PAC Imitation and Model-based Batch Learning of Contextual MDPs

Yash Nair 1  Finale Doshi-Velez 1

Abstract

We consider the problem of learning CMDPs in batch mode. In particular, we study two general classes of learning algorithms: direct policy learning (DPL), an imitation-learning based approach which learns from expert trajectories, and model-based learning. First, we derive sample complexity bounds for DPL, and then show that model-based learning from expert actions can, even with a finite model class, be impossible. After relaxing the conditions under which the model-based approach is expected to learn by allowing greater coverage of state-action space, we provide sample complexity bounds for model-based learning with finite model classes, showing that there exist model classes with sample complexity exponential in their statistical complexity. Our results give formal justification for imitation learning over model-based learning in this setting.

1. Introduction

Families of context-dependent tasks are common in many real-world settings. For example, controlling a UAV might depend on factors such as the parameters of the specific UAV’s weight and wingspan. After successfully controlling several different UAVs, one might hope to be able to control a new UAV quickly. Similarly, managing hypotension well may depend on some specific properties of the patient; after treating many distinct patients, one may hope to manage a new patient well.

The question of efficiently learning a collection of related, context-dependent tasks has been studied in the reinforcement learning (RL) literature under many names such as lifelong RL, multi-task RL, and, more generally, transfer learning (see, e.g., Isele et al. (2017), D’Eramo et al. (2019), and Taylor and Stone (2009)). Even more specifically, the question of learning to generalize from a collection of data has been considered, for example, in Brunskill and Li (2013) and Lazaric and Restelli (2011). These works consider the problem in an online setting and develop algorithmic contributions in the batch setting, respectively.

In this work, we consider the following setting and question: Suppose we are given a batch of trajectories obtained from experts in multiple contexts, where each context’s transition is parametrized by some observed parameter \( \theta \)—a framework called a Contextual MDP (CMDP). Will it be more sample efficient to directly learn a policy from these data (that is, imitate the expert), or to learn the transition function, parametrized by \( \theta \), and then plan according to it? We derive sample complexity bounds for direct policy learning (DPL); our upper bound for DPL is polynomial in all the relevant parameters.

Along the way, we prove impossibility results for learning certain transitions in the model-based paradigm, while our sample complexity upper bound for direct policy learning holds in a more general sense. Next, we show that, under a relaxed data generation process which affords greater coverage of state-action space, their exist hard families of CMDPs for model-based learning. Finally, we derive a distribution-dependent sample complexity upper bound for model-based learning. Our theory provides a formal justification for why imitation may be more successful than model-based learning in these settings, confirming trends observed in several more empirical and application-oriented works, including Yao et al. (2018) and Yang et al. (2019).

2. Related Work

Some papers have considered the problem when the context is observable. In particular, Isele et al. (2017) describe the problem as zero-shot transfer learning and provide an algorithm which, under certain linearity assumptions regarding the task descriptions, is able to perform zero-shot transfer on large classes of MDPs. While they consider the problem both empirically and theoretically, they provide only convergence results and not a finite-sample analysis.

Piot et al. (2013) consider Apprenticeship Learning, whereby the learner has access to a set of states and expert actions and is tasked with learning a policy via imitation. They upper bound the difference in the value function of the
learner with that of the expert in terms of the classification error of the learner’s hypothesis. Our work most notably differs from theirs in that we consider the contextual setting: DPL reduces to this work in the case when \(|\Theta| = 1\), and thus only a single MDP is being considered.

Other works have viewed the problem in the domain where the parameter \(\theta\) is latent. In particular, Lazaric and Restelli (2011) consider the problem of transfer learning between MDPs in the batch setting, where they design \(Q\)-value approximation algorithms designed to learn from a batch of trajectories to plan on a new MDP with potentially different transitions. We approach the problem of model-based learning in a slightly different way, and instead of following off the analysis of Chen and Jiang (2019). Finally, Yao et al. (2018) consider both a direct policy approach to the unobservable problem—Direct Policy Transfer (DPT)—as well as a model-based approach in an empirical setting. They evaluate both learning strategies on a 2D Navigation task, Acrobot, and a simulated HIV treatment domain, and find that, empirically, DPT is more sample-efficient than model-based learning.

