Provable Hierarchical Imitation Learning via EM

Zhiyu Zhang
Boston University

Ioannis Ch. Paschalidis
Boston University

Abstract

Due to recent empirical successes, the options framework for hierarchical reinforcement learning is gaining increasing popularity. Rather than learning from rewards, we consider learning an options-type hierarchical policy from expert demonstrations. Such a problem is referred to as *hierarchical imitation learning*. Converting this problem to parameter inference in a latent variable model, we develop convergence guarantees for the EM approach proposed by Daniel et al. (2016b). The population level algorithm is analyzed as an intermediate step, which is nontrivial due to the samples being correlated. If the expert policy can be parameterized by a variant of the options framework, then, under regularity conditions, we prove that the proposed algorithm converges with high probability to a norm ball around the true parameter. To our knowledge, this is the first performance guarantee for an hierarchical imitation learning algorithm that only observes primitive state-action pairs. [1]

1 Introduction

Recent empirical studies (Kulkarni et al., 2016; Tessler et al., 2017; Vezhnevets et al., 2017; Nachum et al., 2018) have shown that the scalability of Reinforcement Learning (RL) algorithms can be improved by incorporating hierarchical structures. As an example, consider the *options* framework (Sutton et al., 1999) representing a two-level hierarchical policy: with a set of multi-step low level procedures (options), the high level policy selects an option, which, in turn, decides the primitive action applied at each time step until the option terminates. Learning such a hierarchical policy from environmental feedback effectively breaks the overall task into sub-tasks, each easier to solve.

Researchers have investigated the hierarchical RL problem under various settings. Existing theoretical analyses (Brunskill and Li, 2014; Mann and Mannor, 2014; Fruit and Lazaric, 2017; Fruit et al., 2017) typically assume that the options are given. As a result, only the high-level policy needs to be learned. Recent advances in deep hierarchical RL (e.g., Bacon et al. 2017) focus on concurrently learning the full options framework, but still the initialization of the options is critical. A promising practical approach is to learn an initial hierarchical policy from expert demonstrations. Then, deep hierarchical RL algorithms can be applied for policy improvement. The former step is named as *Hierarchical Imitation Learning* (HIL).

Due to its practicality, HIL has been extensively studied within the deep learning and robotics communities. However, existing works typically suffer from the following limitations. First, the considered HIL formulations often lack rigor and clarity. Second, existing works are mostly empirical, only testing on a few specific benchmarks. Without theoretical justification, it remains unclear whether the proposed methods can be generalized beyond their experimental settings.

In this paper, we investigate HIL from a theoretical perspective. Our problem formulation is concise while retaining the essential difficulty of HIL: we need to learn a complete hierarchical policy from an *unsegmented* sequence of state-action pairs. Under this setting, HIL becomes an inference problem in a latent variable model. Such a transformation was first proposed by Daniel et al. (2016b), where the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) was applied for policy learning. Empirical results for this algorithm and its gradient variants (Fox et al., 2017; Krishnan et al., 2017) demonstrate good performance, but the theoretical analysis remains open. By bridging this gap, we aim to solidify the foundation of HIL and provide some high level guidance for its practice.

[1] Complete version available at `https://arxiv.org/abs/2010.03133`.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1.1 Related work

Due to its intrinsic difficulty, existing works on HIL typically consider its easier variants for practicality. If the expert options are observed, standard imitation learning algorithms can be applied to learn the high and low level policies separately (Le et al., 2018). If those are not available, a popular idea (Butterfield et al., 2010; Niekum et al., 2012; Manschitz et al., 2014; Niekum et al., 2015) is to first divide the expert demonstration into segments using domain knowledge or heuristics, learn the individual option corresponding to each segment, and finally learn the high level policy. With additional supervision, these steps can be unified (Shiarlis et al., 2018). In this regard, the EM approach (Daniel et al., 2016b; Fox et al., 2017; Krishnan et al., 2017) is this particular idea pushed to an extreme: without any other forms of supervision, we simultaneously segment the demonstration and learn from it, by exploiting the latent variable structure.

From the theoretical perspective, inference in parametric latent variable models is a long-standing problem in statistics. For many years the EM algorithm has been considered the standard approach, but performance guarantees (McLachlan and Krishnan, 2007; Wu, 1983) were generally weak, only characterizing the convergence of parameter estimates to stationary points of the finite sample likelihood function. Under additional local assumptions, convergence to the Maximum Likelihood Estimate (MLE) can be further established. However, due to the randomness in sampling, the finite sample likelihood function is usually highly non-concave, leading to stringent requirements on initialization. Another weakness is that converging to the finite sample MLE does not directly characterize the distance to the maximizer of the population likelihood function which is the true parameter.

Recent ideas on EM algorithms (Wang et al., 2015; Yi and Caramanis, 2015; Balakrishnan et al., 2017; Yang et al., 2017) focus on the convergence to the true parameter directly, relying on an instrumental object named as the *population EM algorithm*. It has the same two-stage iterative procedure as the standard EM algorithm, but its *Q-function*, the maximization objective in the M-step, is defined as the infinite sample limit of the finite sample Q-function. Under regularity conditions, the population EM algorithm converges to the true parameter. The standard EM algorithm is then analyzed as its perturbed version, converging with high probability to a norm ball around the true parameter. The main advantage of this approach is that the true parameter usually has a large basin of attraction in the population EM algorithm. Therefore, the requirement on initialization is less stringent. See Figure 1 of (Yang et al., 2017) for an illustration.

The Q-function adopted in the population EM algorithm is named as the *population Q-function*. To properly define such a quantity, the stochastic convergence of the finite sample Q-function needs to be constructed. When the samples are i.i.d., such as in Gaussian Mixture Models (GMMs) (Xu et al., 2016; Daskalakis et al., 2017; Balakrishnan et al., 2017), the required convergence follows directly from the law of large numbers. However, this argument is less straightforward in time-series models such as Hidden Markov Models (HMMs) and the model considered in HIL. For HMMs, Yang et al. (2017) showed that the expectation of the Q-function converges, but both the stochastic convergence analysis and the analytical expression of the population Q-function are not provided. The missing techniques could be borrowed from a body of work (Cappé et al., 2006; van Handel, 2008; Le Corff and Fort, 2013; De Castro et al., 2017) analyzing the asymptotic behavior of HMMs. Most notably, Le Corff and Fort (2013) provided a rigorous treatment of the population EM algorithm via sufficient statistics, assuming the HMM is parameterized by an exponential family.

