Causal Imitation Learning with Unobserved Contexts

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider imitation learning problems where the expert has access to a per-episode *context* that is hidden from the learner, both in the demonstrations and at test-time. While the learner might not be able to accurately reproduce expert behavior early on in an episode, by considering the entire history of states and actions, they might be able to eventually identify the context and act as the expert would. We show that on-policy imitation learning algorithms (with or without access to a queryable expert) are better equipped to handle these sorts of *asymptotically realizable* problems than off-policy methods. We conduct experiments in a toy bandit domain that show that there exist sharp *phase transitions* of whether off-policy approaches are able to match expert performance asymptotically, in contrast to the uniformly good performance of on-policy approaches. We demonstrate that on several continuous control tasks, off-policy approaches are not able to effectively use history to reduce their uncertainty over the context, in contrast to on-policy methods.

1 Introduction

An unstated assumption in much of the work in imitation learning (IL) is that the learner and the expert have access to the same state information. With powerful enough statistical models, this assumption places the problem in the *realizable* setting – i.e. the imitator can actually behave like the expert. In practice, however, an expert might have more information than the learner does. For example, an expert policy might be trained in simulation with privileged access to state before being used to supervise a learner policy that has access to a subset of features, a recipe that has enjoyed success in domains from motion planning with obstacles [Choudhury et al., 2018], to autonomous driving [Chen et al., 2019], to legged locomotion [Lee et al., 2020, Kumar et al., 2021].

Recent theoretical work has established "no-go" results for successfully imitating an expert that has access to more information [Zhang et al., 2020, Kumor et al., 2021]. The core of their arguments is that without seeing some feature that influences expert behavior but is not echoed elsewhere in the state, the learner might not be able to properly ground the expert actions in the observed state. In causal inference terms, this hidden information acts as an *unobserved confounder* which prevents identification of the desired treatment effect (the expert action). Despite these results, impressive empirical successes have been achieved even when the learner has a more impoverished state representation than the expert.

In this work, we reconcile theory and practice by considering a broad class of problems where the learner's ability to mimic expert actions increases as more observations are revealed. We study the large-horizon limit to tease out what is key to good performance. We find that off-policy approaches (e.g. behavioral cloning) that ignore the resulting covariate shift from initially sub-optimal decisions can lead to poor results, even when there exists a policy that in the large-horizon limit is optimal. We show that for some problem families, there exists a sharp phase transition in problem parameters where behavior cloning shifts between being consistent to having arbitrarily poor performance.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In contrast, we show that on-policy approaches that leverage interaction with the demonstrator [Ross et al., 2010] or take advantage of interaction with the environment (in the style of Inverse Optimal Control [Bagnell, 2015]) [Ziebart et al., 2008, Ho and Ermon, 2016] are always (i.e. independent of parameters) asymptotically consistent on these problems. This strong separation between on- and off-policy approaches we believe explains empirical findings of poor performance of behavioral cloning even in regimes with large data and powerful model classes [Spencer et al., 2021, Muller et al., 2006, Codevilla et al., 2019, de Haan et al., 2019, Bansal et al., 2018, Kuefler et al., 2017] and also explains the success of on-policy methods above.

We note that, in contrast to the hidden state that is common in real-world problems [Boots et al., 2011, Kumar et al., 2021, Lee et al., 2020], standard benchmarks like the PyBullet suite [Coumans and Bai, 2016] are fully observed, enabling off-policy algorithms like behavioral cloning to match expert performance [Swamy et al., 2021]. Thus, for our experiments, we introduce partial observability to ensure that our results are closer to the real world.

We study in detail Contextual Markov Decision Process (CMDP) [Hallak et al., 2015] that satisfy an *identifiability* condition. Intuitively, this means we can expect that proper utilization of history eventually enables accurate prediction of the context. A key result we show is that ***for identifiable CMDPs where the learner can recover from mistakes early on in an episode, on-policy imitation learning algorithms that operate in the space of histories are able to asymptotically match expert value, while off-policy approaches struggle to do so.***

More concretely, our work makes three contributions:

1. We show that under appropriate identifiability and recoverability conditions, the context-dependent expert policy might become *asymptotically realizable*, enabling on-policy imitation learning algorithms to match (or nearly match) expert performance.
2. We conduct experiments in a simplified bandit domain which show that there exist sharp *phase transitions* in terms of when off-policy imitation learning algorithms equipped with history work in contrast to the uniform value-equivalence of the policies produced by on-policy approaches.
3. We conduct experiments on continuous control tasks that show that on-policy algorithms are able to take advantage of history to correctly identify the context in a way that off-policy methods are not.