3. Notation and Background

CMDPs A CMDP is a tuple \( (S,A,R,T,L,\gamma,\Theta,P_{s_0},P_\theta) \) where \( S \subseteq \mathbb{R}^n \) and \( A \) denote the state and (discrete) action space, respectively; \( R : S \times A \to \mathbb{R} \) is a reward function; \( \gamma \), a discount factor used in evaluating long-term return; \( \Theta \subseteq \mathbb{R}^d \); and \( T : S \times A \times \Theta \to \Delta(S) \), is a conditional density over next states given the current state, action, and value of \( \theta \) parametrizing the transition. The initial state \( s_0 \) is drawn from \( P_{s_0} \); at the start of an episode, the task parameter \( \theta \) is drawn from some \( P_\theta \). This is a slight simplification of the standard definition of CMDP of Hallak et al. (2015) in that we assume the reward and initial distribution remain the same.

Setting In this paper, we consider the case in which the learner is given a batch of \( m \) trajectories of length \( L \), each labeled with its associated context \( \theta \). There is one trajectory per parameter setting \( \theta \), corresponding to real settings in which one only gets to treat each patient once. This is in contrast to the setting of Hidden Parameter MDPs (HiP-MDPs), introduced by Doshi-Velez and Konidaris (2016), in which the parameter \( \theta \) is latent. We will further assume that the learner has access to the reward function, \( R \). Throughout this paper, we will assume that the trajectories are formed via following a deterministic (but possibly time-dependent) expert policy \( \pi \) that is \( \alpha \)-optimal; that is, \( v^L_\pi \geq v^L_\pi - \alpha \), where \( \pi^- \) is the optimal deterministic time-dependent policy and \( v^L_\pi := E[\sum_{l=0}^{L-1} r_l] \) is the expected undiscounted value associated with the first \( L \) rewards, where here the expecta-

tion also includes the randomness with respect to the draw of \( \theta \sim P_\theta \). We also define \( V^L_\pi(s;\theta) = E[\sum_{l=0}^{L-1} r_l|s_0 = s, \theta] \).

Goal Under this data generation process, the learner’s goal is to return a policy \( \hat{\pi} : S \times \Theta \times \{0, \ldots, L - 1\} \to A \), such that its value \( v^L_\pi \) is maximized. Specifically, we will define the error in value of a hypothesis to be the amount by which it is suboptimal. Following learning theory terminology, we shall also refer to \( \hat{\pi} \) as the hypothesis returned by the learner. We make two assumptions throughout: all rewards are in \([0, 1]\) and the hypothesis class \( \mathcal{H} \) contains the expert policy, \( \pi^- \).

4. Sample Complexity Bounds for DPL

We now turn to our learning problems. One approach to learning a hypothesis \( \hat{\pi} \) above is simply to treat the problem as a supervised learning problem and directly learn the association between the inputs—the states \( s \) and the task parameters \( \theta \)—and the expert’s action \( a \). In the following, we assume the learner is allowed to return any hypothesis from some hypothesis class, \( \mathcal{H} \) of functions from \( S \times \Theta \times \{0, \ldots, L - 1\} \to A \). In particular, just as in Yao et al. (2018), who consider DPT, we assume that DPL is agnostic to the reward sequence of the expert.

4.1. DPL Sample Complexity Upper Bound

We now derive a sample complexity upper bound for DPL. Our analysis is similar to that of the standard agnostic PAC learning upper bound, except that, in this setting, the batch of data are not i.i.d. but rather, come from a Markov chain. This, however, can be remedied, by simply replacing one of the key concentration inequalities in the standard setting (McDiarmid’s) with an analogous concentration inequality which applies to Markov chains, shown in Paulin et al. (2015). We state the main result below, whose proof can be found in Appendix Section A.1.

**Theorem 4.1.** Let the concept class \( \mathcal{H} \) have Natarajan dimension \( d \). There exists a learning algorithm \( A \) such that for any distribution over the data, there exists \( m \) that is

\[
O \left( \frac{L^d d}{\epsilon^2} \left( \log \left( \frac{L d}{\epsilon} \right) + \frac{d}{T} (\log(L + \log(|A|)) + L^2 \log(1/\delta)) \right) \right)
\]

receives at least \( m \) \( L \)-long trajectories in the batch, then \( A \) returns a hypothesis in \( \mathcal{H} \) which has error \(\epsilon\) (in terms of undiscounted value) at most \(\epsilon + \alpha\) with probability at least \(1 - \delta\).