Finally, apart from the EM algorithm, a separate line of research (Hsu et al., 2012; Anandkumar et al., 2014) applies spectral methods for tractable inference in latent variable models. However, such methods are mainly complementary to the EM algorithm since better performance can usually be obtained by initializing the EM algorithm with the solution of the spectral methods (Kontorovich et al., 2013).

1.2 Our contributions

In this paper, we establish the first known performance guarantee for a HIL algorithm that only observes primitive state-action pairs. Specifically, we first fix and reformulate the original EM approach by Daniel et al. (2016b) in a rigorous manner. The lack of mixing is identified as a technical difficulty in learning the standard options framework, and a novel *options with failure* framework is proposed to circumvent this issue.

Inspired by Balakrishnan et al. (2017) and Yang et al. (2017), the population version of our algorithm is analyzed as an intermediate step. We prove that if the expert policy can be parameterized by the options with failure framework, then, under regularity conditions, the population version algorithm converges to the true parameter, and the finite sample version converges with high probability to a norm ball around the true parameter. Our analysis directly constructs the stochastic convergence of the finite sample Q-function, and an analytical expression of the resulting population Q-function is provided. Finally, we qualitatively validate our theoretical results using a numerical example.

2 Problem settings

Notation. Throughout this paper, we use uppercase letters (e.g., S_t) for random variables and lowercase letters (e.g., s_t) for values of random variables. Let $[t_1 : t_2]$ be the set of integers t such that $t_1 \leq t \leq t_2$. When used in the subscript, the brackets are removed (e.g., $S_{t_1:t_2} = \{S_t\}_{t_1 \leq t \leq t_2}$).

2.1 Definition of the hierarchical policy

In this section, we first introduce the options framework for hierarchical reinforcement learning (Sutton et al., 1999; Barto and Mahadevan, 2003), captured by the probabilistic graphical model shown in Figure 1. The index t represents the time; (S_t, A_t, O_t, B_t) respectively represent the state, the action, the option and the termination indicator. For all t, S_t, A_t and O_t are defined on the finite state space $\mathcal{S}$, the finite action space $\mathcal{A}$ and the finite option space $\mathcal{O}$; B_t is a binary random variable. Define the parameter $\theta := (\theta_{hi}, \theta_{lo}, \theta_b)$ where $\theta_{hi} \in \Theta_{hi}$, $\theta_{lo} \in \Theta_{lo}$, and $\theta_b \in \Theta_b$. The parameter space $\Theta := \Theta_{hi} \times \Theta_{lo} \times \Theta_b$ is a convex and compact subset of a Euclidean space.

For any $(o_0, s_1) \in \mathcal{O} \times \mathcal{S}$, if we fix $(O_0, S_1) = (o_0, s_1)$ and consider a given θ, the joint distribution on the rest of the graphical model is determined by the following components: an unknown environment transition probability P, a high level policy π_{hi} parameterized by θ_{hi}, a low level policy π_{lo} parameterized by θ_{lo} and a termination policy π_b parameterized by θ_b. Sampling a tuple $(s_{2:T}, a_{1:T}, o_{1:T}, b_{1:T})$ from such a joint distribution, or equivalently, implementing the hierarchical decision process, follows the following procedure. Starting from the first time step, the decision making agent first determines whether or not to terminate the current option o_0. The decision is encoded in a termination indicator b_1 sampled from $\pi_b(\cdot|s_1, o_0; \theta_b)$. $b_1 = 1$ indicates that the option o_0 terminates and the next option o_1 is sampled from $\pi_{hi}(\cdot|s_1; \theta_{hi})$; $b_1 = 0$ indicates that the option o_0 continues and $o_1 = o_0$. Next, the primitive action a_1 is sampled from $\pi_{lo}(\cdot|s_1, o_1; \theta_{lo})$, applying the low level policy associated with the option o_1. Using the environment, the next state s_2 is sampled from $P(\cdot|s_1, a_1)$. The rest of the samples $(s_{3:T}, a_{2:T}, o_{2:T}, b_{2:T})$ are generated analogously.

The options framework corresponds to the above hierarchical policy structure and the policy triple $\{\pi_{hi}, \pi_{lo}, \pi_b\}$. However, due to a technicality identified at the end of this subsection, we consider a novel options with failure framework for the remainder of this paper, which adds an extra *failure* mechanism to the graphical model in the case of $b_t = 0$. Specifically, there exists a constant $0 < \zeta < 1$ such that when the termination indicator $b_t = 0$, with probability $1 - \zeta$ the next option o_t is assigned to o_{t-1}, whereas with probability ζ the next option o_t is sampled uniformly from the set of options $\mathcal{O}$. Notice that if $\zeta = 0$, we recover the standard options framework.

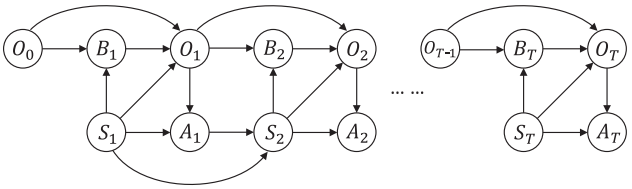

Figure 1: A graphical model for hierarchical reinforcement learning.

