\pi^{\exp}}\left[s_\tau\notin\mathcal{S}_\tau(D_1)\middle|D_1\right] \tag{A.102}$$

$$=\sum_{\tau=1}^{t}\sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}[s_\tau=s]\mathbb{1}(s\notin\mathcal{S}_\tau(D_1)) \tag{A.103}$$

$$\leq\sum_{\tau=1}^{H}\sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}[s_\tau=s]\mathbb{1}(s\notin\mathcal{S}_\tau(D_1)) \tag{A.104}$$

Taking expectation with respect to the expert dataset,

$$\mathbb{E}\left[\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}\middle|D_1\right]\right]\leq\sum_{\tau=1}^{H}\sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}[s_\tau=s]\Pr[s\notin\mathcal{S}_\tau(D_1)] \tag{A.105}$$

The proof of the claim immediately follows by invoking Lemma A.1.1. □

Proof of Lemma A.2.9

Starting from the bound in Lemma A.2.7 and using the fact that at each state s the expert plays a fixed action $\pi_t^{\exp}(s)$ at time t,

$$\mathsf{Gap}(\widehat{\pi}) \le 2\sum_{s\in\mathcal{S}}\sum_{t=1}^{H}\left|\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s))\right] - \frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s)))}{|D_2|}\right| \tag{A.106}$$

Observe that $\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s)))$ is a sub-Gaussian random variable with variance bounded by its expectation. Therefore, by sub-Gaussian concentration [13], for each $s\in\mathcal{S}$ and $t\in[H]$, conditioned on D_1, with probability $\ge 1-\frac{\delta}{2|\mathcal{S}|H}$,

$$\begin{aligned}&\left|\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s)))}{|D_2|} - \Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s))\right]\right| \\ &\le \left(\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s))\right]\right)^{1/2}\sqrt{\frac{2\log(2|\mathcal{S}|H/\delta)}{|D_2|}}\end{aligned} \tag{A.107}$$

By union bounding over $s\in\mathcal{S}$ and $t\in[H]$, conditioned on D_1 with probability $\ge 1-\frac{\delta}{2}$,

$$\begin{aligned}&\sum_{t=1}^{H}\sum_{s\in\mathcal{S}}\left|\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s)))}{|D_2|} - \Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s))\right]\right| \\ &\le \sum_{t=1}^{H}\left(\sum_{s\in\mathcal{S}}\Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s))\right]\right)^{1/2}\sqrt{\frac{2\log(2|\mathcal{S}|H/\delta)}{|D_2|}} &\text{(A.108)}\\ &\le H|\mathcal{S}|^{1/2}\left(\Pr_{\pi^{\exp}}\left[\mathcal{E}_{D_1}\right]\right)^{1/2}\sqrt{\frac{2\log(2|\mathcal{S}|H/\delta)}{|D_2|}} &\text{(A.109)}\end{aligned}$$

Applying Lemma A.1.3, with probability $\ge 1-\delta/2$,

$$\Pr_{\pi^{\exp}}[\mathcal{E}_{D_1}] \le \frac{4|\mathcal{S}|H}{9|D_1|} + \frac{3H\sqrt{|\mathcal{S}|}\log(2H/\delta)}{|D_1|}. \tag{A.110}$$

Therefore union bounding the events of eqs. (A.109) and (A.110), with probability $\ge 1-\delta$,

$$\begin{aligned}&\sum_{t=1}^{H}\sum_{s\in\mathcal{S}}\left|\frac{\sum_{\mathsf{tr}\in D_2}\mathbb{1}(\mathsf{tr}\in\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s)))}{|D_2|} - \Pr_{\pi^{\exp}}\left[\mathcal{T}_t^{D_1}(s,\pi_t^{\exp}(s))\right]\right| \\ &\le H|\mathcal{S}|^{1/2}\left(\frac{4|\mathcal{S}|H}{9|D_1|} + \frac{3H\sqrt{|\mathcal{S}|}\log(2H/\delta)}{|D_1|}\right)^{1/2}\sqrt{\frac{2\log(2|\mathcal{S}|H/\delta)}{|D_2|}} &\text{(A.111)}\\ &\lesssim \frac{|\mathcal{S}|H^{3/2}}{N}\left(1+\frac{3\log(2|\mathcal{S}|H/\delta)}{\sqrt{|\mathcal{S}|}}\right)^{1/2}\sqrt{\log(2|\mathcal{S}|H/\delta)}. &\text{(A.112)}\end{aligned}$$

A.7 Lower bound in the active-interaction setting

In this section we will prove auxiliary results used in the proof of Theorem 2.2.1.

Proof of Lemma A.2.11

Fix some policy $\pi \in \Pi_{\texttt{det}}$. Consider any time $t \in [H]$ and state $s \in \mathcal{S}_t(D)$ which is visited in some trajectory in the dataset at time t. If $\pi_t(s)$ does not match the action $\pi_t^A(s)$ revealed by actively querying the expert in a trajectory in D that visits s at time t, the likelihood of π given D is exactly 0 (since the expert is deterministic). On the other hand, the conditional probability of observing (D, A) does not depend on the expert's action on the states that were not observed in D, since no trajectory visits these states. Since on these states the expert's action marginally follows the uniform distribution over $\mathcal{A}$, the result immediately follows.

Proof of Lemma A.2.12

In order to prove this result, define the auxiliary random time τ_b to be the first time the learner first encounters the state b while rolling out a trajectory. If no such state is encountered, τ is defined as $H + 1$. Formally,

$$\tau_b = \begin{cases} \inf\{t : s_t = b\} & \exists t : s_t = b \\ H + 1 & \text{otherwise.} \end{cases}$$

Conditioning on the learner's dataset (D, A), first observe that

$$H - \mathbb{E}_{(\mathcal{M}, \pi^{\text{exp}}) \sim \mathcal{P}(D,A)} \left[J(\widehat{\pi}) \right] = H - \mathbb{E}_{(\mathcal{M}, \pi^{\text{exp}}) \sim \mathcal{P}(D,A)} \left[\mathbb{E}_{\widehat{\pi}} \left[\sum_{t=1}^{H} r_t(s_t, a_t) \right] \right] \tag{A.113}$$

$$\geq \mathbb{E}_{(\mathcal{M}, \pi^{\text{exp}}) \sim \mathcal{P}(D,A)} \left[\mathbb{E}_{\widehat{\pi}} \left[H - \tau_b + 1 \right] \right] \tag{A.114}$$

where the last inequality follows from the fact that r is bounded in $[0, 1]$, and the state b is absorbing and offers 0 reward irrespective of the choice of action. Fixing the dataset (D, A) and the expert's policy π^{exp} (which determines the MDP $\mathcal{M}[\pi^{\text{exp}}]$), we study $\mathbb{E}_{\widehat{\pi}(D,A)} \left[H - \tau_b + 1 \right]$ and try to relate it to $\mathbb{E}_{\widehat{\pi}(D,A)} \left[H - \tau \right]$.

To this end, first observe that for any $t \leq H - 1$ and state $s \in \mathcal{S}$,

$$\Pr_{\widehat{\pi}} \left[\tau_b = t + 1, \tau = t, s_t = s \right] = \Pr_{\widehat{\pi}} \left[\tau_b = t + 1 | \tau = t, s_t = s \right] \Pr_{\widehat{\pi}} \left[\tau = t, s_t = s \right] \tag{A.115}$$

$$= \left(1 - \widehat{\pi}_t(\pi_t^{\text{exp}}(s) | s) \right) \Pr_{\widehat{\pi}} \left[\tau = t, s_t = s \right]. \tag{A.116}$$

where in the last equation, we use the fact that the learner must play an action other than $\pi_t^{\text{exp}}(s_t)$ to visit b at time $t + 1$. Next we take expectation with respect to the randomness of π^{exp} which conditioned on (D, A) is drawn from $\text{Unif}(\Pi_{\text{det}}^{\texttt{BC}}(D, A))$ which also specifies the

underlying MDP $\mathcal{M}[\pi^{\text{exp}}]$. Observe that the dependence of the second term $\Pr_{\widehat{\pi}}[\tau = t, s_t = s]$ on π^{exp} comes from the probability computed with the underlying MDP chosen as $\mathcal{M}[\pi^{\text{exp}}]$. However observe that it only depends on the characteristics of $\mathcal{M}[\pi^{\text{exp}}]$ till time $t-1$ which are determined by $\pi_1^{\text{exp}}, \cdots, \pi_{t-1}^{\text{exp}}$. On the other hand, the first term $(1 - \widehat{\pi}_t(\pi_t^{\text{exp}}(s)|s))$ depends only on π_t^{exp}. As a consequence the two terms depend on a disjoint set of random variables, which are independent (since conditionally $\pi^{\text{exp}} \sim \Pi_{\text{det}}^{\texttt{BC}}(D, A)$ defined in eq. (A.23))

Therefore taking expectation with respect to the randomness of $\pi^{\text{exp}} \sim \text{Unif}(\Pi_{\text{det}}^{\texttt{BC}}(D, A))$ and $\mathcal{M} = \mathcal{M}[\pi^{\text{exp}}]$ (which defines the joint distribution $\mathcal{P}(D, A)$ in eq. (A.23)),

$$\begin{aligned}
&\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\Pr_{\widehat{\pi}(D,A)}[\tau_b = t+1, \tau = t, s_t = s]\Big] \\
&= \mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[1 - \widehat{\pi}_t(\pi_t^{\text{exp}}(s_t)|s_t)\Big]\ \mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\Pr_{\widehat{\pi}}[\tau = t, s_t = s]\Big] \qquad \text{(A.117)} \\
&\overset{(a)}{=} \left(1 - \frac{1}{|\mathcal{A}|}\right) \mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\Pr_{\widehat{\pi}}[\tau = t, s_t = s]\Big] \qquad \text{(A.118)}
\end{aligned}$$

where in (a), conditioned on (D, A) we use the fact that either (i) $s = b$, in which case $\tau \neq t$ and both sides are 0, or (ii) if $s \neq b$, then $\tau = t$ implies that the state s visited at time t must not be observed in D, so $\pi_t^{\text{exp}}(s) \sim \text{Unif}(\mathcal{A})$. Using the fact that $\Pr_{\widehat{\pi}}[\tau_b = t+1, \tau = t, s_t = s] \leq \Pr_{\widehat{\pi}}[\tau_b = t+1, s_t = s]$ and summing over $s \in \mathcal{S}$ results in the inequality,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\Pr_{\widehat{\pi}}[\tau_b = t+1]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right) \mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\Pr_{\widehat{\pi}}[\tau = t]\Big] \qquad \text{(A.119)}$$

Multiplying both sides by $H - t$ and summing over $t = 1, \cdots, H$,

$$\mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\mathbb{E}_{\widehat{\pi}}[H - \tau_b + 1]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right) \mathbb{E}_{(\mathcal{M},\pi^{\text{exp}})\sim\mathcal{P}(D,A)}\Big[\mathbb{E}_{\widehat{\pi}}[H - \tau]\Big] \qquad \text{(A.120)}$$

here we use the fact that the initial distribution ρ places no mass on the bad state b. Therefore, $\Pr_{\widehat{\pi}(D)}[\tau_b = 1] = \rho(b) = 0$. This equation in conjunction with eq. (A.114) completes the proof.

Proof of Lemma A.2.13

Firstly, in Lemma A.7.2 we show that $\mathbb{E}\left[\Pr_{\widehat{\pi}(D,A)}[\tau \leq \lfloor H/2 \rfloor]\right] \geq 1 - (1-\gamma)^{\lfloor H/2 \rfloor}$ where γ is defined as $\sum_{s\in\mathcal{S}} \rho(s)(1-\rho(s))^N$. Subsequently, in Lemma A.7.3 we show that $\gamma \gtrsim |\mathcal{S}|/N$. Putting these two results together proves the statement of Lemma A.2.13.
Along the way to proving Lemma A.7.2, we introduce an auxiliary result.

Lemma A.7.1. *Fix the dataset (D, A) collected by the learner, and any policy $\pi^{\text{exp}} \in \Pi_{\text{det}}^{\texttt{BC}}(D, A)$ (as defined in eq. (A.23)). Recall that τ as defined in Lemma A.2.12 is the first time t that the learner encounters a state $s_t \neq b$ that has not been visited in D at time t.*
For some $t \in [H]$, consider $\Pr_{\widehat{\pi}(D)}[\tau = t]$ computed with the underlying MDP as $\mathcal{M}[\pi^{\text{exp}}]$. Then,

$$\Pr_{\widehat{\pi}(D,A)}[\tau = t] = (1 - \rho(\mathcal{S}_t(D) \setminus \{b\})) \prod_{t'=1}^{t-1} \rho\Big(\mathcal{S}_{t'}(D) \setminus \{b\}\Big) \qquad \text{(A.121)}$$

Proof. First observe that, the event $\{\tau = t\}$ implies that the learner only visits states in $\mathcal{S}_{t'}(D) \cup \{b\}$ till time $t' < t$, and visits a state in $\mathcal{S}_\tau(D) \cup \{b\}$ at time t. That is,

$$\Pr_{\widehat{\pi}}[\tau = t] = \Pr_{\widehat{\pi}}\Big[s_t \notin \mathcal{S}_t(D) \cup \{b\},\ \forall t' < t, s_{t'} \in \mathcal{S}_{t'}(D) \cup \{b\}\Big] \tag{A.122}$$

$$= \Pr_{\widehat{\pi}}\Big[s_t \notin \mathcal{S}_t(D) \cup \{b\},\ \forall t' < t, s_{t'} \in \mathcal{S}_{t'}(D) \setminus \{b\}\Big] \tag{A.123}$$

where in the last equation, we use the fact that by construction of $\mathcal{M}[\pi^{\exp}]$, the learner is forced to visit the state b at time t if the state b is visited at any time $t' < t$.
Moreover, since the learner never visits b till time $t-1$, this implies that the learner must play the expert's action at each visited state until time $t-1$ (otherwise the state b is visited with probability 1 at time t). Therefore,

$$\Pr_{\widehat{\pi}}[\tau = t] = \Pr_{\pi^{\exp}}\Big[s_t \notin \mathcal{S}_t(D) \cup \{b\},\ \forall t' < t, s_{t'} \in \mathcal{S}_{t'}(D) \setminus \{b\}\Big]. \tag{A.124}$$

Since under the policy $\pi^{\exp}$ rolled out on $\mathcal{M}[\pi^{\exp}]$, the distribution over states induced is i.i.d. across time and drawn from ρ, we have that,

$$\Pr_{\widehat{\pi}}[\tau = t] = (1 - \rho\left(\mathcal{S}_t(D) \cup \{b\}\right)) \prod_{t'=1}^{t-1} \rho(\mathcal{S}_{t'}(D) \setminus \{b\}) \tag{A.125}$$

However the distribution ρ has no mass on the state b. Therefore $\rho\left(\mathcal{S}_t(D) \cup \{b\}\right) = \rho\left(\mathcal{S}_t(D) \setminus \{b\}\right)$ and the proof concludes. □

Corollary A.7.1. $\Pr_{\widehat{\pi}(D,A)}[\tau \le \lfloor H/2 \rfloor] = 1 - \prod_{t=1}^{\lfloor H/2 \rfloor} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)$.

Lemma A.7.2. *Fix some policy $\pi^{\exp} \in \Pi^{\mathit{BC}}_{\det}(D, A)$ and the MDP as $\mathcal{M}[\pi^{\exp}]$. Then,*

$$\mathbb{E}\left[\Pr_{\widehat{\pi}(D,A)}[\tau \le \lfloor H/2 \rfloor]\right] \ge 1 - (1-\gamma)^{\lfloor H/2 \rfloor} \tag{A.126}$$

where $\gamma = \sum_{s \in \mathcal{S}} \rho(s)(1 - \rho(S))^N$.

Proof. Recall that the learner rolls out policies $\pi_1, \cdots, \pi_N$ to generate trajectories $\mathsf{tr}_1, \cdots, \mathsf{tr}_N$. First observe that, conditioned on the learner's dataset truncated till the states visited at time t,

$$\mathbb{E}\left[\prod_{t=1}^{\tau} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)\right] - \mathbb{E}\left[\prod_{t=1}^{\tau+1} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)\right]$$
$$= \mathbb{E}\left[\prod_{t=1}^{\tau} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)\Big(1 - \mathbb{E}\left[\rho\left(\mathcal{S}_{\tau+1}(D) \setminus \{b\}\right) \Big| D_{\le \tau, < \tau}\right]\Big)\right] \tag{A.127}$$

where in the last equation we use the fact $\mathcal{S}_t(D)$ for all $t \le \tau$ is a measurable function of $D_{\le \tau, < \tau}$. Conditioned on $D_{\le \tau, < \tau}$, consider the distribution over actions a^n_τ played by the learner in different trajectories. If $a^n_\tau = \pi^{\exp}_t(s^n_\tau)$, the state $s^n_{\tau+1}$ is renewed in the distribution ρ. If

a_τ^n is any other action, $s_{\tau+1}^n = b$ with probability 1, and does not provide any contribution to $\rho\left(\mathcal{S}_{\tau+1}(D) \setminus \{b\}\right)$. Let the random variable N' denote the number of trajectories that have already visited b prior to time τ or play an action other than the expert's action at time τ. By linearity of expectation,

$$1 - \mathbb{E}\left[\rho\left(\mathcal{S}_{\tau+1}(D) \setminus \{b\}\right) \middle| D_{\leq\tau,<\tau}\right] = \mathbb{E}\left[\sum\nolimits_{s\in\mathcal{S}\setminus\{b\}} \rho(s)\left(1-\rho(s)\right)^{N'} \middle| D_{\leq\tau,<\tau}\right] \tag{A.128}$$

$$\geq \sum\nolimits_{s\in\mathcal{S}\setminus\{b\}} \rho(s)\left(1-\rho(s)\right)^{N} \tag{A.129}$$

Recalling that γ is defined as the constant $\sum_{s\in\mathcal{S}} \rho(s)\left(1-\rho(s)\right)^N$ and $\rho(b) = 0$, from eqs. (A.127) and (A.129),

$$\mathbb{E}\left[\prod\nolimits_{t=1}^{\tau+1} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)\right] \geq (1-\gamma)\mathbb{E}\left[\prod\nolimits_{t=1}^{\tau} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)\right] \tag{A.130}$$

We also have that $\mathbb{E}[\rho(\mathcal{S}_1(D) \setminus \{b\})] = 1 - \sum_{s\in\mathcal{S}\setminus\{b\}} \rho(s)(1-\rho(s))^N = 1-\gamma$ since the initial state s in each trajectory in D is sampled independently and identically from ρ. Using this fact and recursing eq. (A.130) over $\tau = 1, \cdots, \lfloor H/2 \rfloor - 1$ gives,

$$\mathbb{E}\left[\prod\nolimits_{t=1}^{\lfloor H/2 \rfloor} \rho\Big(\mathcal{S}_t(D) \setminus \{b\}\Big)\right] \geq (1-\gamma)^{\lfloor H/2 \rfloor}. \tag{A.131}$$

Invoking Corollary A.7.1 completes the proof. □

Lemma A.7.3. *γ, defined in Lemma A.7.2 as $\sum_{s\in\mathcal{S}} \rho(s)(1-\rho(s))^N$ is $\geq \frac{|\mathcal{S}|-2}{e(N+1)}$.*

Proof. By the definition of ρ, we have that,

$$\gamma = \sum_{s\in\mathcal{S}} \rho(s)(1-\rho(s))^N \overset{(i)}{\geq} \frac{|\mathcal{S}|-2}{N+1}\left(1 - \frac{1}{N+1}\right)^N \geq \frac{|\mathcal{S}|-2}{e(N+1)}. \tag{A.132}$$

where in (i) we lower bound by only considering the $|\mathcal{S}|-2$ states having mass $= \frac{1}{N+1}$ under ρ. □

Lower bound in the known-transition setting

Proof of Lemma A.2.15

The proof of this result closely follows that of Lemma A.2.11. Fix some policy $\pi \in \Pi_{\texttt{det}}$. Consider any time $t \in [H]$ and state $s \in \mathcal{S}_t(D)$ which is visited in some trajectory in the dataset at time t. If $\pi_t(s)$ does not match the unique action $a_t^\star(s)$ played at time t in any trajectory in D that visits s at this time, the likelihood of π given D is exactly 0 (recall we assume that the expert's policy is deterministic). On the contrary, the conditional probability of observing the expert dataset D does not depend on the expert's action on the states that were not observed in D, since no trajectory visits these states. On these states the expert's action marginally follows the uniform distribution over $\mathcal{A}$. Thus the result follows.

Proof of Lemma A.2.16

Observe that,

$$\begin{aligned}
&\mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[H - J_r(\widehat{\pi})\right] \\
&= \mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[\sum\nolimits_{t=1}^{H} 1 - r_t(s_t,a_t)\right]\right] &\text{(A.133)}\\
&\geq \sum\nolimits_{t=1}^{H} \mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[\mathbb{1}(s_1 \notin \mathcal{S}_1(D))\Big(1 - r_t(s_t,a_t)\Big)\right]\right] &\text{(A.134)}
\end{aligned}$$

By construction of the $\mathcal{M}[\pi^{\exp}]$ and P each state $s \in \mathcal{S}$ is absorbing. Therefore, $s_1 \notin \mathcal{S}_1(D) \iff \{\forall t \in [H],\ s_t \notin \mathcal{S}_t(D)\}$. By the structure of the reward function $r[\pi^{\exp}]$, the learner accrues a reward of 1 at some state if and only if the learner plays the expert's action at this state. Therefore, $r_t(s_t,a_t) = \mathbb{1}(a_t = \pi_t^{\exp}(s_t))$ and,

$$\begin{aligned}
&\mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[r_t(s_t,a_t)\middle| s_1 \notin \mathcal{S}_1(D)\right]\right] \\
&= \mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[\mathbb{1}(a_t = \pi_t^{\exp}(s_t))\middle| s_1 \notin \mathcal{S}_1(D)\right]\right] &\text{(A.135)}
\end{aligned}$$

From Lemma A.2.15 observe that conditioned on D, the expert's policy $\pi^{\exp}$ is sampled uniformly from $\Pi_{\mathtt{det}}^{\mathtt{BC}}(D)$. Since we condition on $s_1 \notin \mathcal{S}_1(D) \iff s_t \notin \mathcal{S}_t(D)$ the state s_t is not visited in any trajectory in D at time t. This implies that the expert's action $\pi_t^{\exp}(s_t)$ is uniformly sampled from $\mathcal{A}$. Therefore,

$$\mathbb{E}_{\widehat{\pi}}\left[\mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[\mathbb{1}(a_t = \pi_t^{\exp}(s_t))\middle| s_1 \notin \mathcal{S}_1(D)\right]\right] = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\mathbb{E}_{\widehat{\pi}}\left[\mathbb{1}(a_t = a)\middle| s_1 \notin \mathcal{S}_1(D)\right] = \frac{1}{|\mathcal{A}|}.$$

Plugging this into eq. (A.135) and subtracting 1 from both sides we get that,

$$\mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[\mathbb{E}_{\widehat{\pi}(D,P,\rho)}\left[1 - r_t(s_t,a_t)\middle| s_1 \notin \mathcal{S}_1(D)\right]\right] = 1 - \frac{1}{|\mathcal{A}|}. \tag{A.136}$$

Plugging this back into eq. (A.134) we get that,

$$\mathbb{E}_{(\pi^{\exp},r)\sim\mathcal{P}'(D)}\left[H - J_r(\widehat{\pi})\right] \geq H\left(1 - \frac{1}{|\mathcal{A}|}\right)\Pr_{\widehat{\pi}}\left[s_1 \notin \mathcal{S}_1(D)\right] \tag{A.137}$$

Since s_1 is sampled independently from ρ, the proof of the result concludes.

Proof of Lemma A.2.17

Note that the dataset D follows the posterior distribution generated by rolling out $\pi^{\exp}$ for N episodes when $\pi^{\exp}$ is drawn from the uniform prior $\mathrm{Unif}(\Pi_{\mathtt{det}})$. Irrespective of the choice of $\pi^{\exp}$, note that the initial distribution over states is still ρ. Therefore,

$$\begin{aligned}
\mathbb{E}[1 - \rho(\mathcal{S}_1(D)] &= \sum_{s\in\mathcal{S}}\rho(s)(1-\rho(s))^N &\text{(A.138)}\\
&\overset{(i)}{\geq} \frac{|\mathcal{S}|-1}{N+1}\left(1 - \frac{1}{N+1}\right)^N \geq \frac{|\mathcal{S}|-1}{e(N+1)} &\text{(A.139)}
\end{aligned}$$

where in (i) we lower bound by considering only the $|\mathcal{S}| - 1$ states having mass $\frac{1}{N+1}$ under ρ. Plugging this back into eq. (A.44) completes the proof of the theorem.

Appendix B

Proofs of Results in Chapter 3

We provide proofs for the theorems introduced in Chapter 3 within this appendix.