4.2. DPL Sample Complexity Lower Bound

We derive a DPL lower bound by constructing a family of CMDPs for which the problem reduces to a standard PAC learning problem that must be learned to error \( \epsilon/L \) with confidence \( \delta \). To do so, we put all the decision making power on the action taken at the first state. We state the theorem below: For a proof, see Appendix A.2.
Theorem 4.2. There exist a family of CMDPs, a hypothesis class, $H$, of Natarajan dimension $d$, and value of $m = \Omega \left( \frac{(d+\log(1/\delta))}{\epsilon} \right)$, such that any learning algorithm given at most $m$ $L$-long trajectories returns a policy whose error (in terms of value) is at most $\epsilon$ with probability at most $1 - \delta$.

Our results thus give a $O \left( \frac{L^2 d}{\epsilon} + \frac{\epsilon^5}{\delta} \right)$ separation between upper and lower bounds, in particular, highlighting how poorly the bounds scale with $L$, but how slowly they separate with $\delta$ and $A$; we discuss this further in Section 6.

5. Model-based Approach

In contrast to DPL, the model-based approach does not attempt to directly learn the expert policy, but rather attempts to learn the transition function parametrized by $\theta$, and then plan according to the transition and reward function. One might believe that learning models would be more general than trying to directly learn policies, because one can use them to explore counterfactuals. Indeed, model-based learning is often the go-to approach in low data regimes (see, e.g. Rasmussen (2003) versus Deisenroth et al. (2013), Kamthe and Deisenroth (2017), Kocijan et al. (2004), Ko et al. (2007)). However, we first show that this paradigm of learning from expert actions is, in some finite settings, impossible. Thus, we relax the data generation process under which the model-based method is expected to learn. Throughout our analysis, for simplicity, we will assume the model-based approach has access to an oracle called PLAN, which, upon receiving a transition function, $T$, and reward function, $R$ (which, as mentioned in Section 3, the learner has access to), returns an optimal deterministic (possible time-dependent) policy under $T$ and $R$.

5.1. Impossibility of Model-based Learning via Expert Actions

The primary issue a model-based approach having access only to an expert’s trajectory is the lack of coverage of state-action space. In particular, it may be the case that a large portion of the learner’s hypothesis class always agree on the subset of state-action space traversed by the expert, making it hard—and, in some cases impossible—to output a low-error hypothesis with high confidence. We relegate the construction to Appendix Section A.3 for space.

Theorem 5.1. There exist classes of CMDPs which, if the learner must learn from expert trajectories, require infinite sample complexity even with a finite hypothesis class containing the true model.

Thus, while DPL can achieve suboptimality arbitrarily near that of the expert with sufficiently many samples for finite hypothesis classes, the same cannot be said of model-based learning.

5.2. Hardness of Model-based Learning under strictly-positive visitation distributions

The impossibility result above motivates the use of the following more standard framework under which we expect batch model-based approaches to learn:

Definition 5.1.1 (Model-based Learning Data Generation Process). Let $\mu$ be some distribution over $S \times A$ which assigns non-zero mass/density to every $(s, a) \in S \times A$. For each value of $\theta \sim \mathcal{P}_\Theta$, the model-based approach draws $L$ pairs $(s, a) \sim (s, a) \sim \mu$, and, for each pair, draws $s' \sim T(\cdot | s, a, \theta)$ and $r = R(s, a)$. The model-based approach then has access to each of these one-step trajectories, labelled by $\theta$. We also make the assumption that every element of $S \times A$ is reachable in at most $L$ steps.

Under the framework defined in Definition 5.1.1, we show that there are still classes for which model-based learning is hard. In fact, the construction of Chen and Jiang (2019) for single MDPs gives a lower bound in our setting as well since we can simply consider the CMDP which concentrates all its mass on a single value of $\theta$, thus reducing to the single MDP case. In their construction, states are the nodes of a complete tree with branching factor $|A|$, and all leaf nodes give Bern$(1/2)$ rewards while the leaf node corresponding to the special edge gives Bern$(1/2 + 3\epsilon/2)$ reward. In our setting, the construction of Krishnamurthy et al. (2016) yields a sample complexity lower bound of $\Omega \left( \frac{|A|}{L^{2+\epsilon/2}} \right)$ when active exploration is allowed. In Appendix Section A.4, we give a construction of a family of MDPs, motivated by contextual bandits and the constructions of Krishnamurthy et al. (2016) and Auer et al. (2003), which yields a sample complexity lower bound of $\Omega \left( \frac{|A|}{L^{2+\epsilon/2}} \right)$, allowing exploration for a hypothesis class of cardinality $|A|$. Our bound, on the surface, is asymptotically lower than that of Krishnamurthy et al. (2016); however, since the hypothesis class as well as the size of each CMDP in the class have size only $O(|A|)$, our bound exhibits a stronger dependence on the size of the model class. That is, our second bound gives a $\Omega \left( \frac{|A|}{L^{2+\epsilon/2}} \right)$ dependence on the size of the hypothesis class, rather than the immediate $\Omega \left( \frac{|A|}{L^{2+\epsilon/2}} \right)$ of Krishnamurthy et al. (2016).