To simplify the notation, we define $\bar{\pi}_{hi}$ as the combination of π_{hi} and the failure mechanism. For any θ_{hi}, with any other input arguments,

$$\bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) := \begin{cases} \pi_{hi}(o_t|s_t; \theta_{hi}), & \text{if } b_t = 1, \\ 1 - \zeta + \frac{\zeta}{|\mathcal{O}|}, & \text{if } b_t = 0,\ o_t = o_{t-1}, \\ \frac{\zeta}{|\mathcal{O}|}, & \text{if } b_t = 0,\ o_t \neq o_{t-1}. \end{cases}$$

Formally, the options with failure framework is defined as the class of policy triples $\{\bar{\pi}_{hi}, \pi_{lo}, \pi_b\}$ parameterized by ζ and θ. With $(O_0, S_1) = (o_0, s_1)$ and a given θ, let $\mathbb{P}_{\theta, o_0, s_1}$ be the joint distribution of $\{S_{2:T}, A_{1:T}, O_{1:T}, B_{1:T}\}$. With any input arguments,

$$\mathbb{P}_{\theta,o_0,s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_{1:T} = o_{1:T}, B_{1:T} = b_{1:T}) = \left[\prod_{t=1}^{T-1} P(s_{t+1}|s_t, a_t)\right] \left[\prod_{t=1}^{T} \pi_b(b_t|s_t, o_{t-1}; \theta_b) \times \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t; \theta_{lo})\right].$$

On the policy framework. The options with failure framework is adopted to simplify the construction of the mixing condition (Lemma D.1). It is possible that our analysis could be extended to learn the standard options framework. In that case, instead of constructing the usual one step mixing condition, one could target the multi-step mixing condition similar to (Cappé et al., 2006, Chap. 4.3).

2.2 The imitation learning problem

Suppose an expert uses an options with failure policy with true parameters ζ and $\theta^* = (\theta^*_{hi}, \theta^*_{lo}, \theta^*_b)$; its initial condition (o_0, s_1) is sampled from a distribution ν^*. A finite length observation sequence $\{s_{1:T}, a_{1:T}\} = \{s_t, a_t\}_{t=1}^T$ with $T \geq 2$ is observed from the expert. ζ and the parametric structure of the expert policy are

known, but ν^* is unknown. Our objective is to estimate θ^* from $\{s_{1:T}, a_{1:T}\}$.

On the practicality of our setting. Two comments need to made here. First, it is common in practice to observe not one, but a set of independent observation sequences. In that case, the problem essentially becomes easier. Second, the cardinality of the option space and the parameterization of the expert policy are usually unknown. A popular solution is to assume an expressive parameterization (e.g., a neural network) in the algorithm and select card($\mathcal{O}$) through cross-validation. Theoretical analysis of EM under this setting is challenging, even when samples are i.i.d. (Dwivedi et al., 2018a,b). Therefore, we only consider the domain of *correct-specification*.

Throughout this paper, the following assumptions are imposed for simplicity.

Assumption 1 (Non-degeneracy)**.** *With any other input arguments, the domain of π_{hi}, π_{lo} and π_b as functions of θ can be extended to an open set $\tilde{\Theta}$ that contains Θ. Moreover, for all $\theta \in \tilde{\Theta}$, π_{hi}, π_{lo} and π_b parameterized by θ are strictly positive.*

Assumption 2 (Differentiability)**.** *With any other input arguments, π_{hi}, π_{lo} and π_b as functions of θ are continuously differentiable on $\tilde{\Theta}$.*

Next, consider the stochastic process $\{O_{t-1}, S_t\}_{t=1}^{\infty}$ induced by ν^* and the expert policy. Based on the graphical model, it is a Markov chain with finite state space $\mathcal{O} \times \mathcal{S}$. Let Π_{θ^*} be its set of stationary distributions, which is nonempty and convex.

Assumption 3 (Stationary initial distribution)**.** *ν^* is an extreme point of Π_{θ^*}. That is, $\nu^* \in \Pi_{\theta^*}$, and it cannot be written as the convex combination of two elements of Π_{θ^*}.*

On the assumptions. The first two assumptions are generally mild and therefore hold for many policy parameterizations. The third assumption is a bit more restrictive, but it is essential for our theoretical analysis. In Appendix A, we provide further justification of this assumption in a particular class of environments: $\forall s_t, s_{t+1} \in \mathcal{S}$, there exists $a_t \in \mathcal{A}$ such that $P(s_{t+1}|s_t, a_t) > 0$. In such environments, Π_{θ^*} contains a unique element which is also the limiting distribution. If we start sampling the observation sequence late enough, Assumption 3 is approximately satisfied.

3 A Baum-Welch type algorithm

Adopting the EM approach, we present Algorithm 1 for the estimation of θ^*. It reformulates the algorithm by Daniel et al. (2016b) in a rigorous manner, and an error in the latter is fixed: when defining the posterior distribution of latent variables, at any time $t < T$, the original algorithm neglects the dependency of future states $S_{t+1:T}$ on the current option O_t. A detailed discussion is provided in Appendix B.1.

Algorithm 1 A Baum-Welch type algorithm for provable hierarchical imitation learning

Require: Observation sequence $\{s_{1:T}, a_{1:T}\}$; a probability mass function $\mu(o_0|s_1)$ on $o_0 \in \mathcal{O}$; $N \in \mathbb{N}_+$; $\theta^{(0)} \in \Theta$.

1: **for** $n = 1, \ldots, N$ **do**
2: Compute the forward message $\{\alpha_{\mu,t}^{\theta^{(n-1)}}\}_{t=1}^{T}$ and the backward message $\{\beta_{t|T}^{\theta^{(n-1)}}\}_{t=1}^{T}$ according to (1), (2), (3) and (4).
3: Compute the smoothing distributions $\{\gamma_{\mu,t|T}^{\theta^{(n-1)}}\}_{t=1}^{T}$ and $\{\tilde{\gamma}_{\mu,t|T}^{\theta^{(n-1)}}\}_{t=2}^{T}$ according to (5) and (6).
4: Update the next parameter estimate $\theta^{(n)} \in \arg\max_{\theta\in\Theta} Q_{\mu,T}(\theta|\theta^{(n-1)})$ according to (7).
5: **end for**

Since our graphical model resembles an HMM, Algorithm 1 is intuitively similar to the classical Baum-Welch algorithm (Baum et al., 1970) for HMM parameter inference. Analogously, it iterates between forward-backward smoothing and parameter update. In each iteration, the algorithm first estimates certain marginal distributions of the latent variables $(O_{1:T}, B_{1:T})$ conditioned on the observation sequence $\{s_{1:T}, a_{1:T}\}$, assuming the current estimate of θ is correct. Such conditional distributions are named as *smoothing distributions*, and they are used to compute the Q-function, which is a surrogate of the likelihood function. The next estimate of θ is assigned as one of the maximizing arguments of the Q-function.