B.1 Computational and sample efficiency of (OPT-MD)

In this section we will prove Theorem 3.3.1. We show that solving the objective (OPT-MD) in `Mimic-MD` can be posed as a convex program and is thus computationally tractable. Specifically, any learner's policy π can be represented by a set of joint state-action probabilities $\{f_t^\pi(s_t, a_t)\}_{t\in[H], s_t\in\mathcal{S}, a_t\in\mathcal{A}} \in \Omega$, where Ω is the feasible set of all possible $\{q_t(s_t, a_t)\}$ such that the following constraints hold:

$$\sum_{s_t\in\mathcal{S}}\sum_{a_t\in\mathcal{A}} q_t(s_t, a_t)P_t(s_{t+1} \mid s_t, a_t) = \sum_{a_{t+1}\in\mathcal{A}} q_{t+1}(s_{t+1}, a_{t+1}), \qquad \forall t \in [H-1], s_{t+1} \in \mathcal{S}; \tag{B.1}$$

$$\sum_{a_1\in\mathcal{A}} q_1(s_1, a_1) = \rho(s_1), \qquad \forall s_1 \in \mathcal{S}; \tag{B.2}$$

$$q_t(s_t, a_t) = 0, \qquad \forall t \in [H], s_t \in \mathcal{S}_t(D_1) \text{ and } a_t \neq \pi^{\text{exp}}(s_t); \tag{B.3}$$

$$q_t(s_t, a_t) \geq 0, \qquad \forall t \in [H], s_t \in \mathcal{S}, a_t \in \mathcal{A}. \tag{B.4}$$

These constraints impose that $f_t^\pi(s_t, a_t)$ is a feasible distribution which is consistent under the MDP transition $\{P_t\}$. Given a feasible solution $\{q_t(s_t, a_t)\}$ satisfying eqs. (B.1) to (B.4), the learner's policy $\widehat{\pi}$ is constructed via $\Pr(\widehat{\pi}(s_t) = a_t) = q_t(s_t, a_t)/\sum_{\tilde{a}_t\in\mathcal{A}} q_t(s_t, \tilde{a}_t)$. We prove that $\widehat{\pi} \in \Pi_{\text{det}}^{\texttt{BC}}(D_1)$ and $q_t(s, a) = \Pr_{\widehat{\pi}}[s_t = s, a_t = a]$ for $t \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, thereby establish a one-to-one correspondance between all feasible policies $\Pi_{\text{det}}^{\texttt{BC}}(D_1)$ and the feasible set Ω. First, the non-negativity constraint eq. (B.4) implies that $\widehat{\pi}$ is a valid randomized policy, and eq. (B.3) shows that $\widehat{\pi}$ mimics the expert policy on D_1, i.e. $\widehat{\pi} \in \Pi_{\text{det}}^{\texttt{BC}}(D_1)$. Second, for $t = 1$, the identity eq. (B.2) implies that,

$$\Pr_{\widehat{\pi}}[s_1 = s, a_1 = a] = \rho(s_1) \cdot \frac{q_1(s, a)}{\sum_{\tilde{a}\in\mathcal{A}} q_1(s, \tilde{a})} = \rho(s_1) \cdot \frac{q_1(s, a)}{\rho(s_1)} = q_1(s, a),$$

as claimed. Finally, suppose that $q_t(s,a) = \Pr_{\widehat{\pi}}[s_t = s, a_t = a]$ holds for some $t \in [H-1]$, then for time $t+1$, the compatibility condition eq. (B.1) gives that

$$\begin{aligned}
\Pr_{\widehat{\pi}}[s_{t+1} = s, a_{t+1} = a] &= \Pr_{\widehat{\pi}}[s_{t+1} = s] \cdot \frac{q_{t+1}(s,a)}{\sum_{\tilde{a}} q_{t+1}(s,\tilde{a})} \\
&= \left(\sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \Pr_{\widehat{\pi}}[s_t = s', a_t = a'] \cdot P_t(s \mid s', a') \right) \cdot \frac{q_{t+1}(s,a)}{\sum_{\tilde{a} \in \mathcal{A}} q_{t+1}(s,\tilde{a})} \\
&= \left(\sum_{s_t \in \mathcal{S}} \sum_{a_t \in \mathcal{A}} q_t(s_t, a_t) \cdot P_t(s \mid s_t, a_t) \right) \cdot \frac{q_{t+1}(s,a)}{\sum_{\tilde{a} \in \mathcal{A}} q_{t+1}(s,\tilde{a})} \\
&= q_{t+1}(s,a).
\end{aligned}$$

Therefore by induction, we conclude that any element of the feasible set Ω gives rise to a feasible policy $\widehat{\pi} \in \Pi^{\mathtt{BC}}_{\mathrm{det}}(D_1)$, and the reversed direction is obvious. Consequently, given the feasible set Ω, the `Mimic-MD` objective (OPT-MD) solves the following convex program:

$$\begin{aligned}
\text{minimize} \quad & \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| q_t(s,a) - \frac{1}{|D_2|} \sum_{\mathsf{tr} \in D_2} \mathbb{1}\left(\mathsf{tr}(s_t, a_t) = (s,a)\right) \right|, \\
\text{subject to} \quad & \{q_t(s_t,a_t)\}_{t \in [H], s \in \mathcal{S}, a \in \mathcal{A}} \in \Omega.
\end{aligned} \tag{B.5}$$

It is clear that the convex program eq. (B.5) has $O(|\mathcal{S}||\mathcal{A}|H)$ variables and $O(|\mathcal{S}||\mathcal{A}|H)$ linear constraints, and therefore `Mimic-MD` can be solved approximately in $\mathsf{poly}(|\mathcal{S}|, |\mathcal{A}|, H)$ time. Now we show the upper bound when $|\mathcal{S}| = 2$. The case for $|\mathcal{S}| \geq 3$ was discussed in Chapter 2. Let E_t be the event that there exists one state at time t that has not been visited in the N expert trajectories. Note that we know the policy π^{exp}_t exactly if both states at time t have been visited: it implies that given an estimator $\widehat{f}^{\pi^{\mathrm{exp}}}_t$ for $f^{\pi^{\mathrm{exp}}}_t$, if we have seen both states for $t' = t+1, t+2, \ldots, t+m$, we can estimate $f^{\pi^{\mathrm{exp}}}_{t'}$ via computing the marginal distributions at these time steps t' since we know the conditional distributions exactly, and we consider this estimate in eq. (3.3) in `Mimic-MD`. Using the data processing inequality (Lemma B.1.1) below, we know that $D_{\mathrm{TV}}\left(f^{\pi^{\mathrm{exp}}}_{t'}, \widehat{f}^{\pi^{\mathrm{exp}}}_{t'}\right) \leq D_{\mathrm{TV}}\left(f^{\pi^{\mathrm{exp}}}_{t}, \widehat{f}^{\pi^{\mathrm{exp}}}_{t}\right)$. Hence, we have

$$\sum_{t=1}^{H} D_{\mathrm{TV}}\left(f^{\pi^{\mathrm{exp}}}_{t}, \widehat{f}^{\pi^{\mathrm{exp}}}_{t}\right) \leq H \max_{t: E_t \text{ holds}} D_{\mathrm{TV}}\left(f^{\pi^{\mathrm{exp}}}_{t}, \widehat{f}^{\pi^{\mathrm{exp}}}_{t}\right). \tag{B.6}$$

It follows from the Binomial distribution formula that the marginal probability for the unseen states for each E_i is at most $\log(H/\delta)/N$ with probability at least $1 - \delta/H$ for each i. By union bounding, with probability at least $1-\delta$, for all time steps t such that E_t is true, the unseen state has marginal probability $\lesssim \log(H/\delta)/N$. In other words, with high probability, the state distribution at each time t with an unobserved is of the form $(p, 1-p)$ where $p \lesssim \log(H/\delta)/N$. For such a distribution, using [42, Lemma 4], the empirical

distribution achieves TV error $\lesssim \sqrt{\frac{p}{N}} \lesssim \sqrt{\log(H/\delta)}/N$ for each t, which results in the final $H \times \sqrt{\log(H/\delta)}/N$ bound on $\sum_{t=1}^{H} D_{\mathrm{TV}}\left(f_t^{\pi^{\mathrm{exp}}}, \widehat{f}_t^{\pi^{\mathrm{exp}}}\right)$. Finally, observe that the imitation gap of `Mimic-MD` is upper bounded by this quantity, since $\mathsf{Gap}(\widehat{\pi}) = \sum_{t=1}^{H} \mathbb{E}_{\pi^{\mathrm{exp}}}\left[r_t(s_t, a_t)\right] - \mathbb{E}_{\widehat{\pi}}\left[r_t(s_t, a_t)\right] = \sum_{t=1}^{H} D_{\mathrm{TV}}\left(f_t^{\pi^{\mathrm{exp}}}, f_t^{\widehat{\pi}}\right) \leq 2\sum_{t=1}^{H} D_{\mathrm{TV}}\left(f_t^{\pi^{\mathrm{exp}}}, \widehat{f}_t^{\pi^{\mathrm{exp}}}\right)$ using the definition of `Mimic-MD`. The expected imitation gap bound directly follows from integrating the high probability bound using $\mathbb{E}[X] = \int_0^{\infty} \Pr(X > t)dt$ for nonnegative random variables.

Lemma B.1.1. *Consider any distributions p, q supported on $[n]$. Let P be any Markov kernel from $[n] \to \Delta([n])$. Then $D_{\mathrm{TV}}\left(P \circ p, P \circ q\right) \leq D_{\mathrm{TV}}\left(p, q\right)$.*

Proof. By definition,

$$\begin{aligned} D_{\mathrm{TV}}\left(P \circ p, P \circ q\right) &= \frac{1}{2}\sum_{j=1}^{n}\left|\sum_{i \in n} p_i P_{ij} - \sum_{i \in n} q_i P_{ij}\right| \\ &= \frac{1}{2}\sum_{j=1}^{n}\left|\sum_{i \in n}(p_i - q_i) P_{ij}\right| \\ &\leq \frac{1}{2}\sum_{j=1}^{n}\sum_{i \in n}|p_i - q_i| P_{ij} \\ &= \frac{1}{2}\sum_{i=1}^{n}|p_i - q_i| = D_{\mathrm{TV}}\left(p, q\right) \end{aligned}$$

□

B.2 Statistical lower bounds in the known-transition setting

In this section we will prove the lower bound in Theorem 3.3.2. The first key observation is first that in order to establish a lower bound on the one-sided error probability $\Pr(J_{\mathcal{M}}(\pi^{\mathrm{exp}}) - J_{\mathcal{M}}(\widehat{\pi}) \geq H^{3/2}/N)$ for any learner $\widehat{\pi}$, it suffices to lower bound the two-sided error probability $\Pr(|J_{\mathcal{M}}(\pi^{\mathrm{exp}}) - J_{\mathcal{M}}(\widehat{\pi})| \geq H^{3/2}/N)$. Intuitively this is because the learner gets no reward feedback - a learner which has small one-sided error probability on some MDP $\mathcal{M} = (P, r)$ can potentially have large one-sided error probability on the MDP $\mathcal{M} = (P, 1 - r)$. In the absence of reward feedback, the learner cannot distinguish between these two cases. The only option for the learner is to guarantee small two-sided error probability on all IL instances to ensure a uniform bound on the one-sided error probability. In particular, we show the following result.

Lemma B.2.1. *Suppose there exists an MDP $\mathcal{M}$ with $|\mathcal{S}| = 3$ such that,*

$$\Pr\left(|J_{\mathcal{M}}(\pi^{\exp}) - J_{\mathcal{M}}(\widehat{\pi}(D))| \gtrsim \frac{H^{3/2}}{N}\right) \geq c',$$

for some constant $0 < c' \leq 1$. Then there exists an MDP $\mathcal{M}'$ with $|\mathcal{S}| = 3$ such that, $\Pr\left(J_{\mathcal{M}'}(\pi^{\exp}) - J_{\mathcal{M}'}(\widehat{\pi}(D)) \gtrsim \frac{H^{3/2}}{N}\right) \geq c'/2$.

Proof. Suppose for every MDP $\mathcal{M}$, there exists a learner $\widehat{\pi}$ such that,

$$\Pr\left(J_{\mathcal{M}}(\pi^{\exp}) - J_{\mathcal{M}}(\widehat{\pi}(D)) \gtrsim \frac{H^{3/2}}{N}\right) < \frac{c'}{2}. \tag{B.7}$$

This implies that for every MDP $\mathcal{M}$,

$$\Pr\left(J_{\mathcal{M}}(\widehat{\pi}(D)) - J_{\mathcal{M}}(\pi^{\exp}) \gtrsim \frac{H^{3/2}}{N}\right) < \frac{c'}{2}. \tag{B.8}$$

This follows from the fact that for any MDP $\mathcal{M} = (P, r)$ we can consider an MDP $\mathcal{M}' = (P, r')$ where $r'_t = 1 - r_t$. As a consequence $J_{\mathcal{M}'}(\pi) = H - J_{\mathcal{M}}(\pi)$ for any policy π which gives the equation. By adding together eqs. (B.7) and (B.8) we see that for every MDP $\mathcal{M}$, $\widehat{\pi}$ satsifies the property that

$$\Pr\left(|J_{\mathcal{M}}(\widehat{\pi}(D)) - J_{\mathcal{M}}(\pi^{\exp})| \gtrsim \frac{H^{3/2}}{N}\right) < c'. \tag{B.9}$$

Taking the contrapositive, this implies the required statement. □

In order to furnish the lower bound, we will consider a Bayes IL problem, where the expert's policy $\pi^{\exp}$ and the underlying MDP are sampled from some distribution $\mathcal{D}$.
In order to prove this result, we assume that the underlying MDP $\mathcal{M}$ and the expert policy $\pi^{\exp}$ are jointly sampled from a distribution $\mathcal{D}$ and show that there is a constant c' such that

$$\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{D}}\left[\mathbb{E}\left[\mathbb{1}\left(|J_{\mathcal{M}}(\pi^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{N}\right)\right]\right] < c'. \tag{B.10}$$

This implies the existence of an MDP $\mathcal{M}$ and expert policy $\pi^{\exp}$ with the required property. Next, we use a symmetrization argument to upper bound the LHS of the above formula.

Lemma B.2.2. *For any constant $C > 0$,*

$$\Pr\left(|J_{\mathcal{M}}(\pi^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| \leq \frac{CH^{3/2}}{N}\right) \leq \frac{1}{2} + \frac{1}{2}\mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\pi_2^{\star})| \leq \frac{CH^{3/2}}{N}\middle|\mathcal{M}, D\right)\right] \tag{B.11}$$

where $\pi_1^{\exp}$ and $\pi_2^{\exp}$ are independent copies of the expert's policy drawn from the posterior distribution conditioned on the demonstration dataset D and MDP $\mathcal{M}$.

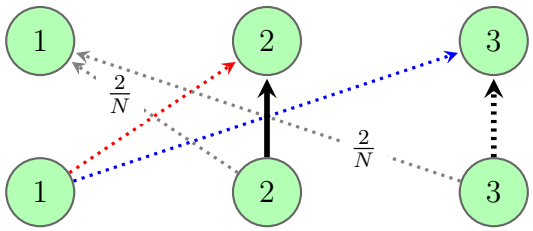


Figure B.1: Lower bound instance for $|\mathcal{S}| = 3$. The dotted transitions offer no reward and solid transitions offer reward 1. State 1 is the only one with 2 actions: red leading to state 2 and blue leading to state 3. The action at state 2 and 3 transitions the learner to state 1 with probability $\frac{1}{N}$ and leaves it unchanged otherwise. The initial distribution is at state 2 with probability 1

Proof. By definition,

$$
\begin{aligned}
&2\Pr\left(|J_{\mathcal{M}}(\pi^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| \le \frac{CH^{3/2}}{N}\right) \\
&\overset{(i)}{=} \mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| \le \frac{CH^{3/2}}{N}\middle|\mathcal{M}, \widehat{\pi}\right)\right] \\
&\qquad + \mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_2^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| \le \frac{CH^{3/2}}{N}\middle|\mathcal{M}, \widehat{\pi}\right)\right] \\
&\overset{(ii)}{\le} 1 + \mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| + |J_{\mathcal{M}}(\pi_2^{\exp}) - J_{\mathcal{M}}(\widehat{\pi})| \le \frac{CH^{3/2}}{N}\middle|\mathcal{M}, \widehat{\pi}\right)\right] \\
&\overset{(iii)}{\le} 1 + \mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\pi_2^{\exp})| \le \frac{CH^{3/2}}{N}\middle|\mathcal{M}, \widehat{\pi}\right)\right] \\
&= 1 + \mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\pi_2^{\exp})| \le \frac{CH^{3/2}}{N}\middle|\mathcal{M}, D\right)\right]
\end{aligned}
$$

where in (i), $\pi_1^{\exp}$ and $\pi_2^{\exp}$ are as defined in the theorem statement. (ii) uses the fact that $\mathbb{1}(x \le a) + \mathbb{1}(y \le b) \le 1 + \mathbb{1}(x + y \le a + b)$ and (iii) follows by triangle inequality. The last inequality follows from the fact that the expert policy is independent of any external randomness employed by $\widehat{\pi}$. □

Lower bound instance for known transition tabular setting

In this section, we describe the prior distribution $\mathcal{D}$ jointly over expert policies and MDPs. The MDP is time invariant. We first describe the transition structure of the MDP. We assume that $N \ge \max\{7, H\}$ and $|\mathcal{S}| \ge 3$.

MDP transition structure. We prove the lower bound for the case of $|\mathcal{S}| = 3$. The transition of the MDP is depicted in fig. B.1. The initial distribution of the MDP is at state 2 with probability 1. We assume that $|\mathcal{A}| = 2$ and furthermore that the states 2 and 3 only have a single action. This is without loss of generality, by assuming that the two actions induce the same distribution over states and constraining the reward to be the same.
At states 2 and 3, playing either action transitions the learner to state 1 with probability $1/N$ and stays put with the remaining probability. On the other hand, at state 1, picking action a_1 deterministically transitions the learner to state 2 while picking actions a_2 transitions the learner to state 3. State 1 is the only one where the choice of action is relevant so we specify a policy by only mentioning the action distribution at state 1 in each group at each time in the episode.

MDP reward structure. In each group g, the reward function of the MDP is chosen to be 1 for the action at state g_2 and 0 for every other state-action combination.

Expert policy. The expert policy is time variant. Recall that state g_1 in each group g is the only state where the choice of action plays a non-trivial role. Define Π_{det} as the set of all time-variant deterministic policies. Recall the assumption that in each group g, states g_2 and g_3 have only a single action.

To finally obtain the lower bound, we simply invoke the symmetrization argument in Lemma B.2.2. First, conditioned on the dataset D, we compute the posterior distribution of the expert policy. To this end, recall the definition of $\Pi^{\texttt{BC}}(D)$ (cf. eq. (2.3)), the set of deterministic policies which are "consistent" with the dataset D and at each state visited in D play the same action as observed in D. Namely,

$$\Pi^{\texttt{BC}}_{\text{det}}(D) \triangleq \left\{\pi \in \Pi_{\text{det}} : \forall t \in [H],\ s \in \mathcal{S}_t(D),\ \pi_t(\cdot|s) = \delta_{\pi^{\text{exp}}_t(s)}\right\},$$

where $\mathcal{S}_t(D)$ denotes the set of states visited at time t in some trajectory in D, and $\pi^{\text{exp}}_t(s)$ is the unique action played by the expert at time t in any trajectory in D that visits the state s at time t. Invoking [70, Lemma A.14], it follows that:

Lemma B.2.3. *Conditioned on the demonstration dataset D, the expert policy is distributed as* $\text{Unif}(\Pi^{\texttt{BC}}_{\text{det}}(D))$. *In other words, at each time t such that state 1 is unvisited in any trajectory in the demonstration dataset,* $\pi^{\text{exp}}_t(a_1|1) \sim \text{Unif}(\{0, 1\})$.

Finally, consider $\pi^{\text{exp}}_t(a_1|1)$, which is an indicator random variable for the event that the expert plays action a_1 at the state 1 at time t. With this notation, we can compute the total reward collected by the expert policy.

Lemma B.2.4. *Consider the expert policy π^{exp}. Then,*

$$J_{\mathcal{M}}(\pi^{\text{exp}}) = \sum_{t=1}^{H-1}\left(\sum_{t'=t+1}^{H}\left(1-\frac{1}{N}\right)^{H-t'}\right)\Pr(s_t = 1)\pi^{\text{exp}}_t(a_1|g_1) + \sum_{t=1}^{H}\left(1-\frac{2}{N}\right)^{t-1}.$$

Proof. Fixing the expert policy π^{exp}, the probability that the expert visits the state 1 at time 2 satisfies the condition:

$$\begin{aligned} \Pr_{\pi^{\text{exp}}}(s_t = 1) &= \frac{2}{N}\left(\Pr_{\pi^{\text{exp}}}(s_{t-1} = 2) + \Pr_{\pi^{\text{exp}}}(s_{t-1} = 3)\right) \\ \implies \Pr_{\pi^{\text{exp}}}(s_t = 1) &= \frac{2}{N}(1 - \Pr(s_{t-1} = 1)). \end{aligned}$$

With the initial condition $\Pr_{\pi^{\text{exp}}}(s_1 = 1) = 0$, the solution to the recurrence relation is, $\Pr_{\pi^{\text{exp}}}(s_t = 1) = \frac{1}{(N/2)+1}\left(1 - \frac{1}{(-N/2)^{t-1}}\right)$. Note that this probability is independent of the actions chosen by the expert at state 2 so henceforth we denote it by $\Pr(s_t = 1)$. Moreover for $t > 1$,

$$\frac{2(N-2)}{N^2} \leq \Pr(s_t = 1) \leq \frac{2}{N} \tag{B.12}$$

with the upper bound for $t = 2$ and the lower bound for $t = 3$. Next observe that,

$$\Pr_{\pi^{\text{exp}}}(s_t = 2) = \left(1 - \frac{2}{N}\right)\Pr_{\pi^{\text{exp}}}(s_{t-1} = 2) + \Pr\left(s_{t-1} = 1\right)\pi_t^{\text{exp}}(a_1|1) \tag{B.13}$$

observe that $\pi_t^{\text{exp}}(a_1|1)$ is a $\mathsf{Bern}(1/2)$ random variable indicating whether the expert picks the action a_1 at state 1 at time t. Finally,

$$\begin{aligned} J(\pi^{\text{exp}}) &= \sum_{t=1}^{H}\Pr_{\pi^{\text{exp}}}(s_t = 2) \\ &= \sum_{t=1}^{H-1}\left(\sum_{t'=t+1}^{H}\left(1 - \frac{2}{N}\right)^{H-t'}\right)\Pr(s_t = 1)\pi_t^{\text{exp}}(a_1|1) + \sum_{t=1}^{H}\left(1 - \frac{2}{N}\right)^{t-1}. \end{aligned}$$

where the last equation uses the recursion for $\Pr_{\pi^{\text{exp}}}(s_t = 2)$ in eq. (B.13). □

Lemma B.2.5. *Conditioned on the demonstration dataset D, sample two instances of the expert policy π_1^{exp} and π_2^{exp}. Then,*

$$J_{\mathcal{M}}(\pi_1^{\text{exp}}) - J_{\mathcal{M}}(\pi_2^{\text{exp}}) = \sum_{t=1}^{H-1}\left(\sum_{t'=t+1}^{H}\left(1 - \frac{2}{N}\right)^{H-t'}\right)\Pr(s_t = 1)X_t\mathbb{1}(1 \in \mathcal{S}_t(D)).$$

where recall that $\mathcal{S}_t(D)$ is the set of states visited in some trajectory in the dataset D at time t, and X_t are i.i.d. random variables distributed as:

$$X_t(i) = \begin{cases} -1, & w.p.\ \frac{1}{4} \\ 0, & w.p.\ \frac{1}{2} \\ +1, & w.p.\ \frac{1}{4} \end{cases}$$

Proof. Invoking Lemmas B.2.3 and B.2.4 for $\pi_1^{\exp}$ and $\pi_2^\star$, the statement follows immediately. □

Lemma B.2.6. *There exists a constant $C > 0$ such that, if $N \geq \max\{7, H\}$,*

$$\mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\pi_2^{\exp})| \leq \frac{CH^{3/2}}{N}\middle| D\right)\right] \leq 0.9.$$

Proof. Define the zero-mean random variable, $Z_D = J(\pi_1^{\exp}) - J(\pi_2^{\exp})$ where $\pi_1^{\exp}$ and $\pi_2^\star$ are sampled from the posterior distribution conditioned on the demonstration dataset D. From Lemma B.2.5, observe that $Z_D = \sum_{t=1}^{H-1} \kappa_t X_t$ where $\kappa_t = \sum_{t'=t+1}^{H} \left(1 - \frac{2}{N}\right)^{H-t'} \Pr(s_t = 1)\mathbb{1}(1 \in \mathcal{S}_t(D))$. By the Paley Zygmund inequality, for $0 \leq \theta \leq 1$,

$$\Pr\left(Z_D^2 \geq \theta \mathrm{Var}(Z_D)\middle| D\right) \geq (1-\theta)^2 \frac{\mathbb{E}\left[Z_D^2|D\right]^2}{\mathbb{E}\left[Z_D^4|D\right]}. \tag{B.14}$$

Then, $\mathrm{Var}(Z_D) = \mathbb{E}\left[Z_D^2|D\right] = \frac{1}{2}\sum_{t=1}^{H} \kappa_t^2$. Furthermore, $\mathbb{E}[Z_D^4] \leq \frac{3}{4}\sum_{t_1 \neq t_2 \in [H]} \kappa_{t_1}^2 \kappa_{t_2}^2 + \frac{1}{2}\sum_{t=1}^{H} \kappa_t^4 \leq \frac{3}{4}(\sum_{t=1}^{H} \kappa_t^2)^2$. Therefore, with $\theta = \frac{1}{10}$,

$$\Pr\left(Z_D^2 \geq \frac{1}{10}\mathbb{E}\left[Z_D^2|D\right]\middle| D\right) \geq \frac{99}{100}\frac{1/4}{3/4} = \frac{33}{100}. \tag{B.15}$$

Now it remains to bound $\mathbb{E}[Z_D^2|D]$. In the sequel we will apply the second moment method yes again to show a lower bound with constant probability. We first lower bound $\mathbb{E}[Z_D^2] \gtrsim H^3/N^2$, and also upper bound $\mathrm{Var}(\mathbb{E}[Z_D^2|D]) \leq c\mathbb{E}[Z_D^2]$ for a small constant $c > 0$. Together, with another application of the second moment method, this implies that $\mathbb{E}[Z_D^2|D] \geq \mathbb{E}[Z_D^2] \gtrsim H^{3/2}/N$ with constant probability.

Lemma B.2.7. $\mathbb{E}[Z_D^2] \gtrsim \frac{H^3}{N^2}$.

