Provable Hierarchical Imitation Learning via EM

Zhiyu Zhang [1] **Ioannis Ch. Paschalidis** [1]

Abstract

Due to recent empirical successes, the *options* framework for hierarchical reinforcement learning is gaining increasing popularity. In this paper, we consider learning an options-type hierarchical policy from expert demonstrations, which is also referred to as hierarchical imitation learning. We present the first theoretical analysis of an approach that uses the EM algorithm to learn a parameterized hierarchical policy in a latent variable model. To that end, we modify a related algorithm, first proposed in (Daniel et al., 2016b), in order to address a problem in its formulation. If the expert policy can be parameterized by a specific variant of the options framework, then, under regularity conditions, we establish that the proposed algorithm converges with high probability to a norm ball around the true parameter.

1. Introduction

Recent empirical studies show that the scalability of RL algorithms can be improved by incorporating hierarchical structures. As an example, consider the *options* framework (Sutton et al., 1999; Barto & Mahadevan, 2003) representing a two-level hierarchical policy: with a set of multi-step low level procedures (a.k.a., options), the high level policy chooses an option, which, in turn, chooses the primitive action applied at each time step until the option terminates. Learning such a two-level hierarchical policy effectively breaks the overall task into sub-tasks, each easier to solve.

Many research questions arise depending on how the options are generated. Existing theoretical analyses (Brunskill & Li, 2014; Fruit & Lazaric, 2017; Fruit et al., 2017) typically assume the options are given. As a result, only the high-level policy needs to be learned through sequential interaction with the environment. On the contrary, deep hierarchical RL approaches (Bacon et al., 2017) focus on concurrently learning the full hierarchical policy, but still the initialization of the options is essential. 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. We name the former step as *Hierarchical Imitation Learning* (HIL).

There are fairly limited existing works on the topic of HIL. Assuming the specific option adopted by the expert is observed at each time step, (Le et al., 2018) extends algorithms in standard imitation learning to HIL. However, the high level decision process is usually intrinsic to the expert, and only the primitive state action pairs can be observed. In such cases, HIL becomes an inference problem in a latent variable model. Such a formulation is first proposed in (Daniel et al., 2016b), where the classical EM algorithm is applied for policy learning. Empirical studies demonstrate good performance, but the theoretical analysis remains open.

In this paper, we establish the first known performance guarantees for the EM approach to HIL. In particular, we address a problem in the algorithm of (Daniel et al., 2016b), and a modified algorithm is proposed instead. We identify the lack of mixing as a technical difficulty in learning the standard options framework. As a circumvention, a novel *options with failure* framework is considered. If the expert policy can be parameterized by this new framework, then under regularity conditions, we prove that the proposed algorithm converges with high probability to a norm ball around the true parameter. Our analysis involves recent theories of EM algorithms (Balakrishnan et al., 2017; Yang et al., 2017) and the classical asymptotic analysis of Hidden Markov Models (HMMs) (Cappé et al., 2006).

2. Problem settings

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 $t_1 \leq t \leq t_2$.

2.1. Definitions of the hierarchical policy

We first introduce the options framework for hierarchical reinforcement learning, 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. S_t, A_t and O_t are defined on finite sets $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{O}$; B_t is binary. Define the

[1]Boston University. Correspondence to: Zhiyu Zhang <zhiyuz@bu.edu>.

Preliminary work. Under review by Theoretical Foundations of Reinforcement Learning @ ICML 2020. Do not distribute.

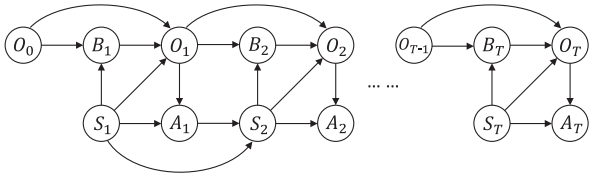

Figure 1. A graphical model for hierarchical RL.

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 an 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}, a low level policy π_{lo} and a termination policy π_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, has the following procedure. Starting at 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 $o_1 = o_0$. Then, 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, it is hard to obtain performance guarantees when the expert policy is parameterized by the standard options framework, due to the construction of Lemma D.1. For simplicity, we consider a novel options with failure framework, which relaxes the options framework by the addition of an extra *failure* mechanism when $b_t = 0$. 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 $\zeta = 0$ recovers the options framework. For clarity, we define $\bar{\pi}_{hi}$ as the combination of π_{hi} and the failure mechanism.

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

Concretely, the options with failure framework is defined as any policy triple $\{\bar{\pi}_{hi}, \pi_{lo}, \pi_b\}$ parameterized by ζ and θ.

With ζ fixed, for any θ and $(O_0, S_1) = (o_0, s_1)$, let $\mathbb{P}_{\theta, o_0, s_1}$ be the joint distribution of $\{S_{2:T}, A_{1:T}, O_{1:T}, B_{1:T}\}$.

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

2.2. The imitation learning problem

Suppose an expert (agent) uses an options with failure policy with true parameters ζ and θ^*; 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. Assume ζ and the parametric structure of the expert policy are known, but θ^* and ν^* are unknown; our objective is to estimate θ^* from $\{s_{1:T}, a_{1:T}\}$. The following technical assumptions are imposed.

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 $\tilde{\Theta}$.*

Assumption 3 (State reachability)**.** *$\forall s_t, s_{t+1} \in \mathcal{S}$, there exists $a_t \in \mathcal{A}$ such that $P(s_{t+1}|s_t, a_t) > 0$.*

The next assumption is motivated by the following result. $\forall \theta \in \Theta$, consider the Markov chain $\{X_t; \theta\}_{t=1}^{\infty} := \{S_t, A_t, O_t, B_t; \theta\}_{t=1}^{\infty}$ generated by any (o_0, s_1) and an options with failure hierarchical policy with parameters ζ and θ. Its state space $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{O} \times \{0, 1\}$ is finite. From Lemma A.1, such a Markov chain is ergodic; that is, the distribution of X_t converges to a unique stationary distribution as t increases. For simplicity, we assume this Markov chain is initialized with the stationary distribution. Specifically, in a stationary Markov Chain $\{X_t; \theta^*\}_{t=1}^{\infty}$, define the marginal distribution of (O_t, S_{t+1}) as $\nu_{\theta^*, OS}$.

Assumption 4 (Stationary initial distribution)**.** *When the expert generates the observation sequence $\{s_{1:T}, a_{1:T}\}$, (o_0, s_1) is sampled from $\nu_{\theta^*, OS}$. Equivalently, $\nu^* = \nu_{\theta^*, OS}$.*

3. A Baum-Welch type algorithm

Adopting the EM approach, we present Algorithm 1 for the inference of θ^*. It is similar to an algorithm in (Daniel et al., 2016b) but fixes a key problem in its formulation; when calculating the posterior distribution of latent variables, at any time $t < T$, (Daniel et al., 2016b) neglects the dependency

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

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

$$\begin{aligned} Q_{\mu,T}(\theta'|\theta) := \frac{1}{T}\Bigg\{ &\sum_{t=2}^{T}\sum_{o_{t-1},b_t} \tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t)\,[\log \pi_b(b_t|s_t, o_{t-1}; \theta'_b)] + \sum_{t=1}^{T}\sum_{o_t,b_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t) \\ &\times [\log \pi_{lo}(a_t|s_t, o_t; \theta'_{lo})] + \sum_{t=1}^{T}\sum_{o_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t = 1)[\log \pi_{hi}(o_t|s_t; \theta'_{hi})]\Bigg\}. \end{aligned} \tag{3}$$

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 smoothing distributions $\{\gamma^{\theta^{(n-1)}}_{\mu,t|T}\}_{t=1}^{T}$ and $\{\tilde{\gamma}^{\theta^{(n-1)}}_{\mu,t|T}\}_{t=2}^{T}$ defined in (1) and (2).
3: Compute $\theta^{(n)} \in \arg\max_{\theta\in\Theta} Q_{\mu,T}(\theta|\theta^{(n-1)})$, with the Q-function on the RHS defined in (3).
4: **end for**

of future states $S_{t+1:T}$ on the current option O_t. A detailed discussion is provided in Appendix B.1.

Since our graphical model resembles an HMM, Algorithm 1 is a variant of the classical Baum-Welch algorithm for HMM parameter inference, which iterates between a latent variable estimation step and a parameter update step. Analogously, we define the *marginal smoothing distributions* in (1) and (2), where $z^{\theta}_{\gamma,\mu}$ is a normalizing constant independent of t such that for all t, both $\gamma^{\theta}_{\mu,t|T}$ and $\tilde{\gamma}^{\theta}_{\mu,t|T}$ are probability mass functions. Then, such quantities are used to compute the maximization objective (*Q-function*) in the parameter update step, which is a surrogate of the likelihood function. Note that the Q-function here follows the notation from EM literature and is different from the usual action value function in RL. For implementation, the marginal smoothing distributions can be computed via forward-backward recursion. Due to limited space, the computational procedure is deferred to Appendix B.2. Finally, for notation, let $\mathcal{M}$ be the set of μ allowed by Algorithm 1.

4. Performance guarantees

The structure of our analysis follows recent theories of EM algorithms (Balakrishnan et al., 2017; Yang et al., 2017). Traditionally, the EM algorithm gained its popularity mainly due to its empirical performance. Its theoretical analysis, however, were generally weak, only characterizing the convergence of parameter estimates to the MLE of the finite sample likelihood function (a.k.a., the *finite sample MLE*). Due to the randomness in sampling, the finite sample likelihood function is usually highly non-convex, leading to stringent requirements on initialization. Moreover, 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 (Balakrishnan et al., 2017; Yang et al., 2017) focus on the convergence to the true parameter directly, relying on the definition of the *population EM algorithm*. It has the same two-stage iterative procedure as the standard EM algorithm, but its Q-function is defined as the infinite sample limit of the finite sample Q-function (a.k.a., the *population 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.

To properly define the population Q-function, the stochastic convergence of the finite sample Q-function needs to be constructed. In the case of i.i.d. samples (Balakrishnan et al., 2017), it follows directly from the law of large numbers. However, this is less obvious in time-series models such as HMMs and the model considered in HIL. For HMMs, (Yang et al., 2017) shows that the expectation of the Q-function converges, but both the full stochastic convergence property and the analytical expression of the population Q-function are not provided. The missing techniques can be borrowed from asymptotic analyses of HMMs (Cappé et al., 2006; De Castro et al., 2017). Notably, a more sensible construction of the population EM algorithm for HMMs is proposed in (Le Corff et al., 2013), under a different setting.

Our analysis of Algorithm 1 has the following steps. 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 a few additional local assumptions, 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 concentration 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 on 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 4, the observation sequence $\{s_{1:T}, a_{1:T}\}$ is generated from a stationary Markov chain. Using Kolmogorov's extension theorem, we can extend this Markov chain to the index set $\mathbb{Z}$ and define a unique probability measure $\mathbb{P}_{\theta^*}$ over the sample space $\mathcal{X}^{\mathbb{Z}}$. Any observation sequence $\{s_{1:T}, a_{1:T}\}$ of length T can be regarded as a subset of an infinite length sample path $\omega \in \mathcal{X}^{\mathbb{Z}}$. If $\{s_{1:T}, a_{1:T}\}$ is not specified, $Q_{\mu,T}(\theta'|\theta)$ is a random variable associated with the probability measure $\mathbb{P}_{\theta^*}$. Its stochastic convergence is characterized in the following theorem.

Theorem 1 (The stochastic convergence of the Q-function)**.** *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 $T \to \infty$,

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

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

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

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

2. (Uniqueness of finite sample updates) For all $\theta \in \Theta_r$, $\omega \in \mathcal{X}^{\mathbb{Z}}$, $T \ge 2$ and $\mu \in \mathcal{M}$, $\arg\max_{\theta' \in \Theta} Q_{\mu,T}(\theta'|\theta;\omega)$ has a unique element $M_{\mu,T}(\theta;\omega)$.

3. (Strong concavity) There exists $\lambda > 0$ such that for all $\theta_1, \theta_2 \in \Theta_r$,

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

In the spirit of (Balakrishnan et al., 2017), the population version algorithm has the following convergence property.

Theorem 2 (Convergence of the population version algorithm)**.** *With all the assumptions,*

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

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

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

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

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 1) 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 3 (Performance guarantee for Algorithm 1)**.** *With all the assumptions,*

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 $T \ge \underline{T}(\Delta, q)$, then with probability at least $1 - q$,

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

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

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

The proof is provided in Appendix C.4. Essentially, we use Theorem 1 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 bound with Theorem 2 and (Balakrishnan et al., 2017, Theorem 2) yields the second part of the theorem. A practical implication is that, under regularity conditions, with large enough T, the algorithm can locally converge with arbitrarily high probability to an arbitrarily small norm ball around the true parameter. Some empirical results are provided in Appendix E.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

Organization of the appendix.

Appendix A presents discussions that motivates Assumption 4, which is omitted from Section 2.2.

Appendix B presents detailed discussions on Algorithm 1, including the comparison with an existing algorithm, the forward-backward implementation and the definition of the Q-function in (3).

Appendix C presents the proofs omitted in Section 4. Technical lemmas involved in the proofs are deferred to Appendix D. Notably, a proof sketch of Theorem 2, which is our main theoretical contribution, is provided at the beginning of Appendix C.

Appendix E presents the empirical results of our algorithm on a simple example. Discussions on the scope of this paper and possible extensions are provided in Appendix F.

A. Ergodicity of the Markov chain

In this section we provide an ergodicity result for the Markov chain considered in Section 2.2, which motivates Assumption 4. As a recap of the settings, we focus on the graphical model in Figure 1. $\forall \theta \in \Theta$, consider the stochastic process $\{X_t;\theta\}_{t=1}^{\infty} := \{S_t, A_t, O_t, B_t;\theta\}_{t=1}^{\infty}$ generated by any (o_0, s_1) and an options with failure hierarchical policy with parameters ζ and θ. The dependency on ζ is dropped since we assume such a parameter adopted by the expert is a known constant. The state space of the process $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{O} \times \{0,1\}$ is finite. From the graphical model, $\{X_t;\theta\}_{t=1}^{\infty}$ is a Markov chain with a transition kernel parameterized by θ. Denote its one step transition kernel as Q_θ and its t step transition kernel as Q_θ^t. Such a Markov chain is uniformly ergodic, as shown in the following lemma.

Lemma A.1 (Ergodicity). *For all $\theta \in \Theta$, a Markov chain with transition kernel Q_θ has a unique stationary distribution ν_θ. There exist constants $\alpha \in (0,1)$ and $C > 0$ such that for all $\theta \in \Theta$ and $t \in \mathbb{N}_+$,*

$$\sup_{\theta\in\Theta} \max_{x\in\mathcal{X}} \left\| Q_\theta^t(x,\cdot) - \nu_\theta \right\|_{\mathrm{TV}} \le C\alpha^t.$$

Proof of Lemma A.1

We start by analyzing the irreducibility of the Markov chain $\{X_t;\theta\}_{t=1}^{\infty}$ with any θ. Denote the probability measure on the natural filtered space as $\mathbb{P}_X$. The dependency on θ is dropped for a cleaner notation, since the following proof holds for all $\theta \in \Theta$. For any $x, \tilde{x} \in \mathcal{X}$, let $x = (s, a, o, b)$ and $\tilde{x} = (\tilde{s}, \tilde{a}, \tilde{o}, \tilde{b})$. For any time t,

$$\mathbb{P}_X(X_{t+2} = \tilde{x} | X_t = x) = \sum_{\bar{s}\in\mathcal{S},\bar{a}\in\mathcal{A}} \mathbb{P}_X(X_{t+2} = \tilde{x} | X_t = x, S_{t+1} = \bar{s}, A_{t+1} = \bar{a}) \mathbb{P}_X(S_{t+1} = \bar{s}, A_{t+1} = \bar{a} | X_t = x).$$

From the non-degeneracy assumption (Assumption 1), there exists a state $\bar{s}$ such that $\forall \bar{a} \in \mathcal{A}$, $\mathbb{P}_X(S_{t+1} = \bar{s}, A_{t+1} = \bar{a} | X_t = x) > 0$. Consider the first factor in the sum,

$$\begin{aligned}\mathbb{P}_X(X_{t+2} = \tilde{x} | X_t = x, S_{t+1} = \bar{s}, A_{t+1} = \bar{a}) &= \mathbb{P}_X(S_{t+2} = \tilde{s} | S_{t+1} = \bar{s}, A_{t+1} = \bar{a}) \\ &\quad \times \mathbb{P}_X(B_{t+2} = \tilde{b}, O_{t+2} = \tilde{o}, A_{t+2} = \tilde{a} | X_t = x, S_{t+1} = \bar{s}, A_{t+1} = \bar{a}, S_{t+2} = \tilde{s}).\end{aligned}$$

From non-degeneracy, the second term on the RHS is positive for all $\bar{s} \in \mathcal{S}$ and $\bar{a} \in \mathcal{A}$. From the reachability assumption (Assumption 3), for any $\bar{s}$ there exists an action $\bar{a}$ such that $\mathbb{P}_X(S_{t+2} = \tilde{s} | S_{t+1} = \bar{s}, A_{t+1} = \bar{a}) > 0$. As a result, $\mathbb{P}_X(X_{t+2} = \tilde{x} | X_t = x) > 0$, and the considered Markov chain is irreducible.