From the structure of our graphical model, a prior distribution of (O_0, S_1) is required to compute the smoothing distributions. Since the true prior distribution ν^* is unknown, $\hat{\nu}$, defined next, is used as its approximation: $\forall o_0 \in \mathcal{O}$, $\hat{\nu}(o_0, s_1) := \mu(o_0|s_1)$; $\forall s_1' \neq s_1$, $\hat{\nu}(o_0, s_1') := 0$. Theorem 2 shows that the additional estimation error introduced by this approximation vanishes as $T \to \infty$, regardless of the choice of μ. Let $\mathcal{M}$ be the set of μ allowed by Algorithm 1.

3.1 Latent variable estimation

In the following, we define the forward message, the backward message and the smoothing distribution for all θ, μ and all $t \in [1:T]$. All of these quantities are probability mass functions over $\mathcal{O} \times \mathcal{S}$, and normalizing constants $z_{\alpha,\mu,t}^{\theta}$, $z_{\beta,t}^{\theta}$ and $z_{\gamma,\mu}^{\theta}$ are adopted to enforce this. With any input arguments o_t and b_t, the forward

message is defined as

$$\alpha^{\theta}_{\mu,t}(o_t, b_t) := z^{\theta}_{\alpha,\mu,t}\mathbb{E}_{O_0\sim\mu(\cdot|s_1)}[\mathbb{P}_{\theta,O_0,s_1}(S_{2:t} = s_{2:t}, A_{1:t} = a_{1:t}, O_t = o_t, B_t = b_t)].$$

On the LHS, the dependency on $\{s_{1:T}, a_{1:T}\}$ is omitted for a cleaner notation. By convention, $\alpha^{\theta}_{\mu,1}$ is equivalent to

$$\alpha^{\theta}_{\mu,1}(o_1, b_1) = z^{\theta}_{\alpha,\mu,1}\mathbb{E}_{O_0\sim\mu(\cdot|s_1)}[\mathbb{P}_{\theta,O_0,s_1}(A_1 = a_1, O_1 = o_1, B_1 = b_1)].$$

The backward message is defined as

$$\beta^{\theta}_{t|T}(o_t, b_t) := z^{\theta}_{\beta,t}\mathbb{P}_{\theta,o_0,s_1}(S_{t+1:T} = s_{t+1:T}, A_{t+1:T} = a_{t+1:T}|S_t = s_t, A_t = a_t, O_t = o_t, B_t = b_t).$$

The value of o_0 on the RHS is arbitrary. By convention, the boundary condition is

$$\beta^{\theta}_{T|T}(o_T, b_T) = (2\,|\mathcal{O}|)^{-1}. \tag{1}$$

The smoothing distribution is defined as

$$\gamma^{\theta}_{\mu,t|T}(o_t, b_t) := z^{\theta}_{\gamma,\mu}\mathbb{E}_{O_0\sim\mu(\cdot|s_1)}[\mathbb{P}_{\theta,O_0,s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_t = o_t, B_t = b_t)].$$

It can be easily verified that the normalizing constant does not depend on t.

Finally, for all θ, μ and all $t \in [2:T]$, with any input arguments o_{t-1} and b_t, we define the two-step smoothing distribution as

$$\tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t) := z^{\theta}_{\gamma,\mu}\mathbb{E}_{O_0\sim\mu(\cdot|s_1)}[\mathbb{P}_{\theta,O_0,s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_{t-1} = o_{t-1}, B_t = b_t)],$$

where $z^{\theta}_{\gamma,\mu}$ is the same normalizing constant as the one for the smoothing distribution $\gamma^{\theta}_{\mu,t|T}$.

The quantities above can be computed using the forward-backward recursion. For conciseness, we replace normalizing constants by the proportional symbol $\propto$. The proof is deferred to Appendix B.2.

Theorem 1 (Forward-backward smoothing)**.** *For all $\theta \in \Theta$ and $\mu \in \mathcal{M}$, with any input arguments on the LHS,*

1. (Forward recursion) $\forall t \in [2:T]$,

$$\alpha^{\theta}_{\mu,t}(o_t, b_t) \propto \sum_{o_{t-1},b_{t-1}} \pi_b(b_t|s_t, o_{t-1}; \theta_b)\bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi})\pi_{lo}(a_t|s_t, o_t; \theta_{lo})\alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1}). \tag{2}$$

When $t = 1$,

$$\alpha^{\theta}_{\mu,1}(o_1, b_1) \propto \mathbb{E}_{O_0\sim\mu(\cdot|s_1)}[\pi_b(b_1|s_1, O_0; \theta_b)\bar{\pi}_{hi}(o_1|s_1, O_0, b_1; \theta_{hi})\pi_{lo}(a_1|s_1, o_1; \theta_{lo})]. \tag{3}$$

2. (Backward recursion) $\forall t \in [1:T-1]$,

$$\beta^{\theta}_{t|T}(o_t, b_t) \propto \sum_{o_{t+1},b_{t+1}} \bar{\pi}_{hi}(o_{t+1}|s_{t+1}, o_t, b_{t+1}; \theta_{hi}) \times \pi_b(b_{t+1}|s_{t+1}, o_t; \theta_b)\pi_{lo}(a_{t+1}|s_{t+1}, o_{t+1}; \theta_{lo}) \times \beta^{\theta}_{t+1|T}(o_{t+1}, b_{t+1}). \tag{4}$$

3. (Smoothing) $\forall t \in [1:T]$,

$$\gamma^{\theta}_{\mu,t|T}(o_t, b_t) \propto \alpha^{\theta}_{\mu,t}(o_t, b_t)\beta^{\theta}_{t|T}(o_t, b_t). \tag{5}$$

4. (Two-step smoothing) $\forall t \in [2:T]$,

$$\tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t) \propto \left[\sum_{o_t} \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) \times \pi_{lo}(a_t|s_t, o_t; \theta_{lo})\beta^{\theta}_{t|T}(o_t, b_t)\right]\pi_b(b_t|s_t, o_{t-1}; \theta_b) \times \left[\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1})\right]. \tag{6}$$

3.2 Parameter update

For all $\theta, \theta' \in \Theta$ and $\mu \in \mathcal{M}$, the (finite sample) Q-function is defined as

$$Q_{\mu,T}(\theta'|\theta) := \frac{1}{T}\left\{\sum_{t=2}^{T}\sum_{o_{t-1},b_t} \tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t)[\log \pi_b(b_t|s_t, o_{t-1}; \theta'_b)] + \sum_{t=1}^{T}\sum_{o_t,b_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t)[\log \pi_{lo}(a_t|s_t, o_t; \theta'_{lo})] + \sum_{t=1}^{T}\sum_{o_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t = 1)[\log \pi_{hi}(o_t|s_t; \theta'_{hi})]\right\}. \tag{7}$$

The parameter update is performed as $\theta^{(n)} \in \arg\max_{\theta\in\Theta} Q_{\mu,T}(\theta|\theta^{(n-1)})$, which may not be unique. Since Θ is compact and $Q_{\mu,T}(\theta'|\theta)$ is continuous with respect to θ', the maximization is well-posed. Note that our definition of $Q_{\mu,T}(\theta'|\theta)$ is an approximation of the standard definition of Q-function in the EM literature. See Appendix B.3 for a detailed discussion.

3.3 Generalization to continuous spaces

Although we require finite state and action space for our theoretical analysis, Algorithm 1 can be readily generalized to continuous $\mathcal{S}$ and $\mathcal{A}$: we only need to replace π_{lo} by a density function. However, generalization to continuous option space requires a substantially different algorithm. The forward-backward smoothing

procedure in Theorem 1 involves integrals rather than sums, and Sequential Monte Carlo (SMC) techniques need to be applied. Fortunately, it is widely accepted that a finite option space is reasonable in the options framework, since the options need to be distinct and separate (Daniel et al., 2016a).

4 Performance guarantee

Our analysis of Algorithm 1 has the following structure. We first prove the stochastic convergence of the Q-function $Q_{\mu,T}(\theta'|\theta)$ to a population Q-function $\bar{Q}(\theta'|\theta)$, leading to a well-posed definition of the population version algorithm. This step is our major theoretical contribution. With additional assumptions, the *first-order stability* condition is constructed, and techniques in (Balakrishnan et al., 2017) can be applied to show the convergence of the population version algorithm. The remaining step is to analyze Algorithm 1 as a perturbed form of its population version, which requires a high probability bound on the distance between their parameter updates. We can establish the strong consistency of the parameter update of Algorithm 1 as an estimator of the parameter update of the population version algorithm. Therefore, the existence of such a high probability bound can be proved for large enough T. However, the analytical expression of this bound requires knowledge of the specific parameterization of $\{\bar{\pi}_{hi}, \pi_{lo}, \pi_b\}$, which is not available in this general context of discussion.

Concretely, we first analyze the asymptotic behavior of the Q-function $Q_{\mu,T}(\theta'|\theta)$ as $T \to \infty$. From Assumption 3, the observation sequence $\{s_{1:T}, a_{1:T}\}$ is generated from a stationary Markov chain $\{X_t\}_{t=1}^{\infty} := \{S_t, A_t, O_t, B_t\}_{t=1}^{\infty}$. Let $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{O} \times \{0,1\}$ be its state space. Using Kolmogorov's extension theorem, we can extend this one-sided Markov chain to the index set $\mathbb{Z}$ and define a unique probability measure $\mathbb{P}_{\theta^*,\nu^*}$ over the sample space $\mathcal{X}^{\mathbb{Z}}$. Any observation sequence $\{s_{1:T}, a_{1:T}\}$ can be regarded as a segment of an infinite length sample path $\omega \in \mathcal{X}^{\mathbb{Z}}$. Therefore, if the observation sequence is not specified, $Q_{\mu,T}(\theta'|\theta)$ is a random variable with underlying probability measure $\mathbb{P}_{\theta^*,\nu^*}$.

One caveat is that the definition of $Q_{\mu,T}(\theta'|\theta)$ from Section 3 fails for some $\omega \in \mathcal{X}^{\mathbb{Z}}$. To fix this issue, define the set of *proper* sample paths as

$$\Omega = \left\{\omega \in \mathcal{X}^{\mathbb{Z}}; P(s_{t+1}|s_t, a_t) > 0, \forall t \in \mathbb{Z}\right\}. \quad (8)$$

Note that $\mathbb{P}_{\theta^*,\nu^*}(\Omega) = 1$; therefore, working on Ω is probabilistically equivalent to working on $\mathcal{X}^{\mathbb{Z}}$. For all $\omega \in \Omega$, $Q_{\mu,T}(\theta'|\theta)$ follows the definition from Section 3; for other sample paths, $Q_{\mu,T}(\theta'|\theta)$ is defined arbitrarily. In this way, $Q_{\mu,T}(\theta'|\theta)$ becomes a well-defined random variable. Its stochastic convergence is characterized in the following theorem.

Theorem 2 (The stochastic convergence of the Q-function)**.** *With Assumption 1, 2 and 3, there exists a real-valued function $\bar{Q}(\theta'|\theta)$ defined on the domain $\theta' \in \tilde{\Theta}$ and $\theta \in \Theta$ such that*

1. *For all $\theta \in \Theta$, $\bar{Q}(\theta'|\theta)$ is continuously differentiable with respect to $\theta' \in \tilde{\Theta}$. Moreover, the set $\arg\max_{\theta' \in \Theta} \bar{Q}(\theta'|\theta)$ is nonempty.*
2. *As $T \to \infty$, P_{θ^*,ν^*}-almost surely,*

$$\sup_{\theta,\theta' \in \Theta} \sup_{\mu \in \mathcal{M}} \left|Q_{\mu,T}(\theta'|\theta;\omega) - \bar{Q}(\theta'|\theta)\right| \to 0.$$

We name $\bar{Q}(\theta'|\theta)$ as the population Q-function. The analytical expressions of $\bar{Q}(\theta'|\theta)$ and $\nabla\bar{Q}(\theta'|\theta)$ are provided in Appendix C.2, where the complete version of the above theorem (Theorem 7) is proved. In the following, we provide a high level sketch of the main idea.

Proof Sketch. The main difficulty of the proof is that, $Q_{\mu,T}(\theta'|\theta)$ defined in (7) is (roughly) the average of T terms, with each term dependent on the entire observation sequence; as $T \to \infty$, all the terms keep changing such that the law of large numbers cannot be applied directly. As a solution, we approximate $\gamma^{\theta}_{\mu,t|T}$ and $\tilde{\gamma}^{\theta}_{\mu,t|T}$ with smoothing distributions in an infinitely extended graphical model independent of T, resulting in an approximated Q-function (still depends on T). The techniques adopted in this step are analogous to *Markovian decomposition* and *uniform forgetting* in the HMM literature (Cappé et al., 2006; van Handel, 2008). The limiting behavior of the approximated Q-function is the same as that of $Q_{\mu,T}(\theta'|\theta)$, since their difference vanishes as $T \to \infty$. For the approximated Q-function, we can apply the ergodic theorem since the smoothing distributions no longer depend on T. □

The population version of Algorithm 1 has parameter updates $\theta^{(n)} \in \arg\max_{\theta \in \Theta} \bar{Q}(\theta|\theta^{(n-1)})$. To characterize the local convergence of Algorithm 1 and its population version, we impose the following assumptions for the remainder of Section 4.

Assumption 4 (Strong concavity)**.** *There exists $\lambda > 0$ such that for all $\theta_1, \theta_2 \in \Theta$,*

$$\bar{Q}(\theta_1|\theta^*) - \bar{Q}(\theta_2|\theta^*) - \langle\nabla\bar{Q}(\theta_2|\theta^*), \theta_1 - \theta_2\rangle \leq -\frac{\lambda}{2}\left\|\theta_1 - \theta_2\right\|_2^2.$$

For any $r > 0$, let $\Theta_r := \{\theta; \theta \in \Theta, \|\theta - \theta^*\|_2 \leq r\}$.

Assumption 5 (Additional local assumptions)**.** *There exists $r > 0$ such that*

1. *(Identifiability) For all $\theta \in \Theta_r$, $\arg\max_{\theta' \in \Theta} \bar{Q}(\theta'|\theta)$ has a unique element $\bar{M}(\theta)$. Moreover, for all $\varepsilon > 0$, with the convention that $\sup_{\theta' \in \varnothing} \bar{Q}(\theta'|\theta) = -\infty$, we have*

$$\inf_{\theta \in \Theta_r} \left[\bar{Q}(\bar{M}(\theta)|\theta) - \sup_{\theta' \in \Theta; \|\theta' - \bar{M}(\theta)\|_2 \geq \varepsilon} \bar{Q}(\theta'|\theta) \right] > 0.$$

2. *(Uniqueness of finite sample parameter updates) For all $\theta \in \Theta_r$, $T \geq 2$ and $\mu \in \mathcal{M}$, P_{θ^*,ν^*}-almost surely, the set $\arg\max_{\theta' \in \Theta} Q_{\mu,T}(\theta'|\theta;\omega)$ has a unique element $M_{\mu,T}(\theta;\omega)$.*

On the additional assumptions. In Assumption 4, we require the strong concavity of $\bar{Q}(\cdot|\theta^*)$ over the entire parameter space since the maximization step in our algorithm is global. Such a requirement could be avoided: if the maximization step is replaced by a gradient update (Gradient EM), then $\bar{Q}(\cdot|\theta^*)$ only needs to be strongly concave in a small region around θ^*. The price to pay is to assume knowledge on structural constants of $\bar{Q}(\cdot|\theta^*)$ (Lipschitz constant and strong concavity constant). See (Balakrishnan et al., 2017) for an analysis of the gradient EM algorithm.

Nonetheless, we expect the following to hold in certain cases of tabular parameterization: for all $\theta \in \Theta$, the function $\bar{Q}(\cdot|\theta)$ is strongly concave over Θ (see the end of Appendix C.2). From this condition, Assumption 4 and 5.1 directly follow. Assumption 5.2 holds as well; in fact, it is a quite mild assumption due to the sample-based nature of $Q_{\mu,T}(\theta'|\theta;\omega)$.

The next step is to characterize the convergence of the population version algorithm.

Theorem 3 (Convergence of the population version algorithm)**.** *With all the assumptions,*

1. *(First-order stability) There exists $\gamma > 0$ such that for all $\theta \in \Theta_r$,*

$$\left\| \nabla \bar{Q}(\bar{M}(\theta)|\theta) - \nabla \bar{Q}(\bar{M}(\theta)|\theta^*) \right\|_2 \leq \gamma \left\| \theta - \theta^* \right\|_2.$$

2. *(Contraction) Let $\kappa = \gamma/\lambda$. For all $\theta \in \Theta_r$,*

$$\left\| \bar{M}(\theta) - \theta^* \right\|_2 \leq \kappa \left\| \theta - \theta^* \right\|_2.$$

If $\kappa < 1$, the population version algorithm converges linearly to the true parameter θ^*.

The proof is given in Appendix C.3, where we also show an upper bound on γ. The idea mirrors that of (Balakrishnan et al., 2017, Theorem 4) with problem-specific modifications. Algorithm 1 can be regarded as a perturbed form of this population version algorithm, with convergence characterized in the following theorem.

Theorem 4 (Performance guarantee for Algorithm 1)**.** *With all the assumptions, if $\kappa < 1$ we have*

1. *For all $\Delta \in (0, (1-\kappa)r]$ and $q \in (0,1)$, there exists $\underline{T}(\Delta, q) \in \mathbb{N}_+$ such that the following statement is true. If the observation length $T \geq \underline{T}(\Delta, q)$, then with probability at least $1-q$,*

$$\sup_{\theta \in \Theta_r} \sup_{\mu \in \mathcal{M}} \left\| M_{\mu,T}(\theta;\omega) - \bar{M}(\theta) \right\|_2 \leq \Delta.$$

2. *If $T \geq \underline{T}(\Delta, q)$, Algorithm 1 with any $\mu \in \mathcal{M}$ has the following performance guarantee. If $\theta^{(0)} \in \Theta_r$, then with probability at least $1-q$, for all $n \in \mathbb{N}_+$,*

$$\|\theta^{(n)} - \theta^*\|_2 \leq \kappa^n \|\theta^{(0)} - \theta^*\|_2 + (1-\kappa)^{-1}\Delta.$$

The proof is provided in Appendix C.4. Essentially, we use Theorem 2 to show the uniform (in θ and μ) strong consistency of $M_{\mu,T}(\theta;\omega)$ as an estimator of $\bar{M}(\theta)$, following the standard analysis of *M-estimators*. A direct corollary of this argument is the high probability bound on the difference between $M_{\mu,T}(\theta;\omega)$ and $\bar{M}(\theta)$, as shown in the first part of the theorem. Combining this high probability bound with Theorem 3 and (Balakrishnan et al., 2017, Theorem 5) yields the final performance guarantee.

Theorem 4 has two practical implications. First, under regularity conditions, with large enough T, Algorithm 1 can converge with arbitrarily high probability to an arbitrarily small norm ball around the true parameter. In other words, with enough samples, the EM approach can recover the true parameter of the expert policy arbitrarily well. Second, the estimation error (upper bound) decreases exponentially in the initial phase of the algorithm. In this regard, a practitioner can allocate his computational budget accordingly.

One limitation of our analysis is that the condition $\kappa < 1$ is hard to verify for a practical parameterization of the expert policy. This is typical in the theory of EM algorithms: even in the case of i.i.d. samples, characterizing the contraction coefficient is intractable except for a few simple parametric models. Nonetheless, such a condition strengthens our intuition on when the EM approach to HIL works: $\bar{Q}(\theta'|\theta)$ should have a large curvature with respect to θ', and the function should not change much with respect to θ around θ^*. In the next section, we present a numerical example to qualitatively demonstrate our result.

5 Numerical example

In this section, we qualitatively demonstrate our theoretical result through an example. Here, we value clarity over completeness, therefore large-scale experiments are deferred to future works.

Consider the Markov Decision Process (MDP) illustrated in Figure 2. There are four states, numbered from left to right as 1 to 4. At any state $s_t \in [1:4]$, there are two allowable actions: LEFT and RIGHT. If $a_t = \text{RIGHT}$, then the next state is sampled uniformly from the states on the right of state s_t (including s_t itself). Symmetrically, if $a_t = \text{LEFT}$, then the next state is sampled uniformly from the states on the left of state s_t (including s_t).

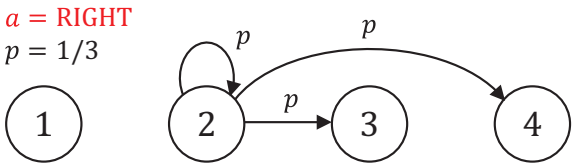

Figure 2: The MDP considered in our example.

Suppose an expert applies the following options with failure policy with parameters $(\theta^*_{hi}, \theta^*_{lo}, \theta^*_b) = (0.6, 0.7, 0.8)$ and $\zeta = 0.1$. The option space has two elements: LEFTEND and RIGHTEND. $\pi_{hi}(o_t = \text{LEFTEND}|s_t; \theta_{hi})$ equals θ_{hi} if $s_t = 1, 2$, and $1 - \theta_{hi}$ if $s_t = 3, 4$. For all s_t, $\pi_{lo}(a_t = \text{LEFT}|s_t, o_t = \text{LEFTEND}; \theta_{lo}) = \pi_{lo}(a_t = \text{RIGHT}|s_t, o_t = \text{RIGHTEND}; \theta_{lo}) = \theta_{lo}$. $\pi_b(b_t = 1|s_t, o_t = \text{LEFTEND}; \theta_b)$ equals θ_b if $s_t = 1$, and $1 - \theta_b$ otherwise. Symmetrically, $\pi_b(b_t = 1|s_t, o_t = \text{RIGHTEND}; \theta_b)$ equals θ_b if $s_t = 4$, and $1 - \theta_b$ otherwise. Intuitively, the high level policy directs the agent to states 1 and 4, and the option terminates with high probability when the corresponding target state is reached.

In our experiment, the parameter spaces Θ_{hi}, Θ_{lo} and Θ_b are all equal to the interval $[0.1, 0.9]$. The initial parameter estimate $(\theta^{(0)}_{hi}, \theta^{(0)}_{lo}, \theta^{(0)}_b) = (0.5, 0.6, 0.7)$. For all s_1, $\mu(o_0 = \text{RIGHTEND}|s_1) = 1$.

We investigate the behavior of $\|\theta^{(n)} - \theta^*\|_2$ as a random variable dependent on n and T. 50 sample paths of length T are sampled from (approximately) the stationary Markov chain induced by the expert policy, with $T \in \{5000, 8000, 10000\}$. After running Algorithm 1 with any sample path ω and any T, we obtain a sequence $\{\|\theta^{(n)} - \theta^*\|_2; \omega, T\}_{n \in [0:N]}$. Let $err(n, T)$ be the average of $\|\theta^{(n)} - \theta^*\|_2$ for fixed n and T, over the 50 sample paths. The result is shown in Figure 3.

Assumption 1, 2, 3 and 5.2 hold in this example, and we speculate that Assumption 4 and 5.1 hold as well. The condition $\kappa < 1$ cannot be verified, but the empirical result exhibits patterns consistent with the performance guarantee, even though rigorously Theorem 4 is not applicable. First, $err(n, T)$ decreases exponentially in the early phase of the algorithm. Second, as T increases, Algorithm 1 achieves better performance.

An observation is worth mentioning as a separate note: for $n > 300$, $err(n, T)$ first slightly increases, then levels off. This is due to the parameter estimate on some sample paths converging to bad stationary points of the finite sample likelihood function, which suggests that early stopping could be helpful in practice. Omitted details and additional experiments are provided in Appendix E, where we also investigate, for example, the effect of μ and random initialization on the performance of Algorithm 1.

6 Conclusion

In this paper, we investigate the EM approach to HIL from a theoretical perspective. We prove that under regularity conditions, the proposed algorithm converges with high probability to a norm ball around the true parameter. To our knowledge, this is the first performance guarantee for an HIL algorithm that only observes primitive state-action pairs. Future works could further investigate the practical performance of this approach, especially its scalability in complicated environments.