Proof. By definition,

$$\begin{aligned}
\mathbb{E}\left[Z_D^2\middle|D\right] &= \frac{1}{2}\sum_{t=1}^{H} \kappa_t^2 \\
&= \frac{1}{2}\sum_{t=1}^{H}\left(\sum_{t'=t+1}^{H}\left(1-\frac{2}{N}\right)^{H-t'}\right)^2 \left(\Pr(s_t=1)\right)^2 \mathbb{1}(1 \in \mathcal{S}_t(D))
\end{aligned}$$

Taking an expectation over D on both sides and simplifying,

$$\begin{aligned}
\mathbb{E}\left[Z_D^2\right] &= \frac{1}{2}\sum_{t=1}^{H-1}\left(\sum_{t'=t+1}^{H}\left(1-\frac{2}{N}\right)^{H-t'}\right)^2 \left(\Pr(s_t=1)\right)^2 \Pr(1 \in \mathcal{S}_t(D)) \\
&\overset{(i)}{\gtrsim} \sum_{t=1}^{H-1} (H-t)^2 \left(\Pr(s_t=1)\right)^2 \Pr(1 \in \mathcal{S}_t(D)) \\
&\overset{(ii)}{\gtrsim} \frac{H^3}{N^2},
\end{aligned} \tag{B.16}$$

Note that (i) follows from the fact that $\sum_{t'=t+1}^{H}\left(1-\frac{1}{N}\right)^{H-t'} \gtrsim H-t$ since $N \geq |\mathcal{S}|H$, while (ii) follows from Eq. (B.12) which shows that $\Pr(s_t = 1) \gtrsim \frac{1}{N}$ and the fact that $\Pr(1 \in \mathcal{S}_t(D)) = 1-(1-\Pr(s_t=1))^N \geq 1-\left(1-\frac{2}{N}\right)^N \geq 4/5$ for $N \geq 7$. □

Lemma B.2.8. $\mathrm{Var}\left(\mathbb{E}\left[Z_D^2|D\right]\right) \leq \sqrt{99/100}\cdot\mathbb{E}[Z_D^2]$

Proof. Observe that,

$$\begin{aligned}\sqrt{\mathrm{Var}\left(\mathbb{E}\left[Z_D^2|D\right]\right)} &\overset{(i)}{\leq} \frac{1}{2}\sum_{t=1}^{H-1}\sqrt{\mathrm{Var}(\kappa_t^2)} \\ &\leq \frac{1}{2}\sum_{t=1}^{H-1}\sqrt{\mathbb{E}\left[\kappa_t^4\right]} \\ &\overset{(ii)}{\leq} \frac{1}{2}\sum_{t=1}^{H-1}\frac{\mathbb{E}[\kappa_t^2]}{\sqrt{4/5}} \leq \frac{\mathbb{E}[Z_D^2]}{\sqrt{4/5}}.\end{aligned} \tag{B.17}$$

In (i), we use the definition $\mathbb{E}\left[Z_D^2|D\right] = \frac{1}{2}\sum_{t=1}^{H-1}\kappa_t^2$. In (ii), we use the fact that $\mathbb{E}[\kappa_t^4]$ is a scaled indicator random variable. Therefore, $\mathbb{E}[\kappa_t^4] = \frac{\mathbb{E}[\kappa_t^2]^2}{\Pr(\kappa_t>0)} \leq \frac{\mathbb{E}[\kappa_t^2]^2}{4/5}$. Here, the last inequality uses the fact that $\Pr(\kappa_t > 0) = \Pr(1 \in \mathcal{S}_t(D)) \geq 1-(1-\Pr(s_t=1))^N \geq 1-\left(1-\frac{2(N-1)}{N^2}\right)^N \geq 4/5$ for $N \geq 7$. □

Finally, we are ready to establish a lower bound on $\mathbb{E}[Z_D^2|D]$. By an application of the second moment method,

$$\Pr\left(\mathbb{E}\left[Z_D^2|D\right] \geq \frac{1}{10}\mathbb{E}\left[Z_D^2\right]\right) \geq \frac{99\mathbb{E}[Z_D^2]^2}{100\mathrm{Var}\left(\mathbb{E}\left[Z_D^2|D\right]\right)} \geq \frac{99}{100}\cdot\frac{4}{5}. \tag{B.18}$$

Putting this together with eq. (B.15), conditioning on the event $\mathcal{E} = \left\{\mathbb{E}\left[Z_D^2|D\right] \geq \frac{1}{10}\mathbb{E}[Z_D^2]\right\}$, which is a measurable function of the dataset D and occurs with large constant probability from eq. (B.18)),

$$\begin{aligned}\mathbb{E}\left[\Pr\left(Z_D^2 \geq \frac{\mathbb{E}\left[Z_D^2\right]}{100}\middle| D\right)\right] &\geq \Pr(\mathcal{E})\mathbb{E}\left[\Pr\left(Z_D^2 \geq \frac{\mathbb{E}[Z_D^2]}{100}\middle|\mathcal{E}, D\right)\right] \\ &\overset{(i)}{\geq} \frac{99}{100}\cdot\frac{4}{5}\mathbb{E}\left[\Pr\left(Z_D^2 \geq \frac{\mathbb{E}[Z_D^2|D]}{10}\middle|\mathcal{E}, D\right)\right] \\ &\geq \frac{99}{100}\cdot\frac{4}{5}\cdot\frac{33}{100} > 0.1\end{aligned}$$

where (i) uses the definition of $\mathcal{E}$ and the last equation follows by eq. (B.14). In particular,

$$\mathbb{E}\left[\Pr\left(Z_D^2 \leq \frac{\mathbb{E}[Z_D^2]}{100}\right)\right] < 0.9$$

Finally, we invoke the lower bound on $\mathbb{E}[Z_D^2] \gtrsim \frac{H^{3/2}}{N}$ from Lemma B.2.7 and use the fact that $Z_D = J(\pi_1^{\mathrm{exp}}) - J(\pi_2^\star)$ to complete the proof. □

Proof of Theorem 3.3.2 From Lemma B.2.5, there exists a constant $C > 0$ such that,

$$\mathbb{E}\left[\Pr\left(|J_{\mathcal{M}}(\pi_1^{\exp}) - J_{\mathcal{M}}(\pi_2^{\exp})| \leq \frac{CH^{3/2}}{N}\middle| D\right)\right] \leq 0.9.$$

Therefore, from Lemma B.2.1 and Lemma B.2.2, we conclude there exists an MDP $\mathcal{M}$ such that for any learner $\widehat{\pi}$,

$$\Pr\left(J_{\mathcal{M}}(\pi^{\exp}) - J_{\mathcal{M}}(\widehat{\pi}) \geq \frac{CH^{3/2}}{N}\right) \geq \frac{1 - 0.95}{2} = 0.025.$$

Appendix C

Proofs of results in Chapter 4

C.1 Proof of Theorem 4.1.1

Consider a Poissonized setting where we receive $\mathrm{Poi}(N)$ trajectories[1]. Let X_t represent the number of trajectories in which the expert visits state 2 at time t in the dataset. Under the Poisson setting, $\{X_t : 1 \le t \le H\}$ are mutually independent with each following distribution $\mathrm{Poi}(Nw_t)$, where $w_t = (1-1/N)^{t-1}/N = \Pr_{\pi^{\exp}}(s_t = 2)$ since this probability does not depend on what $\pi^{\exp}$ is.

Let $\Delta = \lfloor c \log(NH) \rfloor$, where $c > 0$ is some constant to be determined later. Given random observations $\{X_t : 1 \le t \le H\}$, the policy we output is:

$$\widehat{\pi}_t(\mathrm{red} \mid 2) = \begin{cases} \frac{\sum_{i=(k-1)\Delta+1}^{k\Delta} X_i U_i}{\sum_{i=(k-1)\Delta+1}^{k\Delta} X_i} & \text{if } (k-1)\Delta < t \le k\Delta \\ 1 & \sum_{i=(k-1)\Delta+1}^{k\Delta} X_i = 0 \end{cases}$$

At any time t, the expert has marginal probability on state 3, $\Pr_{\pi^{\exp}}(s_t = 3) = \sum_{i=1}^{t-1} w_i U_i$, while our policy $\widehat{\pi}$ has expected marginal probability $\mathbb{E}[\Pr_{\widehat{\pi}}(s_t = 3)] = \sum_{i=1}^{t-1} w_i \mathbb{E}[\widehat{\pi}_t(\mathrm{red} \mid 2)]$. We aim to show that

$$\sum_{t=1}^{H} \left| \sum_{i=1}^{t-1} w_i \mathbb{E}[\widehat{\pi}_t(\mathrm{red} \mid 2)] - \sum_{i=1}^{t-1} w_i U_i \right| \lesssim \frac{H \log(HN)}{N}. \tag{C.1}$$

Using the property that for $X \sim \mathrm{Poi}(\mu)$ and independent $Y \sim \mathrm{Poi}(\lambda)$, $\mathbb{E}\left[\frac{X}{X+Y} \middle| X+Y > 0\right] = \frac{\mu}{\mu+\lambda}$, we know that if $(k-1)\Delta < t \le k\Delta$,

$$\left| \mathbb{E}[\widehat{\pi}_t(\mathrm{red} \mid 2)] - \frac{\sum_{i=(k-1)\Delta+1}^{k\Delta} w_i U_i}{\sum_{i=(k-1)\Delta+1}^{k\Delta} w_i} \right| \le \Pr\left(\sum_{i=(k-1)\Delta+1}^{k\Delta} X_i = 0 \right) \lesssim \frac{1}{N^2 H^2} \tag{C.2}$$

[1] We can always simulate Poisson sampling with $\mathrm{Poi}(N/2)$ trajectories based on N trajectories sampled based on the multinomial distribution, and the failure probability is exponentially small.

if we take c to be large enough constant. Clearly

$$\left|\left(\sum_{i=1}^{t-1} w_i\right)\frac{\sum_{i=1}^{\Delta} w_i U_i}{\sum_{i=1}^{\Delta} w_i} - \sum_{i=1}^{t-1} w_i U_i\right| \lesssim \frac{\log(HN)}{N},$$

for $1 \le t \le \Delta$, where we used the fact that each $0 \le w_i \lesssim 1/N$. Indeed, now the total bias is upper bounded by $H\frac{\log(HN)}{N} + \frac{1}{N^2H^2}H^2 \lesssim H\frac{\log(HN)}{N}$ once we combine it with (C.2).

Since the marginal distributions of states 1 and 2 do not depend on the policy, we have just shown that for every time t and every state s,

$$|\mathbb{E}[\Pr_{\widehat{\pi}}(s_t = s)] - \Pr_{\pi^{\exp}}(s_t = s)| \lesssim \frac{\log(NH)}{N},$$

which implies the final result.

C.2 Proof of Corollary 4.2.1

We use sample splitting and cut N trajectories into two halves. We construct the states $\mathcal{S}_0$ required by the `Mimic-Mixture` algorithm to be the states that are visited in the first half of the dataset, so we know the expert actions there. Then, we search from the last layer backwards for the first time that there exists one state that was not observed in the first half of the dataset. Denote that time as t_0. If there are two states in time t_0 that are unseen, then following the arguments in the binary state case in the proof of Theorem 3.3.1, we know that `Mimic-MD` already works; if there exists only one state that is unseen at time t_0, we use `Mimic-Mixture` to construct the nearly unbiased estimator of the state-action marginal distribution for one of the other two states.

The final result can be proved upon noticing the following two observations. First, by the data processing inequality (Lemma B.1.1), if we have nearly unbiased estimator at time t_0, we have nearly unbiased estimator at time H. Second, the unseen state at time t_0 must have marginal probability at most $\widetilde{O}(1/N)$ due Binomial concentration, which implies that once we construct the π^{L} and π^{S} policies in `Mimic-Mixture` as the two policies that maximize/minimize the marginal probability of the target state while guaranteeing this unseen state has expected visitation probability at most $\widetilde{O}(1/N)$, the overall imitation gap at time H is at most $\widetilde{O}(1/N)$.

C.3 Proof of Lemma 4.2.2

We first show that if the coefficients $\beta^{\star}(\mathsf{tr}), \beta^{\mathrm{S}}(\mathsf{tr}), \beta^{\mathrm{L}}(\mathsf{tr}) \in [0,1]$ such that all the unbiasedness, order, and feasibility properties hold, then we can construct a policy $\widehat{\pi}$ such that whose expected state visitation probability at the terminal state $s^{\star}$ is close to that of the expert up to $O(1/N)$.

The choice of the mixing coefficient $\widehat{\alpha}$ is slightly different from the approach of plugging eq. (4.6) into eq. (4.5): for each tr, we subsample the Poisson random variable $X(\mathsf{tr})$ using the subsampling probability $\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}) \in [0,1]$ to arrive at another Poisson random variable $Y(\mathsf{tr})$, and further subsample $Y(\mathsf{tr})$ with probability $(\beta^\star(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))/(\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr})) \in [0,1]$ to obtain a third Poisson random variable $Z(\mathsf{tr})$. The subsamplings for different tr are mutually independent. Then we construct the mixing coefficient by taking ratios as in eq. (4.4).

By the subsampling property of Poisson random variables and the mutual independence of $\{X(\mathsf{tr})\}$, the Poisson random variables $Z(\mathsf{tr}) \sim \mathsf{Poi}(N/2 \cdot (\beta^\star(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))\Pr_{\pi^{\mathrm{exp}}}(\mathsf{tr}))$ and $Y(\mathsf{tr}) - Z(\mathsf{tr}) \sim \mathsf{Poi}(N/2 \cdot (\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^\star(\mathsf{tr}))\Pr_{\pi^{\mathrm{exp}}}(\mathsf{tr}))$ are independent. Since for independent $X \sim \mathsf{Poi}(\lambda), Y \sim \mathsf{Poi}(\mu)$, it holds that

$$\mathbb{E}\left[\frac{X}{X+Y} \;\middle|\; X+Y \neq 0\right] = \frac{\lambda}{\lambda+\mu},$$

it is clear that the mixing coefficient $\widehat{\alpha}$ constructed in eq. (4.4) satisfies

$$\begin{aligned}
\mathbb{E}\left[\widehat{\alpha} \;\middle|\; \sum_{\mathsf{tr}} Y(\mathsf{tr}) \neq 0\right] &= \mathbb{E}\left[\frac{\sum_{\mathsf{tr}} Z(\mathsf{tr})}{\sum_{\mathsf{tr}} Y(\mathsf{tr})} \;\middle|\; \sum_{\mathsf{tr}} Y(\mathsf{tr}) \neq 0\right] \\
&= \frac{\sum_{\mathsf{tr}} (\beta^\star(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))\Pr_{\pi^{\mathrm{exp}}}(\mathsf{tr})}{\sum_{\mathsf{tr}} (\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))\Pr_{\pi^{\mathrm{exp}}}(\mathsf{tr})} \\
&\overset{(i)}{=} \frac{\Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)}{\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)} = \alpha^\star,
\end{aligned}$$

where (i) is due to the unbiasedness requirement for the coefficients. Consequently,

$$\begin{aligned}
|\mathbb{E}[\widehat{\alpha}] - \alpha^\star| \leq \Pr\left(\sum_{\mathsf{tr}} Y(\mathsf{tr}) = 0\right) &= \exp\left(-\frac{N}{2}\sum_{\mathsf{tr}} (\beta^{\mathrm{L}}(\mathsf{tr}) - \beta^{\mathrm{S}}(\mathsf{tr}))\Pr_{\pi^{\mathrm{exp}}}(\mathsf{tr})\right) \\
&= \exp\left(-\frac{N}{2}(\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star))\right). \qquad \text{(C.3)}
\end{aligned}$$

Using eq. (C.3) and the definition of $\widehat{\pi} = \widehat{\alpha}\pi^{\mathrm{L}} + (1-\widehat{\alpha})\pi^{\mathrm{S}}$, the bias in Theorem 4.2.1 satisfies

$$\begin{aligned}
&|\mathbb{E}[\Pr_{\widehat{\pi}}(s_t = s^\star)] - \Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star)| \\
&= |\mathbb{E}[\widehat{\alpha}](\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)) - (\Pr_{\pi^{\mathrm{exp}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star))| \\
&= |\mathbb{E}[\widehat{\alpha}] - \alpha^\star| \cdot (\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)) \\
&\leq \exp\left(-\frac{N}{2}(\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star))\right) \cdot (\Pr_{\pi^{\mathrm{L}}}(s_t = s^\star) - \Pr_{\pi^{\mathrm{S}}}(s_t = s^\star)) \\
&\leq \frac{2}{eN}, \qquad \text{(C.4)}
\end{aligned}$$

where the last inequality is due to $\sup_x xe^{-tx} = 1/(et)$ for any $t > 0$. According to eq. (C.4), the claimed bias upper bound in Theorem 4.2.1 is proved.

We introduce several useful notations for the proof. Although the state space $\mathcal{S}$ is shared among all times $t \in [H]$, we use the notation $\mathcal{S}_t$ to denote the state space at time t. Given a policy π, for $t_1 < t_2$ and states $s_{t_1} \in \mathcal{S}_{t_1}, s_{t_2} \in \mathcal{S}_{t_2}$, we use $\Pr_\pi(s_{t_1} \to s_{t_2})$ to denote the probability of reaching s_{t_2} from s_{t_1} through *only the states in* $\mathcal{S}_0$ under the policy π; in other words,

$$\Pr_\pi(s_{t_1} \to s_{t_2}) \triangleq \mathbb{E}\left[\sum_{s_{t_1+1} \in \mathcal{S}_{t_1+1} \cap \mathcal{S}_0} \cdots \sum_{s_{t_2-1} \in \mathcal{S}_{t_2-1} \cap \mathcal{S}_0} \prod_{t=t_1}^{t_2-1} P_t(s_{t+1} \mid s_t, \pi(s_t))\right], \tag{C.5}$$

where the expectation is taken with respect to the possible randomness in the policy π. When we start from the initial state distribution ρ, we also write $\Pr_\pi(\mathsf{s} \to s_t)$ to denote

$$\Pr_\pi(\mathsf{s} \to s_t) = \sum_{s_1 \in \mathcal{S}_1 \cap \mathcal{S}_0} \rho(s_1) \cdot \Pr_\pi(s_1 \to s_t), \tag{C.6}$$

where s denotes "start". Similarly, we also define the probability from s_t to the end by

$$\Pr_\pi(s_t \to \mathsf{f}) = \sum_{s_H \in \mathcal{S}_H} \Pr_\pi(s_t \to s_H), \tag{C.7}$$

where f denotes "finish". Note the following difference between eq. (C.6) and our usual notation $\Pr_\pi(s_t = s)$: the latter quantity does not require that the trajectory to s_t only consists of states in $\mathcal{S}_0$. The main motivation behind eq. (C.5), eq. (C.6), and eq. (C.7) is that $\Pr_{\pi^{\exp}}(\rho \to s_t)$ is known to the learner solely based on the publicly known expert actions at states $\mathcal{S}_0$, and the probabilities $\Pr_{\pi^{\exp}}(s_{t_1} \to s_{t_2})$ and $\Pr_{\pi^{\exp}}(s_{t_1} \to \mathsf{f})$ are known as long as the learner knows the first action $\pi^{\exp}(s_{t_1})$.

We also combine and partition all trajectories $\mathsf{tr} \in \mathcal{S}^H$ into several disjoint groups. For a given trajectory $\mathsf{tr} = (s_1, \cdots, s_H)$, we define the following notations:

1. The *characteristic* of tr, or $\mathsf{c}(\mathsf{tr})$, is the set of all times (except for $t = H$) and the corresponding states in the trajectory which are not in $\mathcal{S}_0$. Mathematically, $\mathsf{c}(\mathsf{tr}) = \{(t, s_t) : t \in [H-1], s_t \notin \mathcal{S}_0\}$.

2. The *starting point* of c, or $t_\ell(\mathsf{c})$, is defined to be the smallest $t \in [H]$ with $(t, s_t) \in \mathsf{c}$ for some $s_t \in \mathcal{S}_t$. If $\mathsf{c} = \emptyset$, we define $t_\ell(\mathsf{c}) = \perp$.

3. The *ending point* of c, or $t_r(\mathsf{c})$, is defined to be the largest $t \in [H]$ with $(t, s_t) \in \mathsf{c}$ for some $s_t \in \mathcal{S}_t$. If $\mathsf{c} = \emptyset$, we define $t_r(\mathsf{c}) = \perp$.

4. For each possible characteristic c, let the c-*group*, or $\mathcal{T}_\mathsf{c}$, be the set of all trajectories tr with $\mathsf{c}(\mathsf{tr}) = \mathsf{c}$.

5. For each possible pair (t, s_t) with $t \in [H-1], s_t \in \mathcal{S}_t \backslash \mathcal{S}_0$, let $\mathcal{G}_{t,s_t}$ be the set of all characteristics c such that $t_\ell(\mathsf{c}) = t$ and $(t, s_t) \in \mathsf{c}$. The set $\mathcal{G}_\perp$ is defined analogously.

The main idea behind the above notations is that we require the dependence of coefficients $\beta^\dagger(\mathsf{tr})$ on tr only through $\mathsf{c}(\mathsf{tr})$, for all $\dagger \in \{*, \mathrm{L}, \mathrm{S}\}$, and therefore we only need to specify the coefficients for every c-group. Consequently, we denote by $\beta^\dagger(\mathcal{T}_\mathsf{c})$ the common coefficient $\beta^\dagger(\mathsf{tr})$ for all $\mathsf{tr} \in \mathcal{T}_\mathsf{c}$, and also by $X(\mathcal{T}_\mathsf{c}) = \sum_{\mathsf{tr} \in \mathcal{T}_\mathsf{c}} X(\mathsf{tr})$ the total Poisson count for $\mathcal{T}_\mathsf{c}$. It is clear that for $\mathsf{c} = \{(t_i, s_{t_i}) : i \in [m]\}$, we have

$$\Pr_\pi(\mathcal{T}_\mathsf{c}) = \Pr_\pi(\mathsf{s} \to s_{t_1}) \cdot \prod_{j=1}^{m-1} \Pr_\pi(s_{t_j} \to s_{t_{j+1}}) \cdot \Pr_\pi(s_{t_m} \to \mathsf{f}), \tag{C.8}$$

and $X(\mathcal{T}_\mathsf{c}) \sim \mathsf{Poi}(N/2 \cdot \Pr_{\pi^{\exp}}(\mathcal{T}_\mathsf{c}))$. Note that the first probability term $\Pr_\pi(\mathsf{s} \to s_{t_1})$ of eq. (C.8) in fact does not depend on π, in the sequel the following notation will also be useful:

$$\widetilde{\Pr}_\pi(\mathcal{T}_\mathsf{c}) = \prod_{j=1}^{m-1} \Pr_\pi(s_{t_j} \to s_{t_{j+1}}) \cdot \Pr_\pi(s_{t_m} \to \mathsf{f}). \tag{C.9}$$

The starting point $t_\ell(\mathsf{c})$, as well as the group $\mathcal{G}_{t,s_t}$, is used for further partitioning the c-groups. Specifically, we sequentially assign the coefficients $\beta^\dagger(\mathcal{T}_\mathsf{c})$ to all characteristics in each group $\mathcal{G}$ via an appropriate order, and aim to show that the following three conditions hold for each $\mathcal{G} \in \{\mathcal{G}_{t,s_t} : t \in [H], s_t \in \mathcal{S}_t \backslash \mathcal{S}_0\} \cup \{\mathcal{G}_\perp\}$ (without loss of generality we assume that the target state $s^\star$ belongs to the last layer, i.e. $s^\star \in \mathcal{S}_H$):

1. Unbiasedness: for $\mathcal{G} \neq \mathcal{G}_\perp$, it holds that

$$\sum_{\mathsf{c} \in \mathcal{G}} \beta^\dagger(\mathcal{T}_\mathsf{c}) \cdot \widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_\mathsf{c}) = \sum_{\mathsf{c} \in \mathcal{G}} \widetilde{\Pr}_{\pi^\dagger}(\mathcal{T}_\mathsf{c}) \cdot \frac{\Pr_{\pi^\dagger}(s_{t_r(\mathsf{c})} \to s^\star)}{\Pr_{\pi^\dagger}(s_{t_r(\mathsf{c})} \to \mathsf{f})}, \quad \dagger \in \{*, \mathrm{L}, \mathrm{S}\}, \tag{C.10}$$

   where we recall that $t_r(\mathsf{c})$ is the ending point of c. For $\mathcal{G} = \mathcal{G}_\perp$ and $\mathsf{c} = \emptyset$, the condition eq. (C.10) is replaced by

$$\beta^\dagger(\mathcal{T}_\emptyset) \cdot \Pr_{\pi^{\exp}}(\mathcal{T}_\emptyset) = \Pr_{\pi^\dagger}(\mathsf{s} \to s^\star). \tag{C.11}$$

2. Order: it always holds that $0 \le \beta^{\mathrm{S}}(\mathcal{T}_\mathsf{c}) \le \beta^\star(\mathcal{T}_\mathsf{c}) \le \beta^{\mathrm{L}}(\mathcal{T}_\mathsf{c}) \le 1$ for all $\mathsf{c} \in \mathcal{G}$.

3. Feasibility: for $\dagger \in \{*, \mathrm{L}, \mathrm{S}\}$, the coefficient $\beta^\dagger(\mathcal{T}_\mathsf{c})$ only depends on the public information and $\{\pi^{\exp}(s_{t_j})\}_{j \in [m]}$, where $\mathsf{c} = \{(t_j, s_{t_j}) : j \in [m]\}$.

Note that the order and feasibility conditions are the same as original ones, and below we show that the above unbiasedness condition implies the original unbiasedness property. For a given $\mathcal{G} \neq \mathcal{G}_\perp$, we must have $\mathcal{G} = \mathcal{G}_{t,s_t}$ for some $t \in [H], s_t \in \mathcal{S}_t \backslash \mathcal{S}_0$. Now multiplying

$\Pr_{\pi^{\exp}}(\mathsf{s} \to s_t) = \Pr_{\pi^\dagger}(\mathsf{s} \to s_t)$ to both sides of eq. (C.10), and also using eq. (C.8), eq. (C.9), we arrive at

$$\sum_{\mathsf{c} \in \mathcal{G}} \beta^\dagger(\mathcal{T}_\mathsf{c}) \cdot \Pr_{\pi^{\exp}}(\mathcal{T}_\mathsf{c}) = \sum_{\mathsf{c} \in \mathcal{G}} \Pr_{\pi^\dagger}(\mathcal{T}_\mathsf{c}) \cdot \frac{\Pr_{\pi^\dagger}(s_{t_r(\mathsf{c})} \to s^\star)}{\Pr_{\pi^\dagger}(s_{t_r(\mathsf{c})} \to \mathsf{f})}, \quad \dagger \in \{*, \mathrm{L}, \mathrm{S}\}. \tag{C.12}$$

Using eq. (C.11) and eq. (C.12), we have

$$\begin{aligned}
&\sum_{\mathsf{tr}} \beta^\dagger(\mathsf{tr}) \cdot \Pr_{\pi^{\exp}}(\mathsf{tr}) \\
&= \beta^\dagger(\mathcal{T}_\emptyset) \cdot \Pr_{\pi^{\exp}}(\mathcal{T}_\emptyset) + \sum_{\mathcal{G} \neq \mathcal{G}_\perp} \sum_{\mathsf{c} \in \mathcal{G}} \beta^\dagger(\mathcal{T}_\mathsf{c}) \cdot \Pr_{\pi^{\exp}}(\mathcal{T}_\mathsf{c}) \\
&= \Pr_{\pi^\dagger}(\mathsf{s} \to s^\star) + \sum_{\mathsf{c} \neq \emptyset} \Pr_{\pi^\dagger}(\mathcal{T}_\mathsf{c}) \cdot \frac{\Pr_{\pi^\dagger}(s_{t_r(\mathsf{c})} \to s^\star)}{\Pr_{\pi^\dagger}(s_{t_r(\mathsf{c})} \to \mathsf{f})} \\
&= \Pr_{\pi^\dagger}(\mathsf{s} \to s^\star) + \sum_{\mathsf{c} = \{(t_j, s_{t_j}) : j \in [m]\}} \Pr_{\pi^\dagger}(\mathsf{s} \to s_{t_1}) \cdot \prod_{j=1}^{m-1} \Pr_{\pi^\dagger}(s_{t_j} \to s_{t_{j+1}}) \cdot \Pr_{\pi^\dagger}(s_{t_m} \to s^\star) \\
&= \Pr_{\pi^\dagger}(s_H = s^\star),
\end{aligned}$$

where the last identity follows from the partition of all trajectories to $s^\star$ into disjoint characteristics. This is exactly the original unbiasedness property.

Next we show that for each $\mathcal{G}$ we could fulfill the above conditions. We will first deal with the group $\mathcal{G}_\perp$ in a special way, and then handle other groups $\mathcal{G}_{t,s_t}$ by induction on $t = H-1, H-2, \cdots, 1$.

Remark C.3.1. Note that the condition eq. (C.10) cannot be replaced by eq. (C.12) in general, as it might happen that $\Pr_{\pi^\dagger}(\mathcal{T}_\mathsf{c}) = 0$ while $\widetilde{\Pr}_{\pi^\dagger}(\mathcal{T}_\mathsf{c}) > 0$. For example, in the special case of $\mathcal{S}_0 = \emptyset$, the condition eq. (C.12) is totally non-informative for $\mathcal{G} = \mathcal{G}_{t,s_t}$ with $t \geq 2$.

We also remark that the coefficient $\beta^\dagger(\mathcal{T}_\mathsf{c})$ must be constructed for every characteristic c, even if for certain $\mathcal{S}_0$ there does not exist a trajectory tr such that $\mathsf{c} = \mathsf{c}(\mathsf{tr})$ (e.g. consider $\mathcal{S}_0 = \emptyset$). This is because our construction is sequentially inductive, and therefore must be done step after step.