The above constructions show how the sample complexity for model-based learning can scale poorly with both horizon as well as the size and statistical complexity of the hypothesis class, but fail to show that model-based learning can scale poorly with $|\Theta|$ when it is finite. Modi et al. (2017), who consider online CMDPs, suggest constructing hard CMDPs by making the MDP for each context hard and disallowing any information corresponding to one context be useful to another. While this technique gives a generic way to increase any hard MDP lower bound by a multiplicative factor of $O(|\Theta|)$ in expectation for CMDPs, it makes the
We now derive an upper bound for model-based learning. Let us state our model-based lower bounds. The latter extend off of Krishnamurthy et al. (2016), which yields the same lower bound, but does so in a way that scales linearly with \(|\Theta|\), while having a hypothesis class of cardinality \(|\Theta|\).

We now state our model-based lower bounds. The latter extend off of Krishnamurthy et al. (2016), which yields the same lower bound, but does so in a way that scales linearly with \(|\Theta|\), while having a hypothesis class of cardinality \(|\Theta|\).

We now derive an upper bound for model-based learning to contrast the above lower bound and, in particular, show that the dependence on \(C\) is in fact linear. To do so, we extend off the work of Chen & Jiang (2019), who derive an upper bound for Fitted Q-Iteration (FQI). The FQI sample complexity upper bound then immediately yields a sample complexity upper bound for FQI using techniques of Chen and Jiang (2019), and then derive, as an immediate corollary, a sample complexity upper bound for model-based learning; again, we assume realizability for the hypothesis class \(\mathcal{H}\) and thus of the class \(\mathcal{F}\). We state the corollary below: for the FQI upper bound as well as proofs, see Appendix A.5.

**Corollary 5.2.1** (Model-based Upper Bound). Given the finite hypothesis model class \(\mathcal{H}\), there exists a model-based learning algorithm \(A\) and \(m\) with \(m = O\left(\frac{C |\mathcal{H}|^2 \log |\mathcal{H}|}{\epsilon^2}\right)\), such that if \(A\) receives at least \(m\) samples of \(L\) one-step trajectories under \(\mu\), then it returns a policy with error at most \(\epsilon\) with probability at least \(1 - \delta\).

### 6. Discussion

In this paper we investigate the sample complexities of imitation learning of CMDPs, as well as model-based learning. Our results indicate that DPL is, theoretically, more sound than model-based approaches in that the latter scales with respect to the concentratability coefficient of the distribution \(\mu\). While both upper bounds scale polynomially in all the relevant parameters—and, importantly, in the complexity of hypothesis class—our upper bound for model-based learning scales with \(C\). As our lower bound for model-based learning shows, this additional dependence is, in fact, necessary. This highlights the importance of the data generation process for model-based learning: When data is not sufficient from expert trajectories, model-based learning can be impossible even with finite hypothesis classes, but even when data is drawn i.i.d. from the distribution \(\mu\), model-based learning depends greatly on the coverage of reachable state-action pairs.

We believe the following are primary interests for future work: Deriving general model-based sample complexity upper bounds which do not grow, even logarithmically with \(|\mathcal{H}|\), but rather grow with some other complexity measure of the hypothesis class which can be finite even for infinite \(\mathcal{H}\) (e.g., perhaps with an extension of witness rank introduced in Sun et al. (2018)); investigating a tighter relationship between the upper and lower bounds for DPL, in particular, bounds whose degree of separation scales more slowly with \(L\); and understanding the sample complexity of similar imitation learning and model-based algorithms in the unobserved parameter setting of HiP-MDPs.
References


Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised