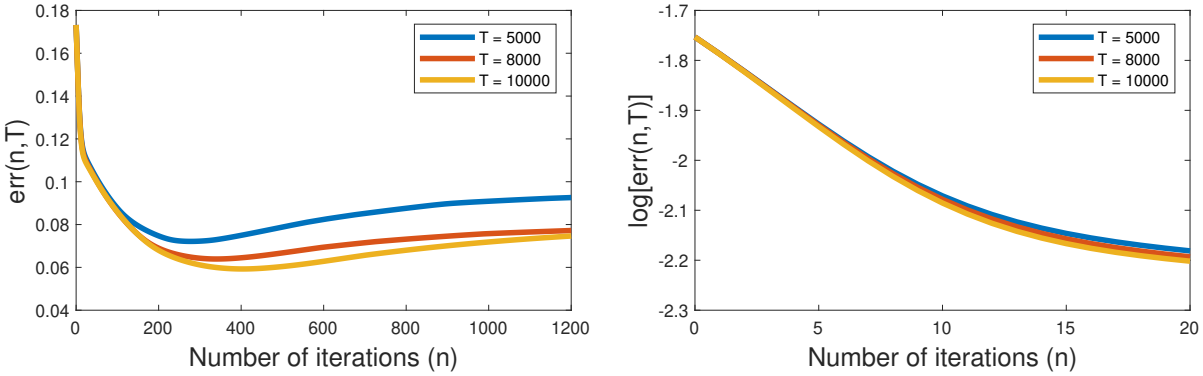

Figure 3: Plots of $err(n, T)$ and $\log err(n, T)$ with varying n and T.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. Z.Z. thanks Tianrui Chen for helpful discussions. The research was partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under grant N00014-19-1-2571, by the NIH under grants R01 GM135930 and UL54 TR004130, and by the DOE under grant DE-AR-0001282.

References

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(80):2773–2832, 2014.

P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1726–1734, 2017.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1-2):41–77, 2003.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

E. Brunskill and L. Li. PAC-inspired option discovery in lifelong reinforcement learning. In *Proceedings of the 31st International Conference on Machine Learning*, pages 316–324, 2014.

J. Butterfield, S. Osentoski, G. Jay, and O. C. Jenkins. Learning from demonstration using a multi-valued function regressor for time-series data. In *2010 10th IEEE-RAS International Conference on Humanoid Robots*, pages 328–333. IEEE, 2010.

O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

C. Daniel, G. Neumann, O. Kroemer, and J. Peters. Hierarchical relative entropy policy search. *Journal of Machine Learning Research*, 17(1):3190–3239, 2016a.

C. Daniel, H. Van Hoof, J. Peters, and G. Neumann. Probabilistic inference for determining options in reinforcement learning. *Machine Learning*, 104(2-3): 337–357, 2016b.

C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Conference on Learning Theory*, pages 704–710, 2017.

J. Davidson. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford, 1994.

Y. De Castro, E. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, 63 (8):4758–4777, 2017.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

R. Dwivedi, N. Ho, K. Khamaru, M. I. Jordan, M. J. Wainwright, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *arXiv preprint arXiv:1810.00828*, 2018a.

R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, and M. I. Jordan. Theoretical guarantees for EM under mis-specified gaussian mixture models. In *Advances in Neural Information Processing Systems 31*, pages 9681–9689, 2018b.

R. Fox, S. Krishnan, I. Stoica, and K. Goldberg. Multi-level discovery of deep options. *arXiv preprint arXiv:1703.08294*, 2017.

R. Fruit and A. Lazaric. Exploration–exploitation in MDPs with options. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 576–584, 2017.

R. Fruit, M. Pirotta, A. Lazaric, and E. Brunskill. Regret minimization in MDPs with options without prior knowledge. In *Advances in Neural Information Processing Systems 30*, pages 3166–3176, 2017.

M. Hairer. Ergodic properties of Markov processes. *Unpublished lecture notes*, 2006. URL `http://www.hairer.org/notes/Markov.pdf`.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

A. Kontorovich, B. Nadler, and R. Weiss. On learning parametric-output HMMs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 702–710, 2013.

S. Krishnan, R. Fox, I. Stoica, and K. Goldberg. DDCO: Discovery of deep continuous options for robot learn-

ing from demonstrations. In *Conference on Robot Learning*, pages 418–437, 2017.

T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems 29*, pages 3675–3683, 2016.

H. Le, N. Jiang, A. Agarwal, M. Dudik, Y. Yue, and H. Daumé. Hierarchical imitation and reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2917–2926, 2018.

S. Le Corff and G. Fort. Online expectation maximization based algorithms for inference in hidden Markov models. *Electronic Journal of Statistics*, 7:763–792, 2013.

T. Mann and S. Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *Proceedings of the 31th International Conference on Machine Learning*, pages 127–135, 2014.

S. Manschitz, J. Kober, M. Gienger, and J. Peters. Learning to sequence movement primitives from demonstrations. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4414–4421. IEEE, 2014.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

O. Nachum, S. S. Gu, H. Lee, and S. Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems 31*, pages 3303–3313, 2018.

S. Niekum, S. Osentoski, G. Konidaris, and A. G. Barto. Learning and generalization of complex tasks from unstructured demonstrations. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5239–5246. IEEE, 2012.

S. Niekum, S. Osentoski, G. Konidaris, S. Chitta, B. Marthi, and A. G. Barto. Learning grounded finite-state representations from unstructured demonstrations. *The International Journal of Robotics Research*, 34(2):131–157, 2015.

K. Shiarlis, M. Wulfmeier, S. Salter, S. Whiteson, and I. Posner. TACO: Learning task decomposition via temporal alignment for control. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4654–4663, 2018.

R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

C. Tessler, S. Givony, T. Zahavy, D. J. Mankowitz, and S. Mannor. A deep hierarchical approach to lifelong learning in minecraft. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1553–1561, 2017.

R. van Handel. Hidden Markov models. *Unpublished lecture notes*, 2008. URL `https://web.math.princeton.edu/~rvan/orf557/hmm080728.pdf`.

A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3540–3549, 2017.

Z. Wang, Q. Gu, Y. Ning, and H. Liu. High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems 28*, pages 2521–2529, 2015.

C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1):95–103, 1983.

J. Xu, D. J. Hsu, and A. Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*, pages 2676–2684, 2016.

F. Yang, S. Balakrishnan, and M. J. Wainwright. Statistical and computational guarantees for the Baum-Welch algorithm. *Journal of Machine Learning Research*, 18(1):4528–4580, 2017.

X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems 28*, pages 1567–1575, 2015.