We begin with a discussion of related work.

2 Related Work

One of the fundamental challenges of imitation learning (or any sequential prediction task where the learner consumes some function of its own prior predictions) is the likelihood of significant covariate shift between training-time data and test-time observations [Ross and Bagnell, 2010]. In short, this happens because the learner might end up in states not seen in the demonstrations and is thus unsure how to act. Early work in this area includes that of Daumé et al. [2009] in the natural language processing and that of Ross et al. [2010] in robotics, both of which come to the preceding conclusion. Recent work by Spencer et al. [2021] shows that there are actually more than one regime of covariate shift. In the "easy" regime where the expert is realizable, off-policy methods like behavior cloning match expert performance when data and model capacity are large enough. However, in harder regimes where the expert is non-realizable due to model misspecification (or effectively so due to data scarcity), off-policy methods compound in error and one must rely on on-policy methods, either those that require an interactive expert [Ross et al., 2010] or an interactive simulator Ziebart et al. [2008], Swamy et al. [2021].

One practically important instance of model misspecification is when the learner is denied state information that the expert uses. For instance, in self-driving, the human expert has richer context about the scene than the limited perception system of the learner. While the standard solution is to add a history of past states and actions to the model, practitioners have often noted that this leads to an "inertia effect" where the learner latches on to the past action. Muller et al. [2006] was the first to note such latching with steering actions, and more more recently Kuefler et al. [2017], Bansal et al. [2018], de Haan et al. [2019] note this with braking actions and Codevilla et al. [2019] with

accelerations. Recent work by Ortega et al. [2021] also points out this latching behavior which they term as "self-delusion".

Once one identifies the downstream effect of missing context as covariate shift, a natural question is whether an extension of prior covariate-shift-robust imitation learning methods to the space of histories would be able to learn effectively in confounded settings. Prior work has answered parts of this question. For example, Choudhury et al. [2018] proves that interactive imitation learning over the space of histories converges to the QMDP approximation of the expert's policy [Littman et al., 1995]. In fact, the proposed algorithm of Ortega et al. [2021] (counterfactual teaching) is identical to the earlier FORWARD algorithm of Ross and Bagnell [2010] under log-loss. Our main contribution provides conditions under which such approaches will converge to a policy that is equivalent in value to that of the expert in the presence of unobserved contexts.

Our focus on the contextual MDP [Hallak et al., 2015] is because, as argued by [Zhang et al., 2020, Kumor et al., 2021], context that is updated throughout the episode and is only reflected for a single step prevents the learner from refining its estimates via considering history. We would expect our results to generalize to settings beyond the CMDP where a similar asymptotic realizability property holds true. Tennenholtz et al. [2021] consider IL in contextual MDPs but give the learner access to the confounder at test time and do not consider policies that operate over the space of histories.

Recently, Swamy et al. [2021] proved that for a wide class of interactive expert imitation learning algorithms, one should expect learner sub-optimality to grow as $O(\epsilon HT)$, where T is the horizon of the problem and H is a recoverability constant that upper bounds the number of steps it takes for the expert to recover from an arbitrary learner mistake Leike et al. [2016], Ross et al. [2010]. We argue that for infinite horizon problems, one requires a finite H to argue value equivalence for the policy that an on-policy learner converges to – otherwise, a mistake early on can doom the learner for the rest of the horizon.

3 Imitation Learning with Unobserved Contexts

We begin with a toy example studied by Ortega et al. [2021].

3.1 Phase Transitions for Off-Policy Imitation Learning

Problem 3.1 (Causal Bandit Problem)**.** Consider a K-armed bandit. At the beginning of each episode, a context $c \in [K]$ is chosen uniformly at random and represents the correct arm for that episode. Pulling an arm leads to binary feedback: $+$ if the arm was correct and $-$ otherwise. This feedback is flipped with probability $\epsilon_{obs} \in [0, 1]$. The learner observes expert demonstrations where at each timestep, the expert plays the correct arm with probability $1 - \epsilon_{exp} \in [0, 1]$ and another arm uniformly at random otherwise. The reward function is 1 for pulling the correct arm and 0 otherwise. We emphasize that the learner does not observe the rewards, just noisy binary feedback as an observation after each pull.