As shown above, for all $\theta \in \Theta$, $\min_{x,\tilde{x}\in\mathcal{X}} Q_\theta^2(x,\tilde{x}) > 0$ where Q_θ^2 is the two step transition kernel of the Markov chain $\{X_t;\theta\}_{t=1}^{\infty}$. Due to Assumption 2, $\min_{x,\tilde{x}\in\mathcal{X}} Q_\theta^2(x,\tilde{x})$ is continuous with respect to θ. Moreover, since Θ is compact, if we let $\delta = \inf_{\theta\in\Theta} \min_{x,\tilde{x}\in\mathcal{X}} Q_\theta^2(x,\tilde{x})$ we have $\delta > 0$. The Doeblin-type condition can be constructed as follows. For all $\theta \in \Theta$ and $x, \tilde{x} \in \mathcal{X}$, with any probability measure ν over the finite sample space $\mathcal{X}$,

$$Q_\theta^2(x,\tilde{x}) \ge \delta\nu(\tilde{x}). \tag{4}$$

A Markov chain convergence result is restated in the following lemma, tailored to our need.

Lemma A.2 ((Cappé et al., 2006), Theorem 4.3.16 restated)**.** *With the Doeblin-type condition in (4), the Markov chain* $\{X_t;\theta\}_{t=1}^{\infty}$ *with any* $\theta\in\Theta$ *has a unique stationary distribution* ν_θ*. Moreover, for all* $\theta\in\Theta$*,* $x\in\mathcal{X}$ *and* $t\in\mathbb{N}_+$*,*

$$\left\|Q_\theta^t(x,\cdot)-\nu_\theta\right\|_{\mathrm{TV}}\le(1-\delta)^{\lfloor t/2\rfloor}.$$

Letting $C=(1-\delta)^{-1}$ and $\alpha=(1-\delta)^{1/2}$, we have

$$\sup_{\theta\in\Theta}\max_{x_1\in\mathcal{X}}\left\|Q_\theta^t(x_1,\cdot)-\nu_\theta\right\|_{\mathrm{TV}}\le(1-\delta)^{\lfloor t/2\rfloor}\le C\alpha^t.$$ □

Note that the proof of Lemma A.1 does not use the failure mechanism imposed on the hierarchical policy, implying that the result also holds for the standard options framework. Loosely speaking, Lemma A.1 shows that in $\{X_t;\theta\}_{t=1}^{\infty}$, the initial distribution (of X_1) is not very important since the distribution of X_t converges to ν_θ uniformly with respect to X_1 and θ. With a long observation sequence, we can always discard a portion in the front such that the rest approximately satisfies Assumption 4.

B. Details on Algorithm 1

B.1. A technical problem of an earlier algorithm

First, we point out a technicality when comparing Algorithm 1 and the algorithm in (Daniel et al., 2016b). The algorithm in (Daniel et al., 2016b) learns a hierarchical policy following the standard options framework, not the options with failure framework considered in Algorithm 1. To draw direct comparison, we need to let $\zeta=0$ in Algorithm 1. However, a problem in the formulation of the existing algorithm can be demonstrated without referring to ζ.

For simplicity, consider O_0 fixed to $o_0\in\mathcal{O}$; let $2\le t\le T-1$. Then, according to the definitions in (Daniel et al., 2016b), the (unnormalized) forward message is defined as

$$\alpha_t^\theta(o_t,b_t)=\mathbb{P}_{\theta,o_0,s_1}(A_{1:t}=a_{1:t},O_t=o_t,B_t=b_t|S_{2:t}=s_{2:t}).$$

The (unnormalized) backward message is defined as

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

The smoothing distribution is defined as

$$\gamma_{t|T}^\theta(o_t,b_t)=\mathbb{P}_{\theta,o_0,s_1}(O_t=o_t,B_t=b_t|S_{2:T}=s_{2:T},A_{1:T}=a_{1:T}).$$

We use the proportional symbol $\propto$ to represent normalizing constants independent of o_t and b_t. (Daniel et al., 2016b) claims that, for any o_t and b_t,

$$\gamma_{t|T}^\theta(o_t,b_t)\propto\alpha_t^\theta(o_t,b_t)\beta_{t|T}^\theta(o_t,b_t).$$

However, applying Bayes' formula, it follows that

$$\gamma_{t|T}^\theta(o_t,b_t)\propto\mathbb{P}_{\theta,o_0,s_1}(A_{1:T}=a_{1:T}|S_{2:T}=s_{2:T},O_t=o_t,B_t=b_t)\mathbb{P}_{\theta,o_0,s_1}(O_t=o_t,B_t=b_t|S_{2:T}=s_{2:T}).$$

Using the Markov property,

$$\begin{aligned}\mathbb{P}_{\theta,o_0,s_1}(A_{1:T}=a_{1:T}|S_{2:T}=s_{2:T},O_t=o_t,B_t=b_t)&=\mathbb{P}_{\theta,o_0,s_1}(A_{1:t}=a_{1:t}|S_{2:T}=s_{2:T},O_t=o_t,B_t=b_t)\\&\quad\times\mathbb{P}_{\theta,o_0,s_1}(A_{t+1:T}=a_{t+1:T}|S_{2:T}=s_{2:T},O_t=o_t,B_t=b_t).\end{aligned}$$

Therefore,

$$\gamma_{t|T}^\theta(o_t,b_t)\propto\mathbb{P}_{\theta,o_0,s_1}(A_{1:t}=a_{1:t},O_t=o_t,B_t=b_t|S_{2:T}=s_{2:T})\beta_{t|T}^\theta(o_t,b_t).$$

Applying Bayes' formula again, it follows that

$$\begin{aligned}&\mathbb{P}_{\theta,o_0,s_1}(A_{1:t}=a_{1:t},O_t=o_t,B_t=b_t|S_{2:T}=s_{2:T})\\\propto\;&\mathbb{P}_{\theta,o_0,s_1}(A_{1:t}=a_{1:t},O_t=o_t,B_t=b_t|S_{2:t}=s_{2:t})\\&\qquad\times\mathbb{P}_{\theta,o_0,s_1}(S_{t+1:T}=s_{t+1:T}|S_{2:t}=s_{2:t},A_{1:t}=a_{1:t},O_t=o_t,B_t=b_t)\\=\;&\alpha_t^\theta(o_t,b_t)\mathbb{P}_{\theta,o_0,s_1}(S_{t+1:T}=s_{t+1:T}|S_t=s_t,A_t=a_t,O_t=o_t,B_t=b_t).\end{aligned}$$

For the claim in (Daniel et al., 2016b) to hold true, $\mathbb{P}_{\theta,o_0,s_1}(S_{t+1:T}=s_{t+1:T}|S_t=s_t,A_t=a_t,O_t=o_t,B_t=b_t)$ should not depend on o_t and b_t. Clearly this requirement does not hold in most cases, since the likelihood of the future observation sequence should depend on the currently applied option.

B.2. Computation of the smoothing distributions

The marginal smoothing distributions $\gamma^{\theta}_{\mu,t|T}$ and $\tilde{\gamma}^{\theta}_{\mu,t|T}$ can be computed via forward-backward recursion, analogous to the Baum-Welch algorithm. To do this, additional quantities are required. In the following, we define the forward message and the backward message for all θ, μ and $t \in [1:T]$. With any input arguments o_t and b_t, the forward message is defined as

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

where $z^{\theta}_{\alpha,\mu,t}$ is a normalizing constant such that $\alpha^{\theta}_{\mu,t}$ is a probability mass function. 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

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

where $z^{\theta}_{\beta,t}$ is a normalizing constant such that $\beta^{\theta}_{t|T}$ is a probability mass function. The value of o_0 on the RHS is arbitrary. By convention, the boundary condition is

$$\beta^{\theta}_{T|T}(o_T, b_T) = (2\,|\mathcal{O}|)^{-1}. \tag{5}$$

The marginal smoothing distributions are defined in (1) and (2). To distinguish these two quantities, we name $\gamma^{\theta}_{\mu,t|T}$ as *smoothing distribution* and $\tilde{\gamma}^{\theta}_{\mu,t|T}$ as *two-step smoothing distribution*. The forward-backward recursion procedure is provided in the following theorem. For ease of notation, we omit normalizing constants by using the proportional symbol $\propto$.

Theorem 4 (Forward-backward smoothing)**.** *For all $\theta \in \Theta$ and $\mu \in \mathcal{M}$, with any input arguments on the LHS,*

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

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

For $t = 1$,

$$\alpha^{\theta}_{\mu,1}(o_1, b_1) \propto \mathbb{E}_{O_0\sim\mu(\cdot|s_1)}[\pi_b(b_1|s_1, O_0; \theta_b)\bar{\pi}_{hi}(o_1|s_1, O_0, b_1; \theta_{hi})\pi_{lo}(a_1|s_1, o_1; \theta_{lo})]. \tag{7}$$

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

$$\beta^{\theta}_{t|T}(o_t, b_t) \propto \sum_{o_{t+1},b_{t+1}} \pi_b(b_{t+1}|s_{t+1}, o_t; \theta_b)\bar{\pi}_{hi}(o_{t+1}|s_{t+1}, o_t, b_{t+1}; \theta_{hi})\pi_{lo}(a_{t+1}|s_{t+1}, o_{t+1}; \theta_{lo})\beta^{\theta}_{t+1|T}(o_{t+1}, b_{t+1}). \tag{8}$$

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

$$\gamma^{\theta}_{\mu,t|T}(o_t, b_t) \propto \alpha^{\theta}_{\mu,t}(o_t, b_t)\beta^{\theta}_{t|T}(o_t, b_t). \tag{9}$$

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

$$\tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t) \propto \pi_b(b_t|s_t, o_{t-1}; \theta_b)\left[\sum_{o_t} \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi})\pi_{lo}(a_t|s_t, o_t; \theta_{lo})\beta^{\theta}_{t|T}(o_t, b_t)\right] \times \left[\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1})\right]. \tag{10}$$

To compute the marginal smoothing distributions in practice, we first run the forward recursion with $t = 2$ to T, using (6) and (7). Then, backward recursion is implemented, using (5) and (8). Finally, the marginal distributions are computed from the forward message and the backward message, using (9) and (10).

Proof of Theorem 4

We drop the dependency on θ, since the following proof holds for all $\theta \in \Theta$. The proportional symbol $\propto$ is used to replace a multiplier term that depends on the context.

1. (Forward recursion) First consider any fixed o_0. For a cleaner notation, we use p as an abbreviation of $\mathbb{P}_{\theta,o_0,s_1}$. Let H_1, H_2 be any two subsets of $\{S_t, A_t, O_t, B_t\}_{t=1}^T$, and let h_1, h_2 be the sets of values generated from H_1 and H_2, respectively, such that the uppercase symbols are replaced by the lowercase symbols. (H_1 and H_2 are two sets of random variables; h_1 and h_2 are two sets of values of random variables.) Then, for all (o_0, s_1), p is defined as

$$p(h_1|h_2, o_0, s_1) := \mathbb{P}_{\theta,o_0,s_1}(H_1 = h_1|H_2 = h_2).$$

If the RHS does not depend on o_0 and s_1, we can omit it on the LHS by using $p(h_1|h_2)$. $\forall t \in [2:T]$,

$$\begin{aligned}
p(s_{2:t}, a_{1:t}, o_t, b_t|o_0, s_1) &= p(s_{2:t}, a_{1:t-1}, o_t, b_t|o_0, s_1)\pi_{lo}(a_t|s_t, o_t) \\
&= \sum_{o_{t-1}} p(s_{2:t}, a_{1:t-1}, o_t, b_t, o_{t-1}|o_0, s_1)\pi_{lo}(a_t|s_t, o_t) \\
&= \sum_{o_{t-1}} p(s_{2:t}, a_{1:t-1}, o_{t-1}|o_0, s_1)\pi_b(b_t|s_t, o_{t-1})\bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t)\pi_{lo}(a_t|s_t, o_t).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
p(s_{2:t}, a_{1:t-1}, o_{t-1}|o_0, s_1) &= p(s_{2:t-1}, a_{1:t-1}, o_{t-1}|o_0, s_1)P(s_t|s_{t-1}, a_{t-1}) \\
&\propto \sum_{b_{t-1}} p(s_{2:t-1}, a_{1:t-1}, o_{t-1}, b_{t-1}|o_0, s_1),
\end{aligned}$$

where $\propto$ replaces a multiplier that does not depend on o_{t-1}. Taking expectation with respect to O_0 gives the desired forward recursion result. For the case of $t = 1$, the proof is analogous.

2. (Backward recursion) For any o_0, $\forall t \in [1:T-1]$,

$$\begin{aligned}
\beta^\theta_{t|T}(o_t, b_t) &\propto p(s_{t+1:T}, a_{t+1:T}|s_t, a_t, o_t, b_t) \\
&= p(s_{t+2:T}, a_{t+1:T}|s_{t+1}, o_t)P(s_{t+1}|s_t, a_t) \\
&\propto \sum_{o_{t+1}, b_{t+1}} p(s_{t+2:T}, a_{t+1:T}|s_{t+1}, o_t, o_{t+1}, b_{t+1})p(o_{t+1}, b_{t+1}|s_{t+1}, o_t),
\end{aligned}$$

where the multipliers replaced by $\propto$ are independent of o_t and b_t. Moreover,

$$\begin{aligned}
p(s_{t+2:T}, a_{t+1:T}|s_{t+1}, o_t, o_{t+1}, b_{t+1}) &= p(s_{t+2:T}, a_{t+2:T}|s_{t+1}, o_t, o_{t+1}, b_{t+1}, a_{t+1})p(a_{t+1}|s_{t+1}, o_t, o_{t+1}, b_{t+1}) \\
&= \beta^\theta_{t+1|T}(o_{t+1}, b_{t+1})p(a_{t+1}|s_{t+1}, o_t, o_{t+1}, b_{t+1}).
\end{aligned}$$

Plugging in the structure of the policy gives the desired result.

3. (Smoothing) Consider any fixed o_0. For any $t \in [2:T]$,

$$\begin{aligned}
p(s_{2:T}, a_{1:T}, o_t, b_t|o_0, s_1) &= p(s_{2:t}, a_{1:t}, o_t, b_t|o_0, s_1)p(s_{t+1:T}, a_{t+1:T}|s_{1:t}, a_{1:t}, o_t, b_t, o_0) \\
&= p(s_{2:t}, a_{1:t}, o_t, b_t|o_0, s_1)p(s_{t+1:T}, a_{t+1:T}|s_t, a_t, o_t, b_t).
\end{aligned}$$

Taking expectation with respect to O_0 on both sides yields the desired result. Notice that the second term on the RHS does not depend on O_0, therefore is not involved in the expectation. For the case of $t = 1$ the proof is analogous.

4. (Two-step smoothing) For any $t \in [3:T]$, consider any fixed o_0,

$$\begin{aligned}
p(s_{2:T}, a_{1:T}, o_{t-1}, b_t|o_0, s_1) &= \sum_{b_{t-1}} p(s_{2:T}, a_{1:T}, o_{t-1}, b_t, b_{t-1}|o_0, s_1) \\
&= \sum_{b_{t-1}} p(s_{2:t-1}, a_{1:t-1}, o_{t-1}, b_{t-1}|o_0, s_1)p(s_{t:T}, a_{t:T}, b_t|s_{1:t-1}, a_{1:t-1}, o_{t-1}, b_{t-1}, o_0) \\
&= \sum_{b_{t-1}} p(s_{2:t-1}, a_{1:t-1}, o_{t-1}, b_{t-1}|o_0, s_1)P(s_t|s_{t-1}, a_{t-1})p(s_{t+1:T}, a_{t:T}, b_t|s_t, o_{t-1}).
\end{aligned}$$

Take expectation with respect to O_0 on both sides. Notice that only the first term on the RHS depends on o_0. We have

$$\begin{aligned}
\tilde{\gamma}_{\mu,t|T}(o_{t-1}, b_t) \propto & \sum_{b_{t-1}} \alpha_{\mu,t-1}(o_{t-1}, b_{t-1}) P(s_t|s_{t-1}, a_{t-1}) p(s_{t+1:T}, a_{t:T}, b_t|s_t, o_{t-1}) \\
\propto & \ \pi_b(b_t|s_t, o_{t-1}) p(s_{t+1:T}, a_{t:T}|s_t, b_t, o_{t-1}) \sum_{b_{t-1}} \alpha_{\mu,t-1}(o_{t-1}, b_{t-1}) \\
= & \ \pi_b(b_t|s_t, o_{t-1}) \left[\sum_{o_t} p(s_{t+1:T}, a_{t:T}, o_t|s_t, b_t, o_{t-1}) \right] \sum_{b_{t-1}} \alpha_{\mu,t-1}(o_{t-1}, b_{t-1}) \\
\propto & \ \pi_b(b_t|s_t, o_{t-1}) \left[\sum_{o_t} \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t) \pi_{lo}(a_t|s_t, o_t) \beta_{t|T}(o_t, b_t) \right] \sum_{b_{t-1}} \alpha_{\mu,t-1}(o_{t-1}, b_{t-1}),
\end{aligned}$$

where the multipliers replaced by $\propto$ are independent of o_{t-1} and b_t. For the case of $t = 2$ the proof is analogous. □

B.3. Discussion on the Q-function

First, we discuss the use of μ in Algorithm 1. Similar to HMMs, since an unknown prior distribution of (O_0, S_1) is required for the statistical model, we need an approximation of it in the definition of the marginal smoothing distributions. We use $\hat{\nu}$ as an approximation of ν^*: $\forall o_0 \in \mathcal{O}$, $\hat{\nu}(o_0, s_1) := \mu(o_0|s_1)$; $\forall s_1' \neq s_1$, $\hat{\nu}(o_0, s_1') := 0$. By doing this, we effectively consider the following statistical model in Figure 1: the prior distribution of (O_0, S_1) is $\hat{\nu}$, and the distribution of the rest of the graphical model is determined by an options with failure policy with parameters ζ and θ. From the EM literature (Balakrishnan et al., 2017; Jain & Kar, 2017), the full likelihood function is

$$L(s_{1:T}, a_{1:T}, o_{0:T}, b_{1:T}; \theta) = \hat{\nu}(o_0, s_1) \mathbb{P}_{\theta, o_0, s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_{1:T} = o_{1:T}, B_{1:T} = b_{1:T}).$$

The marginal likelihood function is

$$L^m(s_{1:T}, a_{1:T}; \theta) = \sum_{o_{0:T}, b_{1:T}} \hat{\nu}(o_0, s_1) \mathbb{P}_{\theta, o_0, s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_{1:T} = o_{1:T}, B_{1:T} = b_{1:T}),$$

where the superscript m means *marginal*. From (1) and (2), we can verify that $L^m(s_{1:T}, a_{1:T}; \theta) = (z^{\theta}_{\gamma,\mu})^{-1}$.