Edge case: $\mathcal{G} = \mathcal{G}_\perp$. In this case, the only element of $\mathcal{G}_\perp$ is $\mathsf{c} = \emptyset$, and we have

$$\begin{aligned}
\Pr_{\pi^{\exp}}(\mathcal{T}_\emptyset) &= \Pr_{\pi^{\exp}}(\mathsf{s} \to \mathsf{f}) = \Pr_{\pi^{\mathrm{L}}}(\mathsf{s} \to \mathsf{f}) = \Pr_{\pi^{\mathrm{S}}}(\mathsf{s} \to \mathsf{f}) \\
\Pr_{\pi^{\exp}}(\mathsf{s} \to s^\star) &= \Pr_{\pi^{\mathrm{L}}}(\mathsf{s} \to s^\star) = \Pr_{\pi^{\mathrm{S}}}(\mathsf{s} \to s^\star),
\end{aligned}$$

and all above quantities are publicly known. By eq. (C.11), all three conditions are fulfilled by choosing

$$\beta^\star(\mathcal{T}_\emptyset) = \beta^{\mathrm{L}}(\mathcal{T}_\emptyset) = \beta^{\mathrm{S}}(\mathcal{T}_\emptyset) = \frac{\Pr_{\pi^{\exp}}(\mathsf{s} \to s^\star)}{\Pr_{\pi^{\exp}}(\mathsf{s} \to \mathsf{f})} \in [0, 1].$$

Base step of induction: $\mathcal{G} = \mathcal{G}_{H-1,s_{H-1}}$ for some $s_{H-1} \in \mathcal{S}_{H-1}$. In this case, the set $\mathcal{G}_{H-1,s_{H-1}}$ has a unique element $\mathsf{c} = \{(H-1, s_{H-1})\}$, and the condition eq. (C.10) is equivalent to

$$\beta^{\dagger}(\mathcal{T}_{\mathsf{c}}) = \Pr_{\pi^{\dagger}}(s_{H-1} \to s^{\star}), \quad \dagger \in \{*, \mathrm{L}, \mathrm{S}\}, \tag{C.13}$$

as $\widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_{\mathsf{c}}) = \widetilde{\Pr}_{\pi^{\dagger}}(\mathcal{T}_{\mathsf{c}}) = \Pr_{\pi^{\dagger}}(s_{H-1} \to \mathsf{f}) = 1$. By definition of the extremal policies π^{L} and π^{S}, Lemma C.3.1 at the end of the section shows that the choice in eq. (C.13) also satisfies the order condition. Finally, the probability in eq. (C.13) for $\dagger \in \{\mathrm{L}, \mathrm{S}\}$ is determined by the known transition and extremal policies, while for $\dagger = *$, the coefficient only requires the additional information $\pi^{\exp}(s_{H-1})$. As the characteristic c is $\{(H-1, s_{H-1})\}$, the choice of eq. (C.13) also satisfies the feasibility condition.

Inductive step: $\mathcal{G} = \mathcal{G}_{t,s_t}$ after handling all $\mathcal{G}_{t',s_{t'}}$ for $t' > t$. We choose the coefficients $\beta^{\dagger}(\mathcal{T}_{\mathsf{c}})$ with $\mathcal{T}_{\mathsf{c}} \in \mathcal{G}_{t,s_t}$ for each $\dagger \in \{*, \mathrm{L}, \mathrm{S}\}$, respectively.
The choice for $\dagger = *$ is the simplest, and is given by

$$\beta^{\star}(\mathcal{T}_{\mathsf{c}}) \triangleq \frac{\Pr_{\pi^{\exp}}(s_{t_r(\mathsf{c})} \to s^{\star})}{\Pr_{\pi^{\exp}}(s_{t_r(\mathsf{c})} \to \mathsf{f})} \in [0,1], \quad \forall \mathsf{c} \in \mathcal{G}_{t,s_t}. \tag{C.14}$$

Plugging eq. (C.14) into eq. (C.10), it is clear that the unbiased condition holds for $\dagger = *$. Moreover, both the numerator and the denominator in eq. (C.14) only require the additional information $\pi^{\exp}(s_{t_r(\mathsf{c})})$, and thus $\beta^{\star}$ satisfies the feasibility condition.
Next we construct $\beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}})$ such that the unbiased and feasibility conditions hold, with $\beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) \in [\beta^{\star}(\mathcal{T}_{\mathsf{c}}), 1]$. An entirely symmetric argument also leads to the claimed construction of $\beta^{\mathrm{S}}(\mathcal{T}_{\mathsf{c}})$. For every $\mathsf{c} \in \mathcal{G}_{t,s_t}$, the coefficient $\beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}})$ is chosen to be

$$\beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) = (1-\alpha)\beta_0^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) + \alpha\beta_1^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}), \tag{C.15}$$

with some scalar $\alpha \in [0,1]$ independent of c, and the candidate coefficients $\beta_0^{\mathrm{L}}, \beta_1^{\mathrm{L}}$ are defined as

$$\beta_0^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) \equiv 1, \qquad \beta_1^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) = \begin{cases} \beta^{\star}(\mathcal{T}_{\mathsf{c}}) & \text{if } \mathsf{c} = \{(t, s_t)\}, \\ \beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}\setminus\{(t,s_t)\}}) & \text{otherwise.} \end{cases}$$

We first show that both $\beta_0^{\mathrm{L}}, \beta_1^{\mathrm{L}}$ satisfy the feasibility condition. This result is trivial for the constant β_0^{L}; the coefficient β_1^{L} is also feasible, for both coefficients $\beta^{\star}(\mathcal{T}_{\mathsf{c}})$ in eq. (C.14) and $\beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}\setminus\{(t,s_t)\}})$ in the inductive hypothesis are feasible. Moreover, whenever $\mathsf{c} \in \mathcal{G}_{t,s_t}$ is not a singleton, by the inductive hypothesis and eq. (C.14) we have

$$\beta^{\star}(\mathcal{T}_{\mathsf{c}}) = \frac{\Pr_{\pi^{\exp}}(s_{t_r(\mathsf{c})} \to s^{\star})}{\Pr_{\pi^{\exp}}(s_{t_r(\mathsf{c})} \to \mathsf{f})} = \frac{\Pr_{\pi^{\exp}}(s_{t_r(\mathsf{c}\setminus\{(t,s_t)\})} \to s^{\star})}{\Pr_{\pi^{\exp}}(s_{t_r(\mathsf{c}\setminus\{(t,s_t)\})} \to \mathsf{f})} = \beta^{\star}(\mathcal{T}_{\mathsf{c}\setminus\{(t,s_t)\}}) \le \beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}\setminus\{(t,s_t)\}}) \le 1.$$

Consequently, we always have $\beta_0^{\mathrm{L}}, \beta_1^{\mathrm{L}} \in [\beta^\star, 1]$, therefore the order condition $\beta^{\mathrm{L}} \in [\beta^\star, 1]$ holds for any mixture in eq. (C.15).
Now it remains to show that there exists $\alpha \in [0,1]$ such that the mixture β^{L} in eq. (C.15) satisfies the condition eq. (C.10) for $\dagger = \mathrm{L}$, and that this common value α is feasible with respect to all $\mathsf{c} \in \mathcal{G}_{t,s_t}$. For the first claim, it suffices to prove that

$$A \triangleq \sum_{\mathsf{c}\in\mathcal{G}_{t,s_t}} \beta_0^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) \cdot \widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_{\mathsf{c}}) \geq \sum_{\mathsf{c}\in\mathcal{G}_{t,s_t}} \widetilde{\Pr}_{\pi^{\mathrm{L}}}(\mathcal{T}_{\mathsf{c}}) \cdot \frac{\Pr_{\pi^{\mathrm{L}}}(s_{t_r(\mathsf{c})} \to s^\star)}{\Pr_{\pi^{\mathrm{L}}}(s_{t_r(\mathsf{c})} \to \mathsf{f})} \triangleq C, \tag{C.16}$$

$$B \triangleq \sum_{\mathsf{c}\in\mathcal{G}_{t,s_t}} \beta_1^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}}) \cdot \widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_{\mathsf{c}}) \leq \sum_{\mathsf{c}\in\mathcal{G}_{t,s_t}} \widetilde{\Pr}_{\pi^{\mathrm{L}}}(\mathcal{T}_{\mathsf{c}}) \cdot \frac{\Pr_{\pi^{\mathrm{L}}}(s_{t_r(\mathsf{c})} \to s^\star)}{\Pr_{\pi^{\mathrm{L}}}(s_{t_r(\mathsf{c})} \to \mathsf{f})} = C. \tag{C.17}$$

Given eq. (C.16) and eq. (C.17), the parameter $\alpha \in [0,1]$ to fulfill the unbiased condition eq. (C.10) is

$$\alpha = \frac{A - C}{A - B}. \tag{C.18}$$

To establish eq. (C.16), eq. (C.17) and show that the parameter α in eq. (C.18) is feasible, we will find simplified expressions for A, B, and C. First we claim that

$$A = 1, \tag{C.19}$$

$$C = \Pr_{\pi^{\mathrm{L}}}(s_H = s^\star \mid s_t). \tag{C.20}$$

To show eq. (C.19), consider all possible trajectories starting from s_t at time t. Partition the trajectories into disjoint sets labeled by different characteristics c, i.e. when and on which states the trajectory hits $\mathcal{S}_0$. It is clear by eq. (C.9) that the probability of the set labeled by c, conditioned on starting from s_t at time t, is precisely $\widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_{\mathsf{c}})$ under the expert policy $\pi^{\exp}$. Summing them up gives $A = 1$. The identity eq. (C.20) could be established in a similar way.
The quantity B is more complicated to deal with, where a key observation is that $\mathcal{G}_{t,s_t} \backslash \{(t, s_t)\} = \mathcal{G}_\perp \cup (\cup_{t'>t, s_{t'}\in\mathcal{S}_{t'}\backslash\mathcal{S}_0} \mathcal{G}_{t',s_{t'}})$. Using the definition of $\beta_1^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}})$, we have

$$\begin{aligned}
B &= \beta^\star(\mathcal{T}_{\{(t,s_t)\}}) \cdot \Pr_{\pi^{\exp}}(s_t \to \mathsf{f}) + \sum_{t'>t} \sum_{s_{t'}\in\mathcal{S}_{t'}\backslash\mathcal{S}_0} \sum_{\mathsf{c}'\in\mathcal{G}_{t',s_{t'}}} \beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}'}) \cdot \Pr_{\pi^{\exp}}(s_t \to s_{t'}) \cdot \widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_{\mathsf{c}'}) \\
&\overset{(i)}{=} \Pr_{\pi^{\exp}}(s_t \to s^\star) + \sum_{t'>t} \sum_{s_{t'}\in\mathcal{S}_{t'}\backslash\mathcal{S}_0} \Pr_{\pi^{\exp}}(s_t \to s_{t'}) \sum_{\mathsf{c}'\in\mathcal{G}_{t',s_{t'}}} \beta^{\mathrm{L}}(\mathcal{T}_{\mathsf{c}'}) \cdot \widetilde{\Pr}_{\pi^{\exp}}(\mathcal{T}_{\mathsf{c}'}) \\
&\overset{(ii)}{=} \Pr_{\pi^{\exp}}(s_t \to s^\star) + \sum_{t'>t} \sum_{s_{t'}\in\mathcal{S}_{t'}\backslash\mathcal{S}_0} \Pr_{\pi^{\exp}}(s_t \to s_{t'}) \sum_{\mathsf{c}'\in\mathcal{G}_{t',s_{t'}}} \widetilde{\Pr}_{\pi^{\mathrm{L}}}(\mathcal{T}_{\mathsf{c}'}) \cdot \frac{\Pr_{\pi^{\mathrm{L}}}(s_{t_r(\mathsf{c}')} \to s^\star)}{\Pr_{\pi^{\mathrm{L}}}(s_{t_r(\mathsf{c}')} \to \mathsf{f})} \\
&\overset{(iii)}{=} \Pr_{\pi^{\exp}}(s_t \to s^\star) + \sum_{t'>t} \sum_{s_{t'}\in\mathcal{S}_{t'}\backslash\mathcal{S}_0} \Pr_{\pi^{\exp}}(s_t \to s_{t'}) \cdot \Pr_{\pi^{\mathrm{L}}}(s_H = s^\star \mid s_{t'}),
\end{aligned}$$

where (i) follows from the definition of $\beta^\star$ in eq. (C.14), (ii) uses the inductive hypothesis eq. (C.10) for $\mathcal{G}_{t',s_{t'}}$, and (iii) follows from eq. (C.20). In other words, we have

$$B = \Pr_{\pi^{\exp}\to\pi^{\mathrm{L}}}(s_H = s^\star \mid s_t), \tag{C.21}$$

where the new policy $\pi^{\exp} \to \pi^{\mathrm{L}}$ means that starting from s_t, the learner initially adopts the policy $\pi^{\exp}$ and switches to π^{L} once he visits a state not in $\mathcal{S}_0$. The expression eq. (C.21) is obtained by distinguishing the first state not in $\mathcal{S}_0$ visited by the learner starting from s_t. By eq. (C.19), eq. (C.20) and eq. (C.21), it is clear from Lemma C.3.1 that $A \geq C \geq B$. Regarding the feasibility, it is clear that A and C are both publicly known, and B only requires the knowledge of $\pi^{\exp}(s_t)$, which is shared among all $\mathsf{c} \in \mathcal{G}_{t,s_t}$. Therefore we have completed the inductive step and and the proof of Lemma 4.2.2.

Lemma C.3.1. *For every $t \in [H]$ and $s \in \mathcal{S}_t$, (proper versions of) the extremal policies $\pi^{\mathrm{L}}, \pi^{\mathrm{S}}$ satisfy*

$$\pi^{\mathrm{L}} \in \operatorname*{argmax}_{\pi\in\Pi^{\mathcal{BC}}_{\mathrm{det}}(\mathcal{S}_0)} \Pr_{\pi}(s_H = s^\star \mid s_t = s),$$
$$\pi^{\mathrm{S}} \in \operatorname*{argmin}_{\pi\in\Pi^{\mathcal{BC}}_{\mathrm{det}}(\mathcal{S}_0)} \Pr_{\pi}(s_H = s^\star \mid s_t = s).$$

Proof. By symmetry we only prove the first claim, and we induct on $t = H-1, H-2, \cdots, 1$. For the base case $t = H-1$, the definition of π^{L} implies that changing the action $\pi^{\mathrm{L}}(s)$ to any $\pi(s)$ cannot decrease $\Pr(s_H = s^\star)$, and therefore the statement holds provided that $\Pr_{\pi^{\mathrm{L}}}(s_{H-1} = s) > 0$; moreover, in the edge case $\Pr_{\pi^{\mathrm{L}}}(s_{H-1} = s) = 0$ we may choose $\pi^{\mathrm{L}}(s)$ arbitrarily, so a proper version of π^{L} would work. For the induction step, the same local adjustment argument yields to

$$\begin{aligned}
\Pr_{\pi^{\mathrm{L}}}(s_H = s^\star \mid s_t = s) &\geq \sum_{s'\in\mathcal{S}_{t+1}} \Pr_{\pi}(s_{t+1} = s' \mid s_t = s) \cdot \Pr_{\pi^{\mathrm{L}}}(s_H = s^\star \mid s_{t+1} = s') \\
&\geq \sum_{s'\in\mathcal{S}_{t+1}} \Pr_{\pi}(s_{t+1} = s' \mid s_t = s) \cdot \Pr_{\pi}(s_H = s^\star \mid s_{t+1} = s') \\
&= \Pr_{\pi}(s_H = s^\star \mid s_t = s)
\end{aligned}$$

provided that $\Pr_{\pi^{\mathrm{L}}}(s_t = s) > 0$, where the second inequality makes use of the induction hypothesis. The edge case is again handled by considering a proper version of π^{L}. □

C.4 Proof of Theorem 4.3.1

Notation. Since states $1, 3$ and 4 have a singular action, we can identify the expert's policy by the action distribution at state 2 at each time t. For simplicity of presentation we assume that state 2 only has 2 actions, i.e. $\mathcal{A} = \{a_-, a_+\}$, although this can easily be extended to

larger action spaces. Order the two actions at each time t as a_- being the one which places the largest mass on state 3 and a_+ being the one which places the largest mass on state 4. In the sequel, we will first prove Theorem 4.3.1 for the case where states 3 and 4 of the MDP are absorbing. The extension to the more case where the action at these states induces an arbitrary next-state distribution on state 3 and 4 is deferred to later.

Constructing a reference policy For $i \in [2\log(NH)]$, define

$$\mathcal{T}_i = \left\{ t \in [H] : \frac{1}{2^{i+1}} < P_t(3|2, a_-) - P_t(3|2, a_+) \leq \frac{1}{2^i} \right\} \tag{C.22}$$

as a partitioning of the timesteps in $[H]$ depending on the range of the probabilities of visiting state 3 across actions. It is not necessary to partition beyond $i = 2\log(NH)$ since for any state where $P_t(3|2, a_-) - P_t(3|2, a_+) \leq 1/H^2N^2$, picking the wrong action at these states incurs suboptimality which, in total, amounts to at most $1/N^2$. Intuitively, timesteps belonging to $\mathcal{T}_i$ for small values of i are very informative - knowing the expert's action at this state reveals a lot about the expert's value function, since the alternative action induces a next-state distributions very different from that of the expert's action. However, for the same reasons, picking the wrong action, which might happen if the states were not seen in the dataset, could incur a large penalty in value. In contrast, timesteps belonging to $\mathcal{T}_i$ for large values of i are not informative about the ground truth reward since all the actions induce nearly the same next-state distributions. For the same reasons, this also means that picking the wrong action at these states is not bound to incur a significant penalty. We will formalize this trade-off between "informativeness" vs. "risk" in the following sections.

The learner's policy can be defined as follows: for any time $t \in \mathcal{T}_i$ for some i, let $\Delta \geq 0$ be the smallest value such that $t + \Delta \in \mathcal{T}_i$ and state 2 was observed in the dataset D at this time. The reference policy is defined as,

$$\pi_{\text{ref}}(\cdot|2) = \delta_a, \text{ where } a = \pi^{\text{exp}}_{t+\Delta}(2) \tag{C.23}$$

In other words, the learner looks at the next time belonging to the same bucket $\mathcal{T}_i$, $t + \Delta$, at which state 2 was observed in the dataset, and mimics the same action at time t. The intuition is that the learner partitions the states according to their informativeness (defined by the bucket $\mathcal{T}_i$), and tries to follow the next observed action within the same bucket to balance its risk.

First we characterize what it means for a state to be informative. Indeed, consider any time $t \in \mathcal{T}_i$ where the expert's action was observed at state 2, and suppose this action was a_- (a similar argument applies when the action was a_+). By optimality of the expert's policy, consider any policy $\widetilde{\pi} \in \Pi^{\texttt{BC}}_{\text{det}}(D)$ which agrees with π^{exp} on the states observed in the dataset and any reward function $\widetilde{r} \in \mathcal{R}_{\text{opt}}(D)$ on which this policy is optimal. Since $\widetilde{\pi}$ and π^{exp} agree on the states observed in the dataset, on the reward function $\widetilde{r}$, the value-to-go of $\widetilde{\pi}$ under the action played by the expert, a_- is at least as much as the value under the alternative

choice, a_+. Namely,

$$P(3|2,a_-)\widetilde{V}_{t+1}(3) + (1-P(3|2,a_-))\widetilde{V}_{t+1}(4) + r_{t+1}(2,a_-) \tag{C.24}$$
$$\geq P(3|2,a_+)\widetilde{V}_{t+1}(3) + (1-P(3|2,a_+))\widetilde{V}_{t+1}(4) + r_{t+1}(2,a_+)$$
$$\implies (\widetilde{V}_{t+1}(3) - \widetilde{V}_{t+1}(4))(P(3|2,a_-) - P(3|2,a_+)) \geq -1$$
$$\implies (\widetilde{V}_{t+1}(3) - \widetilde{V}_{t+1}(4)) \geq -\frac{1}{(P(3|2,a_-) - P(3|2,a_+))} \geq -2^{i+1} \tag{C.25}$$

where the last inequality uses the definition of $t \in \mathcal{T}_i$.
The lower bound in eq. (C.25) serves as a certificate showing that the action expert's action (assumed to be a_-) is not significantly worse than the alternative action (a_+) at time t. If i is large, the RHS is smaller and the time-step is less "informative" (i.e. a weaker inequality). If i is small, the RHS is larger and the time-step is more informative. This will play a crucial role in bounding the imitation gap of the learner.
In order to bound the imitation gap of the reference policy, we will decompose it across the imitation gap incurred across all the timesteps belonging to the same bucket $\mathcal{T}_i$ for some $i \in [2\log(NH)]$. Let $t_1, \cdots, t_k$ denote the time-steps belonging to $\mathcal{T}_i$. Consider any $j \in [k]$ and let $\Delta_j \geq 0$ be the smallest value such that the state 2 was observed at time $t_{j+\Delta_j}$ (if no such time exists, define $t_{j+\Delta_j} = H$). By eq. (C.25), at time $t_j + \Delta_j$, assuming the action a_- was played by $\widetilde{\pi}$,

$$\widetilde{V}_{t_{j+\Delta_j}+1}(3) - \widetilde{V}_{t_{j+\Delta_j}+1}(4) \geq -2^{i+1} \tag{C.26}$$

Recall that states 3 and 4 are absorbing, so we can write the inequality,

$$\widetilde{V}_{t_j+1}(3) - \widetilde{V}_{t_j+1}(4) \geq -2^{i+1} - (t_{j+\Delta_j} - t_j). \tag{C.27}$$

Recall that the learner matches the expert's action played at the time $t_{j+\Delta_j}$, and therefore the action a_- is played at time t_j. If the expert played the same action at time t_j, then the learner incurs no suboptimality. In case $\widetilde{\pi}$ plays the action a_+, we can lower bound the imitation gap of the reference policy as follows,

$$\begin{aligned}
\widetilde{V}_{t_j}(2) - V^{\pi_{\text{ref}}}_{t_j}(2) &\leq \left(Q^\star_{t_j}(2,a_+) - V^{\pi_{\text{ref}}}_{t_j}(2)\right) \\
&= \left(P_{t_j}(3|2,a_+) - P_{t_j}(3|2,a_-)\right)\left(V^\star_{t_j+1}(3) - V^\star_{t_j+1}(4)\right) \\
&\overset{(i)}{\leq} \frac{1}{2^i}\left(2^{i+1} + (t_{j+\Delta_j} - t_j)\right) \\
&= 2 + \frac{t_{j+\Delta_j} - t_j}{2^i},
\end{aligned}$$

where (i) uses eq. (C.25) and the fact that $t_j \in \mathcal{T}_i$.

Denoting $q_t = p_t \prod_{s<t}(1-p_s)$, the overall imitation gap of the reference policy is,

$$\begin{aligned} J_{\tilde{r}}(\widetilde{\pi}) - J_{\tilde{r}}(\pi_{\text{ref}}) &= \sum_{t=1}^{H-1} q_t \left(\widetilde{V}_t(2) - V_t^{\pi_{\text{ref}}}(2)\right) \\ &= \sum_{i \in [2\log(NH)]} \sum_{t_j \in \mathcal{T}_i} q_{t_j} \left(\widetilde{V}_{t_j}(2) - V_{t_j}^{\pi_{\text{ref}}}(2)\right) \\ &\leq \sum_{i \in [2\log(NH)]} \sum_{j=1}^{k} q_{t_j} \left(2 + \frac{t_{j+\Delta_j} - t_j}{2^i}\right) \mathbb{I}(\Delta_j > 0), \end{aligned} \tag{C.28}$$

where the last inequality follows from eq. (C.27) and the fact that if $\Delta_j = 0$, the expert's and learner's policies match at time t_j, so no imitation gap is incurred.
Note that Δ_j is a random variable is defined as the interval to the smallest time such that state 2 was observed at time $t_{j+\Delta_j}$. Consider the set of times in $\{t_1, t_2, \cdots, t_k\} \in \mathcal{T}_i$ and define $t_{k+1} = H+1$. Then, $P(\Delta_j = \ell)$ is upper bounded by the probability that state 2 was never observed in the dataset at time $t_j, t_{j+1}, t_{j+2}, \cdots, t_{j+\ell-1}$ and that state 2 was observed at time $t_{j+\ell}$. Indeed, for $\ell = 0$,

$$P(\Delta_j = 0) = 1 - (1 - q_{t_j})^N \leq N q_{t_j}$$

Likewise, note that $\Delta_j = \ell$ is the event that state 2 is seen in dataset at time $t_{j+\ell}$ in at least one trajectory, but not at time $t_j, t_{j+1}, \cdots, t_{j+\ell-1}$ in any trajectory. By union bounding over the index of the trajectory in which state 2 was observed at time $t_{j+\ell}$,

$$P(\Delta_j = \ell) \leq N q_{t_{j+\ell}} \left(1 - \sum_{\ell'=0}^{\ell-1} q_{t_{j+\ell'}}\right)^{N-1}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\sum_{t_j \in \mathcal{T}_i} q_{t_j} \left(t_{j+\Delta_j} - t_j\right)\right] &= \sum_{j=1}^{k} q_{t_j} \sum_{\ell=0}^{k-j+1} (t_{j+\ell} - t_j) N q_{t_{j+\ell}} \left(1 - \sum_{\ell'=0}^{\ell-1} q_{t_{j+\ell'}}\right)^{N-1} \\ &= \sum_{j=1}^{k} q_{t_j} \sum_{\ell=j}^{k+1} (t_\ell - t_j) N q_{t_\ell} \left(1 - \sum_{\ell'=j}^{\ell-1} q_{t_{\ell'}}\right)^{N-1} \end{aligned} \tag{C.29}$$

For each $l \leq k$, define $\theta_l = t_{l+1} - t_l$ (with $\theta_k = H - t_k$) as the gap between consecutive time-steps in the bucket $\mathcal{T}_i$. Noting that $t_\ell - t_j = \sum_{l=j}^{\ell-1} \theta_l$, eq. (C.29) can be rearranged as,

$$\begin{aligned} \mathbb{E}\left[\sum_{t_j \in \mathcal{T}_i} q_{t_j} \left(t_{j+\Delta_j} - t_j\right)\right] &= N \sum_{l=1}^{k-1} \theta_l \sum_{j=1}^{l} q_{t_j} \sum_{\ell=l+1}^{k+1} q_{t_\ell} \left(1 - \sum_{\ell'=j}^{\ell-1} q_{t_{\ell'}}\right)^{N-1} \\ &= N \sum_{l=1}^{k-1} \theta_l \sum_{j=1}^{l} q_{t_j} \sum_{\ell=l+1}^{k+1} q_{t_\ell} \left(1 - \sum_{\ell'=j}^{l} q_{t_{\ell'}} - \sum_{\ell'=l+1}^{\ell-1} q_{t_{\ell'}}\right)^{N-1}. \end{aligned}$$

Let $j_{\min}$ be the smallest value of j such that $\sum_{\ell'=j}^{l} q_{t_{\ell'}} \le 2\log(NH)/N$, and likewise $\ell_{\max}$ be the largest value of ℓ' such that $\sum_{\ell'=l+1}^{\ell-1} q_{t_{\ell'}} \le 2\log(NH)/N$. Note that for each l, we can effectively truncate the sum over $j \in \{1, \cdots, l\}$ to $j \in \{j_{\min}, \cdots, l\}$ and the sum over $\ell \in \{l+1, \cdots, k\}$ to $\ell \in \{l+1, \cdots, \ell_{\max}\}$ at the cost of an additive $1/NH$ error. This is because for all terms involving a value of j or ℓ lying outside this range, $(1 - \sum_{\ell'=j}^{l} q_{t_{\ell'}} - \sum_{\ell'=l+1}^{\ell-1} q_{t_{\ell'}})^{N-1} \le 1/H^2N^2$ and the overall sum over ℓ, j for such terms evaluates to at most $1/N^2$. Therefore, we have,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t_j \in \mathcal{T}_i} q_{t_j}\left(t_{j+\Delta_j} - t_j\right)\right] &\le N \sum_{l=1}^{k-1} \theta_l \sum_{j=j_{\min}}^{l} q_{t_j} \sum_{\ell=l+1}^{\ell_{\max}} q_{t_\ell} \left(1 - \sum_{\ell'=j}^{l} q_{t_{\ell'}} - \sum_{\ell'=l+1}^{\ell-1} q_{t_{\ell'}}\right)^{N-1} + \frac{H}{N} \\
&\le N \sum_{l=1}^{k-1} \theta_l \sum_{j=j_{\min}}^{l} q_{t_j} \sum_{\ell=l+1}^{\ell_{\max}} q_{t_\ell} + \frac{H}{N} \\
&\overset{(i)}{\lesssim} \frac{\log^2(NH)}{N} \sum_{l=1}^{k-1} \theta_l + \frac{H}{N} \\
&\lesssim \frac{H \log^2(NH)}{N}.
\end{aligned}$$

where (i) uses the definition of $\ell_{\max}$ and $j_{\min}$ and the last inequality uses the fact that $\sum_{l=1}^{k-1} \theta_k = H - t_1 \le H$.
Combining with eq. (C.28) proves that the reference policy π_{ref} satisfies the condition that,

$$J_{\widetilde{r}}(\widetilde{\pi}) - J_{\widetilde{r}}(\pi_{\text{ref}}) \lesssim \frac{H \log^2(NH)}{N}. \tag{C.30}$$

Combining with Lemma 4.3.2 and noting that $\widetilde{\pi} \in \Pi_{\text{det}}^{\texttt{BC}}(D)$ and $\widetilde{r} \in \mathcal{R}_{\text{opt}}(D)$ are arbitrary completes the proof.