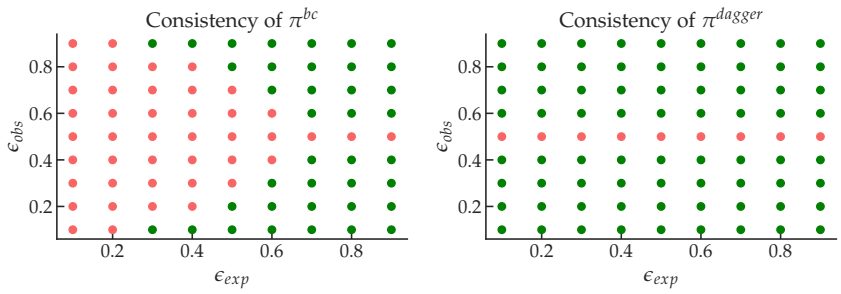

Figure 1: We plot whether the policy has a higher probability of making mistakes than the expert (ϵ_{eps}) after 2000 steps, averaged across 100 trials. We use red dots to indicate when this is true and green dots otherwise. We see DAgger work everywhere the problem is feasible in contrast to BC, for which a slight perturbation of a parameter can lead to a drastically different result in terms of long-term performance.

Intuitively, by tabulating the noisy observations, the learner should be able to narrow down what the correct arm is. We focus on this problem as it is trivially recoverable – the learner is reset to the same state after each action so they are free to explore without consequences. In Fig. 1, we plot, for a variety of settings of the problem parameters (ϵ_{exp}, ϵ_{obs}) whether the learner makes mistakes with a significantly higher probability than the expert for behavioral cloning (an off-policy algorithm) and DAgger (an on-policy algorithm which utilizes a queryable expert). We give all learners access to the full history of interactions (i.e. arms pulled and noisy feedback observed).

With exactly $\epsilon_{obs} = 0.5$, the learner gets no information from observation, preventing any algorithm from learning properly. This means that under this particular setting, the problem instance is not *identifiable*, a concept we formalize below. In contrast to the uniform consistency of DAgger whenever the problem is identifiable, we see sharp *phase transitions* in terms of where where BC is consistent – a small change to either problem parameter can lead to a drastically different result. This means that it can be rather hard to predict whether BC will work for a given problem, rendering it dangerous to apply in practice.

As we increase ϵ_{exp}, the expert has a higher chance of making a mistake, bringing its trajectory distribution closer to that of the uniform policy and lowering the worst-case density ratio between learner and expert trajectory distributions. As was established by Spencer et al. [2021], a low density ratio is a sufficient condition for behavioral cloning to be consistent. Our experiments echo this condition as a higher ϵ_{exp} leads to greater flexibility in terms of choice of ϵ_{obs} while maintaining consistency. The symmetry around the $\epsilon_{obs} = 0.5$ axis is due to the fact that the learner can still figure which arm is different from the rest on the basis of negated feedback.

The main take-away from this experiment is that on-policy approaches appear to be able to handle hidden context given access to history, in contrast to off-policy approaches, which can do so but in a way that depends heavily on problem parameters. To begin to explain this difference in behavior of on-policy and off-policy algorithms, we take a brief detour into the structural causal model (SCM) each of these classes of algorithms is implicitly assuming.

3.2 A Graphical Formalism for Imitation Learning

Consider a finite-horizon Contextual Markov Decision Process (CMDP) parameterized by $\langle \mathcal{S}, \mathcal{A}, \mathcal{C}, \mathcal{T}, r, T \rangle$ where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{C}$ are the state, action, and context spaces, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{C} \to \Delta(\mathcal{S})$ is the transition operator, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{C} \to [-1, 1]$ is the reward function, and T is the horizon. At the beginning of each episode, a context is sampled from $p(c)$ and is held fixed until the next reset. Intuitively, there exists a family of reward and transition functions that are indexed by the context for each episode. We use $h_t \in \mathcal{H}$ to denote a t-step history: $(s_1, a_1, \ldots, s_t)$. We see trajectories generated by an expert policy $\pi^E : \mathcal{S} \times \mathcal{C} \to \Delta(\mathcal{A})$ and search over time-varying policies $\pi : \mathcal{H} \to \Delta(\mathcal{A})$. We use $\pi_{1\ldots t}$ to refer to the sequence of policies that comprise this time-varying policy. We assume that our policy class Π is convex and compact.