The Q-function defined in (3) is not exactly the Q-function in EM literature, but rather an approximation of it. We provide an explanation in the following.

We start by deriving the usual definition of Q-function in EM literature. The standard MLE approach maximizes the logarithm of the marginal likelihood function (marginal log-likelihood) with respect to θ. However, such an optimization objective is hard to evaluate for time series models (e.g., HMMs and the graphical model for HIL). As an alternative, the marginal log-likelihood can be lower bounded (Jain & Kar, 2017, Chap. 5.4) as

$$\log L^m(s_{1:T}, a_{1:T}; \theta') \geq \sum_{o_{0:T}, b_{1:T}} \frac{L(s_{1:T}, a_{1:T}, o_{0:T}, b_{1:T}; \theta)}{L^m(s_{1:T}, a_{1:T}; \theta)} \log L(s_{1:T}, a_{1:T}, o_{0:T}, b_{1:T}; \theta'),$$

where θ on the RHS is arbitrary. The RHS is usually called the (unnormalized) Q-function. For our graphical model, it is denoted as $\tilde{Q}_{\mu,T}(\theta'|\theta)$.

$$\begin{aligned}
\tilde{Q}_{\mu,T}(\theta'|\theta) = \sum_{o_{0:T}, b_{1:T}} & \hat{\nu}(o_0, s_1) \mathbb{P}_{\theta, o_0, s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_{1:T} = o_{1:T}, B_{1:T} = b_{1:T}) \\
& \times z^{\theta}_{\gamma,\mu} \log[\hat{\nu}(o_0, s_1) \mathbb{P}_{\theta', o_0, s_1}(S_{2:T} = s_{2:T}, A_{1:T} = a_{1:T}, O_{1:T} = o_{1:T}, B_{1:T} = b_{1:T})].
\end{aligned}$$

The RHS is well-defined from the non-degeneracy assumption. From the classical monotonicity property of EM updates (Jain & Kar, 2017, Chap. 5.7), maximizing the (unnormalized) Q-function $\tilde{Q}_{\mu,T}(\theta'|\theta)$ with respect to θ' guarantees non-negative improvement on the marginal log-likelihood. Therefore, improvements on parameter inference can be achieved via iteratively maximizing the (unnormalized) Q-function.

Using the structure of the hierarchical policy, $\tilde{Q}_{\mu,T}$ can be rewritten as

$$\begin{aligned}\tilde{Q}_{\mu,T}(\theta'|\theta) = &\sum_{t=2}^{T}\sum_{o_{t-1},b_t}\tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1},b_t)[\log\pi_b(b_t|s_t,o_{t-1};\theta'_b)]\\ &+\sum_{t=1}^{T}\sum_{o_t,b_t}\gamma^{\theta}_{\mu,t|T}(o_t,b_t)[\log\pi_{lo}(a_t|s_t,o_t;\theta'_{lo})]+\sum_{t=1}^{T}\sum_{o_t}\gamma^{\theta}_{\mu,t|T}(o_t,b_t=1)[\log\pi_{hi}(o_t|s_t;\theta'_{hi})]\\ &+z^{\theta}_{\gamma,\mu}\sum_{o_0,b_1}\mu(o_0|s_1)\mathbb{P}_{\theta,o_0,s_1}(S_{2:T}=s_{2:T},A_{1:T}=a_{1:T},B_1=b_1)[\log\pi_b(b_1|s_1,o_0;\theta'_b)]+C,\end{aligned}$$

where C contains terms unrelated to θ'. Consider the first term on the last line, which partially captures the effect of assuming $\hat{\nu}$ on the parameter inference. Since this term is upper bounded by $\max_{b_1,s_1,o_0}|\log\pi_b(b_1|s_1,o_0;\theta'_b)|$, when T is large enough this term becomes negligible. The precise argument is similar to the proof of Lemma C.2. Therefore, after dropping the last line and normalizing, we arrive at our definition of the (normalized) Q-function in (3). Maximizing the (normalized) Q-function in (3) is approximately equivalent to maximizing $\tilde{Q}_{\mu,T}$.

C. Details on the performance guarantees

For better communication of the idea, we present the proof sketch of Theorem 2, which is our main theoretical contribution, as follows. Detailed proof is developed later.

Proof sketch of Theorem 2. Our target is to show the stochastic convergence of $Q_{\mu,T}(\theta'|\theta)$ defined in (3). The main difficulty for the proof is that, $Q_{\mu,T}(\theta'|\theta)$ is (roughly) the average of T terms, with each term dependent on the whole observation sequence; as $T\to\infty$, all the terms are changed 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.

C.1. Smoothing in an extended graphical model

Before providing the proofs, we first introduce a few definitions. Consider the extended graphical model shown in Figure 2 with a parameter k; $k\in\mathbb{N}_+$.

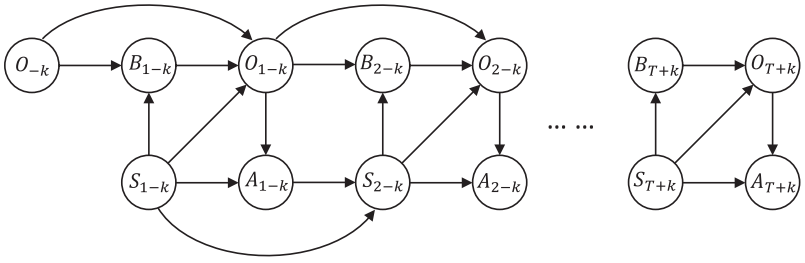

Figure 2. An extended graphical model for hierarchical imitation learning.

Let the joint distribution of (O_{-k},S_{1-k}) be ν^*. Define the distribution of the rest of the graphical model using an options with failure hierarchical policy with parameters ζ and θ, analogous to our settings so far. With these two components, the joint distribution on the graphical model is determined; let $\mathbb{P}_{\theta,k}$ be the corresponding joint distribution.

The comparison between $\mathbb{P}_{\theta,k}$ and $\mathbb{P}_{\theta,o_0,s_1}$ should be emphasized. Notice that the sample space of $\mathbb{P}_{\theta,k}$ is the domain of $\{S_{1-k:T+k},A_{1-k:T+k},O_{-k:T+k},B_{1-k:T+k}\}$, whereas the sample space of $\mathbb{P}_{\theta,o_0,s_1}$ is the domain of $\{S_{2:T},A_{1:T},O_{1:T},B_{1:T}\}$ since (O_0,S_1) is fixed to (o_0,s_1).

Consider any given infinite length observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$. Analogous to the unextended model (Figure 1), we can define smoothing distributions for the extended model with any parameter k. For all $\theta \in \Theta$ and $t \in [1:T]$, with any input arguments o_t and b_t, the forward message is defined as

$$\alpha^{\theta}_{k,t}(o_t, b_t) := z^{\theta}_{\alpha,k,t}\mathbb{P}_{\theta,k}(S_{1-k:t} = s_{1-k:t}, A_{1-k:t} = a_{1-k:t}, O_t = o_t, B_t = b_t).$$

The backward message is defined as

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

The smoothing distribution is defined as

$$\gamma^{\theta}_{k,t}(o_t, b_t) := z^{\theta}_{\gamma,k}\mathbb{P}_{\theta,k}(S_{1-k:T+k} = s_{1-k:T+k}, A_{1-k:T+k} = a_{1-k:T+k}, O_t = o_t, B_t = b_t).$$

The two-step smoothing distribution is defined as

$$\tilde{\gamma}^{\theta}_{k,t}(o_{t-1}, b_t) := z^{\theta}_{\gamma,k}\mathbb{P}_{\theta,k}(S_{1-k:T+k} = s_{1-k:T+k}, A_{1-k:T+k} = a_{1-k:T+k}, O_{t-1} = o_{t-1}, B_t = b_t).$$

The quantities $z^{\theta}_{\alpha,k,t}$, $z^{\theta}_{\beta,k,t}$ and $z^{\theta}_{\gamma,k}$ are normalizing constants such that the LHS of the expressions above are probability mass functions. In particular, since $k > 0$, we can define $\alpha^{\theta}_{k,t}$ for $t = 0$ in the same way as $t \in [1:T]$. Note that the dependency on T in the smoothing distributions is dropped for a cleaner noation.

Recursive results similar to Theorem 4 can be established; the proof is analogous and therefore omitted. As in Theorem 4, we make extensive use of the proportional symbol $\propto$ which stands for, *the LHS equals the RHS multiplied by a normalizing constant*. Moreover, the normalizing constant does not depend on the input arguments of the LHS.

Corollary 5 (Forward-backward smoothing for the extended model)**.** *For all $\theta \in \Theta$ and $k \in \mathbb{N}_+$, with any input arguments,*

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

$$\alpha^{\theta}_{k,t}(o_t, b_t) \propto \sum_{o_{t-1},b_{t-1}} \pi_b(b_t|s_t, o_{t-1}; \theta_b)\bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi})\pi_{lo}(a_t|s_t, o_t; \theta_{lo})\alpha^{\theta}_{k,t-1}(o_{t-1}, b_{t-1}). \tag{11}$$

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

$$\beta^{\theta}_{k,t}(o_t, b_t) \propto \sum_{o_{t+1},b_{t+1}} \pi_b(b_{t+1}|s_{t+1}, o_t; \theta_b)\bar{\pi}_{hi}(o_{t+1}|s_{t+1}, o_t, b_{t+1}; \theta_{hi})\pi_{lo}(a_{t+1}|s_{t+1}, o_{t+1}; \theta_{lo})\beta^{\theta}_{k,t+1}(o_{t+1}, b_{t+1}). \tag{12}$$

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

$$\gamma^{\theta}_{k,t}(o_t, b_t) \propto \alpha^{\theta}_{k,t}(o_t, b_t)\beta^{\theta}_{k,t}(o_t, b_t). \tag{13}$$

4. (Two-step smoothing) $\forall t \in [1:T]$,

$$\tilde{\gamma}^{\theta}_{k,t}(o_{t-1}, b_t) \propto \pi_b(b_t|s_t, o_{t-1}; \theta_b)\left[\sum_{o_t} \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi})\pi_{lo}(a_t|s_t, o_t; \theta_{lo})\beta^{\theta}_{k,t}(o_t, b_t)\right] \times \left[\sum_{b_{t-1}} \alpha^{\theta}_{k,t-1}(o_{t-1}, b_{t-1})\right]. \tag{14}$$

The following lemma characterizes the limiting behavior of $\gamma^{\theta}_{k,t}$ and $\tilde{\gamma}^{\theta}_{k,t}$ as $k \to \infty$.

Lemma C.1 (Limits of smoothing distributions)**.** *For all $T \geq 2$, $\theta \in \Theta$ and $t \in [1:T]$, with any infinite length observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$, the limits of both the sequences $\{\gamma^{\theta}_{k,t}\}_{k\in\mathbb{N}_+}$ and $\{\tilde{\gamma}^{\theta}_{k,t}\}_{k\in\mathbb{N}_+}$ as $k \to \infty$ exist with respect to the total variation distance. Let $\gamma^{\theta}_{\infty,t} := \lim_{k\to\infty} \gamma^{\theta}_{k,t}$ and $\tilde{\gamma}^{\theta}_{\infty,t} := \lim_{k\to\infty} \tilde{\gamma}^{\theta}_{k,t}$. They have the following properties:*

1. $\gamma^{\theta}_{\infty,t}$ and $\tilde{\gamma}^{\theta}_{\infty,t}$ do not depend on T.

2. $\gamma^{\theta}_{\infty,t}$ and $\tilde{\gamma}^{\theta}_{\infty,t}$ are entry-wise Lipschitz continuous with respect to $\theta \in \Theta$.

The proof is given in Appendix D.4. The dependency of $\gamma^{\theta}_{\infty,t}$ and $\tilde{\gamma}^{\theta}_{\infty,t}$ on $\{s_t, a_t\}_{t\in\mathbb{Z}}$ is omitted for a cleaner notation.

C.2. The stochastic convergence of the Q-function

Using the definitions from Section 4, the quantities defined in Appendix C.1 can also be analyzed in the *infinitely extended* probability space. Remember that $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{O} \times \{0, 1\}$. Formally, let $\Omega = \mathcal{X}^{\mathbb{Z}}$ and consider the probability space $(\Omega, 2^{\Omega}, \mathbb{P}_{\theta^*})$. From this perspective, $\gamma^{\theta}_{\infty,t}$ and $\tilde{\gamma}^{\theta}_{\infty,t}$ are both functions of the observation sequence, therefore also functions of a sample path $\omega \in \Omega$. For any sample path ω, let $\omega(s_t)$ and $\omega(a_t)$ be the values of S_t and A_t corresponding to the sample path ω. With a slight overload of notation, let $\omega(t) = \{\omega(s_t), \omega(a_t), \omega(o_t), \omega(b_t)\}$, which is the set of elements in ω corresponding to time t.

For all $\theta \in \Theta$, $\theta' \in \tilde{\Theta}$, $\omega \in \Omega$ and $t \in \mathbb{N}_+$, define

$$f_t(\theta'|\theta;\omega) := \sum_{o_{t-1},b_t} \tilde{\gamma}^{\theta}_{\infty,t}(o_{t-1}, b_t; \omega) \left[\log \pi_b(b_t|\omega(s_t), o_{t-1}; \theta'_b)\right] + \sum_{o_t,b_t} \gamma^{\theta}_{\infty,t}(o_t, b_t; \omega) \left[\log \pi_{lo}(\omega(a_t)|\omega(s_t), o_t; \theta'_{lo})\right] + \sum_{o_t} \gamma^{\theta}_{\infty,t}(o_t, b_t = 1; \omega) \left[\log \pi_{hi}(o_t|\omega(s_t); \theta'_{hi})\right],$$

where the dependency of the RHS on ω is shown explicitly for clarity. $|f_t(\theta'|\theta;\omega)|$ is upper bounded by a constant that does not depend on θ, θ', ω and t, due to Assumption 1 and Assumption 2. Moreover, for all θ, ω and t, $f_t(\theta'|\theta;\omega)$ is continuously differentiable with respect to $\theta' \in \tilde{\Theta}$. For all θ', ω and t, $f_t(\theta'|\theta;\omega)$ is Lipschitz continuous with respect to $\theta \in \Theta$, due to Lemma C.1.

Next, define

$$\bar{Q}(\theta'|\theta) := \mathbb{E}_{\theta^*}[f_1(\theta'|\theta;\omega)]. \tag{15}$$

The subscript θ^* in $\mathbb{E}_{\theta^*}$ means the expectation is taken with respect to the probability measure $\mathbb{P}_{\theta^*}$.

With the above definitions, we state the complete version of Theorem 1. The Q-function defined in (3) is written as $Q_{\mu,T}(\theta'|\theta;\omega)$, showing its dependency on the sample path.

Theorem 6 (The complete version of Theorem 1)**.** *For $\bar{Q}(\theta'|\theta)$ defined in (15), we have*

1. For any $\theta \in \Theta$, $\bar{Q}(\theta'|\theta)$ is continuously differentiable with respect to $\theta' \in \tilde{\Theta}$, where $\tilde{\Theta}$ is defined in Assumption 1. The gradient is

$$\nabla \bar{Q}(\theta'|\theta) = \mathbb{E}_{\theta^*}[\nabla f_1(\theta'|\theta;\omega)].$$

Moreover, as the set of maximizing arguments, $\arg\max_{\theta' \in \Theta} \bar{Q}(\theta'|\theta)$ is nonempty.