Arbitrary next-state distributions at state 3 and 4 Note that in the previous section, we assume that the states 3 and 4 are absorbing. It remains to prove Theorem 4.3.1 in the case where the next-state distributions under the action at this state follow an arbitrary distribution supported on the same two states. This can be proved with a small modification to the reference policy defined in eq. (C.23). As before, for any time $t \in \mathcal{T}_i$ for some i, with $\Delta \ge 0$ being the smallest value such that $t + \Delta \in \mathcal{T}_i$, and such that state 2 was observed in the dataset D at this time. Recall that a_- is defined as the action at state 2 which places maximum mass on state 3. Define $\widetilde{a}$ as the action at state 2 at time t which induces the distribution with maximum mass on the state 3 at time $t + \Delta$.

$$\widetilde{a} = \underset{a \in \{a_+, a_-\}}{\arg\max}\, P(s_{t+\Delta} = 3 | s_t = 2, a_t = a)$$

With this definition, we play the reference policy which picks the action with highest probability of visiting the state 3 if the expert was observed to pick the action a_- at time $t + \Delta$ (i.e. the action which also maximizes the probability of visiting the state 3). Namely,

$$\pi_{\text{ref}}(\cdot|2) = \begin{cases} \delta_{\widetilde{a}}, & \text{if } \pi^{\text{exp}}_{t+\Delta}(2) = a_- \\ 1 - \delta_{\widetilde{a}}, & \text{otherwise.} \end{cases} \tag{C.31}$$

When states 3 and 4 are absorbing, this reference policy matches that defined in eq. (C.23). Under this new reference policy, the previous proof largely carries through, with the exception of eq. (C.27), where the instead get the inequality,

$$\widetilde{V}_{t_j+1}(s) - \widetilde{V}_{t_j+1}(s') \geq -2^{i+1} - (t_{j+\Delta_j} - t_j) \tag{C.32}$$

where $\{s, s'\} = \{3, 4\}$ and s is the state having higher probability of reaching state 3 at time $t_j + \Delta_j$ under the MDP transition. This inequality follows by noting that the action at s and s' induce a mixture distribution on states 3 and 4 at time t (with s having the larger component on state 3, by definition). By another application of the data-processing inequality in Lemma B.1.1, the largest value gap at time t_j is achieved when the dynamics at states 3 and 4 are deterministic, at which point it is lower bounded by the value gap at time $t_j + \Delta_j + 1$ up to an additive error of 2^{i+1}.

Appendix D

Proof of main results in Chapter 5

Supplementary notation. Within this appendix, we will use $\lll$ and $\ggg$ to indicate the partial ordering on vectors where $a \lll b$ if a is not larger than b co-ordinate wise ($\ggg$ is defined analogously). We also use $\mathbf{0}$ to denote the all 0's vector and $\mathbf{1}$ denote the all 1's vector (where the dimension is inferred from context). For a vector $\mathbf{0} \lll w \in \mathbb{R}^d$, the norm $\|x\|_w^2 \triangleq \sum_{i=1}^d w_i x_i^2$ is the weighted-L_2 norm. For a function $g : \mathcal{X} \to \mathbb{R}$, the norm $\|g\|_\infty \triangleq \sup_{x \in \mathcal{X}} |g(x)|$.

D.1 Proofs of results invoked in Theorem 5.0.1

Recall from eq. (5.1), the population expected 0-1 loss of a policy π is defined as

$$\mathcal{L}_{0\text{-}1}(\pi; f^\pi) = \frac{1}{H} \sum_{t=1}^{H} \mathbb{E}_{S_t \sim f_t^\pi(\cdot)} \left[\mathbb{E}_{a \sim \widehat{\pi}_t(\cdot|s)} \left[\mathbb{1}(a \neq \pi_t^{\exp}(s)) \right] \right].$$

Lemma D.1.1. *Suppose there exists an online learning algorithm which outputs policies $\{\widehat{\pi}^1, \cdots, \widehat{\pi}^N\}$ sequentially, according to any procedure where the learner samples the policy $\widehat{\pi}_i$ from some distribution conditioned on $\mathsf{tr}_1, \cdots, \mathsf{tr}_{i-1}$, subsequently samples a trajectory tr_i by rolling out $\widehat{\pi}_i$, repeating this process for N iterations. Denote $\widehat{f}^i = \{\widehat{f}_1^i, \cdots, \widehat{f}_H^i\}$ where $\widehat{f}_t^i$ denotes the empirical distribution over states at time t induced by the single trajectory tr_i drawn from $\widehat{\pi}^i$. Denote $\widehat{\pi} = \frac{1}{N} \sum_{i=1}^N \widehat{\pi}^i$ as the mixture policy. Then,*

$$\mathbb{E}\left[\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f^{\widehat{\pi}})\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; \widehat{f}^i)\right].$$

Proof. Since the trajectory tr_i is rolled out using $\widehat{\pi}^i$, conditioned on $\widehat{\pi}^i$, $\widehat{f}^i$ is conditionally unbiased and in expectation equal to $f^{\widehat{\pi}^i}$ (conditioned on $\widehat{\pi}^i$). Therefore, for each i, since $\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f)$ is a linear function in f,

$$\mathbb{E}\left[\mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; \widehat{f}^i) \middle| \widehat{\pi}^i\right] = \mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; f^{\widehat{\pi}^i}).$$

Summing across $i = 1, \cdots, N$ and using the fact that $\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f^{\widehat{\pi}}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; f^{\widehat{\pi}^i})$ and taking expectation completes the proof. □

The conclusion of this lemma is that is it suffices to minimize the empirical 0-1 loss under the learner's own *one-trajectory* empirical state distribution $\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; \widehat{f^i})$. Note that for any policy π, the loss

$$\mathcal{L}_{0\text{-}1}(\pi; \widehat{f^i}) = \frac{1}{H}\sum_{t=1}^{H}\sum_{s\in\mathcal{S}} \left\langle \pi_t(\cdot|s), z_t^i(s) \right\rangle \tag{D.1}$$

where $z_t^i(s) = \left\{ \widehat{f_t^i}(s)(1 - \pi^{\exp}(a|s)) \right\}_{a\in\mathcal{A}} \in \Delta_{\mathcal{A}}^1$ is a linear function in the policy π. The constraint on the policy variable π is that for each $t \in [H]$ and $s \in \mathcal{S}$, $\pi_t(\cdot|s) \in \Delta_{\mathcal{A}}^1$.

Define the loss $\ell_{i,s,t}(\pi) = \sum_{s\in\mathcal{S}} \left\langle \pi_t(\cdot|s), z_t^i(s) \right\rangle$. Then the variable $\pi_t(\cdot|s)$ lies in the simplex $\Delta_{\mathcal{A}}^1$ and the vector $z_t^i(s)$ is co-ordinate wise ≥ 0 and ≤ 1.

To learn the sequence of policies returned by the learner, we use the normalized-EG algorithm of [84] which is also known as Follow-the-regularized-leader / Online Mirror Descent with entropy regularization for online learning. Formally, the online learning problem and the algorithm are as defined in Section 2 of [84].

Theorem D.1.2 (Adapted from Theorem 2.22 in [84]). *Assume that the normalized EG algorithm is run on a sequence of linear loss functions* $\{\langle z_i, \cdot\rangle : i = 1, \cdots, T\}$, *with* $\eta = 1/2$ *to return a sequence of distributions* $w_1, \cdots, w_T \in \Delta_{\mathcal{A}}^1$. *Assume that for all* $t \in [H]$, $\mathbf{0} \lll z_t \lll \mathbf{1}$. *For any* u *such that* $\sum_{i=1}^{T}\langle z_i, u\rangle = 0$,

$$\sum_{t=1}^{T}\langle w_i - u, z_i\rangle \leq 4\log(|\mathcal{A}|).$$

This result is adapted from Theorem 2.22 in [84] by invoking the condition that $\mathbf{0} \lll z_t \lll \mathbf{1}$, so the local norm $\|z_t\|_{w_t}^2$ can be upper bounded by $\langle z_t, w_t\rangle$. Choosing $\eta = \frac{1}{2}$, using the assumption that $\sum_{i=1}^{T}\langle z_i, u\rangle = 0$ and simplifying results in the statement of Theorem D.1.2. Suppose the learner returns the sequence of policies $\widehat{\pi}^i, \cdots, \widehat{\pi}^N$ by running the normalized EG algorithm on the sequence of losses $\ell_{1,s,t}, \cdots, \ell_{N,s,t}$ for each $s \in \mathcal{S}$ and $t \in [H]$ to return a sequence of distributions $\widehat{\pi}_t^1(\cdot|s), \cdots, \widehat{\pi}_t^N(\cdot|s) \in \Delta_{\mathcal{A}}^1$. Finally, for $i = 1, \cdots, N$, the learner returns the policy $\widehat{\pi}^i$ as $\{\{\widehat{\pi}_t^i(\cdot|s) : s \in \mathcal{S}\} : t \in [H]\}$.

Invoking the guarantee in Theorem D.1.2 for the sequence of policies $\widehat{\pi}_t^1(\cdot|s), \cdots, \widehat{\pi}_t^N(\cdot|s)$ returned by a single instance of the normalized-EG algorithm,

$$\sum_{i=1}^{T}\langle z_t^i(s), \widehat{\pi}_t^i(\cdot|s)\rangle \leq 4\log(|\mathcal{A}|)$$

Averaging across $t \in [H]$, summing across $s \in \mathcal{S}$ and recalling the definition of $\mathcal{L}_{0\text{-}1}(\pi; f)$ in eq. (D.1) results in the bound,

$$\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{0\text{-}1}(\widehat{\pi}^i; \widehat{f}^i) \leq \frac{4|\mathcal{S}|\log(|\mathcal{A}|)}{N}.$$

Finally invoking Lemma D.1.1 shows that the resulting sequence of policies $\widehat{\pi}^1, \cdots, \widehat{\pi}^N$ and their mixtures $\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}^i$ satisfies,

$$\mathbb{E}\left[\mathcal{L}_{0\text{-}1}(\widehat{\pi}; f^{\widehat{\pi}})\right] \leq \frac{4|\mathcal{S}|\log(|\mathcal{A}|)}{N}.$$

Invoking [78, Theorem 2], under μ-recoverability shows that the resulting policy $\widehat{\pi}$ satisfies,

$$\mathbb{E}[\mathsf{Gap}(\widehat{\pi})] \leq \frac{4|\mathcal{S}|\log(|\mathcal{A}|)}{N}.$$

This completes the proof of Theorem 5.0.1.

D.2 Proof of Theorem 5.0.2

For each active learner $\widehat{\pi}$ and the worst-case IL instance $(\pi^{\text{exp}}, \mathcal{M})$ considered in the lower bound in the active-interaction setting (Theorem 2.2.1), consider the IL instance $(\pi^{\text{exp}}, \mathcal{M}_\mu)$ where the only difference between $\mathcal{M}$ and $\mathcal{M}_\mu$ is that each reward is scaled by a factor of $\mu/H \leq 1$. Note that $\mathcal{M}_\mu$ satisfies μ-recoverability. Indeed, consider any state s. Since the rewards in $\mathcal{M}_\mu$ are in the interval $[0, \mu/H]$, the total reward of any trajectory in $\mathcal{M}_\mu$ lies in the interval $[0, \mu]$. Therefore, trivially, for each $(s, a, t) \in \mathcal{S}\times\mathcal{A}\times[H]$ tuple, $Q_t^{\pi^{\text{exp}}}(s, \pi_t^{\text{exp}}(s)) - Q_t^{\pi^{\text{exp}}}(s, a) \leq \mu - 0 = \mu$ and the IL instance satisfies μ-recoverability. More importantly the imitation gap of $\widehat{\pi}$ on the IL instance $(\pi^{\text{exp}}, \mathcal{M}_\mu)$ is $\frac{\mu}{H}$ times the imitation gap under $(\pi^{\text{exp}}, \mathcal{M})$. In other words,

$$\begin{aligned}
\mathbb{E}\left[J_{\mathcal{M}_\mu}(\pi^{\text{exp}}) - J_{\mathcal{M}_\mu}(\widehat{\pi})\right] &= \frac{\mu}{H}\mathbb{E}\left[J_{\mathcal{M}}(\pi^{\text{exp}}) - J_{\mathcal{M}}(\widehat{\pi})\right] \\
&\gtrsim \frac{\mu}{H}\min\left\{H, \frac{|\mathcal{S}|H^2}{N}\right\} \\
&= \min\left\{\mu, \frac{\mu|\mathcal{S}|H}{N}\right\},
\end{aligned}$$

where the last inequality uses [70, Theorem 6.1]. This concludes the proof of Theorem 5.0.2.

D.3 Proof of Theorem 5.0.3

Define $\mathcal{S}_t(D)$ as the set of states observed in at least one trajectory at time t in the demonstration dataset D. In particular, the learner exactly knows the expert's policy $\pi_t^{\text{exp}}(\cdot|s)$ at all states $s \in \mathcal{S}_t(D)$ for each $t = 1, \cdots, H$.

The expert policy is deterministic in the lower bound instances we construct. As originally defined in Eq. (2.3), define $\Pi^{\mathtt{BC}}_{\mathrm{det}}(D)$ as the family of policies which mimics the expert on the states visited in D. Namely,

$$\Pi^{\mathtt{BC}}_{\mathrm{det}}(D) \triangleq \left\{\pi : \forall t \in [H], s \in \mathcal{S}_t(D),\ \pi_t(\cdot|s) = \pi^{\mathrm{exp}}_t(\cdot|s)\right\},$$

$\Pi^{\mathtt{BC}}_{\mathrm{det}}(D)$ is the family of policies which are consistent with the demonstration dataset D.

In order to prove the lower bound on the worst-case expected imitation gap of any learner $\widehat{\pi}(D)$, it suffices to lower bound the Bayes expected imitation gap and find a joint distribution $\mathcal{P}$ over MDPs and expert policies satisfying μ-recoverability, such that,

$$\mathbb{E}_{(\pi^{\mathrm{exp}},\mathcal{M})\sim\mathcal{P}}\left[J_{\mathcal{M}}(\pi^{\mathrm{exp}}) - \mathbb{E}\left[J_{\mathcal{M}}(\widehat{\pi}(D))\right]\right] \gtrsim \min\left\{H, \frac{|\mathcal{S}|H^2}{N}\right\}.$$

Construction of $\mathcal{P}$. First the expert's policy is sampled uniformly from Π_{det}: for each $t \in [H]$ and $s \in \mathcal{S}$, $\pi^{\mathrm{exp}}_t(s) \sim \mathrm{Unif}(\mathcal{A})$. Conditioned on π^{exp}, the distribution over MDPs induced by $\mathcal{P}$ is deterministic and given by the MDP $\mathcal{M}[\pi^{\mathrm{exp}}]$ in fig. D.1. $\mathcal{M}[\pi^{\mathrm{exp}}]$ has a fixed initial distribution over states $\rho = \{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-2)\zeta, 0\}$ where $\zeta = \frac{1}{N+1}$. There is a special state $b \in \mathcal{S}$ in the MDP which has behavior different from the remaining states. At each state $s \in \mathcal{S}$, choosing the expert's action renews the state in the initial distribution ρ providing a reward of 1 (except at state b it provides a reward of 0), while every other action deterministically transitions the learner to the bad state and provides no reward. That is,

$$P_t(\cdot|s,a) = \begin{cases} \rho, & s \in \mathcal{S},\ a = \pi^{\mathrm{exp}}_t(s) \\ \delta_b, & \text{otherwise,} \end{cases}$$

and the reward function of the MDP is given by,

$$r_t(s,a) = \begin{cases} 1, & s \in \mathcal{S} \setminus \{b\},\ a = \pi^{\mathrm{exp}}_t(s) \\ 0, & \text{otherwise.} \end{cases}$$

We first state a simple consequence of the construction of the MDP instances and $\mathcal{P}$. Note that the expert never visits the bad state b by virtue of the distribution ρ placing no mass on b. Therefore, the value of π^{exp} on the MDP $\mathcal{M}[\pi^{\mathrm{exp}}]$ is H.

Lemma D.3.1. *The value of π^{exp} on the MDP $\mathcal{M}[\pi^{\mathrm{exp}}]$ is H. Namely $J_{\mathcal{M}[\pi^{\mathrm{exp}}]}(\pi^{\mathrm{exp}}) = H$.*

Proof. Playing the expert's action at any state in $\mathcal{S}$ is the only way to accrue non-zero reward, and in fact accrues a reward of 1. Thus the expert collects a reward of 1 at each time in any trajectory. □

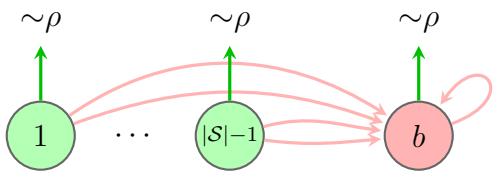


Figure D.1: Upon playing the expert's (green) action at any state, the learner is renewed in the initial distribution $\rho = \{\zeta, \cdots, \zeta, 1-(|\mathcal{S}|-2)\zeta, 0\}$ where $\zeta = \frac{1}{N+1}$. Any other choice of action (red) deterministically transitions the learner to b.

At the states unvisited in the dataset D, the learner cannot infer the expert's policy or even the transitions induced under different actions. Intuitively, the learner cannot guess the expert's action with probability $\geq 1/|\mathcal{A}|$ at such states, a statement which we prove by leveraging the Bayesian construction. In turn, the learner is forced to visit the bad state b at the next point in the episode. Since the bad state is never observed in the dataset, the learner is forced to guess the expert's action to be able to recover in the distribution ρ over the remaining states (lest it collects a reward of 0 for the rest of the episode). However by making $|\mathcal{A}|$ large ($\gtrsim H$), any learner, with constant probability fails to guess the expert's action at b at at least a constant fraction of the episode.

Using Lemma A.2.11 the conditional distribution of the expert's policy given the demonstration dataset D can be characterized, and is distributed $\sim \text{Unif}(\Pi^{\texttt{BC}}_{\text{det}}(D))$.

Definition D.3.1. *Define $\mathcal{P}(D)$ as the joint distribution of $(\pi^{\text{exp}}, \mathcal{M})$ conditioned on the demonstration dataset D. Conditionally, $\pi^{\text{exp}} \sim \text{Unif}(\Pi^{\texttt{BC}}_{\text{det}}(D))$ and $\mathcal{M} = \mathcal{M}[\pi^{\text{exp}}]$.*

From Lemma A.2.11 and the definition of $\mathcal{P}(D)$ in Definition D.3.1, applying Fubini's theorem gives,

$$\mathbb{E}_{(\pi^{\text{exp}},\mathcal{M})\sim\mathcal{P}}\Big[J_{\mathcal{M}}(\pi^{\text{exp}}) - \mathbb{E}\left[J_{\mathcal{M}}(\widehat{\pi})\right]\Big] = \mathbb{E}\left[\mathbb{E}_{(\pi^{\text{exp}},\mathcal{M})\sim\mathcal{P}(D)}\left[H - J_{\mathcal{M}}(\widehat{\pi}(D))\right]\right]. \tag{D.2}$$

Next we relate this to the first time the learner visits a state unobserved in D.

Lemma D.3.2. *Define the stopping time τ as the first time t that the learner encounters a state $s_t \neq b$ that has not been visited in D at time t. That is,*

$$\tau = \begin{cases} \inf\{t : s_t \notin \mathcal{S}_t(D) \cup \{b\}\} & \exists t : s_t \notin \mathcal{S}_t(D) \cup \{b\} \\ H+1 & \textit{otherwise.} \end{cases}$$

Then, conditioned on the demonstration dataset D,

$$\mathbb{E}_{(\pi^{\text{exp}},\mathcal{M})\sim\mathcal{P}(D)}\Big[J(\pi^{\text{exp}}) - \mathbb{E}\left[J(\widehat{\pi})\right]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)^{H+1} \mathbb{E}_{(\pi^{\text{exp}},\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}(D)}\left[H - \tau\right]\right].$$

We defer the proof of this result to the end of this section.

Finishing proof of Theorem 5.0.3. Plugging the result of Lemma D.3.2 into eq. (D.2), and recalling the assumption that $|\mathcal{A}| \geq H+1$,

$$\begin{aligned}\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}}\left[J(\pi^{\exp}) - \mathbb{E}\left[J(\widehat{\pi})\right]\right] &\geq \frac{1}{4}\mathbb{E}\left[\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[H-\tau\right]\right]\right], \\ &\overset{(i)}{\geq} \frac{H}{8}\mathbb{E}\left[\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right], \\ &= \frac{H}{8}\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right],\end{aligned}$$

where (i) uses Markov's inequality, and the last equation uses Fubini's theorem.
The last remaining element of he proof is to indeed bound the probability that the learner visits a state unobserved in the dataset before time $\lfloor H/2 \rfloor$ which immediately follows from Lemma A.2.13 shows that for any learner $\widehat{\pi}$, $\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}}\left[\tau \leq \lfloor H/2 \rfloor\right]\right]\right]$ is lower bounded by $\gtrsim \min\{1, |\mathcal{S}|H/N\}$. Therefore,

$$\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}}\left[J(\pi^{\exp}) - \mathbb{E}\left[J(\widehat{\pi})\right]\right] \gtrsim H \min\left\{1, \frac{|\mathcal{S}|H}{N}\right\}.$$

as long as $|\mathcal{A}| \geq H+1$. This completes the proof.

Finally, we prove Lemma D.3.2.

Proof of Lemma D.3.2. Define the random time τ_b to be the first time the learner encounters the state b while rolling out a trajectory. Formally,

$$\tau_b = \begin{cases}\inf\{t : s_t = b\} & \exists t : s_t = b \\ H+1 & \text{otherwise.}\end{cases}$$

Furthermore, define Γ_b as the random variable which counts the number of time steps the trajectory stays in the state b after visiting it for the first time. Namely,

$$\Gamma_b = \begin{cases}\inf\{\Delta \geq 0 : s_{\tau_b+\Delta+1} \neq b\} & \tau_b \leq H \\ 0 & \text{otherwise.}\end{cases}$$

Since the state b always dispenses 0 reward and since r is bounded in $[0,1]$, conditioned on the demonstration dataset D,

$$\begin{aligned}H - \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[J(\widehat{\pi})\right] &= H - \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\right]\right] \\ &\geq \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[\Gamma_b\right]\right]\end{aligned}$$

Fixing the demonstration dataset D and the expert's policy $\pi^{\exp}$ (which determines the MDP $\mathcal{M}[\pi^{\exp}]$), we under the distribution of Γ_b.

To this end, first observe that for any $t \le H-1$ and state $s \in \mathcal{S}$,

$$\begin{aligned}
&\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta+1, \Gamma_b \ge \Delta, \tau_b = t\right] \\
&= \Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta+1 | \Gamma_b \ge \Delta, \tau_b = t\right] \Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta, \tau_b = t\right] \\
&= \left(1 - \widehat{\pi}_{t+\Delta}(\pi^{\exp}_{t+\Delta}(b)|b)\right) \Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta, \tau_b = t\right].
\end{aligned}$$

where in the last equation, we use the fact that the learner must play an action other than $\pi^{\exp}_{t+\Delta}(b)$ to stay in state b at time $t+\Delta$. Next, we take expectation with respect to the randomness of $\pi^{\exp}$. Conditioned on D, $\pi^{\exp}$ is sampled uniformly from the set of policies $\Pi^{\texttt{BC}}_{\text{det}}(D)$ (Lemma A.2.11). In particular, conditioned on D, the expert policy is sampled independently at states. Conditioned on $\pi^{\exp}$, the underlying MDP is $\mathcal{M}[\pi^{\exp}]$. Observe that the dependence of the second term $\Pr_{\widehat{\pi}}[\tau = t, s_t = s]$ on $\pi^{\exp}$ comes from the probability computed with the underlying MDP chosen as $\mathcal{M}[\pi^{\exp}]$. Observe that it only depends on the characteristics of $\mathcal{M}[\pi^{\exp}]$ till time $t-1$ which are determined by $\pi^{\exp}_1, \cdots, \pi^{\exp}_{t+\Delta-1}$. On the other hand, the first term $\left(1 - \widehat{\pi}_t(\pi^{\exp}_{t+\Delta}(b)|b)\right)$ depends only on the random variable $\pi^{\exp}_{t+\Delta}$. As a consequence, the two terms depend on a disjoint set of random variables which are independent.

Taking expectation with respect to the randomness of $\pi^{\exp} \sim \text{Unif}(\Pi^{\texttt{BC}}_{\text{det}}(D))$ and $\mathcal{M} = \mathcal{M}[\pi^{\exp}]$ (together which define the joint distribution $\mathcal{P}(D)$), for $0 \le \Delta \le H - t$ and $t \in [H]$,

$$\begin{aligned}
&\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta+1, \Gamma_b \ge \Delta, \tau_b = t\right]\right] \\
&= \mathbb{1}(t+\Delta \le H) \cdot \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[1 - \widehat{\pi}_{t+\Delta}(\pi^{\exp}_{t+\Delta}(b)|b)\right] \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta, \tau_b = t\right]\right] \\
&= \mathbb{1}(t+\Delta \le H) \cdot \left(1 - \frac{1}{|\mathcal{A}|}\right) \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta, \tau_b = t\right]\right],
\end{aligned}$$

where in the last equation we use the fact that the state b is never observed in the demonstration dataset. So conditioned on D, $\pi^{\exp}_{t+\Delta}(b)$ is sampled uniformly from $\mathcal{A}$. By upper bounding $\Pr_{\widehat{\pi}}[\Gamma_b \ge \Delta+1, \Gamma_b \ge \Delta, \tau_b = t] \le \Pr_{\widehat{\pi}}[\Gamma_b \ge \Delta+1, \tau_b = t]$ results in the inequality,

$$\begin{aligned}
&\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta+1, \tau_b = t\right]\right] \\
&\ge \mathbb{1}(t+\Delta \le H) \cdot \left(1 - \frac{1}{|\mathcal{A}|}\right) \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta, \tau_b = t\right]\right]
\end{aligned}$$

Unrolling the equation, for each $\Delta = 0, 1, \cdots, H-t+1$ we have,

$$\begin{aligned}
&\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\Gamma_b \ge \Delta, \tau_b = t\right]\right] \\
&\ge \left(1 - \frac{1}{|\mathcal{A}|}\right)^{\Delta} \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}}\left[\tau_b = t\right]\right]
\end{aligned}$$

Summing up over $\Delta = 0, 1, \cdots, H - t + 1$,

$$\begin{aligned}
&\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\Big[\mathbb{E}_{\widehat{\pi}}\left[\Gamma_b \mathbb{1}(\tau_b = t)\right]\Big] \\
&\geq \sum_{\Delta=0}^{H-t+1}\left(1 - \frac{1}{|\mathcal{A}|}\right)^{\Delta} \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\Big[\Pr_{\widehat{\pi}}\left[\tau_b = t\right]\Big] \\
&\geq (H - t + 1)\left(1 - \frac{1}{|\mathcal{A}|}\right)^{H} \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\Big[\Pr_{\widehat{\pi}}\left[\tau_b = t\right]\Big].
\end{aligned}$$

Finally summing across $t = 1, \cdots, H + 1$,

$$\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\Big[\mathbb{E}_{\widehat{\pi}}\left[\Gamma_b\right]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)^{H} \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[H - \tau_b + 1\right]\right].$$

Finally, we invoke the same inequality used in the proof of Lemma A.2.12 (specifically, eq. (A.120)) to arrive at the desired bound,

$$\mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\Big[\mathbb{E}_{\widehat{\pi}}\left[\Gamma_b\right]\Big] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)^{H+1} \mathbb{E}_{(\pi^{\exp},\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[H - \tau\right]\right]$$

Note that although the MDP family considered in Lemma A.2.12 is different, until the state b is visited the two MDPs are identical and therefore τ and τ_b are distributed identically under either MDP family for the same policy. □

Appendix E

Proofs of results in Chapter 6

E.1 Imitation gap of Empirical Moment Matching

Below we state an upper bound on the imitation gap of MM and provide a proof of this result.

Definition E.1.1 (Instantiation of $\mathcal{F}$ in MM)**.** *In the tabular setting, we instantiate the discriminator class as* $\mathcal{F}_t = \{f_t : \|f_t\|_\infty \leq 1\}$ *for each* t*, as the set of all* 1*-bounded functions, and the policy class* Π *as the set of all tabular policies. eq.* (6.2) *corresponds to finding a policy which best matches the empirical state-action visitation measure observed in the dataset* D *in total variation (TV) distance.*

Theorem E.1.1. *The policy* $\pi^{\texttt{MM}}$ *returned by empirical moment matching (Definition 6.0.1) satisfies the following upper bound on its imitation gap in the tabular setting,*

$$\mathbb{E}[\mathit{Gap}(\pi^{\texttt{MM}})] \lesssim H\sqrt{\frac{|\mathcal{S}|}{N}}.$$

The key observation is that since the learner $\pi^{\texttt{MM}}$ best matches the empirical distribution in the dataset, which is in turn close to the population visitation measure induced by π^{exp}, we can expect the visitation measure induced by π^{exp} and $\pi^{\texttt{MM}}$ to be close. This in turns implies that both policies will collect a similar value under any reward function. Precisely characterizing the rates at which these distributions converge to one another results in the final bound.

Proof. Recall that the learner $\pi^{\texttt{MM}}$ is the solution to the following optimization problem,

$$\arg\min_{\pi} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_\pi \left[\frac{\sum_{t=1}^H f_t(s_t, a_t)}{H} \right] - \mathbb{E}_D \left[\frac{\sum_{t=1}^H f_t(s_t, a_t)}{H} \right] \right\}$$

Exchanging the summation and maximization operators and recalling from Definition E.1.1 that in the tabular setting, the discriminator class $\mathcal{F}$ is instantiated as the set of all 1-bounded

functions $\bigoplus_{t=1}^{H}\{f_t : \|f_t\|_\infty \leq 1\}$, $\pi^{\texttt{MM}}$ is a solution to

$$\arg\min_{\pi} \frac{1}{H}\sum_{t=1}^{H}\left(\sup_{f:\|f\|_\infty\leq 1} \mathbb{E}_{\pi}\left[f_t(s_t,a_t)\right] - \mathbb{E}_{D}\left[f_t(s_t,a_t)\right]\right) = \arg\min_{\pi}\frac{1}{H}\sum_{t=1}^{H} D_{\text{TV}}\left(d_t^{\pi}, d_t^{D}\right) \quad \text{(E.1)}$$

where the equation follows by the variational definition of the total variation distance, and where d_t^{π} is the state-action visitation measure induced by π^{exp} and d_t^{D} is the empirical state-action visitation measure in the dataset D. The imitation gap of this policy can be upper bounded by,

$$\begin{aligned}
\mathsf{Gap}(\pi^{\texttt{MM}}) &= \mathbb{E}_{\pi^{\text{exp}}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\right] - \mathbb{E}_{\pi^{\texttt{MM}}}\left[\sum_{t=1}^{H} r_t(s_t,a_t)\right] \\
&\overset{(i)}{\leq} \sum_{t=1}^{H}\sup_{r_t:\|r_t\|_\infty\leq 1}\left(\mathbb{E}_{\pi^{\text{exp}}}\left[r_t(s_t,a_t)\right] - \mathbb{E}_{\pi^{\texttt{MM}}}\left[r_t(s_t,a_t)\right]\right) \\
&\overset{(ii)}{=} \sum_{t=1}^{H} D_{\text{TV}}\left(d_t^{\pi^{\text{exp}}}(\cdot,\cdot), d_t^{\pi^{\texttt{MM}}}(\cdot,\cdot)\right)
\end{aligned}$$

where (i) maximizes over the reward function which is assumed to lie in the interval $[0,1]$ pointwise. (ii) again follows from the variational definition of total variation distance. This goes to show that in the tabular setting, MM is equivalent to finding the policy which best matches (in TV-distance) the empirical state-action distribution observed in the dataset. By an application of triangle inequality,

$$\begin{aligned}
\mathsf{Gap}(\pi^{\texttt{MM}}) &\leq \sum_{t=1}^{H} D_{\text{TV}}\left(d_t^{\pi^{\text{exp}}}(\cdot,\cdot), d_t^{D}(\cdot,\cdot)\right) + D_{\text{TV}}\left(d_t^{D}(\cdot,\cdot), d_t^{\pi^{\texttt{MM}}}(\cdot,\cdot)\right) \\
&\leq 2\sum_{t=1}^{H} D_{\text{TV}}\left(d_t^{\pi^{\text{exp}}}(\cdot,\cdot), d_t^{D}(\cdot,\cdot)\right) \quad \text{(E.2)}
\end{aligned}$$

where (i) follows from eq. (E.1) which shows that $\pi^{\texttt{MM}}$ is the policy which best approximates the empirical state-action visitation measure in total variation distance, and therefore $D_{\text{TV}}\left(d_t^{\pi^{\texttt{MM}}}(\cdot,\cdot), d_t^{D}(\cdot,\cdot)\right) \leq D_{\text{TV}}\left(d_t^{\pi^{\text{exp}}}(\cdot,\cdot), d_t^{D}(\cdot,\cdot)\right)$. The final element is to identify the rate of convergence of the empirical visitation measure d_t^{D}, to the population $d_t^{\pi^{\text{exp}}}$ in total variation distance. This result is known from Theorem 1 of [42], which shows that,

$$\mathbb{E}\left[D_{\text{TV}}\left(d_t^{\pi^{\text{exp}}}(\cdot,\cdot), d_t^{D}(\cdot,\cdot)\right)\right] \lesssim \sqrt{\frac{|\mathcal{S}|}{N}},$$

noting that $d_t^{\pi^{\text{exp}}}$ is a distribution with support size $|\mathcal{S}|$ since π^{exp} is deterministic. Putting it together with eq. (E.2) after taking expectations on both sides gives,

$$\mathsf{Gap}(\pi^{\texttt{MM}}) \lesssim \sum_{t=1}^{H}\sqrt{\frac{|\mathcal{S}|}{N}} = H\sqrt{\frac{|\mathcal{S}|}{N}}.$$

This completes the proof of the result. □

E.2 Lower bounding the imitation gap of MM

In this section, we show that in the tabular setting, empirical moment matching is suboptimal compared for Imitation Learning in the worst-case. The main result we prove in this section is Theorem 6.1.1.

First note that the learner $\pi^{\texttt{MM}}$ carries out empirical moment matching (eq. (6.2)), with the discriminator class $\mathcal{F}$ as initialized in Definition E.1.1. As shown in eq. (E.1), the empirical moment matching learner can be redefined as the solution to a distribution matching problem,

$$\arg\min_{\pi} \frac{1}{H} \sum_{t=1}^{H} D_{\text{TV}} \left(d_t^{\pi}(\cdot, \cdot), d_t^{D}(\cdot, \cdot) \right) \tag{E.3}$$

Consider an MDP instance with 2 states and 2 actions with a non-stationary transition and reward structure as described in Figure fig. E.1. State 1 effectively has a single action (i.e. two actions, a_1 and a_2 with both inducing the same next-state distribution and reward). One of the actions at state 2 induces the uniform distribution over next states. The other action deterministically keeps the learner at state 2. The reward function is 0 at $t = 1$, and the action a_2 at state 2 is the only one which offers a reward of 1. The initial state distribution is highly skewed toward the state $s = 1$ and places approximately $1/\sqrt{N}$ mass on $s = 2$ and the remaining on $s = 1$.

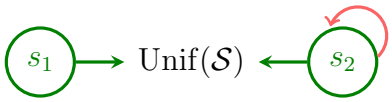


Figure E.1: MDP instance which shows that L_1 distribution matching is suboptimal. Here the transition structure is illustrated for $t = 1$. Both states have one action which reinitializes in the uniform distribution. State 2 has an additional action which keeps the state the same. The reward function is 0 for $t = 1$. For $t \geq 2$ the transition function is absorbing at both states; the reward function equals 1 at the state $s = 1$ for any action and is 0 everywhere else.

MDP transition. The state 2 is the only one with two actions. Action a_1 induces the uniform distribution over states, while action a_2 transitions the learner to state 2 with probability 1. Namely,

$$\begin{aligned} P_1(\cdot|s = 1, a) &= \text{Unif}(\mathcal{S}) \text{ for all } a \in \mathcal{A} \\ P_1(\cdot|s = 2, a_1) &= \text{Unif}(\mathcal{S}) \\ P_1(\cdot|s = 2, a_2) &= \delta_2 \end{aligned}$$

From time $t = 2$ onward, the actions are all absorbing. Namely, for all $t \geq 2$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$P_t(\cdot|s, a) = \delta_s.$$

Initial state distribution. The initial state distribution $\rho = \left(1 - \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}\right)$.

MDP reward function. The reward function of the MDP encourages the learner to stay at the state $s = 1$ from time $t = 2$ onward. Namely,

$$r_t(s, a) = \begin{cases} 1, & \text{if } t \geq 2 \text{ and } s = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{E.4}$$

Expert policy. At both states in the MDP, the expert picks the action a_1 to play, which induces the uniform distribution over actions at the next state. Namely, for each $t \in [H]$ and $s \in \mathcal{S}$,

$$\pi_t^{\text{exp}}(\cdot|s) = \delta_{a_1}$$

The intuition behind the lower bound is as follows. The only action which affects the value of a policy is the choice made at $s = 2$ at time $t = 1$. At all other states, we may assume that there is effectively only a single action.
By the absorbing nature of states for $t \geq 2$, it turns out that if the observed empirical distribution in the dataset at time 2 is skewed toward state 2 (which is possible because of the inherent randomness in the data generation process), the learner's behavior at time 1 may be to *ignore* the expert's action observed at state $s = 2$, and instead pick the action a_2 which moves the learner to the state $s = 2$ deterministically. The learner is willing to choose a different action because the loss function eq. (E.3) encourages the state-action distribution at time $t = 2$ also to be well matched with what is observed in the dataset. Even if it comes at the cost of picking an action different from what the expert plays. By exploiting this fact, we are able to show that the error incurred by a learner which solves eq. (E.3) in this simple 2 state example must be $\Omega(H/\sqrt{N})$.
Formally, we define 3 events,

1. $\mathcal{E}_1$: All states in the MDP are visited in the dataset at each time $t = 1, 2, \cdots, H$.
2. $\mathcal{E}_2$: State 2 is visited at most $\sqrt{N}$ times at time 1 in the dataset D. In other words, $d_1^D(s = 2) = \delta'$ where $\delta' \leq \frac{1}{\sqrt{N}}$.
3. $\mathcal{E}_3$: At time 2 in the dataset D, the empirical distribution over states is of the form $\left(\frac{1}{2} - \delta, \frac{1}{2} + \delta\right)$ for some $\delta \geq \frac{2}{\sqrt{N}}$.

Lemma E.2.1. *Jointly, the events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ occur with at least constant probability.*

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq C,$$

for some constant $C > 0$.

Proof. By the absorbing nature of states for $t \geq 2$, it suffices for both states of the MDP to be visited in the dataset at time $t = 1, 2$. At time $t = 2$, the marginal state distribution under $\pi^{\exp}$ is the uniform distribution. By binomial concentration, both states are observed in the dataset at time $t = 2$ with probability $\geq 1 - e^{-C_1 N}$ for some constant $C_1 > 0$. On the other hand, at time $t = 1$, the marginal state distribution is $\rho = \left(1 - \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}\right)$. Yet again, by binomial concentration, both states are observed with probability $\geq 1 - e^{-C_2\sqrt{N}}$ for some constant $C_2 > 0$. By union bounding,

$$\Pr(\mathcal{E}_1) \geq 1 - e^{-C_1 N} - e^{-C_2\sqrt{N}}.$$

Next we study $\mathcal{E}_2$ and $\mathcal{E}_3$ together. First of all, note that the state observed at $t = 1$ and $t = 2$ in a rollout of the expert policy are independent. This is because at both states at $t = 1$, the next state distribution under $\pi^{\exp}$ is uniform. Because of this fact, $\mathcal{E}_2$ and $\mathcal{E}_3$ are independent. Next we individually bound the probability of the two events.

$\mathcal{E}_2$: The number of times $s = 2$ is the initial state in trajectories the dataset D is distributed as a binomial random variable with distribution $\mathrm{Bin}(N, q)$ with $q = \rho(s = 2) = \frac{1}{\sqrt{N}}$. A median of a binomial random variable is $Nq = \sqrt{N}$ (in fact any number in the interval $[\lfloor Nq \rfloor, \lceil Nq \rceil]$ is a median). Therefore, the probability that $s = 2$ is visited $\leq \sqrt{N}$ times in the dataset at time 1 is at least $1/2$. In summary,

$$\Pr(\mathcal{E}_2) \geq \frac{1}{2}$$

$\mathcal{E}_3$: The marginal distribution over states at time 2 in the dataset is uniform. Therefore, we expect the states 1 and 2 to be visited roughly $N/2$ times each in the dataset, but with a random variation of $\approx \sqrt{N}$ around this average. In other words, the empirical distribution fluctuates as $\left(\frac{1}{2} - \delta, \frac{1}{2} + \delta\right)$ with $\delta \geq \frac{2}{\sqrt{N}}$ with constant probability.

By the independence of $\mathcal{E}_2$ and $\mathcal{E}_3$ and union bounding to account for $\mathcal{E}_1$, the statement of the lemma follows. □

Lemma E.2.2. *For each $t \geq 2$,*

$$D_{\mathrm{TV}}\left(d_t^{\pi}(\cdot,\cdot), d_t^{D}(\cdot,\cdot)\right) \geq D_{\mathrm{TV}}\left(d_2^{\pi}(\cdot), d_2^{D}(\cdot)\right) \tag{E.5}$$

The RHS is the TV distance between the state-visitation measure at time $t = 2$ under π and that empirically observed in the dataset D. Conditioned on the events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ occuring, equality is met in eq. (E.5) if any only if $\pi_t(\cdot|s) = \pi_t^{\exp}(\cdot|s)$ for all states $s \in \mathcal{S}$.

Proof. For any state $s \in \mathcal{S}$ and $t \geq 2$, observe that,

$$\begin{aligned}
&\sum_{a \in \mathcal{A}} \left| d_t^{\pi}(s,a) - d_t^{D}(s,a) \right| \\
&= d_t^{\pi}(s)(1 - \pi_t(a^\star|s)) + \left| d_t^{\pi}(s)\pi_t(a^\star|s) - d_t^{D}(s,a^\star) \right|, \text{ where } a^\star = \pi_t^{\exp}(s), \\
&\overset{(i)}{=} d_2^{\pi}(s)(1 - \pi_t(a^\star|s)) + \left| d_2^{\pi}(s)\pi_t(a^\star|s) - d_2^{D}(s,a^\star) \right| \\
&\overset{(ii)}{\geq} \left| d_2^{\pi}(s) - d_2^{D}(s) \right|,
\end{aligned}$$

where (i) follows by the fact that the states of the MDP are absorbing under π for $t \geq 2$. (ii) follows by triangle inequality and using the fact that π^{exp} is deterministic, so $d_t^D(s, a^\star) = d_t^D(s)$. Equality is met only if $\pi_t(a^\star|s) = 1$ (since $d_t^D(s, a^\star) > 0$ conditioned on $\mathcal{E}_1$). □

The above lemma asserts the behavior of $\pi^{\texttt{MM}}$ in eq. (E.3) for $t \geq 2$. Namely, conditioned on the event $\mathcal{E}_1$ which happens with very high probability, all states are visited in the MDP and therefore, $\pi_t^{\texttt{MM}}(\cdot|s) = \pi_t^{\text{exp}}(\cdot|s)$ for each state $s \in \mathcal{S}$ and time $t \geq 2$.

The only thing left to study is the MM learner's behavior at $t = 1$. We wish to show that with constant probability, the learner may choose to deviate from the expert policy in order to better match empirical state-action visitation measures. Conditioned on $\mathcal{E}_1$, the learner's policy at time $t = 1$ can be computed by solving the following optimization problem,

$$D_{\text{TV}}\left((,d)_1^\pi(\cdot,\cdot), d_1^D(\cdot,\cdot)\right) + (H-1)D_{\text{TV}}\left(d_2^\pi(\cdot), d_2^D(\cdot)\right).$$

This follows directly by simplifying the learner's objective using Lemma E.2.2.
Now, conditioned on the event $\mathcal{E}_1$, at time $t = 1$, the learner policy only needs to be optimized at the state $s = 2$. At the state $s = 1$, we may assume that the learner picks the expert's action $\pi_1^{\text{exp}}(s = 1)$. To this end, suppose the learner picks the action a_1 with probability p and the action a_2 with probability $1 - p$.

$$\begin{aligned} D_{\text{TV}}\left(d_1^\pi(\cdot,\cdot), d_1^D(\cdot,\cdot)\right) &= \sum_{a\in\mathcal{A}} \left|d_1^{\pi^{\text{exp}}}(s=2,a) - d_1^D(s=2,a)\right| \\ &= |\rho(2)p - \delta'| + |\rho(2)(1-p) - 0| \\ &= \left|\frac{p}{\sqrt{N}} - \delta'\right| + \frac{1-p}{\sqrt{N}}. \end{aligned} \tag{E.6}$$

which follows by plugging in $\rho(2) = \frac{1}{\sqrt{N}}$. And,

$$D_{\text{TV}}\left(d_2^\pi(\cdot), d_2^D(\cdot)\right) = \left|\left(\frac{1}{2} - \delta\right) - \frac{\rho(1)}{2} - \rho(2)\frac{p}{2}\right| + \left|\left(\frac{1}{2} + \delta\right) - \frac{\rho(1)}{2} - \rho(2)\left((1-p) + \frac{p}{2}\right)\right|. \tag{E.7}$$

Plugging in $\rho(2) = \frac{1}{\sqrt{N}}$ and $\rho(1) = 1 - \frac{1}{\sqrt{N}}$, we get,

$$D_{\text{TV}}\left(d_2^\pi(\cdot), d_2^D(\cdot)\right) = \left|\frac{1}{2\sqrt{N}} - \delta - \frac{p}{2\sqrt{N}}\right| + \left|\frac{p}{2\sqrt{N}} - \frac{1}{2\sqrt{N}} + \delta\right|. \tag{E.8}$$

Summing up eqs. (E.6) and (E.8), p minimizes,

$$\underbrace{\left|\frac{p}{\sqrt{N}} - \delta'\right| + \frac{1-p}{\sqrt{N}}}_{(i)} + \underbrace{(H-1)\left(\left|\frac{p}{2\sqrt{N}} + \delta - \frac{1}{2\sqrt{N}}\right| + \left|\frac{1}{2\sqrt{N}} - \delta - \frac{p}{2\sqrt{N}}\right|\right)}_{(ii)}. \tag{E.9}$$

Intuitively, term (i) captures the error incurred by the learner in the loss eq. (E.3) by deviating from π^{exp} at the first time step. Term (ii) captures the decrease in the error at every subsequent time step because of the same deviation, since the learner is able to better match the state distribution at future time steps. In the next lemma we show that under events that hold with at least constant probability, the empirical moment matching learner chooses to play the wrong action at time $t = 1$ at the state $s = 2$.

Lemma E.2.3. *Conditioned on the events $\mathcal{E}_2$ and $\mathcal{E}_3$, for $H \geq 4$, the unique minimizer of eq. (E.9) for $p \in [0, 1]$ is $p = 0$.*

Proof. The first term of eq. (E.9) is $\left|p/\sqrt{N} - \delta'\right| + (1-p)/\sqrt{N}$, the error from not picking the expert's action at state 1 at time 1 decreases at most linearly with a slope of $2/\sqrt{N}$. Conditioned on the event $\mathcal{E}_3$, $\delta \geq 2/\sqrt{N}$. Therefore, $\left|p/2\sqrt{N} + \delta - 1/2\sqrt{N}\right| = p/2\sqrt{N} + \delta - 1/2\sqrt{N}$. Therefore, the decrease in error at future steps by deviating from π^{exp} at the time $t = 1$, term (ii) in eq. (E.9) is,

$$2(H-1)\left(\frac{p}{2\sqrt{N}} + \delta - \frac{1}{2\sqrt{N}}\right) \tag{E.10}$$

which is an increasing function of p with slope $\frac{H-1}{\sqrt{N}}$. For $H \geq 4$ and the argument from the previous paragraph, this implies that term (ii) increases more rapidly in p than the rate at which term (i) decreases. Therefore, the minimizer must be $p = 0$. □

Thus, we conclude from Lemmas E.2.2 and E.2.3 that conditioned on the events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$, the learner $\pi^{\texttt{MM}}$ perfectly mimics π^{exp} at each time $t \geq 2$, but deviates from the action played by π^{exp} at the state $s = 1$ at time $t = 1$. Finally, we bound the difference in value between π^{exp} and $\pi^{\texttt{MM}}$ induced because of this deviation under the reward eq. (E.4).

Lemma E.2.4. *Under the events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$, under the reward eq. (E.4), the empirical moment matching learner $\pi^{\texttt{MM}}$ incurs imitation gap,*

$$\mathsf{Gap}(\pi^{\texttt{MM}}) = \frac{H}{2\sqrt{N}}.$$

Proof. Recall that under the events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$, the learner $\pi^{\texttt{MM}}$ is identical to π^{exp} except at the state $s = 2$ where they perfectly deviate from each other. The state distribution induced by π^{exp} at each time $t \geq 2$ is the uniform distribution over states $\left(\frac{1}{2}, \frac{1}{2}\right)$. On the other hand, for $t \geq 2$, the state distribution induced by $\pi^{\texttt{MM}}$ at each time $t \geq 2$ is $\left(\rho(1)\frac{1}{2}, \rho(1)\frac{1}{2} + \rho(2)\right) = \left(\frac{1-1/\sqrt{N}}{2}, \frac{1+1/\sqrt{N}}{2}\right)$. Since the reward function is 1 on state 1, the difference in value between the expert and learner policy is,

$$\mathsf{Gap}(\pi^{\texttt{MM}}) = \frac{H}{2} - H\left(\frac{1-1/\sqrt{N}}{2}\right) = \frac{H}{2\sqrt{N}}.$$

This completes the proof. □

Since $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ jointly occur with constant probability by Lemma E.2.1, this completes the proof of Theorem 6.1.1.

E.3 Imitation gap of RE: Proof of Theorem 6.2.1

In this section, we discuss a proof of a more general version of Theorem 6.2.1, where N_{replay} can be finite. We prove the following result,

Theorem E.3.1. *Consider the policy $\pi^{\texttt{RE}}$ returned by Algorithm 6. Assume that $\pi^{\text{exp}} \in \Pi$ and the ground truth reward function $r_t \in \mathcal{F}_t$, which is assumed to be symmetric ($f_t \in \mathcal{F}_t \iff -f_t \in \mathcal{F}_t$) and bounded (For all $f_t \in \mathcal{F}_t$, $\|f_t\|_\infty \le 1$). Choose $|D_1|, |D_2| = \Theta(N)$. With probability $\ge 1 - 3\delta$,*

$$\textsf{Gap}(\pi^{\texttt{RE}}) \lesssim \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \frac{\log\left(F_{\max} H/\delta\right)}{N}$$

where $F_{\max} \triangleq \max_{t\in[H]} |\mathcal{F}_t|$, and,

$$\mathcal{L}_1 \triangleq H^2\, \mathbb{E}_{\pi^{\text{exp}}}\left[\frac{\sum_{t=1}^H \texttt{MEM}(s_t,t) D_{\text{TV}}\left(\pi_t^{\text{exp}}(\cdot|s_t), \pi_t^{\texttt{BC}}(\cdot|s_t)\right)}{H}\right],$$

$$\mathcal{L}_2 \triangleq H^{3/2}\sqrt{\frac{\log\left(F_{\max}H/\delta\right)}{N}\frac{\sum_{t=1}^H \mathbb{E}_{\pi^{\text{exp}}}\left[1 - \texttt{MEM}(s_t,t)\right]}{H}},$$

And,

$$\mathcal{L}_3 \triangleq H\sqrt{\frac{\log(F_{\max}H/\delta)}{N_{\text{replay}}}} + \frac{H\log(F_{\max}H/\delta)}{N_{\text{replay}}}.$$

Recall that the learner carrying out replay estimation returns the policy which minimizes the loss $\sup_{f\in\mathcal{F}} J_f(\pi) - \widehat{\mathbb{E}}(f)$ over policies π, where $J_f(\pi) \triangleq \mathbb{E}_\pi\left[\frac{1}{H}\sum_{t=1}^H f_t(s_t,a_t)\right]$ where $f = (f_1, \cdots, f_H)$. Note that,

$$\begin{aligned}
\textsf{Gap}(\pi^{\texttt{RE}}) &\overset{(i)}{\le} \sup_{f\in\mathcal{F}} J_f(\pi^{\text{exp}}) - J_f(\pi^{\texttt{RE}}) \\
&\le \sup_{f\in\mathcal{F}}\left|J_f(\pi^{\text{exp}}) - \widehat{\mathbb{E}}(f)\right| + \sup_{f\in\mathcal{F}}\left|\widehat{\mathbb{E}}(f) - J_f(\pi^{\texttt{RE}})\right| \\
&\overset{(ii)}{\le} 2\sup_{f\in\mathcal{F}}\left|J_f(\pi^{\text{exp}}) - \widehat{\mathbb{E}}(f)\right|. \qquad \text{(E.11)}
\end{aligned}$$

where (i) uses the realizability assumption that the ground truth reward lies in $\mathcal{F}$, and (ii) uses the fact that $\pi^{\texttt{RE}}$ is a minimizer of eq. (6.4) and the fact that $\mathcal{F}$ is symmetric (this implies that $\sup_{f\in\mathcal{F}} J_f(\pi^{\text{exp}}) - \widehat{\mathbb{E}}(f) = \sup_{f\in\mathcal{F}}\left|J_f(\pi^{\text{exp}}) - \widehat{\mathbb{E}}(f)\right|$).