Expert trajectories of length t are generated according to

$$p(\tau; \pi^E) \triangleq p(c)p(s_1) \prod_{i=1}^{t-1} \pi^E(a_i|c, s_i)\mathcal{T}(s_{i+1}|s_i, a_i, c), \tag{1}$$

while learner trajectories are generated according to

$$p(\tau; \pi) \triangleq p(c)p(s_1) \prod_{i=1}^{t-1} \pi_t(a_i|h_i)\mathcal{T}(s_{i+1}|s_i, a_i, c). \tag{2}$$

We use $\tau \sim \pi^E$ and $\tau \sim \pi$ to denote T-step trajectories sampled according to the above distributions.

The key difference between on-policy and off-policy imitation learning algorithms is the difference between the center and right SCMs of Fig. 2: (center) on-policy assume the data they have seen thus far is generated by executing learner π while (right) off-policy assumes it is generated by executing the expert policy π^E. The off-policy approximation is what leads to the latching behavior observed empirically: even if the first action was chosen at random, by treating it as though it was produced by the expert π^E (who sees the context c and picks actions that are correlated across time), the learner is likely to continue to repeat the past action. This is also what we observe empirically in the causal

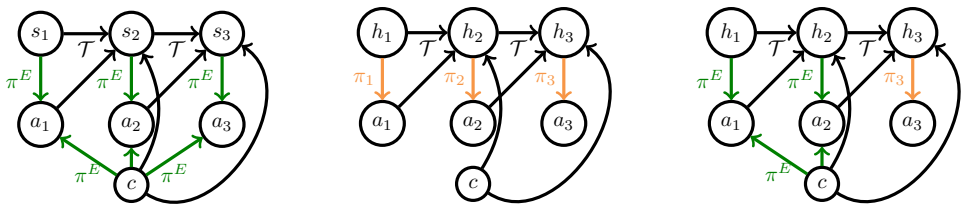

Figure 2: **Left**: The SCM that corresponds to the generative process for expert trajectories. **Center**: The SCM corresponds to the generative process for learner trajectories in reality. **Right**: The SCM that corresponds to the generative process that off-policy algorithms assume – intuitively, it corresponds to the expert taking all actions up till this point.

bandit problem in the cases where behavioral cloning does not work: the learner continues to play the first arm it chose, ignoring the feedback it gets that it is repeatedly making a mistake.

Another way of seeing this point is by considering what the posterior over the confounder would be under samples from each of these SCMs. Via Bayes Rule and assuming a uniform prior over contexts, we know that

$$p(c|h_t) \propto p(c, h_t). \tag{3}$$

Under the off-policy graphical model, $p(c, h_t) \propto p(\tau; \pi^E)$ – the probability of a history h_t under the expert's distribution (the left SCM in Fig. 2). Expanding terms, we arrive at

$$p_{\text{off}}(c, h_t) \propto p(\tau; \pi^E) \propto p(c)p(s_1) \prod_{i=1}^{t-1} \pi^E(a_i|c, s_i)\mathcal{T}(s_{i+1}|s_i, a_i, c) \tag{4}$$

Notice how under the expert's generative model, the context c directly influences the actions so it is useful to condition on them when attempting to predict c. We highlight this term that encodes this dependence in green. In contrast, in the on-policy graphical model (Fig. 2, center), we could treat the actions as *interventions* rather than as variables with a path to c and still end up with the same posterior over contexts. Mathematically,

$$p_{\text{on}}(c, h_t) \propto p(\tau; \pi) = p(c|do(a_1) \dots do(a_{t-1}), s_1 \dots s_t) \propto p(c)p(s_1) \prod_{i=1}^{t-1} \mathcal{T}(s_{i+1}|s_i, a_i, c), \tag{5}$$

where the equality follows from the fact that

$$(c \perp a_{1\dots t}|s_{1\dots t})_{\mathcal{G}_{\underline{a_{1:t}}}} \tag{6}$$

and the standard $do()$-calculus rules [Pearl et al., 2016]. Intuitively, we are leveraging the fact that the learner's actions have all their dependence on the context mediated through the history of states to ignore them in our posterior calculation. Notice that this expression matches the off-policy posterior except for the term in green – because the on-policy learner knows that the actions in the history were not produced by the expert, it does not weight them by the expert's probability of playing them in it's posterior update. Graphically, as far as the posterior over the context is concerned, we could be in the following SCM:

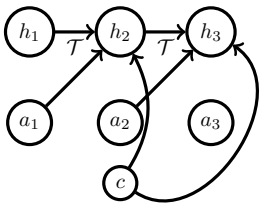

Figure 3: The SCM for a CMDP that treats actions as interventions that provide no useful information about the context. This SCM is what the on-policy learner uses when performing posterior updates.

