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 θ—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 θ, 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

[1]Harvard University, Cambridge, Massachusetts, USA. Correspondence to: Yash Nair <yashnair@college.harvard.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

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 θ 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 extend 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 $\langle S, A, R, T, L, \gamma, \Theta, \mathcal{P}_{s_0}, \mathcal{P}_\Theta \rangle$ where $S \subset \mathbb{R}^n$ and A denote the state and (discrete) action space, respectively; $R : S \times A \to \mathbb{R}$ is a reward function; γ, a discount factor used in evaluating long-term return; $\Theta \subset \mathbb{R}^d$; 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 θ parametrizing the transition. The initial state s_0 is drawn from $\mathcal{P}_{s_0}$; at the start of an episode, the task parameter θ is drawn from some $\mathcal{P}_\theta$. 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 θ. There is one trajectory per parameter setting θ, 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 θ 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 π that is α-optimal; that is, $v_\pi^L \geq v_{\pi^*}^L - \alpha$, where π^* is the optimal deterministic time-dependent policy and $v_\pi^L := \mathbb{E}[\sum_{l=0}^{L-1} r_l]$ is the expected undiscounted value associated with the first L rewards, where here the expectation also includes the randomness with respect to the draw of $\theta \sim \mathcal{P}_\Theta$. We also define $V_\pi^L(s;\theta) = \mathbb{E}[\sum_{l=0}^{L-1} r_l | s_0 = s, \theta]$.

Goal Under this data generation process, the learner's goal is to return a policy $\hat{\pi} : S \times \Theta \times \{0, \dots, L-1\} \to A$, such that its value $v_{\hat{\pi}}^l$ 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, π.

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 θ—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 $S \times \Theta \times \{0, \dots, L-1\} \to A$. 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 $\mathcal{A}$ such that for any distribution over the data, there exists m that is*

$$O\left(\frac{L^4 d}{\epsilon^2}\left(\log\left(\frac{Ld}{\epsilon}\right) + \frac{d}{L}(\log(L) + \log(|A|)) + L^2 \log(1/\delta)\right)\right)$$

receives at least m L-long trajectories in the batch, then $\mathcal{A}$ returns a hypothesis in $\mathcal{H}$ which has error (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 ϵ/L with confidence δ. 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, $\mathcal{H}$ of Natarajan dimension d, and value of $m = \Omega\left(\frac{L(d+\log(1/\delta))}{\epsilon}\right)$, such that any learning algorithm given at most m L-long trajectories returns a policy whose error (in terms of value) is at most ϵ with probability at most $1-\delta$.*

Our results thus give a $\tilde{O}\left(\frac{L^2 d}{\epsilon} + \frac{L^5}{\epsilon}\right)$ separation between upper and lower bounds, in particular, highlighting how poorly the bounds scale with L, but how slowly they separate with δ 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 θ, 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 μ 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 $(s,a) \overset{i.i.d}{\sim} \mu$ 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 θ. 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 θ, 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 $\Omega\left(\frac{|A|^L}{L\epsilon^2}\right)$ 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 $\Omega\left(\frac{|A|}{\epsilon^2}\right)$, 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 $\Omega\left(\frac{|\mathcal{H}|}{\epsilon^2}\right)$ dependence on the size of the hypothesis class, rather than the immediate $\Omega\left(\frac{|\mathcal{H}|}{L\epsilon^2}\right)$ 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

hypothesis class have cardinality exponential in $|\Theta|$, thus explaining away the factor of $O(|\Theta|)$. We give a construction of a class of CMDPs in Appendix Section A.4, extending 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 bound is in terms of, C, the concentratability coefficient of μ, which measures how much μ covers reachable state-action pairs (for a definition, see Section 5.3).

Theorem 5.2. *There exist hard families of CMDPs which are subject to the following sample complexity lower bounds (all are asymptotically at most $\Omega\left(\frac{|A|^L}{L\epsilon^2}\right)$): $\Omega\left(\frac{|A|^L}{L\epsilon^2}\right)$, $\Omega\left(\frac{|\mathcal{H}|}{\epsilon^2}\right)$, and $\Omega\left(\frac{|\Theta|}{L\epsilon^2}\right)$ with $|\Theta| = |\mathcal{H}|$. Furthermore, when active exploration is not allowed, we have the following lower bound, in expectation, with respect to the randomness of drawing a leaf state from μ: $\Omega\left(\frac{C|A|^L}{L\epsilon^2}\right)$*

5.3. Model-based Learning Sample Complexity Upper Bound

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 any model-based approach with a finite model class.

We give a brief outline of finite-horizon FQI on a CMDP below; it is essentially the same as FQI on a single MDP except that Bellman backups are done with respect to the context, θ. We first define this back-up and give the algorithm below:

Definition 5.2.1. *Define the lth Bellman backup of $f : S \times \Theta \times \{0, \dots, L\} \to \mathbb{R}$ with respect to θ to be $(\mathcal{T}_l(\theta)f)(s,a) = R(s,a) + \mathbb{E}_{s' \sim T(\cdot|s,a,\theta)} V_f(s', \theta, l)$, where $V_f(s, \theta, l) = \max_{a \in A} f(s, a, \theta, l)$.*

We will also assume FQI has access to a family of time-indexed Q functions. That is we have a set $\mathcal{F}$ which contains Q-value functions of the form $Q : S \times A \times \Theta \times \{0, \dots, L\}$, where $Q(s, a, \theta, l) = (\mathcal{T}_{l-1}(\theta)Q)(s, a)$ for $l \geq 1$, and $Q(s, a, \theta, 0) = 0$. FQI on CMDPs operates in essentially the same way as the finite horizon case except that all backups and value functions are additionally parametrized by θ (see Appendix Section A.5 for pseudocode). We now give the definition of admissible distribution and the assumption of concentratability of the data disitribution μ which extends that of Chen & Jiang (2019).

Definition 5.2.2 (Admissible Distribution)**.** *A conditional distribution ν over $S \times A$ given $\theta \in \Theta$ is said to be admissible if there exists $0 \leq l \leq L-1$ if there exists a possibly time-dependent stochastic policy π such that $(\nu(\theta))(s,a) = P[s_l = s, a_h = a | \theta, s_0 \sim \mathcal{P}_{s_0}, \pi]$*

Assumption 5.1 (Concentratability)**.** *We assume that there exists some $C < \infty$ such that, for any admissible distribution ν, $\frac{(\nu(\theta))(s,a)}{\mu(s,a)} \leq C, \forall (s, a, \theta) \in S \times A \times \Theta$.*

With these definitions, we are able to derive 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 $\mathcal{A}$ and m with $m = O\left(\frac{CL^6 \log(L|\mathcal{H}|)}{\epsilon^2}\right)$, such that if $\mathcal{A}$ receives at least m samples of L one-step trajectories under μ, then it returns a policy with error at most ϵ 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 μ. 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 gotten 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 μ, 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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):408–423, 2013.

Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2019.

Finale Doshi-Velez and George Konidaris. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *IJCAI: proceedings of the conference*, volume 2016, page 1432. NIH Public Access, 2016.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes, 2015.

David Isele, Mohammad Rostami, and Eric Eaton. Using task descriptions in lifelong machine learning for improved performance and zero-shot transfer. *CoRR*, abs/1710.03850, 2017. URL http://arxiv.org/abs/1710.03850.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

Sanket Kamthe and Marc Peter Deisenroth. Data-efficient reinforcement learning with probabilistic model predictive control. *arXiv preprint arXiv:1706.06491*, 2017.

Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. In *Advances in neural information processing systems*, pages 6250–6261, 2017.

Jonathan Ko, Daniel J Klein, Dieter Fox, and Dirk Haehnel. Gaussian processes and reinforcement learning for identification and control of an autonomous blimp. In *Proceedings 2007 ieee international conference on robotics and automation*, pages 742–747. IEEE, 2007.

Juš Kocijan, Roderick Murray-Smith, Carl Edward Rasmussen, and Agathe Girard. Gaussian process model based predictive control. In *Proceedings of the 2004 American control conference*, volume 3, pages 2214–2219. IEEE, 2004.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Alessandro Lazaric and Marcello Restelli. Transfer from multiple mdps. *CoRR*, abs/1108.6211, 2011. URL http://arxiv.org/abs/1108.6211.

Aditya Modi and Ambuj Tewari. Contextual markov decision processes using generalized linear models. *arXiv preprint arXiv:1903.06187*, 2019.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. *arXiv preprint arXiv:1711.05726*, 2017.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*, 2018.

Daniel Paulin et al. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

Bilal Piot, Matthieu Geist, and Olivier Pietquin. Learning from demonstrations: Is it worth estimating a reward function? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 17–32. Springer, 2013.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

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, 2010.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Sungryull Sohn, Junhyuk Oh, and Honglak Lee. Hierarchical reinforcement learning for zero-shot generalization with subtask dependencies. In *Advances in Neural Information Processing Systems*, pages 7156–7166, 2018.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. *arXiv preprint arXiv:1811.08540*, 2018.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

Jiachen Yang, Brenden Petersen, Hongyuan Zha, and Daniel Faissol. Single episode policy transfer in reinforcement learning. *arXiv preprint arXiv:1910.07719*, 2019.

Jiayu Yao, Taylor Killian, George Konidaris, and Finale Doshi-Velez. Direct policy transfer via hidden parameter markov decision processes. In *LLARLA Workshop, FAIM*, volume 2018, 2018.