2. As $T \to \infty$,

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

Before proving Theorem 6, we state the following definition and an auxiliary lemma required for the proof.

For all $\theta, \theta' \in \Theta$, $\omega \in \Omega$ and $T \geq 2$, the sample-path-based population Q-function $Q^s_{\infty,T}(\theta'|\theta;\omega)$ is defined as

$$Q^s_{\infty,T}(\theta'|\theta;\omega) := \frac{1}{T} \sum_{t=1}^{T} f_t(\theta'|\theta;\omega). \tag{16}$$

The superscript s in $Q^s_{\infty,T}$ stands for *sample-path-based*. If the sample path ω is not specified, $Q^s_{\infty,T}(\theta'|\theta)$ is a random variable associated with probability measure $\mathbb{P}_{\theta^*}$. Note that due to stationarity, for any θ, θ' and T, $\bar{Q}(\theta'|\theta) = \mathbb{E}_{\theta^*}[Q^s_{\infty,T}(\theta'|\theta;\omega)]$.

The difference between $Q^s_{\infty,T}$ and $Q_{\mu,T}$ is bounded in the following lemma.

Lemma C.2 (Bounding the difference between the Q-function and the sample-path-based population Q-function)**.** *For all $T \geq 2$ and $\omega \in \Omega$,*

$$\sup_{\theta,\theta' \in \Theta} \sup_{\mu \in \mathcal{M}} \left|Q^s_{\infty,T}(\theta'|\theta;\omega) - Q_{\mu,T}(\theta'|\theta;\omega)\right| \leq const \cdot T^{-1},$$

where const is a constant independent of T and ω.

The proof is provided in Appendix D.5. Now we are ready to present the proof of Theorem 6 step-by-step. The structure of this proof is similar to the standard analysis of HMM maximum likelihood estimators (Cappé et al., 2006, Chap. 12).

Proof of Theorem 6

1. For all $\theta' \in \tilde{\Theta}$, there exists $\delta_{\theta'} > 0$ such that the set $\{\tilde{\theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta_{\theta'}\} \subseteq \tilde{\Theta}$. For all $\theta \in \Theta$ and $\omega \in \Omega$, due to the differentiability of $f_1(\theta'|\theta;\omega)$ with respect to θ', there exists a gradient $\nabla f_1(\theta'|\theta;\omega)$ at any $\theta' \in \tilde{\Theta}$ such that

$$\lim_{\delta \to 0} \sup_{\tilde{\theta} \in \tilde{\Theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta} \frac{|f_1(\tilde{\theta}|\theta;\omega) - f_1(\theta'|\theta;\omega) - \langle \nabla f_1(\theta'|\theta;\omega), \tilde{\theta} - \theta' \rangle|}{\|\tilde{\theta} - \theta'\|_2} = 0.$$

We need to transform the above pointwise (in ω) convergence to the convergence of expectation, using the dominated convergence theorem. As a requirement, the quantity inside the limit on the LHS needs to be upper-bounded. For all $\theta \in \Theta$, $\theta' \in \tilde{\Theta}$, $\omega \in \Omega$ and $0 < \delta \le \delta_{\theta'}$,

$$\begin{aligned}\sup_{\tilde{\theta} \in \tilde{\Theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta} \frac{|f_1(\tilde{\theta}|\theta;\omega) - f_1(\theta'|\theta;\omega) - \langle \nabla f_1(\theta'|\theta;\omega), \tilde{\theta} - \theta' \rangle|}{\|\tilde{\theta} - \theta'\|_2} \le & \sup_{\tilde{\theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta_{\theta'}} \frac{|f_1(\tilde{\theta}|\theta;\omega) - f_1(\theta'|\theta;\omega)|}{\|\tilde{\theta} - \theta'\|_2} \\ & + \sup_{\tilde{\theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta_{\theta'}} \frac{|\langle \nabla f_1(\theta'|\theta;\omega), \tilde{\theta} - \theta' \rangle|}{\|\tilde{\theta} - \theta'\|_2}. \end{aligned} \tag{17}$$

Since continuously differentiable functions are Lipschitz continuous on convex and compact subsets, π_{hi}, π_{lo} and π_b as functions of $\tilde{\theta} \in \tilde{\Theta}$ are Lipschitz continuous on $\{\tilde{\theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta_{\theta'}\}$, with any other input arguments. Therefore from the expression of f_1, we can verify that for any fixed θ and ω, $f_1(\tilde{\theta}|\theta;\omega)$ as a function of $\tilde{\theta}$ is Lipschitz continuous on $\{\tilde{\theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta_{\theta'}\}$, and the Lipschitz constant only depends on θ' and $\delta_{\theta'}$. Consequently, the RHS of (17) can be upper-bounded uniformly in $\omega \in \Omega$. Applying the dominated convergence theorem, we have

$$\lim_{\delta \to 0} \mathbb{E}_{\theta^*} \left[\sup_{\tilde{\theta} \in \tilde{\Theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta} \frac{|f_1(\tilde{\theta}|\theta;\omega) - f_1(\theta'|\theta;\omega) - \langle \nabla f_1(\theta'|\theta;\omega), \tilde{\theta} - \theta' \rangle|}{\|\tilde{\theta} - \theta'\|_2} \right] = 0. \tag{18}$$

On the other hand, notice that for all $\theta \in \Theta$, $\theta' \in \tilde{\Theta}$ and $\delta > 0$,

$$\begin{aligned} & \sup_{\tilde{\theta} \in \tilde{\Theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta} \frac{|\bar{Q}(\tilde{\theta}|\theta) - \bar{Q}(\theta'|\theta) - \langle \mathbb{E}_{\theta^*}[\nabla f_1(\theta'|\theta;\omega)], \tilde{\theta} - \theta' \rangle|}{\|\tilde{\theta} - \theta'\|_2} \\ = & \sup_{\tilde{\theta} \in \tilde{\Theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta} \frac{|\mathbb{E}_{\theta^*}[f_1(\tilde{\theta}|\theta;\omega) - f_1(\theta'|\theta;\omega) - \langle \nabla f_1(\theta'|\theta;\omega), \tilde{\theta} - \theta' \rangle]|}{\|\tilde{\theta} - \theta'\|_2} \\ \le & \ \mathbb{E}_{\theta^*} \left[\sup_{\tilde{\theta} \in \tilde{\Theta}; \|\tilde{\theta} - \theta'\|_2 \le \delta} \frac{|f_1(\tilde{\theta}|\theta;\omega) - f_1(\theta'|\theta;\omega) - \langle \nabla f_1(\theta'|\theta;\omega), \tilde{\theta} - \theta' \rangle|}{\|\tilde{\theta} - \theta'\|_2} \right]. \end{aligned}$$

Combining with (18) proves the differentiability of $\bar{Q}(\theta'|\theta)$ with respect to $\theta' \in \tilde{\Theta}$ for any fixed θ. The gradient is

$$\nabla \bar{Q}(\theta'|\theta) = \mathbb{E}_{\theta^*}[\nabla f_1(\theta'|\theta;\omega)].$$

Analogously, using the dominated convergence theorem we can also show that the gradient $\nabla \bar{Q}(\theta'|\theta)$ is continuous with respect to $\theta' \in \tilde{\Theta}$. Details are omitted due to the similarity with the above procedure. It is worth noting that we let $\theta' \in \tilde{\Theta}$ instead of Θ. In this way, the gradient $\nabla \bar{Q}(\theta'|\theta)$ can be naturally defined when θ' is not an interior point of Θ.

From differentiability and $\Theta \subseteq \tilde{\Theta}$, $\bar{Q}(\theta'|\theta)$ is also continuous with respect to $\theta' \in \Theta$. Since Θ is compact, the set of maximizing arguments $\arg\max_{\theta' \in \Theta} \bar{Q}(\theta'|\theta)$ is nonempty.

2. We need to prove the uniform (in $\theta, \theta' \in \Theta$ and $\mu \in \mathcal{M}$) almost sure convergence of the Q-function $Q_{\mu,T}(\theta'|\theta;\omega)$ to the population Q-function $\bar{Q}(\theta'|\theta)$. The proof is separated into three steps. First, we show the almost sure convergence of $Q^s_{\infty,T}(\theta'|\theta;\omega)$ to $\bar{Q}(\theta'|\theta)$ for all $\theta, \theta' \in \Theta$ using the ergodic theorem. Second, we extend this pointwise convergence

to uniform (in θ, θ') convergence using a version of the Arzelà-Ascoli theorem (Davidson, 1994, Chap. 21). Finally, from Lemma C.2, the difference between $Q_{\mu,T}(\theta'|\theta;\omega)$ and $Q^s_{\infty,T}(\theta'|\theta;\omega)$ vanishes uniformly in μ as $T \to \infty$.

Concretely, for the pointwise (in θ, θ') almost sure convergence of $Q^s_{\infty,T}(\theta'|\theta;\omega)$ as $T \to \infty$, we apply Birkhoff's ergodic theorem. Let $\mathcal{T} : \Omega \to \Omega$ be the standard shift operator. That is, for any $t \in \mathbb{Z}$, $\mathcal{T}\omega(t) = \omega(t+1)$. Due to stationarity, $\mathcal{T}$ is a measure-preserving map, i.e., $\mathbb{P}_{\theta^*}(\mathcal{T}^{-1}F) = \mathbb{P}_{\theta^*}(F)$ for all $F \in 2^\Omega$. Therefore, the quadruple $\{\Omega, 2^\Omega, \mathbb{P}_{\theta^*}, \mathcal{T}\}$ defines a dynamical system.

Since $\mathbb{P}_{\theta^*}$ is extended from the unique stationary distribution of the Markov chain with the true parameter θ^*, the dynamical system $\{\Omega, 2^\Omega, \mathbb{P}_{\theta^*}, \mathcal{T}\}$ is ergodic (Hairer, 2006, Corollary 5.12). For our case, Birkhoff's ergodic theorem is restated as follows.

Lemma C.3 ((Hairer, 2006), Corollary 5.3 restated)**.** *If a dynamical system $\{\Omega, 2^\Omega, \mathbb{P}_{\theta^*}, \mathcal{T}\}$ is ergodic and $f : \Omega \to \mathbb{R}$ satisfies $\mathbb{E}_{\theta^*}[f(\omega)] < \infty$, then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathcal{T}^t\omega) \to \mathbb{E}_{\theta^*}[f(\omega)],\ P_{\theta^*}\text{-a.s.}$$

For our purpose, observe that for any $\theta, \theta' \in \Theta$, $f_t(\theta'|\theta;\omega) = f_1(\theta'|\theta;\mathcal{T}^{t-1}\omega)$. Therefore, applying the ergodic theorem to $Q^s_{\infty,T}(\theta'|\theta)$, as $T \to \infty$,

$$Q^s_{\infty,T}(\theta'|\theta;\omega) \to \bar{Q}(\theta'|\theta),\ P_{\theta^*}\text{-a.s.} \tag{19}$$

To extend the pointwise convergence in (19) to uniform (in θ, θ') convergence, the following concept is required. The sequence $\{Q^s_{\infty,T}(\theta'|\theta)\}$ indexed by T as functions of θ and θ' is *strongly stochastically equicontinuous* (Davidson, 1994, Equation 21.43) if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\limsup_{T\to\infty} \sup_{\theta_1,\theta_1',\theta_2,\theta_2' \in \Theta; \|\theta_1-\theta_2\|_2 + \|\theta_1'-\theta_2'\|_2 \le \delta} \left|Q^s_{\infty,T}(\theta_1'|\theta_1;\omega) - Q^s_{\infty,T}(\theta_2'|\theta_2;\omega)\right| < \varepsilon,\ P_{\theta^*}\text{-a.s.} \tag{20}$$

Indeed this property holds for $\{Q^s_{\infty,T}(\theta'|\theta)\}$, as shown in Appendix D.6. The version of the Arzelà-Ascoli theorem we use is restated as follows, tailored to our need.

Lemma C.4 ((Davidson, 1994), Theorem 21.8 restated)**.** *Given (19) and (20) hold, as $T \to \infty$ we have*

$$\sup_{\theta,\theta' \in \Theta} \left|Q^s_{\infty,T}(\theta'|\theta;\omega) - \bar{Q}(\theta'|\theta)\right| \to 0,\ P_{\theta^*}\text{-a.s.}$$

Combining Lemma C.2 and Lemma C.4 concludes the proof of the second part. □

C.3. The convergence of the population version algorithm

We first present the complete version of Theorem 2, where an upper bound on γ is also shown. Notice that we assume all the assumptions, including Assumption 5.

Theorem 7 (The complete version of Theorem 2)**.** *With all the assumptions,*

1. (First-order stability) There exists $0 < \gamma \le \bar{\gamma}$ such that for all $\theta \in \Theta_r$,

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

Specifically, the upper bound $\bar{\gamma}$ is given by

$$\bar{\gamma} = \frac{4|\mathcal{O}|L_{\theta^*,r}}{\varepsilon_b^2\zeta}\left(\sup_{\theta'\in\Theta_r} z_{\theta',\theta^*}\right)\left(2\max_{o_0,s_1,b_1}\sup_{\theta_b'\in\Theta_b}\|\nabla\log\pi_b(b_1|s_1,o_0;\theta_b')\|_2 + \max_{s_1,a_1,o_1}\sup_{\theta_{lo}'\in\Theta_{lo}}\|\nabla\log\pi_{lo}(a_1|s_1,o_1;\theta_{lo}')\|_2 + \max_{s_1,o_1}\sup_{\theta_{hi}'\in\Theta_{hi}}\|\nabla\log\pi_{hi}(o_1|s_1;\theta_{hi}')\|_2\right).$$

ζ is the failure parameter in the options with failure framework; ε_b is a mixing constant defined in Lemma D.1; $L_{\theta^*,r}$ is a Lipschitz constant defined in Lemma D.2; z_{θ',θ^*} is defined in Lemma D.5.

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^*.

Proof of Theorem 7

1. For convenience of notation, let $\nabla\bar{Q}(\theta'|\theta) = [\nabla_b\bar{Q}(\theta'|\theta), \nabla_{lo}\bar{Q}(\theta'|\theta), \nabla_{hi}\bar{Q}(\theta'|\theta)]$ such that, for example, $\nabla_b\bar{Q}(\theta'|\theta)$ is the gradient of $\bar{Q}(\theta'|\theta)$ with respect to θ'_b. Using the expressions of $\nabla\bar{Q}(\theta'|\theta)$ from Theorem 6, we have

$$\begin{aligned}\left\|\nabla\bar{Q}(\bar{M}(\theta)|\theta) - \nabla\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2 \leq{}& \left\|\nabla_b\bar{Q}(\bar{M}(\theta)|\theta) - \nabla_b\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2 + \left\|\nabla_{lo}\bar{Q}(\bar{M}(\theta)|\theta) - \nabla_{lo}\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2 \\ &+ \left\|\nabla_{hi}\bar{Q}(\bar{M}(\theta)|\theta) - \nabla_{hi}\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2\end{aligned}$$

Consider the first term,

$$\begin{aligned}&\left\|\nabla_b\bar{Q}(\bar{M}(\theta)|\theta) - \nabla_b\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2 \\ ={}& \left\|\mathbb{E}_{\theta^*}\left\{\sum_{o_0,b_1}\left[\tilde{\gamma}^{\theta}_{\infty,1}(o_0,b_1;\omega) - \tilde{\gamma}^{\theta^*}_{\infty,1}(o_0,b_1;\omega)\right]\left[\nabla\log\pi_b(b_1|\omega(s_1),o_0;\bar{M}(\theta)_b)\right]\right\}\right\|_2 \\ \leq{}& \sum_{o_0,b_1}\left\|\mathbb{E}_{\theta^*}\left\{\left[\tilde{\gamma}^{\theta}_{\infty,1}(o_0,b_1;\omega) - \tilde{\gamma}^{\theta^*}_{\infty,1}(o_0,b_1;\omega)\right]\left[\nabla\log\pi_b(b_1|\omega(s_1),o_0;\bar{M}(\theta)_b)\right]\right\}\right\|_2 \\ \leq{}& \sum_{o_0,b_1}\mathbb{E}_{\theta^*}\left\{\left|\tilde{\gamma}^{\theta}_{\infty,1}(o_0,b_1;\omega) - \tilde{\gamma}^{\theta^*}_{\infty,1}(o_0,b_1;\omega)\right|\left\|\nabla\log\pi_b(b_1|\omega(s_1),o_0;\bar{M}(\theta)_b)\right\|_2\right\} \\ \leq{}& \max_{o_0,s_1,b_1}\sup_{\theta'_b\in\Theta_b}\left\|\nabla\log\pi_b(b_1|s_1,o_0;\theta'_b)\right\|_2\mathbb{E}_{\theta^*}\left\{\sum_{o_0,b_1}\left|\tilde{\gamma}^{\theta}_{\infty,1}(o_0,b_1;\omega) - \tilde{\gamma}^{\theta^*}_{\infty,1}(o_0,b_1;\omega)\right|\right\} \\ \leq{}& 2\max_{o_0,s_1,b_1}\sup_{\theta'_b\in\Theta_b}\left\|\nabla\log\pi_b(b_1|s_1,o_0;\theta'_b)\right\|_2 \times \sup_{\omega\in\Omega}\left\|\tilde{\gamma}^{\theta}_{\infty,1}(\omega) - \tilde{\gamma}^{\theta^*}_{\infty,1}(\omega)\right\|_{\mathrm{TV}} \\ \leq{}& \frac{8|\mathcal{O}|L_{\theta^*,r}}{\varepsilon_b^2\zeta}\left(\sup_{\theta'\in\Theta_r} z_{\theta',\theta^*}\right)\left(\max_{o_0,s_1,b_1}\sup_{\theta'_b\in\Theta_b}\left\|\nabla\log\pi_b(b_1|s_1,o_0;\theta'_b)\right\|_2\right)\left\|\theta - \theta^*\right\|_2 .\end{aligned}$$

We use the triangle inequality and the Jensen's inequality in the third and the fourth line respectively. The fifth line is finite due to θ_b being compact and the continuity of the gradient (Assumption 2). The last line is due to the limit form of Lemma D.7, similar to the argument in Appendix D.4. Notice that the coefficient of $\|\theta - \theta^*\|_2$ on the last line does not depend on θ.