By ignoring the incorrect actions early on, the on-policy learner does not learn to merely repeat the first thing it tried. While this single-term might seem rather unimportant, it can lead to markedly different results. In fact, the learners we used to generate Fig. 1 had policies given by

$$\pi^{bc}(a_t|h_t) = \sum_{c \in \mathcal{C}} p_{\text{off}}(c|h_t)\pi^E(a_t|c, s_t), \tag{7}$$

and

$$\pi^{dagger}(a_t|h_t) = \sum_{c \in \mathcal{C}} p_{\text{on}}(c|h_t)\pi^E(a_t|c, s_t). \tag{8}$$

Note that these policies are exactly equivalent except for the green term in the posterior of the off-policy π^{bc}, yet have markedly different consistency properties as the parameters of the problem are changed. We now turn our attention to generalizing this argument beyond likelihood-based learners (i.e. those that match sufficient statistics of expert behavior – *moments* – rather than maintaining explicit posteriors) and providing value equivalence guarantees for on-policy learners in the limit of infinite timesteps.

3.3 Asymptotic Realizability in Imitation Learning

One might hope that with the use of history and the on-policy graphical model, one might be able to eventually match expert performance in contextual problems. This is an example of a more general phenomenon we term *asymptotic realizability* (AR), in which the learner is able to perform as well as the expert does with high probability after observing an arbitrary history of some length. We begin by defining our goal in AR problems:

Definition 3.2 (Asymptotic Value Equivalence (AVE))**.** We say that policy π is asymptotically value-equivalent (AVE) when the following condition holds true as $T \to \infty$:

$$\mathbb{E}_{\tau \sim \pi^E}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] = \mathbb{E}_{\tau \sim \pi}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)].$$

Intuitively, this condition means that the learner performs as well as the expert does on average, given enough time. In the finite horizon, we use *average imitation gap* (AIG) to refer to the following quantity:

$$\mathbb{E}_{\tau \sim \pi^E}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] - \mathbb{E}_{\tau \sim \pi}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)]. \tag{9}$$

A sufficient condition for AVE is that the AIG goes to 0 as $T \to \infty$. We will now study the AIG properties of on-policy and off-policy imitation learning algorithms. While up to this point we have focused on *likelihood-based* methods (e.g. behavioral cloning or history-based methods that maintain an explicit posterior and perform Bayesian filtering), our theory generalizes to the broader class of moment-matching algorithms, which includes, beyond the previously discussed approaches, AdVIL/AdRIL/DAeQuIL [Swamy et al., 2021], GAIL [Ho and Ermon, 2016], and MaxEnt IRL [Ziebart et al., 2008].

First, as in Swamy et al. [2021], we define $\mathcal{F}_{Q_E}$ as the set of *on-Q moments*, with $f \in \mathcal{F}_{Q_E}$ satisfying type signature $\mathcal{S} \times \mathcal{A} \times \mathcal{C} \to [-T, T]$. Intuitively, $\mathcal{F}_{Q_E}$ spans the set of possible expert Q-functions. We require the actual expert Q-function to be contained: $Q^{\pi_E} \in \mathcal{F}_{Q_E}$. We assume that $\mathcal{F}_{Q_E}$ is convex, compact, closed under negation, and finite dimensional. We use $\mathcal{F}_{\text{on}}(t) = \{f/2(T-t) : f \in \mathcal{F}_{Q_E}\}$ to refer to a scaled version of the on-Q moments.

We define two key quantities:

$$\epsilon_{\text{on}}(t) = \sup_{f \in \mathcal{F}_{\text{on}}(t)} \mathbb{E}_{\tau \sim \pi}[f(s_t, a_t, c) - \mathbb{E}_{a' \sim \pi^E(h_t)}[f(s_t, a', c)]] \tag{10}$$

$$H = \sup_{\substack{s \in \mathcal{S}, a \in \mathcal{A} \\ c \in \mathcal{C}, f \in \mathcal{F}_{Q_E}}} f(s, a, c) - \mathbb{E}_{a' \sim \pi^E(s,c)}[f(s, a', c)] \tag{11}$$

where $\epsilon_{\text{on}}(t)$ is the moment matching error at time t between actions selected by the expert and the learner on states visited by the learner, H is a moment recoverability constant which bounds how

much total cost is incurred for the expert to recover from an arbitrary mistake Leike et al. [2016], Ross et al. [2010], Swamy et al. [2021]. These can be seen as the natural history-space generalizations of the analogous concepts in Swamy et al. [2021] and govern the performance of interactive imitation learning algorithms.