Note that $\widehat{\mathbb{E}}(f)$ can be decomposed into a sum of two parts,

$$\widehat{\mathbb{E}}^{(1)}(f) = \mathbb{E}_{D_{\text{replay}}}\left[\frac{1}{H}\sum_{t=1}^{H} f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\ldots t-1})\right)\right], \text{ and,}$$
$$\widehat{\mathbb{E}}^{(2)}(f) = \mathbb{E}_{D_2}\left[\frac{1}{H}\sum_{t=1}^{H} f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\ldots t-1})\right)\right]$$

Likewise, we can decompose $J_f(\pi^{\text{exp}})$ into two terms,

$$J_f^{(1)}(\pi^{\text{exp}}) \triangleq \mathbb{E}_{\pi^{\text{exp}}}\left[\sum_{t=1}^{H} f_t(s_t, a_t)\mathcal{P}(s_{1\ldots t-1})\right], \text{ and}$$
$$J_f^{(2)}(\pi^{\text{exp}}) \triangleq \mathbb{E}_{\pi^{\text{exp}}}\left[\sum_{t=1}^{H} f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\ldots t-1})\right)\right]$$

Then, from eq. (E.11),

$$\begin{aligned} J(\pi^{\text{exp}}) - J(\pi^{\texttt{RE}}) &\leq 2\sup_{f\in\mathcal{F}}\left|J_f(\pi^{\text{exp}}) - \widehat{\mathbb{E}}(f)\right| \\ &\leq \underbrace{2\sup_{f\in\mathcal{F}}\left|J_f^{(1)}(\pi^{\text{exp}}) - \mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right]\right|}_{\text{(I)}} + \underbrace{2\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right] - \widehat{\mathbb{E}}^{(1)}(f)\right|}_{\text{(II)}} \\ &\quad + \underbrace{2\sup_{f\in\mathcal{F}}\left|J_f^{(2)}(\pi^{\text{exp}}) - \widehat{\mathbb{E}}^{(2)}(f)\right|}_{\text{(III)}}. \end{aligned} \tag{E.12}$$

where the last line follows by triangle inequality. We bound each of these terms in the next 3 lemmas, starting with (I).

Lemma E.3.2 (Bounding term (I))**.**

$$\sup_{f\in\mathcal{F}}\left|J_f^{(1)}(\pi^{\text{exp}}) - \mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right]\right| \leq H\sum_{h=1}^{H}\mathbb{E}_{\pi^{\text{exp}}}\left[\textit{MEM}(s_h, h)D_{\text{TV}}\left(\pi_h^{\text{exp}}(\cdot|s_h), \pi_h^{\textit{BC}}(\cdot|s_h)\right)\right] \tag{E.13}$$

Proof. The proof of this result closely follows the supervised learning reduction of BC (cf. [75]). Note that,

$$\mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right] - J_f^{(1)}(\pi^{\text{exp}}) = \sum_{t=1}^{H}\mathbb{E}_{\pi^{\texttt{BC}}}\left[f_t(s_t, a_t)\mathcal{P}(s_{1\cdots t-1})\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\mathcal{P}(s_{1\cdots t-1})\right]. \tag{E.14}$$

Define $\pi^{(h)}$ as the policy which plays $\pi^{\exp}$ until (and including) time h and $\pi^{\texttt{BC}}$ after time h. Then, by cascading,

$$\begin{aligned}
&\mathbb{E}_{\pi^{\texttt{BC}}}\left[f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})\right] - \mathbb{E}_{\pi^{\exp}}\left[f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})\right] \\
&= \sum_{h=0}^{t-1} \mathbb{E}_{\pi^{(h)}}\left[f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})\right] - \mathbb{E}_{\pi^{(h+1)}}\left[f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})\right] \qquad \text{(E.15)}
\end{aligned}$$

Define, the uncertainty weighted state visitation measure $d^{\texttt{MEM}}$ and the uncertainty weighted look-forward reward $\rho^{\texttt{MEM}}$ as follows,

$$\begin{aligned}
d^{\texttt{MEM}}_{h+1}(s') &\triangleq \mathbb{E}_{\pi^{\exp}}\left[\mathbb{1}(s_{h+1}=s')\prod_{t'=1}^{h}\texttt{MEM}(s_{t'},t')\right] \\
\rho^{\texttt{MEM}}_{h+1}(s',a') &\triangleq \mathbb{E}_{\pi^{\texttt{BC}}}\left[f_t(s_t,a_t)\prod_{t'=h+2}^{t}\texttt{MEM}(s_{t'},t')\middle| s_{h+1}=s', a_{h+1}=a'\right]
\end{aligned}$$

By decomposing expectations along trajectories, using the fact that $\mathcal{P}(s_{1\cdots t-1}) = \prod_{t'=1}^{H}\texttt{MEM}(s_{t'},t')$ some simplification results in the following equation,

$$\begin{aligned}
&\left|\mathbb{E}_{\pi^{(h)}}\left[f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})\right] - \mathbb{E}_{\pi^{(h+1)}}\left[f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})\right]\right| \\
&= \left|\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} d^{\texttt{MEM}}_{h+1}(s')\texttt{MEM}(s',h+1)\left(\pi^{\exp}_{h+1}(a'|s') - \pi^{\texttt{BC}}_{h+1}(a'|s')\right)\rho^{\texttt{MEM}}_{h+1}(s',a')\right| \\
&\overset{(i)}{\leq} \sum_{s'\in\mathcal{S}} d^{\texttt{MEM}}_{h+1}(s')\texttt{MEM}(s',h+1)D_{\mathrm{TV}}\left(\pi^{\exp}_{h+1}(\cdot|s'),\pi^{\texttt{BC}}_{h+1}(\cdot|s')\right) \\
&= \mathbb{E}_{\pi^{\exp}}\left[\texttt{MEM}(s_{h+1},h+1)D_{\mathrm{TV}}\left(\pi^{\exp}_{h+1}(\cdot|s_{h+1}),\pi^{\texttt{BC}}_{h+1}(\cdot|s_{h+1})\right)\right].
\end{aligned}$$

where (i) uses the fact that the membership oracle is a function $\in[0,1]$ and f is bounded and lies in the interval $[0,1]$ (which implies that $\rho^{\texttt{MEM}}$ also lies in $[0,1]$ pointwise). Plugging into eq. (E.15) and subsequently into eq. (E.14) completes the proof. □

Next we bound the 3rd term, (III). This follows by an application of Bernstein's inequality.

Lemma E.3.3 (Bounding term (III))**.** *With probability* $\geq 1-\delta$,

$$\sup_{f\in\mathcal{F}}\left|J_f^{(2)}(\pi^{\exp}) - \widehat{\mathbb{E}}^{(2)}(f)\right| \leq H\sqrt{\frac{\log(F_{\max}H/\delta)\sum_{t=1}^{H-1}\mathbb{E}_{\pi^{\exp}}\left[1-\texttt{MEM}(s_t,t)\right]}{N}} + \frac{H\log(F_{\max}H/\delta)}{N}$$

Proof. First observe that,

$$\begin{aligned}
&J_f^{(2)}(\pi^{\exp}) - \widehat{\mathbb{E}}^{(2)}(f) \\
&= \sum_{t=1}^{H}\mathbb{E}_{\mathsf{tr}\sim\mathrm{Unif}(D_2)}\left[f_t(s_t,a_t)\left(1-\mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\exp}}\left[f_t(s_t,a_t)\left(1-\mathcal{P}(s_{1\cdots t-1})\right)\right] \qquad \text{(E.16)}
\end{aligned}$$

For each t, note that $f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)$ is bounded in the range $[0, 1]$. Therefore, invoking Bernstein's inequality, with probability $\geq 1 - \delta$,

$$\begin{aligned}
&\left|\mathbb{E}_{\text{Unif}(D_2)}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\lesssim \sqrt{\frac{\text{Var}_{\pi^{\text{exp}}}\left(f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right)\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{N} \\
&\leq \sqrt{\frac{\mathbb{E}_{\pi^{\text{exp}}}[\left(f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right)^2]\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{N} \\
&\overset{(i)}{\leq} \sqrt{\frac{\mathbb{E}_{\pi^{\text{exp}}}[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)]\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{N} \\
&\overset{(ii)}{\leq} \sqrt{\frac{\mathbb{E}_{\pi^{\text{exp}}}\left[1 - \mathcal{P}(s_{1\cdots t-1})\right]\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{N}
\end{aligned} \tag{E.17}$$

where (i) uses the fact that $f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)$ is bounded in the range $[0, 1]$, and (ii) uses the fact that $0 \leq f_t(s_t, a_t) \leq 1$. Assuming $0 \leq x_i \leq 1$ for all $i \in [n]$, we have the inequality,

$$1 - \prod_{i=1}^{n} x_i \leq \sum_{i=1}^{n} 1 - x_i \tag{E.18}$$

Applying this to eq. (E.17) for $1 - \mathcal{P}(s_{1\cdots t-1}) = 1 - \prod_{t'=1}^{t-1} \texttt{MEM}(s_{t'}, t')$, we have,

$$\begin{aligned}
&\left|\mathbb{E}_{\text{Unif}(D_2)}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\leq \sqrt{\frac{\sum_{t'=1}^{t-1} \mathbb{E}_{\pi^{\text{exp}}}\left[1 - \texttt{MEM}(s_{t'}, t')\right]\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{N}
\end{aligned} \tag{E.19}$$

Therefore, by union bounding, with probability $\geq 1 - \delta/H$, simultaneously for every $f_t \in \mathcal{F}_t$,

$$\begin{aligned}
&\left|\mathbb{E}_{\text{Unif}(D_2)}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\lesssim \sqrt{\frac{\log(|\mathcal{F}_t|H/\delta)\sum_{t'=1}^{t-1} \mathbb{E}_{\pi^{\text{exp}}}\left[1 - \texttt{MEM}(s_{t'}, t')\right]}{N}} + \frac{\log(|\mathcal{F}_t|H/\delta)}{N}.
\end{aligned} \tag{E.20}$$

This implies that the maximum over f_t of the LHS is upper bounded by the RHS. Union bounding over $t = 1, \cdots, H$ and plugging into eq. (E.16), we have that with probability $\geq 1 - \delta$,

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \left| J_f^{(2)}(\pi^{\text{exp}}) - \widehat{\mathbb{E}}^{(2)}(f) \right| \\
&\leq \sum_{t=1}^{H} \left|\mathbb{E}_{\text{Unif}(D_2)}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\lesssim H\sqrt{\frac{\log(F_{\max}H/\delta)\sum_{t=1}^{H-1} \mathbb{E}_{\pi^{\text{exp}}}\left[1 - \texttt{MEM}(s_t, t)\right]}{N}} + \frac{H\log(F_{\max}H/\delta)}{N}.
\end{aligned} \tag{E.21}$$

□

Finally, we are ready to bound term II.

Lemma E.3.4 (Bounding term (II))**.** *With probability* $\geq 1-\delta$,

$$\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right]-\widehat{\mathbb{E}}^{(1)}(f)\right|\lesssim H\sqrt{\frac{\log(F_{\max}H/\delta)}{N_{\text{replay}}}}+\frac{H\log(F_{\max}H/\delta)}{N_{\text{replay}}}.$$

Proof. The proof follows essentially the same structure as Lemma E.3.3 by decomposing $\widehat{\mathbb{E}}^{(1)}(f)$ into a sum of H terms of the form $f_t(s_t,a_t)\mathcal{P}(s_{1\cdots t-1})$, applying Bernstein's inequality to bound the deviation of each term from its mean and finally union bounding over the rewards $f_t\in\mathcal{F}_t$ to get the uniform bound over all discriminators $f\in\mathcal{F}$. □

Putting together Lemmas E.3.2 to E.3.4 with eq. (E.12) completes the proof of Theorem 6.2.1.

Recovering bounds in the tabular setting

In this section, we provide an upper bound on the imitation gap of RE in the tabular setting when the expert is a deterministic policy. This recovers the bound on the imitation gap for RE we proved in Chapter 2.

Theorem E.3.5. *Consider an appropriately initialized version of RE, and let the size of the replay dataset* $N_{\text{replay}}\to\infty$*. For any tabular IL instance with* $H\geq 10$*, with probability* $\geq 1-3\delta$,

$$\mathsf{Gap}(\pi^{\texttt{RE}})\lesssim\min\left\{\frac{|\mathcal{S}|H^{3/2}}{N},H\sqrt{\frac{|\mathcal{S}|}{N}}\right\}\log\left(\frac{|\mathcal{S}|H}{\delta}\right).$$

Below we describe the implementation of RE corresponding to Theorem E.3.5 in more detail. The membership oracle we use in this setting for RE is defined below,

$$\texttt{MEM}(s,t)=\begin{cases}1 & \text{if } s \text{ is visited in } D_1 \text{ at time } t\\ 0 & \text{otherwise.}\end{cases}\tag{E.22}$$

The function class $\mathcal{F}$ which we use is identical to that for empirical moment matching, which is described in Definition E.1.1.

Note that in the tabular setting, BC simply mimics the deterministic expert's actions at states visited in the dataset D_1 and plays an arbitrary deterministic action on the remaining states. As a consequence of this definition, if $\texttt{MEM}(s,t)=1\iff\pi_t^{\texttt{BC}}(\cdot|s)=\pi_t^{\text{exp}}(\cdot|s)$ and $\texttt{MEM}(s,t)=0$ otherwise. We instantiate the family of discriminators as in Definition E.1.1, as $\mathcal{F}=\bigoplus_{t=1}^{H}\{f_t:\|f_t\|_\infty\leq 1\}$ and the set of policies Π optimized over is chosen as the set of all deterministic policies. While the guarantee of Theorem 6.2.1 depends on $F_{\max}=\max_{t\in[H]}|\mathcal{F}_t|$ which is unbounded (or $\exp(|\mathcal{S}||\mathcal{A}|)$ by using a discretization of the reward space), note that we can improve the guarantee to effectively have $F_{\max}\approx\exp(|\mathcal{S}|)$ noting the structure of the

set of discriminators. Looking into the proof of Theorem 6.2.1 we bring out this dependence below. We note that there are many ways of bringing out this dependence, including a careful net argument directly on top of the guarantee of Theorem 6.2.1. We simply present one such argument below.

The critical step where the finiteness of the set of discriminators $\mathcal{F}$ is used, is in union bounding the gap between the population and the empirical estimate of $f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)$ in eq. (E.19).

$$\left|\mathbb{E}_{\text{Unif}(D_2)}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \tag{E.23}$$

In the next step of the proof of Theorem 6.2.1, we union bound over all $f_t \in \mathcal{F}_t$. However, note that for $\mathcal{F}_t = \{f_t : \|f_t\|_\infty \leq 1\}$, we have that,

$$\begin{aligned}
&\sup_{f_t:\|f_t\|_\infty \leq 1} \left|\mathbb{E}_{\text{Unif}(D_2)}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[f_t(s_t, a_t)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\overset{(i)}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left|\mathbb{E}_{\text{Unif}(D_2)}\left[\mathbb{I}(s_t = s, a_t = a)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{I}(s_t = s, a_t = a)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\overset{(ii)}{=} \sum_{s\in\mathcal{S}} \left|\mathbb{E}_{\text{Unif}(D_2)}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\overset{(iii)}{\leq} \sum_{s\in\mathcal{S}} \left|\mathbb{E}_{\text{Unif}(D_2)}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right|
\end{aligned} \tag{E.24}$$

where (i) follows similar to the equivalence between the variational representation of TV distance ($D_{\text{TV}}\left((, P), Q\right) = \frac{1}{2}\sup_{f:\|f\|_\infty\leq 1} \mathbb{E}_P[f] - \mathbb{E}_Q[f]$) and the relationship to the L_1 distance, $D_{\text{TV}}\left((, P), Q\right) = \frac{1}{2}L_1(P, Q)$. On the other hand, (ii) follows by noting that the expert is a deterministic policy (and D_2 is generated by rolling out π^{exp}). (iii) follows by an application of Holder's inequality. By subgaussian concentration, for each $s \in \mathcal{S}$, with probability $\geq 1 - \frac{\delta}{|\mathcal{S}|H}$,

$$\begin{aligned}
&\left|\mathbb{E}_{\text{Unif}(D_2)}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\lesssim \sqrt{\frac{\text{Var}_{\pi^{\text{exp}}}\left(\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right)\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}} + \frac{\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|} \\
&\overset{(i)}{\leq} \sqrt{\frac{\mathbb{E}_{\pi^{\text{exp}}}\left[\mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right)\right]\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}} + \frac{\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}
\end{aligned}$$

where (i) uses the fact that $0 \leq \mathbb{I}(s_t = s)\left(1 - \mathcal{P}(s_{1\cdots t-1})\right) \leq 1$. Combining with eq. (E.24),

union bounding and applying Cauchy Schwarz inequality, with probability $\geq 1 - \frac{\delta}{H}$,

$$\begin{aligned}
&\sup_{f_t:\|f_t\|_\infty\leq 1} \left|\mathbb{E}_{\mathrm{Unif}(D_2)}\left[f_t(s_t,a_t)\left(1-\mathcal{P}(s_{1\cdots t-1})\right)\right] - \mathbb{E}_{\pi^{\mathrm{exp}}}\left[f_t(s_t,a_t)\left(1-\mathcal{P}(s_{1\cdots t-1})\right)\right]\right| \\
&\lesssim \sqrt{|\mathcal{S}|}\sqrt{\sum_{s\in\mathcal{S}} \frac{\mathbb{E}_{\pi^{\mathrm{exp}}}\left[\mathbb{I}(s_t=s)\left(1-\mathcal{P}(s_{1\cdots t-1})\right)\right]\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}} + \frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|} \\
&= \sqrt{|\mathcal{S}|}\sqrt{\frac{\mathbb{E}_{\pi^{\mathrm{exp}}}\left[1-\mathcal{P}(s_{1\cdots t-1})\right]\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}} + \frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|} \\
&\overset{(i)}{\leq} \min\left\{\sqrt{\frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}}, \sqrt{|\mathcal{S}|\frac{\sum_{t=1}^{H-1}\mathbb{E}_{\pi^{\mathrm{exp}}}\left[1-\texttt{MEM}(s_t,t)\right]\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}}\right\} + \frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}
\end{aligned} \tag{E.25}$$

where (i) follows by the same simplification as in eq. (E.18). Comparing with eq. (E.20), this roughly corresponds to setting $F_{\max} \approx \exp(|\mathcal{S}|)$. All in all, summing eq. (E.25) over $t \in [H]$ and plugging into eq. (E.21), with probability $\geq 1-\delta$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|J_f^{(2)}(\pi^{\mathrm{exp}}) - \widehat{\mathbb{E}}^{(2)}(f)\right| \\
&\lesssim H\left\{\sqrt{\frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}}, \sqrt{|\mathcal{S}|\frac{\sum_{t=1}^{H-1}\mathbb{E}_{\pi^{\mathrm{exp}}}\left[1-\texttt{MEM}(s_t,t)\right]\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}}\right\} + \frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}
\end{aligned}$$

Finally, we plug this into eq. (E.12), which is restated below,

$$\begin{aligned}
\mathsf{Gap}(\pi^{\texttt{RE}}) &\leq 2\sup_{f\in\mathcal{F}}\left|J_f(\pi^{\mathrm{exp}}) - \widehat{\mathbb{E}}(f)\right| \\
&\leq \underbrace{2\sup_{f\in\mathcal{F}}\left|J_f^{(1)}(\pi^{\mathrm{exp}}) - \mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right]\right|}_{\text{(I)}} + \underbrace{2\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right] - \widehat{\mathbb{E}}^{(1)}(f)\right|}_{\text{(II)}} \\
&\qquad + \underbrace{2\sup_{f\in\mathcal{F}}\left|J_f^{(2)}(\pi^{\mathrm{exp}}) - \widehat{\mathbb{E}}^{(2)}(f)\right|}_{\text{(III)}}.
\end{aligned}$$

For the chosen membership oracle in eq. (E.22), the term (I) is 0, since by Lemma E.3.2 it is upper bounded by $H\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathrm{exp}}}\left[\texttt{MEM}(s_h,h)D_{\mathrm{TV}}\left(\pi_h^{\mathrm{exp}}(\cdot|s_h), \pi_h^{\texttt{BC}}(\cdot|s_h)\right)\right]$. This is equal to 0 since $\texttt{MEM}(s,t) = 0$ wherever $\pi_t^{\mathrm{exp}}(\cdot|s) \neq \pi_t^{\texttt{BC}}(\cdot|s)$. On the other hand, $N_{\mathrm{replay}} \to \infty$ ensures

that the term (III) goes to 0 by the strong law of large numbers. Therefore, with probability $\geq 1-2\delta$,

$$\begin{aligned}
&\mathsf{Gap}(\pi^{\texttt{RE}}) \\
&\leq 2\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\widehat{\mathbb{E}}^{(1)}(f)\middle|D_1\right]-\widehat{\mathbb{E}}^{(1)}(f)\right| \\
&\lesssim H\left\{\sqrt{\frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}},\sqrt{|\mathcal{S}|\frac{\sum_{t=1}^{H-1}\mathbb{E}_{\pi^{\exp}}\left[1-\texttt{MEM}(s_t,t)\right]\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}}\right\}+\frac{|\mathcal{S}|H\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_2|}
\end{aligned} \tag{E.26}$$

Finally, we bound $\mathbb{E}_{\pi^{\exp}}[1-\texttt{MEM}(s_t,t)]$ for the membership oracle defined in eq. (E.22). By definition, this quantity is the same as $\Pr_{\pi^{\exp}}(s_t \text{ not visited in } D_1 \text{ at time t })$. This is the probability that given N samples from a distribution (the state visited at time t in an expert rollout), the probability that a new sample from the same distribution is not in the support of the observed samples. This is known as the missing mass [55]. In Lemma A.3 [71] it is shown that with probability $\geq 1-\delta$,

$$\sum_{t=1}^{H-1}\Pr_{\pi^{\exp}}(s_t \text{ not visited in } D_1 \text{ at time t }) \lesssim \frac{|\mathcal{S}|H}{|D_1|}+\frac{\sqrt{|\mathcal{S}|}H\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{|D_1|}$$

Finally, combining with eq. (E.26) and using the fact that that $|D_1|,|D_2|=\Theta(N)$, with probability $\geq 1-3\delta$,

$$\mathsf{Gap}(\pi^{\texttt{RE}}) \lesssim \min\left\{H\sqrt{\frac{|\mathcal{S}|\log\left(\frac{|\mathcal{S}|H}{\delta}\right)}{N}},\frac{|\mathcal{S}|H^{3/2}}{N}\log\left(\frac{|\mathcal{S}|H}{\delta}\right)\right\}.$$

This completes the proof of Theorem E.3.5.

Appendix F

Proofs of results in Chapter 7

Notation

In this section, we use the notation $d_t^\pi(\cdot,\cdot)$ to indicate the state-action visitation measure induced by the policy π at time t. We overload the notation $d_t^\pi(\cdot)$ to denote the state-visitation measure induced by the policy π at time t. Likewise, the notations $d_t^D(\cdot,\cdot)$ and $d_t^D(\cdot)$ indicate the empirical visitation measures in the dataset D.

F.1 IL in the linear-expert setting: Proofs of Theorems 7.1.1 and 7.1.3

Proof of Lemma 7.1.2. Conditioned on the expert and the learner playing the same actions in the state, the error of the learner is exactly 0 since in such trajectories both policies collect the same reward. On the other hand, when the learner plays an action different from the expert at a visited state (and thus the 0-1 loss for this trajectory is 1), the maximum error the learner can incur is H. □

Proof of Theorem 7.1.3. Consider the compression based multi-class linear classification algorithm of [25]. This algorithm admits the following guarantee for multi-class sequence classification.

Theorem F.1.1 (Theorem 5 in [25])**.** *Consider any linear multi-class classification problem with features $\phi : X \times Y \to \mathbb{R}$. The learner is provided samples $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$: each x_i is sampled from an unknown distribution ρ and with label $y_i = \arg\max_{y \in Y} \langle \phi(x, y), \theta^\star \rangle$ for an unknown $\theta^\star \in \mathbb{R}^d$. Then, if $n \geq \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$, with probability $\geq 1-\delta$ the compression algorithm of [25] returns a linear function $\widehat{\theta} \in \mathbb{R}^d$ such that, the expected 0-1 loss is bounded by ϵ. Namely,*

$$\mathbb{E}_{x \sim \rho}\left[\mathbb{1}\left(\arg\max_{y \in Y} \left\langle \phi(x, y), \widehat{\theta} \right\rangle \neq \arg\max_{y \in Y} \langle \phi(x, y), \theta^\star \rangle\right)\right] \leq \epsilon$$

Consider the dataset as tuples of sequences of states and sequences of actions $\mathcal{S}^H \to \mathcal{A}^H$. In addition, the expert policy can be thought of as a classifier which takes sequences and outputs sequences. Namely, it is a mapping from $\mathcal{S}^H \to \mathcal{A}^H$, in the sense: for $\mathcal{S}^H \ni (s_1, \cdots, s_H) \mapsto (\pi_1^{\exp}(s_1), \cdots, \pi_H^{\exp}(s_H)) \in \mathcal{A}^H$. For $\theta \in \mathbb{R}^d$, the corresponding linear sequence classifier is

$$(s_1, \cdots, s_H) \mapsto \arg\max_{a_1, \cdots, a_H \in \mathcal{A}} \theta \mapsto \left\langle \theta, \sum_{t=1}^{H} \phi_t(s_t, a_t) \right\rangle.$$

Define the set of linear sequence classifiers corresponding to $\theta \in \mathbb{R}^d$. Then, the following two propositions are true:

1. The expert policy $\pi^{\exp}$ is a linear sequence classifier under the linear-expert assumption. At any time t and state s_t, the expert chooses the action $a_t = \arg\max_{a \in \mathcal{A}} \langle \theta_t^\star, \phi_t(s, a) \rangle$. Summing across any sequence of states $(s_1, \cdots, s_H)$ shows that the sequence of actions played by the expert satisfies: $(a_1, \cdots, a_H) = \arg\max_{a'_1, \cdots, a'_H \in \mathcal{A}} \langle \theta, \phi_t(s_t, a_t) \rangle$ which proves the claim.

2. Every sequence linear classifier corresponds to a meaningful Markovian policy. Indeed, for some $\theta \in \mathbb{R}^d$, consider the sequence linear classifier corresponding to θ. If at a state s_t at time t, the classifier does not choose the action $a_t = \arg\max_{a \in \mathcal{A}} \langle \theta, \phi_t(s_t, a) \rangle$, then on any sequence that visits the state s_t at time t, $(a_1, \cdots, a_H) \neq \arg\max_{a'_1, \cdots, a'_H \in \mathcal{A}} \langle \theta, \phi_t(s_t, a'_t) \rangle$ which leads to a contradiction. Therefore, the sequence linear classifier plays the action $a_t = \arg\max_{a \in \mathcal{A}} \langle \theta, \phi_t(s_t, a) \rangle$ at each state s_t at each time t. It is therefore a Markovian policy.

The implication of these two points is that it suffices to find a sequence linear classification algorithm from $\mathcal{S}^H \to \mathcal{A}^H$ with small expected 0-1 error, given a dataset of trajectories from the expert policy. Invoking the algorithm of [25] for linear multi-class classification and Theorem F.1.1 completes the proof shows that indeed there is a linear sequence classifier with expected 0-1 loss upper bounded by $\frac{(d+\log(1/\delta)\log(N)}{N}$ which completes the proof, invoking Lemma 7.1.2.

The proof of Theorem 7.1.1 follows immediately as a corollary of Theorem 7.1.3, by invoking Remark 7.1.2. □

F.2 Parametric function approximation under Lipschitzness

In this section, we provide an upper bound on the imitation gap of `RE` in the presence of parametric function approximation under a Lipschitzness assumption on the function classes, and assuming access to a parameter estimation oracle for offline classification. To aid in our presentation, we will define concretely the notion of a policy "induced" by a multi-class classifier.

Definition F.2.1 (Policy induced by a classifier)**.** *Consider a set of parameters* $\theta = \{\theta_1, \cdots, \theta_H\}$ *where* $\theta_t \in \Theta_t$ *for each* t*. A policy* π^θ *is said to be induced by the set of classifiers defined by* θ *if for all* $s \in \mathcal{S}$ *and* $t \in [H]$*,*

$$\pi_t^\theta(s) = \arg\max_{a \in \mathcal{A}} f_{\theta_t}(s, a).$$

By this definition, $\pi^{\exp} = \pi^{\theta^E}$ *where* $\theta^E = \{\theta_1^E, \cdots, \theta_H^E\}$*.*

In order to prove the main result we establish in this setting (Theorem 7.2.1), we first discuss the implementation of `RE`.

Implementation of `RE` (Algorithm 6) We discuss the instantiation of `RE` in the Lipschitz parameterization setting below. The underlying function class $\mathcal{F}$ is chosen arbitrarily (note that the guarantee we prove depends on this function class, and the only constraints on $\mathcal{F}$ are those in Theorem 6.2.1 - the ground truth reward must belong in $\mathcal{F} = \otimes_{t=1}^H \mathcal{F}_t$, the function class is symmetric, i.e., $f_t \in \mathcal{F}_t \iff -f_t \in \mathcal{F}_t$ for each t and for all $f_t \in \mathcal{F}_t$, $\|f_t\|_\infty \leq 1$) This requires specifying the choice of the membership oracle `MEM` and describing the instantiation of `BC`.

Implementation of `BC`. Recall that in Algorithm 6, the learner trains `BC` on the dataset D_1. In particular, under the offline classification oracle condition, Assumption 7.2.2, the learner trains H classifiers, one for each t, trained on the state-action pairs (i.e. state is the input, and the action at this state is the corresponding class) observed in the demonstration dataset at time t using the offline classifier in Assumption 7.2.2. We assume that each classifier is trained with the failure probability chosen as δ/H. Denoting the resulting set of H classifiers as,

$$\widehat{\theta}^{\mathtt{BC}} = \left\{\widehat{\theta}_1^{\mathtt{BC}}, \cdots, \widehat{\theta}_H^{\mathtt{BC}}\right\},$$

this corresponds to a policy $\pi^{\mathtt{BC}} = \pi^{\widehat{\theta}^{\mathtt{BC}}}$ induced by the classifier $\widehat{\theta}^{\mathtt{BC}}$ (in the sense of Definition F.2.1). In particular, by a union bound, the classifiers $\widehat{\theta}^{\mathtt{BC}}$ satisfy with probability $\geq 1 - \delta$ simultaneously for each time $t \in [H]$,

$$\|\theta_t^E - \widehat{\theta}_t\|_2 \leq \mathcal{E}_{\Theta_t, N, \delta/H}. \tag{F.1}$$

Membership oracle. Fix a time-step $t \in [H]$. The membership oracle `MEM` is defined in eq. (7.4) as,

$$\mathtt{MEM}(s,t) = \begin{cases} +1 & \text{if } \exists a \in \mathcal{A} \text{ such that, } \forall a' \in \mathcal{A},\ f_{\widehat{\theta}_t^{\mathtt{BC}}}(s,a) - f_{\widehat{\theta}_t^{\mathtt{BC}}}(s,a') \geq 2L\mathcal{E}_{\Theta_t, N, \delta/H} \\ 0 & \text{otherwise.} \end{cases}$$

We first show that on the states such that the membership oracle is 1, the expert policy perfectly matches the learner's policy.

Lemma F.2.1. *At every state s such that $\texttt{MEM}(s,t) = +1$, $\pi_t^{\exp}(s) = \pi_t^{\texttt{BC}}(s)$.*

Proof. Note that θ_t^E satisfies $\|\theta_t^E - \widehat{\theta}_t^{\texttt{BC}}\|_2 \leq \mathcal{E}_{\Theta_t,N,\delta/H}$ with probability $1-\delta$. Consider the action a played by the learner, for any $a' \in \mathcal{A}$,

$$\begin{aligned} f_{\theta_t^E}(s,a) - f_{\theta_t^E}(s,a') &\geq f_{\widehat{\theta}_t^{\texttt{BC}}}(s,a) - \mathcal{E}_{\Theta_t,N,\delta/H} L - f_{\widehat{\theta}_t^{\texttt{BC}}}(s,a') - \mathcal{E}_{\Theta_t,N,\delta/H} L \\ &\geq 0 \end{aligned}$$

where the first inequality follows by Lipschitzness of $f_\cdot(s,a)$ and the last inequality follows by definition of the set of states where $\texttt{MEM}(s,t) = +1$: $\forall a' \in \mathcal{A}$, $f_{\widehat{\theta}_t^{\texttt{BC}}}(s,a) - f_{\widehat{\theta}_t^{\texttt{BC}}}(s,a') \geq 2\mathcal{E}_{\Theta_t,N,\delta/H} L$.

Since for this action a, $f_{\theta_t^E}(s,a) - f_{\theta_t^E}(s,a') \geq 0$ for all other actions $a' \in \mathcal{A}$, a must be the action played by the expert policy. This completes the proof. □

Note that $\pi^{\texttt{BC}}$ always matches $\pi^{\exp}$ wherever the membership oracle $\texttt{MEM}$ is non-zero. We run Algorithm 6. Therefore, from Theorem 6.2.1, with probability $\geq 1-4\delta$, the imitation gap of the learner is bounded by,

$$\mathsf{Gap}(\pi^{\texttt{RE}}) \lesssim H^{3/2}\sqrt{\frac{\log\left(F_{\max}H/\delta\right)}{N}\frac{\sum_{t=1}^{H}\mathbb{E}_{\pi^{\exp}}\left[1-\texttt{MEM}(s_t,t)\right]}{H}} + \frac{\log\left(F_{\max}H/\delta\right)}{N}. \tag{F.2}$$

To complete the proof, we must bound $\mathbb{E}_{\pi^{\exp}}\left[1-\texttt{MEM}(s_t,t)\right]$, which is the measure of states s such that $\forall a \in \mathcal{A}, \exists a' \in \mathcal{A} : f_{\widehat{\theta}_t}(s,a) - f_{\widehat{\theta}_t}(s,a') \leq 2L\mathcal{E}_{\Theta_t,N,\delta/H}$, i.e. the mass of states which are very close to a decision boundary. The probability of this set of states is upper bounded by the weak margin condition. Indeed, for each $t \in [H]$, defining $a_s^\star = \arg\max_{a\in\mathcal{A}} f_{\widehat{\theta}_t^{\texttt{BC}}}(s,a)$,

$$\begin{aligned} \Pr_{\pi^{\exp}}\left(f_{\widehat{\theta}_t^{\texttt{BC}}}(s_t,a_{s_t}^\star) - \max_{a\neq a_{s_t}^\star} f_{\widehat{\theta}_t^{\texttt{BC}}}(s_t,a) \geq 2L\mathcal{E}_{\Theta_t,N,\delta/H}\right) &\geq e^{-\mu L\mathcal{E}_{\Theta_t,N,\delta/H}} \\ &\geq 1-\mu L\mathcal{E}_{\Theta_t,N,\delta/H}. \end{aligned} \tag{F.3}$$

Therefore,

$$\mathbb{E}_{\pi^{\exp}}\left[1-\texttt{MEM}(s_t,t)\right] \lesssim \mu L\mathcal{E}_{\Theta_t,N,\delta/H}.$$

Putting it together with eq. (F.2), and simplifying, with probability $\geq 1-4\delta$,

$$\mathsf{Gap}(\pi^{\texttt{RE}}) \lesssim H^{3/2}\sqrt{\frac{\mu L\log\left(F_{\max}H/\delta\right)}{N}\frac{\sum_{t=1}^{H}\mathcal{E}_{\Theta_t,N,\delta/H}}{H}} + \frac{\log\left(F_{\max}H/\delta\right)}{N}.$$

Note that in Eq. (F.3), we only use the fact that the probability mass of states which are η-close to any decision boundary is not too high. Similar to [8], we may consider relaxations of the weak margin condition, as below.

Assumption F.2.1 (α-weak margin condition)**.** *Consider any $t \in [H]$ and $\theta \in \Theta_t$. For each $s \in \mathcal{S}$, define $a_s^\star = \arg\max_{a \in \mathcal{A}} f_\theta(s,a)$ as the classifier output under f_θ. The α weak margin condition with parameter μ assumes that, for any $\eta \leq 1/\mu$,*

$$\forall \theta \in \Theta_t, \quad \Pr_{\pi^{\exp}}\left(f_\theta(s_t, a_{s_t}^\star) - \max_{a \neq a_{s_t}^\star} f_\theta(s_t, a) \geq \eta \right) \geq 1 - (\mu\eta)^\alpha. \tag{F.4}$$

When $\alpha = 1$, this condition is effectively equivalent to the weak margin condition in Assumption 7.2.3.

Following the proof of Theorem 7.2.1, we may obtain the following result under the α weak margin condition for $\alpha \neq 1$.

Theorem F.2.2. *For IL with parametric function approximation, under Assumptions 7.2.1, 7.2.2 and F.2.1, appropriately instatiating* `RE` *ensures that with probability $\geq 1 - 4\delta$,*

$$\mathsf{Gap}(\pi^{\texttt{RE}}) \lesssim H^{3/2} \sqrt{\frac{(\mu L)^\alpha \log\left(F_{\max} H/\delta\right)}{N} \frac{\sum_{t=1}^H (\mathcal{E}_{\Theta_t, N, \delta/H})^\alpha}{H}} + \frac{\log\left(F_{\max} H/\delta\right)}{N}. \tag{F.5}$$

Once again, we assume the same conditions on $\mathcal{F}$ as required in Theorem 6.2.1.

Extension to unbounded discriminator families

Note that when the family of discriminators $\mathcal{F}$ does not have finite cardinality, it in fact suffices to just bound the imitation gap against a finite covering of $\mathcal{F}$. We spell out the details explicitly below.

In particular, we can replace $F_{\max}$ by $\max_{t \in [H]} \mathcal{N}(\mathcal{F}_t, 1/N, \|\cdot\|_\infty)$, where $\mathcal{N}(\mathcal{G}, \xi, \|\cdot\|)$ denotes the covering number of $\mathcal{G}$ in the norm $\|\cdot\|$ as defined below.

Definition F.2.2 (Covering number)**.** *For a function class $\mathcal{G}$, tolerance ξ and norm $\|\cdot\|$, the covering number $\mathcal{N}(\mathcal{G}, \xi, \|\cdot\|)$ is defined as the cardinality of the smallest set of functions $\mathcal{G}^\xi$ such that for each $g \in \mathcal{G}$, there exists a $g' \in \mathcal{G}^\xi$,*

$$\|g - g'\| \leq \xi.$$

Corollary F.2.1. *When $\mathcal{G}$ is chosen as the set of 1-bounded linear functions, $\mathcal{G} = \{\{\langle x, \theta\rangle : x \in \mathbb{B}_2^d\} : \theta \in \mathbb{B}_2^d\}$, where $\mathbb{B}_2^d$ denotes the L_2 unit ball in $\mathbb{R}^d$, $\mathcal{N}(\mathcal{G}, \xi, \|\cdot\|_\infty) \leq \left(\frac{2\sqrt{d}}{\xi} + 1\right)^d$.*

Proof. For any $g, g' \in \mathcal{G}$, where g and g' correspond to parameters $\theta, \theta' \in \mathbb{B}_2^d$,

$$\begin{aligned} \|g - g\|_\infty &\leq \max_{x \in \mathcal{X}} \langle x, \theta - \theta' \} \\ &\leq \|x\|_2 \|\theta - \theta'\|_2 \\ &\leq \|\theta - \theta'\|_2. \end{aligned}$$

Since the L_2 covering number of $\mathbb{B}_2^d$ is bounded by $\left(2\sqrt{d}/\xi + 1\right)^d$, the result immediately follows by defining the covering of $\mathcal{G}$ as $\{\langle\theta, \cdot\rangle : \theta \in \mathcal{K}\}$ where $\mathcal{K}$ is the optimal covering of $\mathbb{B}_2^d$ in L_2 norm. □

Definition F.2.3 (Discretization of discriminator space)**.** *Define $\mathcal{F}_t^\xi$ as the optimal covering of $\mathcal{F}_t$ under the L_∞ norm in the sense of Definition F.2.2. The discretized family of discriminators we consider is, $\mathcal{F}^\xi = \otimes_{t=1}^H \mathcal{F}_t^\xi$.*

Lemma F.2.3. *Suppose for all functions $f' \in \mathcal{F}_t^{\xi_1/H}$, simultaneously $J_{f'}(\pi^{\text{exp}}) - J_{f'}(\pi^{\texttt{RE}}) \le \xi_2$. Then, for all discriminators $f \in \mathcal{F}$, $J_f(\pi^{\text{exp}}) - J_f(\pi^{\texttt{RE}}) \le 2\xi_1 + \xi_2$.*

Proof. Consider any discriminator $f \in \mathcal{F}$. By construction, there exists an $f' \in \mathcal{F}^{\xi_1/H}$ such that,

$$\|f - f'\|_\infty \le \xi_1/H.$$

Since for any policy π, the value $J_f(\pi)$ under a discriminator $f \in \mathcal{F}$ is an H-Lipschitz function of f, we can make a statement about how well $J_{f'}(\pi)$ approximates $J_f(\pi)$ for an appropriately chosen $f' \in \mathcal{F}^{\xi_1/H}$. In particular, the nearest (in L_∞ norm) $f' \in \mathcal{F}^{\xi_1/H}$ to $f \in \mathcal{F}$ satisfies that for any policy π,

$$|J_f(\pi) - J_{f'}(\pi)| \le H \times \frac{\xi_1}{H}.$$

As a consequence, for any discriminator $f \in \mathcal{F}$,

$$\begin{aligned} J_f(\pi^{\text{exp}}) - J_f(\pi^{\texttt{RE}}) &\le |J_f(\pi^{\text{exp}}) - J_{f'}(\pi^{\text{exp}})| + J_{f'}(\pi^{\text{exp}}) - J_{f'}(\pi^{\texttt{RE}}) + |J_{f'}(\pi^{\texttt{RE}}) - J_f(\pi^{\texttt{RE}})| \\ &\le \xi_1 + \xi_2 + \xi_1 = 2\xi_1 + \xi_2. \end{aligned}$$

□

In particular, this means that if we minimize $J_{f'}(\pi^{\text{exp}}) - J_{f'}(\pi^{\texttt{RE}}) \le \xi_2$ for all $f' \in \mathcal{F}^{1/NH}$, then we can ensure that for all $f \in \mathcal{F}$,

$$J_f(\pi^{\text{exp}}) - J_f(\pi^{\texttt{RE}}) \le \frac{2}{N} + \xi_2.$$

This implies the following theorem,

Theorem F.2.4. *Consider the policy $\pi^{\texttt{RE}}$ returned by Algorithm 6 where $\mathcal{F}$ is instead chosen as $\mathcal{F}^{\frac{1}{HN}}$ (as defined in Definition F.2.3). Assume that $\pi^{\text{exp}} \in \Pi$, the ground truth reward function $r_t \in \mathcal{F}_t$ which is assumed to be bounded (For all $f_t \in \mathcal{F}_t$, $\|f_t\|_\infty \le 1$). Choose $|D_1|, |D_2| = \Theta(N)$ and suppose $N_{\text{replay}} \to \infty$. With probability $\ge 1 - 3\delta$,*

$$\textsf{Gap}(\pi^{\texttt{RE}}) \lesssim \mathcal{L}_1 + \mathcal{L}_2 + \frac{\log\left(\mathcal{N}_{\max} H/\delta\right) + 1}{N}$$

where $\mathcal{N}_{\max} \triangleq \max_{t\in[H]} \mathcal{N}(\mathcal{F}_t, 1/HN, \|\cdot\|_\infty)$ *corresponds to the maximal covering number of the function classes* $\mathcal{F}_t$*, and,*

$$\mathcal{L}_1 \triangleq H^2\, \mathbb{E}_{\pi^{\exp}}\left[\frac{\sum_{t=1}^{H} \texttt{MEM}(s_t,t) D_{\mathrm{TV}}\left(\pi_t^{\exp}(\cdot|s_t), \pi_t^{\texttt{BC}}(\cdot|s_t)\right)}{H}\right],$$

$$\mathcal{L}_2 \triangleq H^{3/2}\sqrt{\frac{\log\left(\mathcal{N}_{\max} H/\delta\right)}{N}\frac{\sum_{t=1}^{H}\mathbb{E}_{\pi^{\exp}}\left[1-\texttt{MEM}(s_t,t)\right]}{H}}.$$

Remark F.2.1. Note that this line of reasoning can be extended to Theorem 7.2.1 and Theorem F.2.2 to show that the same guarantees as eq. (7.6) and eq. (F.5) respectively hold, but with $F_{\max}$ replaced by $\mathcal{N}_{\max}$.

F.3 Bounds on RE in the linear-expert setting

Next we switch tracks and look at the known-transition setting and prove the upper bound on the imitation gap of RE established in Theorem 7.3.1. At a high level, the proof will follow by showing that under Assumption 7.3.1, both the weak margin condition (Assumption 7.2.3) is satisfied, and the classification oracle (Assumption 7.2.2) can be constructed. We begin by showing the former.

Lemma F.3.1. *Under Assumption 7.3.1, the* α*-weak margin condition is satisfied with* $\alpha = 1$ *and* $\mu = 2c_{\max}\sqrt{d}$*. In particular, for all* $\theta \in \mathbb{S}^{d-1}$*,*

$$\Pr_{\pi^{\exp}}\left(\langle\theta, \phi_t(s, a^\theta_{s_t})\rangle - \max_{a\neq a^\theta_{s_t}}\langle\theta,\phi_t(s,a)\rangle \geq \eta\right) \geq 1 - \left(2c_{\max}\sqrt{d}\right)\eta.$$

where $a^\theta_{s_t} \triangleq \arg\max_{a\in\mathcal{A}}\langle\theta,\phi_t(s,a)\rangle$*.*

Proof. Observe that,

$$\begin{aligned}
&\Pr_{\pi^{\exp}}\left(\exists a \neq a^\theta_{s_t} : \langle\theta,\phi_t(s_t,a^\theta_{s_t})\rangle - \langle\theta,\phi_t(s_t,a)\rangle \leq \eta\right)\\
&\quad= \Pr_{\pi^{\exp}}\left(\exists a \neq a^\theta_{s_t} : \phi_t(s_t,a^\theta_{s_t}) - \phi_t(s_t,a) \in \{x\in\mathbb{H}^d_\theta : \langle x,\theta\rangle \leq \eta\}\right)\\
&\quad\overset{(i)}{=} \Pr_{\pi^{\exp}}\left(\exists a \neq a^\theta_{s_t} : \overline{\phi}_t(s_t,a) \in \{x\in\mathbb{H}^d_\theta : \langle x,\theta\rangle \leq \eta\}\right)\\
&\quad\overset{(ii)}{\leq} c_{\max}\Pr(\langle U,\theta\rangle \leq \eta)
\end{aligned} \tag{F.6}$$

where in (i), $\overline{\phi}_t$ is as defined in Assumption 7.3.1 and in (ii), U is uniformly distributed on the unit hemisphere, $\mathbb{H}^d_\theta$. Note that (ii) follows from the bounded density condition, Assumption 7.3.1. Note that the RHS essentially corresponds to the volume (probability

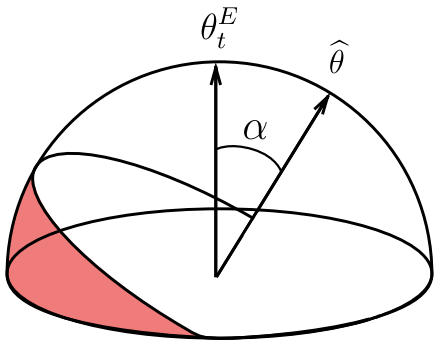


Figure F.1: If at any time $t \in [H]$ and state s, $\phi_t(s, \pi_t^{\exp}(s)) - \phi_t(s, a)$ lies in the red shaded region for some action a, then, the action played by $\pi^{\texttt{BC}}$ and $\pi^{\exp}$ at this state are different.

measure) of a disc of height η cut out of a sphere from the center. Up to normalization factors, this can be upper bounded by the surface area of the base of the disc, multiplied by the height of the disc. Namely,

$$\frac{\eta \times \frac{\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d-1}{2}+1\right)}}{\frac{1}{2}\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}}$$

Using Gautschi's inequality, for any $x \geq 0$ and $\ell \in (0,1)$ $x^{1-\ell} \leq \frac{\Gamma(x+1)}{\Gamma(x+\ell)} \leq (1+x)^{1-\ell}$. With $\ell = \frac{1}{2}$, $\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})} \leq \sqrt{d}$. Combining with eq. (F.6) results in,

$$\Pr_{\pi^{\exp}}\left(\exists a \neq a_{s_t}^{\theta} : \langle \theta, \phi_t(s_t, a_{s_t}^{\theta})\rangle - \langle \theta, \phi_t(s_t, a)\rangle \leq \eta\right) \leq 2c_{\max}\sqrt{d}\eta$$

Therefore the probability of the complement event is lower bounded by $1 - 2c_{\max}\sqrt{d}\eta$, completing the proof. □

The final thing to show is that the bounded density assumption can also be used to construct a classification oracle in the sense of Assumption 7.2.2. In particular, we will that under this assumption, any algorithm which achieves low test error can be used to construct a classification oracle in this sense. The compression based algorithm of [26] provides a guarantee on the generalization error. From Theorem 5 of [26], in the realizable setting, for linear classification, the resulting classifier $\widehat{\theta}$ has expected 0-1 loss upper bounded by $(d + \log(1/\delta))\log(n)/n$, given n classification examples. Namely, in the notation of Assumption 7.2.2, the resulting classifier $\widehat{\theta}$ satisfies with probability $\geq 1 - \delta$,

$$\Pr_{s\sim\mathcal{D}}\left(\arg\max_{a\in\mathcal{A}} f_{\theta^\star}(s^i, a) \neq \arg\max_{a\in\mathcal{A}} f_{\widehat{\theta}}(s^i, a)\right) \leq \frac{(d + \log(1/\delta)\log(n)}{n}. \tag{F.7}$$

Next we show that under Assumption 7.3.1, this equation can be used to bound the error in the parameter space, $\|\theta^\star - \widehat{\theta}\|_2$. Namely, in Assumption 7.2.2, we may choose $\mathcal{E}_{\mathbb{B}_2^d,n,\delta}$ as $\asymp \frac{(d+\log(1/\delta)\log(n)}{n}$, up to constants depending on $c_{\min}$.

Lemma F.3.2. *Consider the compression based learner $\widehat{\theta}_t^{\mathsf{BC}} = \widehat{\theta}_t$ of [26] for multi-class linear classification. Then, under Assumption 7.3.1, with probability $\geq 1-\delta$,*

$$\|\widehat{\theta}_t^{\mathsf{BC}} - \theta_t^\star\|_2 \leq \frac{2\pi}{c_{\min}} \frac{(d + \log(1/\delta)\log(N)}{N}$$

Proof. Fix $t \in [H]$. The generalization error of $\widehat{\theta}_t^{\mathsf{BC}} = \widehat{\theta}_t$ can be written as,

$$\begin{aligned} &\Pr_{\pi^{\exp}}\left(\arg\max_{a\in\mathcal{A}}\langle\theta_t^\star, \phi_t(s_t,a)\rangle \neq \arg\max_{a\in\mathcal{A}}\langle\widehat{\theta}_t^{\mathsf{BC}}, \phi_t(s_t,a)\rangle\right) \\ &\quad = \Pr_{\pi^{\exp}}\left(\exists a \neq \pi_t^{\exp}(s_t) : \phi_t(s_t, \pi_t^{\exp}(s_t)) - \phi_t(s_t,a) \in \mathcal{C}\right), \end{aligned} \tag{F.8}$$

where $\mathcal{C}$ is illustrated in fig. F.1 and is formally defined as,

$$\mathcal{C} \triangleq \{x \in \mathbb{H}_t^d : \langle x, \widehat{\theta}_t^{\mathsf{BC}}\rangle \leq 0\}.$$

On the states which "belong" to $\mathcal{C}$ (i.e. at those states s where $\exists a \neq \pi_t^{\exp}(s_t) : \phi_t(s, \pi_t^{\exp}(s_t)) - \phi_t(s,a) \in \mathcal{C}$), there exists an action a such that $\widehat{\theta}_t^{\mathsf{BC}}$ is better correlated with this action than $a_s^\star$. In other words, $\widehat{\theta}_t^{\mathsf{BC}}$ and $\theta^\star$ play different actions at this state. Note that $\mathcal{C}$ is essentially the set difference of two hemispheres with different poles. By the bounded density condition, Assumption 7.3.1, and eq. (F.8),

$$\Pr_{\pi^{\exp}}\left(\pi_t^{\exp}(s_t) \neq \arg\max_{a\in\mathcal{A}}\langle\widehat{\theta}_t^{\mathsf{BC}}, \phi(s,a)\rangle\right) \geq c_{\min}\Pr\left(U \in \mathcal{C}\right), \tag{F.9}$$

where U is uniformly distributed over $\mathbb{H}_\theta^d$. Referring to fig. F.1, we have that,

$$\Pr\left(U \in \mathcal{C}\right) = \frac{\alpha}{\pi}$$

where α is the angle between $\widehat{\theta}_t^{\mathsf{BC}}$ and θ_t^E. In particular, from eq. (F.9),

$$\Pr_{\pi^{\exp}}\left(a_s^\star \neq \arg\max_{a\in\mathcal{A}}\langle\widehat{\theta}_t^{\mathsf{BC}}, \phi(s,a)\rangle\right) \geq c_{\min}\frac{\alpha}{\pi} \geq c_{\min}\frac{\|\theta^\star - \widehat{\theta}_t^{\mathsf{BC}}\|_2}{\pi},$$

where in the last inequality, we use the fact that $\|\theta^\star\|_2 = \|\widehat{\theta}_t^{\mathsf{BC}}\|_2 = 1$ without loss of generality. By the generalization error bound on $\widehat{\theta}_t^{\mathsf{BC}} = \widehat{\theta}_t$ in eq. (F.7), with probability $\geq 1-\delta$,

$$\|\theta^\star - \widehat{\theta}_t^{\mathsf{BC}}\|_2 \leq \frac{\pi}{c_{\min}} \frac{(d + \log(1/\delta)\log(N)}{N}$$

□

Lemma F.3.2 shows that under the bounded density condition Assumption 7.3.1, the compression based learner $\widehat{\theta}$ of [26] essentially induces a classification oracle for linear classification with $\mathcal{E}_{\mathbb{B}_2^d,n,\delta} = \frac{\pi}{c_{\min}} \frac{(d+\log(1/\delta)\log(n)}{n}$. Finally, from Corollary F.2.1, we have a bound on the covering number of linear families. Putting together all of these results with Theorem 7.2.1 (noting that we can replace $F_{\max}$ by $\mathcal{N}_{\max}$ from Remark F.2.1) results in Theorem 7.3.1.