Analogously, we have

$$\begin{aligned}&\left\|\nabla_{lo}\bar{Q}(\bar{M}(\theta)|\theta) - \nabla_{lo}\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2 \leq \\ &\qquad\frac{4|\mathcal{O}|L_{\theta^*,r}}{\varepsilon_b^2\zeta}\left(\sup_{\theta'\in\Theta_r} z_{\theta',\theta^*}\right)\left(\max_{s_1,a_1,o_1}\sup_{\theta'_{lo}\in\Theta_{lo}}\left\|\nabla\log\pi_{lo}(a_1|s_1,o_1;\theta'_{lo})\right\|_2\right)\left\|\theta - \theta^*\right\|_2 ,\end{aligned}$$

and

$$\begin{aligned}&\left\|\nabla_{hi}\bar{Q}(\bar{M}(\theta)|\theta) - \nabla_{hi}\bar{Q}(\bar{M}(\theta)|\theta^*)\right\|_2 \leq \\ &\qquad\frac{4|\mathcal{O}|L_{\theta^*,r}}{\varepsilon_b^2\zeta}\left(\sup_{\theta'\in\Theta_r} z_{\theta',\theta^*}\right)\left(\max_{s_1,o_1}\sup_{\theta'_{hi}\in\Theta_{hi}}\left\|\nabla\log\pi_{hi}(o_1|s_1;\theta'_{hi})\right\|_2\right)\left\|\theta - \theta^*\right\|_2 .\end{aligned}$$

Combining everything, we have the upper bound on γ.

2. The proof of the second part mirrors the proof of (Balakrishnan et al., 2017, Theorem 1). The main difference is the construction of the following self-consistency (*a.k.a.* fixed-point) condition.

Lemma C.5 (Self-consistency)**.** *With all the assumptions,* $\theta^* = \bar{M}(\theta^*)$.

The proof of this lemma is presented in Appendix D.7. Such a condition is used without proof in (Balakrishnan et al., 2017) since it only considers i.i.d. samples, and the self-consistency condition for EM with i.i.d. samples is a well-established result. However, for the case of dependent samples like our graphical model, such a condition results from the stochastic convergence of the Q-function which is not immediate.

For the rest of the proof, we present a brief sketch here for completeness. Due to concavity, we have the first order optimality conditions: for any $\theta, \theta' \in \Theta_r$, $\langle \nabla \bar{Q}(\bar{M}(\theta^*)|\theta^*), \theta - \bar{M}(\theta^*)\rangle \le 0$ and $\langle \nabla \bar{Q}(\bar{M}(\theta)|\theta), \theta' - \bar{M}(\theta)\rangle \le 0$. Using $\theta^* = \bar{M}(\theta^*)$, we can combine the two optimality conditions together and obtain the following. For any $\theta \in \Theta_r$,

$$\langle \nabla \bar{Q}(\bar{M}(\theta)|\theta^*) - \nabla \bar{Q}(\theta^*|\theta^*), \theta^* - \bar{M}(\theta)\rangle \le \langle \nabla \bar{Q}(\bar{M}(\theta)|\theta^*) - \nabla \bar{Q}(\bar{M}(\theta)|\theta), \theta^* - \bar{M}(\theta)\rangle.$$

From the strong concavity assumption, LHS $\ge \lambda \|\theta^* - \bar{M}(\theta)\|_2^2$. From Cauchy-Schwarz and the first order stability assumption, RHS $\le \gamma \|\theta^* - \bar{M}(\theta)\|_2 \|\theta - \theta^*\|_2$. Canceling $\|\theta^* - \bar{M}(\theta)\|_2$ on both sides completes the proof. □

C.4. Proof of Theorem 3

1. We first show the strong consistency of $M_{\mu,T}(\theta;\omega)$, the parameter update of Algorithm 1, as an estimator of $\bar{M}(\theta)$. This follows from standard techniques in the analysis of M-estimators. In particular, for all $\theta \in \Theta$, $\omega \in \Omega$, $T \ge 2$ and $\mu \in \mathcal{M}$,

$$\begin{aligned}
0 &\le \bar{Q}(\bar{M}(\theta)|\theta) - \bar{Q}(M_{\mu,T}(\theta;\omega)|\theta) \\
&\le \bar{Q}(\bar{M}(\theta)|\theta) - Q_{\mu,T}(\bar{M}(\theta)|\theta;\omega) + Q_{\mu,T}(\bar{M}(\theta)|\theta;\omega) - Q_{\mu,T}(M_{\mu,T}(\theta;\omega)|\theta;\omega) \\
&\qquad\qquad + Q_{\mu,T}(M_{\mu,T}(\theta;\omega)|\theta;\omega) - \bar{Q}(M_T(\theta;\omega)|\theta) \\
&\le 2 \sup_{\theta' \in \Theta} \left|\bar{Q}(\theta'|\theta) - Q_{\mu,T}(\theta'|\theta;\omega)\right|.
\end{aligned}$$

From Theorem 6, $\mathbb{P}_{\theta^*}$ almost surely, $\sup_{\theta,\theta' \in \Theta} \sup_{\mu \in \mathcal{M}} |\bar{Q}(\theta'|\theta) - Q_{\mu,T}(\theta'|\theta;\omega)| \to 0$ as $T \to \infty$. Therefore,

$$\sup_{\theta \in \Theta_r} \sup_{\mu \in \mathcal{M}} \left[\bar{Q}(\bar{M}(\theta)|\theta) - \bar{Q}(M_{\mu,T}(\theta;\omega)|\theta)\right] \to 0, \ P_{\theta^*}\text{-a.s.}$$

An equivalent argument is the following. $\mathbb{P}_{\theta^*}$ almost surely, for any $\delta > 0$ there exists $T_\omega \in \mathbb{N}_+$ such that for all $T \ge T_\omega$, $\sup_{\theta \in \Theta_r} \sup_{\mu \in \mathcal{M}} [\bar{Q}(\bar{M}(\theta)|\theta) - \bar{Q}(M_{\mu,T}(\theta;\omega)|\theta)] \le \delta$. In particular, for any $\varepsilon > 0$, let

$$\delta = \frac{1}{2} \inf_{\theta \in \Theta_r} \left[\bar{Q}(\bar{M}(\theta)|\theta) - \sup_{\theta' \in \Theta; \|\theta' - \bar{M}(\theta)\|_2 \ge \varepsilon} \bar{Q}(\theta'|\theta)\right].$$

From the identifiability assumption (Assumption 5), the RHS is positive. Therefore, such an assignment of δ is valid. Consequently, for all $T \ge T_\omega$, $\theta \in \Theta_r$ and $\mu \in \mathcal{M}$,

$$\bar{Q}(\bar{M}(\theta)|\theta) - \bar{Q}(M_{\mu,T}(\theta;\omega)|\theta) < \bar{Q}(\bar{M}(\theta)|\theta) - \sup_{\theta' \in \Theta; \|\theta' - \bar{M}(\theta)\|_2 \ge \varepsilon} \bar{Q}(\theta'|\theta),$$

which means that $\|M_{\mu,T}(\theta;\omega) - \bar{M}(\theta)\|_2 < \varepsilon$. Taking supremum over $\theta \in \Theta_r$ and $\mu \in \mathcal{M}$, we summarize the argument as the following. $\mathbb{P}_{\theta^*}$ almost surely, for any $\varepsilon > 0$ there exists $T_\omega \in \mathbb{N}_+$ such that for all $T \ge T_\omega$,

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

Such a result is equivalent to the uniform (in θ and μ) strong consistency of $M_{\mu,T}(\theta;\omega)$ as an estimator of $\bar{M}(\theta)$. As $T \to \infty$,

$$\sup_{\theta \in \Theta_r} \sup_{\mu \in \mathcal{M}} \left\|M_{\mu,T}(\theta;\omega) - \bar{M}(\theta)\right\|_2 \to 0, \ P_{\theta^*}\text{-a.s.}$$

This result is insufficient for Part 1, since T_ω is sample path dependent. To get rid of this sample path dependency, we use the dominated convergence theorem. Notice that for all $T \ge 2$ and $\omega \in \Omega$, $\sup_{\theta \in \Theta_r} \sup_{\mu \in \mathcal{M}} \|M_{\mu,T}(\theta;\omega) - \bar{M}(\theta)\|_2$ is bounded due to the compactness of Θ. Therefore we have

$$\lim_{T \to \infty} \mathbb{E}_{\theta^*} \left[\sup_{\theta \in \Theta_r} \sup_{\mu \in \mathcal{M}} \left\|M_{\mu,T}(\theta;\omega) - \bar{M}(\theta)\right\|_2\right] = 0.$$

For any $q > 0$, there exists $\underline{T}(q) \in \mathbb{N}_+$ such that for all $T \geq \underline{T}(q)$,

$$\mathbb{E}_{\theta^*}\left[\sup_{\theta\in\Theta_r}\sup_{\mu\in\mathcal{M}}\left\|M_{\mu,T}(\theta;\omega)-\bar{M}(\theta)\right\|_2\right]\leq q.$$

Applying Markov's inequality, for any $\Delta > 0$,

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta_r}\sup_{\mu\in\mathcal{M}}\left\|M_{\mu,T}(\theta;\omega)-\bar{M}(\theta)\right\|_2\geq\Delta\right)\leq\frac{1}{\Delta}\mathbb{E}_{\theta^*}\left[\sup_{\theta\in\Theta_r}\sup_{\mu\in\mathcal{M}}\left\|M_{\mu,T}(\theta;\omega)-\bar{M}(\theta)\right\|_2\right]\leq\frac{q}{\Delta}.$$

Scaling q yields the desired result.

2. The proof of Part 2 is the same as (Balakrishnan et al., 2017, Theorem 2). We present a sketch for completeness. For all $T \geq \underline{T}(\Delta, q)$, condition the following proof on the high probability event that $\sup_{\theta\in\Theta_r}\sup_{\mu\in\mathcal{M}}\left\|M_{\mu,T}(\theta;\omega)-\bar{M}(\theta)\right\|_2 \leq \Delta$.

Assume $\|\theta^{(n-1)}-\theta^*\|_2 \leq r$, which holds for $n = 1$. Then, using the triangle inequality, the result from Theorem 2, the above concentration and $\Delta \leq (1-\kappa)r$, we have the following for any μ.

$$\left\|\theta^{(n)}-\theta^*\right\|_2\leq\left\|\bar{M}(\theta^{(n-1)})-\theta^*\right\|_2+\left\|M_{\mu,T}(\theta^{(n-1)})-\bar{M}(\theta^{(n-1)})\right\|_2\leq\kappa\|\theta^{(n-1)}-\theta^*\|_2+\Delta, \tag{21}$$

and $\|\theta^{(n)}-\theta^*\|_2 \leq \kappa r + (1-\kappa)r = r$. From induction, the one step relation (21) holds for all $n \in \mathbb{N}_+$. Unrolling (21) and regrouping the terms completes the proof. □

D. Proofs of auxiliary lemmas

This section presents proofs omitted in earlier sections.

In particular, the first three subsections develop a few essential lemmas required for the proofs in later subsections. Assumptions 1, 2, 3 and 4 are assumed. Concretely, in Appendix D.1 we show an important mixing property of the options with failure framework. In Appendix D.2, such a mixing property is used to prove a general contraction result of our forward-backward smoothing procedure (Theorem 4 and Corollary 5), similar to the concept of *filtering stability* in the HMM literature. At a high level, considering the forward-backward recursion in the extended graphical model (Corollary 5), this result characterizes the effect of changing θ and the boundary conditions $\alpha^{\theta}_{k,0}$ and $\beta^{\theta}_{k,T}$ on the smoothing distribution $\gamma^{\theta}_{k,t}$, given any observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$. Due to this high level reasoning, we name this result as the *smoothing stability* lemma. Appendix D.3 provides concrete applications of this lemma to quantities defined in earlier sections.

D.1. Mixing

Remember that ζ is the auxiliary parameter in the options with failure framework.

Lemma D.1 (Mixing)**.** *There exists a constant $\varepsilon_b > 0$ and a conditional distribution $\bar{\pi}_{o,b}(o_t, b_t|s_t;\theta)$ parameterized by θ such that for all $\theta \in \Theta$, with any input arguments b_t, s_t, o_{t-1} and o_t,*

$$0<\varepsilon_b\zeta\bar{\pi}_{o,b}(o_t,b_t|s_t;\theta)\leq\pi_b(b_t|s_t,o_{t-1};\theta_b)\bar{\pi}_{hi}(o_t|s_t,o_{t-1},b_t;\theta_{hi})\leq\varepsilon_b^{-1}|\mathcal{O}|\bar{\pi}_{o,b}(o_t,b_t|s_t;\theta).$$

Proof of Lemma D.1

The proof is separated into two parts.

1. We first show an intermediate result: there exists a constant $\varepsilon_b > 0$ and a conditional distribution $\bar{\pi}_b(b_t|s_t;\theta_b)$ parameterized by θ_b such that for all $\theta_b \in \Theta_b$, with any input arguments b_t, s_t and o_{t-1},

$$0<\varepsilon_b\bar{\pi}_b(b_t|s_t;\theta_b)\leq\pi_b(b_t|s_t,o_{t-1};\theta_b)\leq\varepsilon_b^{-1}\bar{\pi}_b(b_t|s_t;\theta_b).$$

This can be proved as follows. Let $c_b = \inf_{\theta_b\in\Theta_b}\min_{b_t,s_t,o_{t-1}}\pi_b(b_t|s_t,o_{t-1};\theta_b)$. Similar to the procedure in Appendix A, from the non-degeneracy assumption, the differentiabiilty assumption and Θ being compact, we have $c_b > 0$. For

any $\theta_b \in \Theta_b$, with any input arguments b_t and s_t, let $f(b_t, s_t; \theta_b) = \min_{o_{t-1} \in \mathcal{O}} \pi_b(b_t|s_t, o_{t-1}; \theta_b)$. Observe that $c_b \leq f(b_t, s_t; \theta_b) \leq 1$. Let $\varepsilon_b = c_b/2$ and

$$\bar{\pi}_b(b_t|s_t; \theta_b) = \frac{f(b_t, s_t; \theta_b)}{\sum_{b'_t \in \{0,1\}} f(b'_t, s_t; \theta_b)}.$$

Clearly $\varepsilon_b \bar{\pi}_b(b_t|s_t; \theta_b) > 0$. Moreover, for any o_{t-1}, $\varepsilon_b \bar{\pi}_b(b_t|s_t; \theta_b) < 2c_b \bar{\pi}_b(b_t|s_t; \theta_b) \leq f(b_t, s_t; \theta_b) \leq \pi_b(b_t|s_t, o_{t-1}; \theta_b)$.

On the other hand, with any input arguments,

$$\varepsilon_b^{-1} \bar{\pi}_b(b_t|s_t; \theta_b) \geq \varepsilon_b^{-1} c_b/2 = 1 \geq \pi_b(b_t|s_t, o_{t-1}; \theta_b),$$

which completes the proof of the first part.