A similar moment-matching error can be defined for off-policy algorithms, this time computed on expert rollouts and as a function of learner Q moments. Let $\mathcal{F}_Q$ be the class of learner Q moments and satisfy the same assumptions as $\mathcal{F}_{Q_E}$. Analogously to above, we define $\mathcal{F}_{\text{off}}(t) = \{f/2(T-t) : f \in \mathcal{F}_Q\}$ to refer to a scaled version of the off-Q moments. We can now define our off-policy moment matching error:

$$\epsilon_{\text{off}}(t) = \sup_{f \in \mathcal{F}_{\text{off}}(t)} \mathbb{E}_{\tau \sim \pi^E}[f(s_t, a_t, c) - \mathbb{E}_{a' \sim \pi^E(h_t)}[f(s_t, a', c)]] \tag{12}$$

We can also define a moment-matching error for online imitation learning algorithms (i.e. those that have access to the environment but not a queryable expert like GAIL [Ho and Ermon, 2016] or MaxEnt IRL [Ziebart et al., 2008]) in terms of *reward moments* that span the class of reward functions. Let $f \in \mathcal{F}_r$ satisfy type signature $\mathcal{S} \times \mathcal{A} \times \mathcal{C} \to [-1, 1]$ and assume that $r \in \mathcal{F}_r$. We define the reward moment-matching error as

$$\epsilon_{\text{rew}}(t) = \sup_{f \in \mathcal{F}_r} \mathbb{E}_{\tau \sim \pi}[f(s_t, a_t, c) - \mathbb{E}_{a' \sim \pi^E(h_t)}[f(s_t, a', c)]]. \tag{13}$$

We are now ready to state our main results:

Theorem 3.3 (Interactive AVE)**.** *For all MDPs and π s.t. $\limsup_{t\to\infty} \epsilon_{on}(t) = \epsilon_{on}(\infty)$,*

$$\mathbb{E}_{\tau \sim \pi^E}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] - \mathbb{E}_{\tau \sim \pi}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] \leq \epsilon_{on}(\infty) H. \tag{14}$$

Thus, if $\epsilon_{\text{on}}(\infty) = 0$, the on-policy learner will achieve AVE. Even if it does not, the AIG will be bounded by a constant. In contrast, we have the following result for off-policy algorithms:

Theorem 3.4 (Off-Policy AVE)**.** *For all MDPs and π s.t. $\limsup_{t\to\infty} \epsilon_{off}(t) = \epsilon_{off}(\infty)$,*

$$\mathbb{E}_{\tau \sim \pi^E}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] - \mathbb{E}_{\tau \sim \pi}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] \leq \epsilon_{off}(\infty) T. \tag{15}$$

This result indicates that unless an off-policy learner is able to reduce its moment-matching error at a rate $\propto \frac{1}{T}$, the learner will have a growing AIG and therefore not achieve AVE. We see a similar result for online IL algorithms to that for interactive IL algorithms:

Theorem 3.5 (Online AVE)**.** *For all MDPs and π s.t. $\limsup_{t\to\infty} \epsilon_{rew}(t) = \epsilon_{rew}(\infty)$,*

$$\mathbb{E}_{\tau \sim \pi^E}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] - \mathbb{E}_{\tau \sim \pi}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t, c)] \leq \epsilon_{rew}(\infty). \tag{16}$$

We prove these results in the appendix. To come full circle, we now argue that a subset of CMDPs (including our causal bandit problem) satisfies a condition that makes it possible to drive $\epsilon_{\text{off}}(\infty)$, $\epsilon_{\text{on}}(\infty)$, and $\epsilon_{\text{rew}}(\infty)$ to 0. We now make two assumptions that characterize this subset:

Assumption 3.6 (Asymptotic Identifiability)**.** There exists a function g such that for any context c, and for all $t \in [T]$, and any history of length t generated by any $\pi \in \Pi$, $g(h_t) = c$ with probability $1 - \delta(t)$, where $\lim_{t\to\infty} \delta(t) = 0$.

This assumption means that regardless of the sequence of observations the learner receives, they will eventually be able to pin down the confounder. One way to ensure this condition holds is to add exploration noise to the learner's actions such that they try all actions with sufficiently high probability and assuming an underlying ergodicity of the MDP.

Assumption 3.7 (Realizability)**.** Π contains $\pi_E(s_t, g(h_t))$, where g is defined as above.

This is a relatively weak assumption which simply means that the learner can act as the expert would once they have nailed down the confounder. Under these two assumptions, a CMDP is really just an example of the more general phenomenon of AR. We prove the following result in the appendix:

Theorem 3.8. *For any contextual MDP and policy class that satisfy the preceding two assumptions, any no-regret learner over* Π *will be able to drive* $\epsilon_{on}(\infty)$ *and* $\epsilon_{rew}(\infty)$ *to 0.*

We also note that our causal bandit problem satisfies these assumptions:

Corollary 3.9. *The Causal Bandit Problem satisfies Asymptotic Identifiability and Realizability and hence achieve AVE for any on-policy moment matching algorithm.*

In short, for asymptotically realizable problems, on-policy imitation learning algorithms have stronger guarantees than their off-policy counterparts with respect to asymptotic value equivalence. Contextual MDPs are a specific example of AR problems if they satisfy an identifiability condition. One specific example of such a CMDP is our Causal Bandit Problem. We now turn our attention to another CMDP that satisfies our assumptions to further validate our theory empirically.

4 Experiments

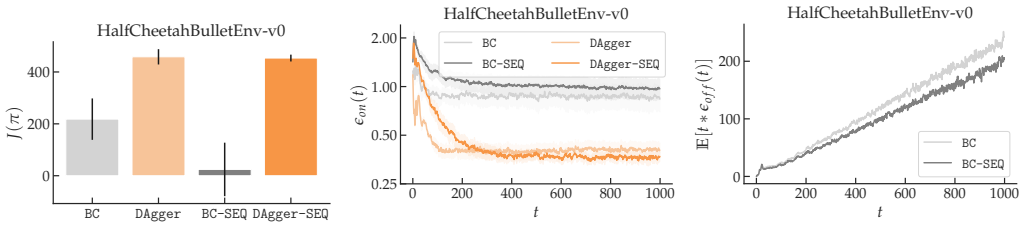

Figure 4: We consider a modification of the HalfCheetah task where the goal is for the agent to run at a particular velocity. The expert sees this velocity while the learner observes an indicator of whether their current velocity is above or below the target. We use the suffix `-SEQ` to refer to models that have access to history (of length 5 for all experiments). Standard errors are computed across 4 runs. **Left:** We see that adding history to `BC` actually *reduces* the performance of the learned policies. In contrast, for `DAgger`, using a sequence model does not harm performance. **Center:** We see that off-policy approaches are far worse at driving down on-policy error than on-policy approaches. We see `DAgger-SEQ` eventually out-perform DAgger in terms of error. **Right:** As argued by our theory, if off-policy classification error drops at a rate faster than $1/t$, we would expect value equivalence for off-policy methods. However, we do not see this for either method, which matches the results from the left plot. Notice that while adding history enables `BC-SEQ` to drive down off-policy classification error better than `BC`, it is not able to do sufficiently fast enough to guarantee value equivalence.

We conduct experiments in a CMDP extension of the standard PyBullet tasks Coumans and Bai [2016] that draws heavily on the multi-task reinforcement learning setups of Finn et al. [2017] and Kumar et al. [2020]. In these tasks, the agent is rewarded for running at a particular velocity that is randomly sampled at the beginning of each episode. We train expert policies that have access to this privileged information. We append an indicator to the learner's state of whether their current velocity is below or above the target velocity for the episode – by observing a few such indicators, one might hope that the learner is able to hone in on what their velocity should be via a sort of binary search. We compare two algorithms: the off-policy behavioral cloning (`BC`) [Pomerleau, 1989] and the on-policy DAgger (`DAgger`) [Ross et al., 2010]. We use two model classes. The base class is a two-layer MLP with 256 hidden units and ReLu activations. The history-dependent class (for which we use the suffix `-SEQ`) is an LSTM with a hidden unit size of 256 that is preceded by a 1-layer MLP with 256 units and a ReLu activation. We give the learner access to the last 5 timesteps of history. We emphasize that the expert operates only using the current state.

In Fig. 4 Left, we see that without access to history, `BC` performs poorly. Interestingly, when we equip the learner with history, we actually see it perform *worse* on average, reflecting the short of latching-style behavior that has been observed repeatedly in practice. In contrast, we see that equipping our on-policy learner with access to the last few observations and actions does not lead to a sharp decline in terms of performance. For the center and right plots, we use MSE as a surrogate for moment-matching error. In the center plot of Fig. 4, we see that, as expected, on-policy methods are far better at matching expert actions on their own rollout distribution. We see that with enough

time, `DAgger-SEQ` is able to predict actions marginally better than `DAgger`. In Fig. 4 right, we plot the product of $\epsilon_{\text{off}}(t)$ and t. If this curve was roughly constant or decaying (reflecting an off-policy moment matching error that decays at or faster than $1/t$), our theory would indicate that the learner would perform as well as the expert under an arbitrary reward function. However, we actually see a *linear growth*, indicating that our off-policy learners should not achieve value equivalence, which echoes what we see in the left plot. Thus, we find that our theory accurately predicts these results.

5 Conclusion

We study imitation learning problems for which the expert has access to more information than the learner. We find that as long as the learner can *identify* this missing information and *recover* from the mistakes it made during identification, on-policy algorithms are able to match time-averaged expert value, in contrast to off-policy algorithms. We believe our results point towards a unified explanation of both empirical successes and failures.

Limitations. Our theoretical analyses primarily focus on the so-called asymptotic regime where T goes to infinity. For many applications, finite time bounds are equally important. There will also be value in deriving more refined *finite-sample* bounds. In particular, it will be interesting to see if there are concrete structure in the problem instances that can replace our asymptotic identifiability assumption. Moreover, our framework currently relies on the inherent randomness in the environment to establish identifiability of the context. While this is a natural assumption for many environments, we believe there is also a problem space of designing more sample-efficient exploration algorithms that leverage the structure of the underlying MDPs.

Societal impacts. Our paper aims to address an important issue in data-driven decision-making processes in the real world—that is, the data collection in many environments may miss contextual information that is critical to the problem. Our work focuses on this issue in the domain of imitation learning, and establishes the theoretical guarantees for each type of algorithms. We believe that our studies will promote appropriate reliance on different types of algorithms when we deploy imitation learning technologies in the real world.

References

J Andrew Bagnell. An invitation to imitation. Technical report, Carnegie-Mellon Univ Pittsburgh Pa Robotics Inst, 2015.

Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *CoRR*, abs/1812.03079, 2018. URL `http://arxiv.org/abs/1812.03079`.

Byron Boots, Sajid M Siddiqi, and Geoffrey J Gordon. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7):954–966, 2011.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating, 2019. URL `https://arxiv.org/abs/1912.12294`.

Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. Data-driven planning via imitation learning. *The International Journal of Robotics Research*, 37(13-14):1632–1672, 2018.

Felipe Codevilla, Eder Santana, Antonio M. López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. *CoRR*, abs/1904.08980, 2019. URL `http://arxiv.org/abs/1904.08980`.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016.

Hal Daumé, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine learning*, 75(3):297–325, 2009.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32:11698–11709, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. URL `https://arxiv.org/abs/1703.03400`.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes, 2015. URL `https://arxiv.org/abs/1502.02259`.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Alex Kuefler, Jeremy Morton, Tim Wheeler, and Mykel Kochenderfer. Imitating driver behavior with generative adversarial networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 204–211. IEEE, 2017.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots, 2021. URL `https://arxiv.org/abs/2107.04034`.

Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. One solution is not all you need: Few-shot extrapolation via structured maxent rl. *Advances in Neural Information Processing Systems*, 33:8198–8210, 2020.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 34, 2021.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

Jan Leike, Tor Lattimore, Laurent Orseau, and Marcus Hutter. Thompson sampling is asymptotically optimal in general environments. *arXiv preprint arXiv:1602.07905*, 2016.

Michael L Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*, pages 362–370. Elsevier, 1995.

Urs Muller, Jan Ben, Eric Cosatto, Beat Flepp, and Yann L Cun. Off-road obstacle avoidance through end-to-end learning. In *Advances in neural information processing systems*, pages 739–746. Citeseer, 2006.

Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. 1989.

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

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning, 2010. URL `https://arxiv.org/abs/1011.0686`.

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J. Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift, 2021. URL `https://arxiv.org/abs/2102.02872`.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

Guy Tennenholtz, Assaf Hallak, Gal Dalal, Shie Mannor, Gal Chechik, and Uri Shalit. On covariate shift of latent confounders in imitation and reinforcement learning. *arXiv preprint arXiv:2110.06539*, 2021.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in neural information processing systems*, 33:12263–12274, 2020.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]