2. Define $\bar{\pi}_{o,b}(o_t, b_t|s_t; \theta)$ as follows. With any input arguments, let

$$\begin{aligned}
\bar{\pi}_{o,b}(o_t, b_t = 0|s_t; \theta) &:= \bar{\pi}_b(b_t = 0|s_t; \theta_b)/|\mathcal{O}|, \\
\bar{\pi}_{o,b}(o_t, b_t = 1|s_t; \theta) &:= \bar{\pi}_b(b_t = 1|s_t; \theta_b) \pi_{hi}(o_t|s_t; \theta_{hi}).
\end{aligned}$$

Clearly $\varepsilon_b \zeta \bar{\pi}_{o,b}(o_t, b_t|s_t; \theta) > 0$. Omit the dependency on θ for a cleaner notation since every term is parameterized by θ. When $b_t = 1$, with any other input arguments,

$$\varepsilon_b \bar{\pi}_b(b_t = 1|s_t) \pi_{hi}(o_t|s_t) \leq \pi_b(b_t = 1|s_t, o_{t-1}) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t = 1) \leq \varepsilon_b^{-1} \bar{\pi}_b(b_t = 1|s_t) \pi_{hi}(o_t|s_t).$$

Similarly, when $b_t = 0$ and $o_t = o_{t-1}$,

$$\begin{aligned}
\varepsilon_b \bar{\pi}_b(b_t = 0|s_t) \zeta/|\mathcal{O}| &\leq \varepsilon_b \bar{\pi}_b(b_t = 0|s_t) \left(1 - \frac{|\mathcal{O}| - 1}{|\mathcal{O}|} \zeta\right) \\
&\leq \pi_b(b_t = 0|s_t, o_{t-1}) \bar{\pi}_{hi}(o_t = o_{t-1}|s_t, o_{t-1}, b_t = 0) \leq \varepsilon_b^{-1} \bar{\pi}_b(b_t = 0|s_t).
\end{aligned}$$

Finally, when $b_t = 0$ and $o_t \neq o_{t-1}$,

$$\varepsilon_b \bar{\pi}_b(b_t = 0|s_t) \zeta/|\mathcal{O}| \leq \pi_b(b_t = 0|s_t, o_{t-1}) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t = 0) \leq \varepsilon_b^{-1} \bar{\pi}_b(b_t = 0|s_t) \zeta/|\mathcal{O}|.$$

Combining the above cases and the definition of $\bar{\pi}_{o,b}(o_t, b_t|s_t; \theta)$ completes the proof. □

D.2. Smoothing stability

Before stating the smoothing stability lemma, we introduce a few definitions. The quantities defined in this subsection depend on an observation sequence $\{s_t, a_t\}_{t \in \mathbb{Z}}$, but such a dependency is usually omitted for simplifying the notation, unless specified otherwise. Consistent with our notations so far, in the following we make extensive use of the proportional symbol $\propto$.

D.2.1. Forward and backward recursion operators

With any given observation sequence $\{s_t, a_t\}_{t \in \mathbb{Z}}$ and any $\theta \in \Theta$, define the filtering operator F_t^θ as the following. For any probability measure φ on $\mathcal{O} \times \{0, 1\}$, $F_t^\theta \varphi$ is also a probability measure such that with any input arguments o_t and b_t,

$$F_t^\theta \varphi(o_t, b_t) \propto \sum_{o_{t-1}, b_{t-1}} \pi_b(b_t|s_t, o_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t; \theta_{lo}) \varphi(o_{t-1}, b_{t-1}). \tag{22}$$

The RHS has exactly the form of the forward recursion, therefore the recursion on both $\alpha_{k,t}^\theta$ in (6) and $\alpha_{\mu,t}^\theta$ in (11) can be expressed using F_t^θ. For generality, let $\{\varphi_t^\theta\}_{t \in \mathbb{Z}}$ and $\{\hat{\varphi}_t^{\hat{\theta}}\}_{t \in \mathbb{Z}}$ be any two indexed sets of probability measures such that $F_t^\theta \varphi_{t-1}^\theta = \varphi_t^\theta$ and $F_t^{\hat{\theta}} \hat{\varphi}_{t-1}^{\hat{\theta}} = \hat{\varphi}_t^{\hat{\theta}}$. We restrict $\{\varphi_t^\theta\}_{t \in \mathbb{Z}}$ and $\{\hat{\varphi}_t^{\hat{\theta}}\}_{t \in \mathbb{Z}}$ for any t to be strictly positive. Due to Assumption 1, such a restriction is valid. Notice that θ and $\hat{\theta}$ here can be equal. We use the seemingly more complicated notation $\{\hat{\varphi}_t^{\hat{\theta}}\}_{t \in \mathbb{Z}}$ because even if $\theta = \hat{\theta}$, $\{\varphi_t^\theta\}_{t \in \mathbb{Z}}$ and $\{\hat{\varphi}_t^{\hat{\theta}}\}_{t \in \mathbb{Z}}$ are still different; in this case they are just two different sets of probability measures satisfying the same recursion (F_t^θ).

Similarly, we define the backward recursion operator B_t^θ as follows. For any probability measure ρ on $\mathcal{O} \times \{0,1\}$, $B_t^\theta \rho$ is also a probability measure such that with any input arguments o_t and b_t,

$$B_t^\theta \rho(o_t, b_t) \propto \sum_{o_{t+1}, b_{t+1}} \pi_b(b_{t+1}|s_{t+1}, o_t; \theta_b) \bar{\pi}_{hi}(o_{t+1}|s_{t+1}, o_t, b_{t+1}; \theta_{hi}) \pi_{lo}(a_{t+1}|s_{t+1}, o_{t+1}; \theta_{lo}) \rho(o_{t+1}, b_{t+1}). \quad (23)$$

The recursion on both $\beta_{t|T}^\theta$ in (8) and $\beta_{k,t}^\theta$ in (12) can be expressed using B_t^θ. Let $\{\rho_t^\theta\}_{t\in\mathbb{Z}}$ and $\{\hat{\rho}_t^{\hat{\theta}}\}_{t\in\mathbb{Z}}$ be any two indexed sets of probability measures such that $B_t^\theta \rho_{t+1}^\theta = \rho_t^\theta$ and $B_t^{\hat{\theta}} \hat{\rho}_{t+1}^{\hat{\theta}} = \hat{\rho}_t^{\hat{\theta}}$. We restrict $\{\rho_t^\theta\}_{t\in\mathbb{Z}}$ and $\{\hat{\rho}_t^{\hat{\theta}}\}_{t\in\mathbb{Z}}$ for any t to be strictly positive.

The operation $\otimes$ is defined as follows: $\{(\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_t\}_{t\in\mathbb{Z}}$ is an indexed set of probability measures such that for any input arguments o_t and b_t,

$$(\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_t(o_t, b_t) \propto \varphi_t^\theta(o_t, b_t) \hat{\rho}_t^{\hat{\theta}}(o_t, b_t). \quad (24)$$

Finally, we clarify the use of $\propto$ in the above definitions. In (22), (23) and (24), the normalizing constants replaced by $\propto$ are independent of the input arguments (o_t, b_t).

D.2.2. Forward and backward smoothing operators

For any $\theta, \hat{\theta} \in \Theta$ and any t, with any observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$ and any input arguments o_t and b_t, observe that

$$(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_t(o_t, b_t) \propto \sum_{o_{t-1}, b_{t-1}} \pi_b(b_t|s_t, o_{t-1}; \hat{\theta}_b) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \hat{\theta}_{hi}) \pi_{lo}(a_t|s_t, o_t; \hat{\theta}_{lo}) \times \rho_t^\theta(o_t, b_t) \frac{(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1}(o_{t-1}, b_{t-1})}{\rho_{t-1}^\theta(o_{t-1}, b_{t-1})},$$

and

$$\rho_{t-1}^\theta(o_{t-1}, b_{t-1}) \propto \sum_{o_t', b_t'} \pi_b(b_t'|s_t, o_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t'|s_t, o_{t-1}, b_t'; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t'; \theta_{lo}) \rho_t^\theta(o_t', b_t').$$

As an abbreviation, let

$$h(\theta; o_{t-1}, s_t, a_t, o_t, b_t) = \pi_b(b_t|s_t, o_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t; \theta_{lo}). \quad (25)$$

Then,

$$(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_t(o_t, b_t) = C_F^{\hat{\theta},\theta} \sum_{o_{t-1}, b_{t-1}} \frac{h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) \rho_t^\theta(o_t, b_t) (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1}(o_{t-1}, b_{t-1})}{\sum_{o_t', b_t'} h(\theta; o_{t-1}, s_t, a_t, o_t', b_t') \rho_t^\theta(o_t', b_t')}, \quad (26)$$

where $C_F^{\hat{\theta},\theta}$ is a normalizing constant such that

$$\left(C_F^{\hat{\theta},\theta}\right)^{-1} = \sum_{o_{t-1}, b_{t-1}} \frac{\sum_{o_t, b_t} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) \rho_t^\theta(o_t, b_t)}{\sum_{o_t', b_t'} h(\theta; o_{t-1}, s_t, a_t, o_t', b_t') \rho_t^\theta(o_t', b_t')} (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1}(o_{t-1}, b_{t-1}).$$

From (26), we define the forward smoothing operator $K_{F,t}^{\hat{\theta},\theta}$ on the probability measure $(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1}$ such that as probability measures,

$$(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1} K_{F,t}^{\hat{\theta},\theta} = (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_t.$$

The subscript F in $K_{F,t}^{\hat{\theta},\theta}$ stands for *forward*. $K_{F,t}^{\hat{\theta},\theta}$ depends on the the parameters θ and $\hat{\theta}$, the observation $\{s_t, a_t\}_{t\in\mathbb{Z}}$, and the specific choice of $\{\rho_t^\theta\}_{t\in\mathbb{Z}}$. In the general case of $\theta \neq \hat{\theta}$, $K_{F,t}^{\hat{\theta},\theta}$ is a nonlinear operator which requires rather sophisticated analysis. However, when $\theta = \hat{\theta}$, it is straightforward to verify that the normalizing constant $C_F^{\theta,\theta} = 1$, and $K_{F,t}^{\theta,\theta}$ becomes a linear operator.

In fact, the linear operator $K_{F,t}^{\theta,\theta}$ can be regarded as the standard operation of a Markov transition kernel on probability measures. With a slight overload of notation, define such a Markov transition kernel on $\mathcal{O} \times \{0,1\}$, entry-wise, as the following. For any (o_t, b_t) and (o_{t-1}, b_{t-1}) in $\mathcal{O} \times \{0,1\}$,

$$K_{F,t}^{\theta,\theta}(o_t, b_t | o_{t-1}, b_{t-1}) := \frac{h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)}{\sum_{o_t', b_t'} h(\theta; o_{t-1}, s_t, a_t, o_t', b_t')\rho_t^\theta(o_t', b_t')}. \tag{27}$$

We name this Markov transition kernel as the forward smoothing kernel. Such a definition is analogous to Markovian decomposition in the HMM literature (Cappé et al., 2006). The only caveat here is that we also allow perturbations on the parameter. The resulting operator $K_{F,t}^{\hat{\theta},\theta}$ is nonlinear and no longer corresponds to a Markov transition kernel.

To proceed, we characterize the difference between operators $K_{F,t}^{\hat{\theta},\theta}$ and $K_{F,t}^{\theta,\theta}$ when $\hat{\theta}$ and θ are close. First, we show a version of Lipschitz continuity for the options with failure framework.

Lemma D.2 (Lipschitz continuity)**.** *For all $\theta \in \Theta$ and $\delta > 0$, there exists a real number $L_{\theta,\delta}$ such that with any input arguments o_{t-1}, s_t, a_t, o_t and b_t, the function $h(\tilde{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)$ defined in (25) is $L_{\theta,\delta}$-Lipschitz with respect to $\tilde{\theta}$ on the set $\{\tilde{\theta}; \tilde{\theta} \in \Theta, \|\tilde{\theta} - \theta\|_2 \le \delta\}$. Moreover, $L_{\theta,\delta}$ is upper bounded by a constant that does not depend on θ and δ.*

Proof of Lemma D.2

Due to Assumption 2, with any input arguments o_{t-1}, s_t, a_t, o_t and b_t, $h(\tilde{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)$ is continuously differentiable with respect to $\tilde{\theta} \in \tilde{\Theta}$. As continuously differentiable functions are Lipschitz continuous on convex and compact subsets, $h(\tilde{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)$ is Lipschitz continuous on Θ, hence also on $\{\tilde{\theta}; \tilde{\theta} \in \Theta, \|\tilde{\theta} - \theta\|_2 \le \delta\}$. The Lipschitz constants depend on the choice of input arguments o_{t-1}, s_t, a_t, o_t and b_t.

We can let $L_{\theta,\delta}$ be the smallest Lipschitz constant on $\{\tilde{\theta}; \tilde{\theta} \in \Theta, \|\tilde{\theta} - \theta\|_2 \le \delta\}$ that holds for all input arguments o_{t-1}, s_t, a_t, o_t and b_t. Clearly $L_{\theta,\delta}$ is upper bounded by any Lipschitz constant on Θ that holds for all input arguments, which does not depend on θ and δ. □

Next, we bound the difference between operators $K_{F,t}^{\hat{\theta},\theta}$ and $K_{F,t}^{\theta,\theta}$.

Lemma D.3 (Perturbation on the forward smoothing kernel)**.** *Let φ be any probability measure on $\mathcal{O} \times \{0,1\}$. Let $K_{F,t}^{\hat{\theta},\theta}$ and $K_{F,t}^{\theta,\theta}$ be defined with the same observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$ and the same choice of $\{\rho_t^\theta\}_{t\in\mathbb{Z}}$. Their difference is only in the first entry of the superscript ($\hat{\theta}$ in $K_{F,t}^{\hat{\theta},\theta}$; θ in $K_{F,t}^{\theta,\theta}$). Then, for all t, φ, θ, $\hat{\theta}$, $\{s_t, a_t\}_{t\in\mathbb{Z}}$ and $\{\rho_t^\theta\}_{t\in\mathbb{Z}}$,*

$$\left\|\varphi K_{F,t}^{\hat{\theta},\theta} - \varphi K_{F,t}^{\theta,\theta}\right\|_{\mathrm{TV}} \le \frac{\max_{o_{t-1},o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)}{\min_{o_{t-1},o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)} \frac{L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta} - \theta\|_2}{\min_{o_{t-1},o_t,b_t} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)}.$$

Proof of Lemma D.3

From the definitions, for any t, φ, θ, $\hat{\theta}$, $\{s_t, a_t\}_{t\in\mathbb{Z}}$ and $\{\rho_t^\theta\}_{t\in\mathbb{Z}}$,

$$\begin{aligned}
&\left\|\varphi K_{F,t}^{\hat{\theta},\theta} - \varphi K_{F,t}^{\theta,\theta}\right\|_{\mathrm{TV}} \\
=\ & \frac{1}{2}\sum_{o_t,b_t}\left|\sum_{o_{t-1},b_{t-1}} \frac{\left[C_F^{\hat{\theta},\theta} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) - h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\right]}{\sum_{o_t',b_t'} h(\theta; o_{t-1}, s_t, a_t, o_t', b_t')\rho_t^\theta(o_t', b_t')}\rho_t^\theta(o_t, b_t)\varphi(o_{t-1}, b_{t-1})\right| \\
\le\ & \frac{1}{2}\sum_{o_{t-1},b_{t-1}} \frac{\sum_{o_t,b_t}\left|C_F^{\hat{\theta},\theta} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) - h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\right|\rho_t^\theta(o_t, b_t)}{\sum_{o_t',b_t'} h(\theta; o_{t-1}, s_t, a_t, o_t', b_t')\rho_t^\theta(o_t', b_t')}\varphi(o_{t-1}, b_{t-1}).
\end{aligned}$$

From the definition of the normalizing constant $C_F^{\hat{\theta},\theta}$, we have

$$\left(C_F^{\hat{\theta},\theta}\right)^{-1} = \sum_{o_{t-1},b_{t-1}} \frac{\sum_{o_t,b_t} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)}{\sum_{o_t',b_t'} h(\theta; o_{t-1}, s_t, a_t, o_t', b_t')\rho_t^\theta(o_t', b_t')}\varphi(o_{t-1}, b_{t-1}).$$

Therefore,

$$C_F^{\hat{\theta},\theta} \le \max_{o_{t-1}} \frac{\sum_{o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)}{\sum_{o_t,b_t} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)},$$

and

$$\begin{aligned}\left|C_F^{\hat{\theta},\theta} - 1\right| &= \left|\sum_{o_{t-1},b_{t-1}} \frac{\sum_{o_t,b_t}[h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) - h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)]\rho_t^\theta(o_t, b_t)}{\sum_{o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)}\varphi(o_{t-1}, b_{t-1})\right| C_F^{\hat{\theta},\theta} \\ &\le \frac{L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2 C_F^{\hat{\theta},\theta}}{\min_{o_{t-1}} \sum_{o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)}.\end{aligned}$$

As a result, for any given o_{t-1}, o_t and b_t,

$$\begin{aligned}&\left|C_F^{\hat{\theta},\theta} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) - h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\right| \\ \le\ & C_F^{\hat{\theta},\theta}\left|h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t) - h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\right| + \left|C_F^{\hat{\theta},\theta} - 1\right| h(\theta; o_{t-1}, s_t, a_t, o_t, b_t) \\ \le\ & \left[1 + \frac{h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)}{\min_{o'_{t-1}} \sum_{o'_t,b'_t} h(\theta; o'_{t-1}, s_t, a_t, o'_t, b'_t)\rho_t^\theta(o'_t, b'_t)}\right] L_{\theta,\|\hat{\theta}-\theta\|_2}\left\|\hat{\theta}-\theta\right\|_2 C_F^{\hat{\theta},\theta}.\end{aligned}$$

Combining everything together,

$$\begin{aligned}\left\|\varphi K_{F,t}^{\hat{\theta},\theta} - \varphi K_{F,t}^{\theta,\theta}\right\|_{\mathrm{TV}} &\le L_{\theta,\|\hat{\theta}-\theta\|_2}\left\|\hat{\theta}-\theta\right\|_2 C_F^{\hat{\theta},\theta} \times \max_{o_{t-1}} \frac{1 + \frac{\sum_{o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)\rho_t^\theta(o_t, b_t)}{\min_{o'_{t-1}} \sum_{o'_t,b'_t} h(\theta; o'_{t-1}, s_t, a_t, o'_t, b'_t)\rho_t^\theta(o'_t, b'_t)}}{2\sum_{o'_t,b'_t} h(\theta; o_{t-1}, s_t, a_t, o'_t, b'_t)\rho_t^\theta(o'_t, b'_t)} \\ &= \frac{L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2 C_F^{\hat{\theta},\theta}}{\min_{o'_{t-1}} \sum_{o'_t,b'_t} h(\theta; o'_{t-1}, s_t, a_t, o'_t, b'_t)\rho_t^\theta(o'_t, b'_t)} \\ &\le \frac{\max_{o_{t-1},o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)}{\min_{o_{t-1},o_t,b_t} h(\theta; o_{t-1}, s_t, a_t, o_t, b_t)} \frac{L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2}{\min_{o_{t-1},o_t,b_t} h(\hat{\theta}; o_{t-1}, s_t, a_t, o_t, b_t)}.\end{aligned}$$

□

On the other hand, we can formulate a backward smoothing recursion as

$$(\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_t(o_t, b_t) = C_B^{\theta,\hat{\theta}} \sum_{o_{t+1},b_{t+1}} \frac{h(\hat{\theta}; o_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})\varphi_t^\theta(o_t, b_t)(\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_{t+1}(o_{t+1}, b_{t+1})}{\sum_{o'_t,b'_t} h(\theta; o'_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})\varphi_t^\theta(o'_t, b'_t)}, \tag{28}$$

where $C_B^{\theta,\hat{\theta}}$ is a normalizing constant such that

$$\left(C_B^{\theta,\hat{\theta}}\right)^{-1} = \sum_{o_{t+1},b_{t+1}} \frac{\sum_{o_t,b_t} h(\hat{\theta}; o_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})\varphi_t^\theta(o_t, b_t)}{\sum_{o'_t,b'_t} h(\theta; o'_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})\varphi_t^\theta(o'_t, b'_t)}(\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_{t+1}(o_{t+1}, b_{t+1}).$$

The subscript B in $K_{B,t}^{\theta,\hat{\theta}}$ stands for *backward*. Similar to the forward smoothing operator $K_{F,t}^{\hat{\theta},\theta}$, we can define the backward smoothing operator $K_{B,t}^{\theta,\hat{\theta}}$ from (28) such that as probability measures,

$$(\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_{t+1} K_{B,t}^{\theta,\hat{\theta}} = (\varphi^\theta \otimes \hat{\rho}^{\hat{\theta}})_t.$$

Analogous to $K_{F,t}^{\hat{\theta},\theta}$, in the general case of $\theta \neq \hat{\theta}$, $K_{B,t}^{\theta,\hat{\theta}}$ is a nonlinear operator. However, if $\theta = \hat{\theta}$, $K_{B,t}^{\theta,\hat{\theta}}$ becomes a linear operator and induces a Markov transition kernel.

The following lemma is similar to Lemma D.3. we state it without proof.

Lemma D.4 (Perturbation on the backward smoothing kernel). *Let ρ be any probability measure on $\mathcal{O} \times \{0,1\}$. Let $K_{B,t}^{\theta,\hat{\theta}}$ and $K_{B,t}^{\theta,\theta}$ be defined with the same observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$ and the same choice of $\{\varphi_t^\theta\}_{t\in\mathbb{Z}}$. Then, for any t, ρ, θ, $\hat{\theta}$, $\{s_t, a_t\}_{t\in\mathbb{Z}}$ and $\{\varphi_t^\theta\}_{t\in\mathbb{Z}}$,*

$$\left\|\rho K_{B,t}^{\theta,\hat{\theta}} - \rho K_{B,t}^{\theta,\theta}\right\|_{\mathrm{TV}} \leq \frac{\max_{o_t,o_{t+1},b_{t+1}} h(\theta; o_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})}{\min_{o_t,o_{t+1},b_{t+1}} h(\theta; o_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})} \frac{L_{\hat{\theta},\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2}{\min_{o_t,o_{t+1},b_{t+1}} h(\hat{\theta}; o_t, s_{t+1}, a_{t+1}, o_{t+1}, b_{t+1})}.$$

Notice that the bounds in both Lemma D.3 and Lemma D.4 depend on the observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$.

D.2.3. A PERTURBED CONTRACTION RESULT FOR SMOOTHING STABILITY

For any $t_1, t_2 \in \mathbb{Z}$ with $t_1 \leq t_2$, let $\mathbb{I} = [t_1 : t_2]$. Remember the following definition from Appendix D.2.1, with the index set restricted to $\mathbb{I}$: For any $\theta, \hat{\theta} \in \Theta$, $\{\varphi_t^\theta\}_{t\in\mathbb{I}}$ and $\{\hat{\varphi}_t^{\hat{\theta}}\}_{t\in\mathbb{I}}$ are two indexed sets of probability measures defined on $\mathcal{O} \times \{0,1\}$ such that, for all $t \in \mathbb{I}$, (1) if $t \neq t_1$, $F_t^\theta \varphi_{t-1}^\theta = \varphi_t^\theta$ and $F_t^{\hat{\theta}} \hat{\varphi}_{t-1}^{\hat{\theta}} = \hat{\varphi}_t^{\hat{\theta}}$; (2) φ_t^θ and $\hat{\varphi}_t^{\hat{\theta}}$ are strictly positive on their domains. $\{\rho_t^\theta\}_{t\in\mathbb{I}}$ and $\{\hat{\rho}_t^{\hat{\theta}}\}_{t\in\mathbb{I}}$ are two indexed sets of probability measures defined on $\mathcal{O} \times \{0,1\}$ such that for all $t \in \mathbb{I}$, (1) if $t \neq t_2$, $B_t^\theta \rho_{t+1}^\theta = \rho_t^\theta$ and $B_t^{\hat{\theta}} \hat{\rho}_{t+1}^{\hat{\theta}} = \hat{\rho}_t^{\hat{\theta}}$; (2) ρ_t^θ and $\hat{\rho}_t^{\hat{\theta}}$ are strictly positive on their domains. θ and $\hat{\theta}$ are allowed to be equal.

The smoothing stability lemma is stated as follows.

Lemma D.5 (Smoothing stability). *With $\{\varphi_t^\theta\}_{t\in\mathbb{I}}$, $\{\hat{\varphi}_t^{\hat{\theta}}\}_{t\in\mathbb{I}}$, $\{\rho_t^\theta\}_{t\in\mathbb{I}}$ and $\{\hat{\rho}_t^{\hat{\theta}}\}_{t\in\mathbb{I}}$ defined above,*

$$\left\|(\varphi^\theta \otimes \rho^\theta)_{t_2} - (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t_2}\right\|_{\mathrm{TV}} \leq \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t_2-t_1} + \frac{|\mathcal{O}| z_{\theta,\hat{\theta}} L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2,$$

and

$$\left\|(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t_1} - (\hat{\varphi}^{\hat{\theta}} \otimes \hat{\rho}^{\hat{\theta}})_{t_1}\right\|_{\mathrm{TV}} \leq \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t_2-t_1} + \frac{|\mathcal{O}| z_{\theta,\hat{\theta}} L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2,$$

where $z_{\theta,\theta'}$ is a positive real number dependent only on θ and $\hat{\theta}$. Specifically,

$$z_{\theta,\theta'} = \max_{s_t',a_t'} \frac{[\max_{o_{t-1},o_t,b_t} h(\theta; o_{t-1}, s_t', a_t', o_t, b_t)] \vee [\max_{o_{t-1},o_t,b_t} h(\hat{\theta}; o_{t-1}, s_t', a_t', o_t, b_t)]}{[\min_{o_{t-1},o_t,b_t} h(\theta; o_{t-1}, s_t', a_t', o_t, b_t)][\min_{o_{t-1},o_t,b_t} h(\hat{\theta}; o_{t-1}, s_t', a_t', o_t, b_t)]}.$$

Intuitively, if $\hat{\theta} = \theta$, Lemma D.5 has the form of an exact contraction, which is similar to the standard filtering stability result for HMMs. Indeed, our proof uses the classical techniques of uniform forgetting from the HMM literature (Cappé et al., 2006). If $\hat{\theta}$ is different from θ, such a contraction is perturbed. For HMMs, similar results are provided in (De Castro et al., 2017, Proposition 2.2, Theorem 2.3).

Proof of Lemma D.5

1. Consider the first bound. It holds trivially when $t_2 = t_1$. Now consider only $t_2 > t_1$.

Using the forward smoothing operators, for any $t_1 < t \leq t_2$,

$$(\varphi^\theta \otimes \rho^\theta)_{t-1} K_{F,t}^{\theta,\theta} - (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1} K_{F,t}^{\hat{\theta},\theta} = (\varphi^\theta \otimes \rho^\theta)_t - (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_t.$$

Therefore,

$$\begin{aligned}\left\|(\varphi^\theta \otimes \rho^\theta)_t - (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_t\right\|_{\mathrm{TV}} \leq{}& \left\|\left[(\varphi^\theta \otimes \rho^\theta)_{t-1} - (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1}\right] K_{F,t}^{\theta,\theta}\right\|_{\mathrm{TV}} \\ &+ \left\|(\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1} K_{F,t}^{\theta,\theta} - (\hat{\varphi}^{\hat{\theta}} \otimes \rho^\theta)_{t-1} K_{F,t}^{\hat{\theta},\theta}\right\|_{\mathrm{TV}},\end{aligned}$$

where the first term is due to $K_{F,t}^{\theta,\theta}$ being a linear operator.

From Lemma D.3, the second term on the RHS is upper bounded by $z_{\theta,\hat{\theta}} L_{\theta,\|\hat{\theta}-\theta\|_2} \|\hat{\theta}-\theta\|_2$. As for the first term, we can construct the classical Doeblin-type minorization condition (Cappé et al., 2006, Chap. 4.3). Applying Lemma D.1 in the definition of the Markov transition kernel $K_{F,t}^{\theta,\theta}$ (27), we have

$$K_{F,t}^{\theta,\theta}(o_t,b_t|o_{t-1},b_{t-1}) \geq \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|} \frac{\bar{\pi}_{o,b}(o_t,b_t|s_t;\theta)\pi_{lo}(a_t|s_t,o_t;\theta_{lo})\rho_t^\theta(o_t,b_t)}{\sum_{o_t',b_t'}\bar{\pi}_{o,b}(o_t',b_t'|s_t;\theta)\pi_{lo}(a_t|s_t,o_t';\theta_{lo})\rho_t^\theta(o_t',b_t')} =: \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\bar{\pi}_{F,t}^\theta(o_t,b_t). \tag{29}$$

Observe that $\bar{\pi}_{F,t}^\theta$ just defined is a probability measure. Further define $\bar{K}_{F,t}^{\theta,\theta}$ entry-wise as

$$\bar{K}_{F,t}^{\theta,\theta}(o_t,b_t|o_{t-1},b_{t-1}) := \left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{-1}\left(K_{F,t}^{\theta,\theta}(o_t,b_t|o_{t-1},b_{t-1}) - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\bar{\pi}_{F,t}^\theta(o_t,b_t)\right).$$

We can verify that $\bar{K}_{F,t}^{\theta,\theta}$ is also a Markov transition kernel. Moreover,

$$\left[(\varphi^\theta\otimes\rho^\theta)_{t-1} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t-1}\right]K_{F,t}^{\theta,\theta} = \left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)\left[(\varphi^\theta\otimes\rho^\theta)_{t-1} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t-1}\right]\bar{K}_{F,t}^{\theta,\theta}.$$

To proceed, the standard approach is to use the fact that the Dobrushin coefficient of $\bar{K}_{F,t}^{\theta,\theta}$ is upper bounded by one. For clarity, we avoid such definitions and take a more direct approach here, which requires the extension of the total variation distance for two probability measures to the total variation norm for a finite signed measure. For a finite signed measure ν over a finite set Ω, let the total variation norm of ν be

$$\|\nu\|_{\mathrm{TV}} := \frac{1}{2}\sum_{\omega\in\Omega}|\nu(\omega)|.$$

When ν is the difference between two probability measures $\nu_1-\nu_2$, the total variation norm of ν coincides with the total variation distance between ν_1 and ν_2. Therefore, the same notation $\|\cdot\|_{\mathrm{TV}}$ is adopted here.

Let $\bar{\mathcal{M}}(\mathcal{O}\times\{0,1\})$ be the set of finite signed measures over the finite set $\mathcal{O}\times\{0,1\}$. From (Cappé et al., 2006, Chap. 4.3.1), $\bar{\mathcal{M}}(\mathcal{O}\times\{0,1\})$ is a Banach space. Define an operator norm $\|\cdot\|_{\mathrm{op}}$ for $\bar{K}_{F,t}^{\theta,\theta}$ as

$$\left\|\bar{K}_{F,t}^{\theta,\theta}\right\|_{\mathrm{op}} := \sup\left\{\left\|\nu\bar{K}_{F,t}^{\theta,\theta}\right\|_{\mathrm{TV}}; \|\nu\|_{\mathrm{TV}}=1, \nu\in\bar{\mathcal{M}}(\mathcal{O}\times\{0,1\})\right\}.$$

Since $\bar{K}_{F,t}^{\theta,\theta}$ is a Markov transition kernel, $\|\bar{K}_{F,t}^{\theta,\theta}\|_{\mathrm{op}}=1$ (Cappé et al., 2006, Lemma 4.3.6). Therefore,

$$\begin{aligned}
&\left\|(\varphi^\theta\otimes\rho^\theta)_{t_2} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t_2}\right\|_{\mathrm{TV}} \\
\leq\ &\left\|\left[(\varphi^\theta\otimes\rho^\theta)_{t_2-1} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t_2-1}\right]K_{F,t_2}^{\theta,\theta}\right\|_{\mathrm{TV}} + \left\|(\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t_2-1}\left(K_{F,t_2}^{\theta,\theta}-K_{F,t_2}^{\hat{\theta},\theta}\right)\right\|_{\mathrm{TV}} \\
=\ &\left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)\left\|\left[(\varphi^\theta\otimes\rho^\theta)_{t_2-1} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t_2-1}\right]\bar{K}_{F,t_2}^{\theta,\theta}\right\|_{\mathrm{TV}} + z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2 \\
\leq\ &\left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)\left\|(\varphi^\theta\otimes\rho^\theta)_{t_2-1} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t_2-1}\right\|_{\mathrm{TV}}\left\|\bar{K}_{F,t_2}^{\theta,\theta}\right\|_{\mathrm{op}} + z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2 \\
=\ &\left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)\left\|(\varphi^\theta\otimes\rho^\theta)_{t_2-1} - (\hat{\varphi}^{\hat{\theta}}\otimes\rho^\theta)_{t_2-1}\right\|_{\mathrm{TV}} + z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}\|\hat{\theta}-\theta\|_2.
\end{aligned}$$

The second inequality is due to the sub-multiplicativity of the operator norm. Finally, the desired result follows from unrolling the summation and identifying the geometric series.

2. The proof of the second bound is analogous, using the backward smoothing operators instead of the forward smoothing operators. Details are omitted. □

Note that Lemma D.5 only holds when considering the options with failure framework. For the standard options framework, the one-step Doeblin-type minorization condition (29) we construct in the proof does not hold anymore, due to the failure of Lemma D.1. Instead, one could target the two-step minorization condition: define a two step smoothing kernel similar to $K_{F,t}^{\theta,\theta}$ and lower bound it similar to (29). Notations are much more complicated. For simplicity, this extension is not considered in this paper.

D.3. The approximation lemmas

This subsection applies Lemma D.5 to quantities defined in earlier sections.

First, we bound the difference of smoothing distributions in the unextended graphical model (as in Theorem 4) and the extended one with parameter k (as in Corollary 5). The parameter θ in the two models can be different. The bounds use quantities defined in Appendix D.1 and Appendix D.2.

Lemma D.6 (Bounding the difference of smoothing distributions, Part I). *For all $\theta, \hat{\theta} \in \Theta$, $k \in \mathbb{N}_+$ and $\mu \in \mathcal{M}$, with any observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$, we have*

1. $\forall t \in [1:T]$,

$$\left\|\gamma^{\theta}_{\mu,t|T} - \gamma^{\hat{\theta}}_{k,t}\right\|_{\mathrm{TV}} \le \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t-1} + \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{T-t} + \frac{2|\mathcal{O}|z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2.$$

2. $\forall t \in [2:T]$,

$$\left\|\tilde{\gamma}^{\theta}_{\mu,t|T} - \tilde{\gamma}^{\hat{\theta}}_{k,t}\right\|_{\mathrm{TV}} \le 2\left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t-2} + \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{T-t} + \frac{4|\mathcal{O}|z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2.$$

Similarly, we can bound the difference of smoothing distributions in two extended graphical models with different k and different parameter θ.

Lemma D.7 (Bounding the difference of smoothing distributions, Part II). *For all $\theta, \hat{\theta} \in \Theta$ and $t \in [1:T]$, with any two integers $k_2 > k_1 > 0$ and any observation sequence $\{s_t, a_t\}_{t\in\mathbb{Z}}$, we have*

$$\left\|\gamma^{\theta}_{k_1,t} - \gamma^{\hat{\theta}}_{k_2,t}\right\|_{\mathrm{TV}} \le \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t+k_1-1} + \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{T+k_1-t} + \frac{2|\mathcal{O}|z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2.$$

and

$$\left\|\tilde{\gamma}^{\theta}_{k_1,t} - \tilde{\gamma}^{\hat{\theta}}_{k_2,t}\right\|_{\mathrm{TV}} \le 2\left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t+k_1-2} + \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{T+k_1-t} + \frac{4|\mathcal{O}|z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2.$$

It can be easily verified that in Lemma D.6 and Lemma D.7, the bounds still hold if θ and $\hat{\theta}$ on the LHS are interchanged. We only present the proof of Lemma D.6. As for Lemma D.7, the proof is analogous therefore omitted. Our proof essentially relies on the smoothing stability lemma (Lemma D.5).

Proof of Lemma D.6

1. For a cleaner noation, let

$$\Delta_{\theta,\hat{\theta}} = \frac{|\mathcal{O}|z_{\theta,\hat{\theta}}L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2\zeta}\left\|\hat{\theta}-\theta\right\|_2.$$

Apply Lemma D.5 as follows: $\forall t \in [1:T]$, let $\varphi^{\theta}_t = \alpha^{\theta}_{\mu,t}$ and $\hat{\varphi}^{\hat{\theta}}_t = \alpha^{\hat{\theta}}_{k,t}$; let $\rho^{\theta}_t = \beta^{\theta}_{t|T}$ and $\hat{\rho}^{\hat{\theta}}_t = \beta^{\hat{\theta}}_{k,t}$. Due to Assumption 1, the strictly positive requirement is satisfied. Then, we have

$$\left\|\frac{\alpha^{\theta}_{\mu,t}\cdot\beta^{\theta}_{t|T}}{\langle\alpha^{\theta}_{\mu,t},\beta^{\theta}_{t|T}\rangle} - \frac{\alpha^{\hat{\theta}}_{k,t}\cdot\beta^{\theta}_{t|T}}{\langle\alpha^{\hat{\theta}}_{k,t},\beta^{\theta}_{t|T}\rangle}\right\|_{\mathrm{TV}} \le \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t-1} + \Delta_{\theta,\hat{\theta}},$$

and

$$\left\|\frac{\alpha^{\hat{\theta}}_{k,t}\cdot\beta^{\theta}_{t|T}}{\langle\alpha^{\hat{\theta}}_{k,t},\beta^{\theta}_{t|T}\rangle} - \frac{\alpha^{\hat{\theta}}_{k,t}\cdot\beta^{\hat{\theta}}_{k,t}}{\langle\alpha^{\hat{\theta}}_{k,t},\beta^{\hat{\theta}}_{k,t}\rangle}\right\|_{\mathrm{TV}} \le \left(1 - \frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{T-t} + \Delta_{\theta,\hat{\theta}}.$$

where $\cdot$ denotes element-wise product and $\langle \cdot, \cdot \rangle$ denotes Euclidean inner product. Therefore,

$$\begin{aligned}
\left\|\gamma^{\theta}_{\mu,t|T} - \gamma^{\hat{\theta}}_{k,t}\right\|_{\mathrm{TV}} &= \left\| \frac{\alpha^{\theta}_{\mu,t} \cdot \beta^{\theta}_{t|T}}{\langle \alpha^{\theta}_{\mu,t}, \beta^{\theta}_{t|T} \rangle} - \frac{\alpha^{\hat{\theta}}_{k,t} \cdot \beta^{\hat{\theta}}_{k,t}}{\langle \alpha^{\hat{\theta}}_{k,t}, \beta^{\hat{\theta}}_{k,t} \rangle} \right\|_{\mathrm{TV}} \\
&\le \left\| \frac{\alpha^{\theta}_{\mu,t} \cdot \beta^{\theta}_{t|T}}{\langle \alpha^{\theta}_{\mu,t}, \beta^{\theta}_{t|T} \rangle} - \frac{\alpha^{\hat{\theta}}_{k,t} \cdot \beta^{\theta}_{t|T}}{\langle \alpha^{\hat{\theta}}_{k,t}, \beta^{\theta}_{t|T} \rangle} \right\|_{\mathrm{TV}} + \left\| \frac{\alpha^{\hat{\theta}}_{k,t} \cdot \beta^{\theta}_{t|T}}{\langle \alpha^{\hat{\theta}}_{k,t}, \beta^{\theta}_{t|T} \rangle} - \frac{\alpha^{\hat{\theta}}_{k,t} \cdot \beta^{\hat{\theta}}_{k,t}}{\langle \alpha^{\hat{\theta}}_{k,t}, \beta^{\hat{\theta}}_{k,t} \rangle} \right\|_{\mathrm{TV}} \\
&\le \left(1 - \frac{\varepsilon_b^2 \zeta}{|\mathcal{O}|}\right)^{t-1} + \left(1 - \frac{\varepsilon_b^2 \zeta}{|\mathcal{O}|}\right)^{T-t} + 2\Delta_{\theta,\hat{\theta}}.
\end{aligned}$$

2. Next, we bound the difference of two-step smoothing distributions $\|\tilde{\gamma}^{\theta}_{\mu,t|T} - \tilde{\gamma}^{\hat{\theta}}_{k,t}\|_{\mathrm{TV}}$. Although the idea is straightforward, the details are tedious. For any $t \in [2:T]$, from (10) we have

$$\begin{aligned}
&\tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t) \\
\propto\ & \pi_b(b_t|s_t, o_{t-1}; \theta_b) \left[\sum_{o_t} \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t; \theta_{lo}) \frac{\gamma^{\theta}_{\mu,t|T}(o_t, b_t)}{\alpha^{\theta}_{\mu,t}(o_t, b_t)} \right] \left[\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1}) \right] \\
\propto\ & \sum_{o_t} \frac{\bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t; \theta_{lo}) \gamma^{\theta}_{\mu,t|T}(o_t, b_t) [\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1})]}{\sum_{o'_{t-1}, b_{t-1}} \pi_b(b_t|s_t, o'_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t|s_t, o'_{t-1}, b_t; \theta_{hi}) \pi_{lo}(a_t|s_t, o_t; \theta_{lo}) \alpha^{\theta}_{\mu,t-1}(o'_{t-1}, b_{t-1})} \pi_b(b_t|s_t, o_{t-1}; \theta_b) \\
=\ & \sum_{o_t} \frac{\pi_b(b_t|s_t, o_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}) [\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1})]}{\sum_{o'_{t-1}} \pi_b(b_t|s_t, o'_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t|s_t, o'_{t-1}, b_t; \theta_{hi}) [\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o'_{t-1}, b_{t-1})]} \gamma^{\theta}_{\mu,t|T}(o_t, b_t).
\end{aligned}$$

The denominators are all positive due to the non-degeneracy assumption. It can be easily verified that the normalizing constants involved in the second and the third line cancel each other. As abbreviations, define

$$\begin{aligned}
g^{\theta}(o_{t-1}, s_t, o_t, b_t) &:= \pi_b(b_t|s_t, o_{t-1}; \theta_b) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \theta_{hi}), \\
g^{\hat{\theta}}(o_{t-1}, s_t, o_t, b_t) &:= \pi_b(b_t|s_t, o_{t-1}; \hat{\theta}_b) \bar{\pi}_{hi}(o_t|s_t, o_{t-1}, b_t; \hat{\theta}_{hi}), \\
f^{\theta}_{\mu,t}(o_{t-1}, s_t, o_t, b_t) &:= \frac{g^{\theta}(o_{t-1}, s_t, o_t, b_t) [\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o_{t-1}, b_{t-1})]}{\sum_{o'_{t-1}} g^{\theta}(o'_{t-1}, s_t, o_t, b_t) [\sum_{b_{t-1}} \alpha^{\theta}_{\mu,t-1}(o'_{t-1}, b_{t-1})]}, \\
f^{\hat{\theta}}_{k,t}(o_{t-1}, s_t, o_t, b_t) &:= \frac{g^{\hat{\theta}}(o_{t-1}, s_t, o_t, b_t) [\sum_{b_{t-1}} \alpha^{\hat{\theta}}_{k,t-1}(o_{t-1}, b_{t-1})]}{\sum_{o'_{t-1}} g^{\hat{\theta}}(o'_{t-1}, s_t, o_t, b_t) [\sum_{b_{t-1}} \alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1}, b_{t-1})]}.
\end{aligned}$$

Then,

$$\begin{aligned}
\left\|\tilde{\gamma}^{\theta}_{\mu,t|T} - \tilde{\gamma}^{\hat{\theta}}_{k,t}\right\|_{\mathrm{TV}} &= \frac{1}{2} \sum_{o_{t-1}, b_t} \left| \sum_{o_t} [f^{\theta}_{\mu,t}(o_{t-1}, s_t, o_t, b_t) \gamma^{\theta}_{\mu,t|T}(o_t, b_t) - f^{\hat{\theta}}_{k,t}(o_{t-1}, s_t, o_t, b_t) \gamma^{\hat{\theta}}_{k,t|T}(o_t, b_t)] \right| \\
&\le \frac{1}{2} \sum_{o_{t-1}, b_t, o_t} \left| f^{\theta}_{\mu,t}(o_{t-1}, s_t, o_t, b_t) - f^{\hat{\theta}}_{k,t}(o_{t-1}, s_t, o_t, b_t) \right| \gamma^{\theta}_{\mu,t|T}(o_t, b_t) \\
&\qquad + \frac{1}{2} \sum_{o_{t-1}, b_t, o_t} f^{\hat{\theta}}_{k,t}(o_{t-1}, s_t, o_t, b_t) \left| \gamma^{\theta}_{\mu,t|T}(o_t, b_t) - \gamma^{\hat{\theta}}_{k,t|T}(o_t, b_t) \right|. \qquad (30)
\end{aligned}$$

Now, we bound the two terms on the RHS separately. Consider the first term in (30),

$$\begin{aligned}
&\frac{1}{2}\sum_{o_{t-1},o_t,b_t}\left|f^{\theta}_{\mu,t}(o_{t-1},s_t,o_t,b_t)-f^{\hat{\theta}}_{k,t}(o_{t-1},s_t,o_t,b_t)\right|\gamma^{\theta}_{\mu,t|T}(o_t,b_t)\\
\le\ &\frac{1}{2}\max_{o_t,b_t}\sum_{o_{t-1},b_{t-1}}\left|\frac{g^{\theta}(o_{t-1},s_t,o_t,b_t)\alpha^{\theta}_{\mu,t-1}(o_{t-1},b_{t-1})}{\sum_{o'_{t-1},b'_{t-1}}g^{\theta}(o'_{t-1},s_t,o_t,b_t)\alpha^{\theta}_{\mu,t-1}(o'_{t-1},b'_{t-1})}\right.\\
&\qquad\left.-\frac{g^{\theta}(o_{t-1},s_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o_{t-1},b_{t-1})}{\sum_{o'_{t-1},b'_{t-1}}g^{\theta}(o'_{t-1},s_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1},b'_{t-1})}\right|\\
&+\frac{1}{2}\max_{o_t,b_t}\sum_{o_{t-1},b_{t-1}}\alpha^{\hat{\theta}}_{k,t-1}(o_{t-1},b_{t-1})\left|\frac{g^{\theta}(o_{t-1},s_t,o_t,b_t)}{\sum_{o'_{t-1},b'_{t-1}}g^{\theta}(o'_{t-1},s_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1},b'_{t-1})}\right.\\
&\qquad\left.-\frac{g^{\hat{\theta}}(o_{t-1},s_t,o_t,b_t)}{\sum_{o'_{t-1},b'_{t-1}}g^{\hat{\theta}}(o'_{t-1},s_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1},b'_{t-1})}\right|. \qquad (31)
\end{aligned}$$

Denote the two terms on the RHS of (31) as Δ_1 and Δ_2 respectively. To bound Δ_1, we can apply Lemma D.5 on the index set $[1:t-1]$ as follows, assuming $t>2$. For any $t'\in[1:t-1]$, let $\varphi^{\theta}_{t'}=\alpha^{\theta}_{\mu,t'}$ and $\hat{\varphi}^{\hat{\theta}}_{t'}=\alpha^{\hat{\theta}}_{k,t'}$. For any (o_t,b_t), let $\rho^{\theta}_{t-1}(o_{t-1},b_{t-1})=z_{\theta}^{-1}g^{\theta}(o_{t-1},s_t,o_t,b_t)$, where z_θ is a normalizing constant. For $1\le t'<t-1$, let $\rho^{\theta}_{t'}=B^{\theta}_{t'}\rho^{\theta}_{t'+1}$. Then,

$$\Delta_1\le\left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t-2}+\Delta_{\theta,\hat{\theta}}.$$

Such a bound holds trivially if $t\le 2$.

Next, we bound Δ_2 as follows. Straightforward computation yields the following result.

$$\begin{aligned}
\Delta_2&=\frac{1}{2}\max_{o_t,b_t}\sum_{o_{t-1},b_{t-1}}\alpha^{\hat{\theta}}_{k,t-1}(o_{t-1},b_{t-1})\left|\frac{h(\theta;o_{t-1},s_t,a_t,o_t,b_t)}{\sum_{o'_{t-1},b'_{t-1}}h(\theta;o'_{t-1},s_t,a_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1},b'_{t-1})}\right.\\
&\qquad\left.-\frac{h(\hat{\theta};o_{t-1},s_t,a_t,o_t,b_t)}{\sum_{o'_{t-1},b'_{t-1}}h(\hat{\theta};o'_{t-1},s_t,a_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1},b'_{t-1})}\right|\\
&\le\max_{o_t,b_t}\frac{\sum_{o_{t-1},b_{t-1}}\left|h(\theta;o_{t-1},s_t,a_t,o_t,b_t)-h(\hat{\theta};o_{t-1},s_t,a_t,o_t,b_t)\right|\alpha^{\hat{\theta}}_{k,t-1}(o_{t-1},b_{t-1})}{\sum_{o'_{t-1},b'_{t-1}}h(\theta;o'_{t-1},s_t,a_t,o_t,b_t)\alpha^{\hat{\theta}}_{k,t-1}(o'_{t-1},b'_{t-1})}\\
&\le\frac{\max_{o_{t-1},o_t,b_t}\left|h(\theta;o_{t-1},s_t,a_t,o_t,b_t)-h(\hat{\theta};o_{t-1},s_t,a_t,o_t,b_t)\right|}{\min_{o_{t-1},o_t,b_t}h(\theta;o_{t-1},s_t,a_t,o_t,b_t)}\le\Delta_{\theta,\hat{\theta}},
\end{aligned}$$

where we use the definition of $h(\theta;o_{t-1},s_t,a_t,o_t,b_t)$ in (25).

As for the second term in (30),

$$\begin{aligned}
\frac{1}{2}\sum_{o_{t-1},b_t,o_t}f^{\theta}_{k,t}(o_{t-1},s_t,o_t,b_t)\left|\gamma^{\theta}_{\mu,t|T}(o_t,b_t)-\gamma^{\theta}_{k,t|T}(o_t,b_t)\right|&=\left\|\gamma^{\theta}_{\mu,t|T}-\gamma^{\theta}_{k,t}\right\|_{\mathrm{TV}}\\
&\le\left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{t-1}+\left(1-\frac{\varepsilon_b^2\zeta}{|\mathcal{O}|}\right)^{T-t}+2\Delta_{\theta,\hat{\theta}}.
\end{aligned}$$

Combining the above gives the desired result. □

D.4. Proof of Lemma C.1

Based on Lemma D.7, for all $T\ge 2$, $\theta\in\Theta$ and $t\in[1:T]$, with any observation sequence, both the sequences $\{\gamma^{\theta}_{k,t}\}_{k\in\mathbb{N}_+}$ and $\{\tilde{\gamma}^{\theta}_{k,t}\}_{k\in\mathbb{N}_+}$ are Cauchy sequences associated with the total variation distance. Moreover, the set of probability measures

over the finite sample space $\mathcal{O} \times \{0, 1\}$ is complete. Therefore, the limits of both $\{\gamma^\theta_{k,t}\}_{k\in\mathbb{N}_+}$ and $\{\tilde{\gamma}^\theta_{k,t}\}_{k\in\mathbb{N}_+}$ as $k \to \infty$ exist with respect to the total variation distance. From the definitions of $\{\gamma^\theta_{k,t}\}_{k\in\mathbb{N}_+}$ and $\{\tilde{\gamma}^\theta_{k,t}\}_{k\in\mathbb{N}_+}$ in Appendix C.1, it is clear that their limits as $k \to \infty$ do not depend on T.

The Lipschitz continuity of $\gamma^\theta_{\infty,t}$ also follows from Lemma D.7. Specifically, for all $\theta, \hat{\theta} \in \Theta$ and $t \in [1 : T]$, with any observation sequence,

$$\left\|\gamma^\theta_{\infty,t} - \gamma^{\hat{\theta}}_{\infty,t}\right\|_{\mathrm{TV}} \le \frac{2|\mathcal{O}| z_{\theta,\hat{\theta}} L_{\theta,\|\hat{\theta}-\theta\|_2}}{\varepsilon_b^2 \zeta} \left\|\hat{\theta} - \theta\right\|_2.$$

The coefficient of $\|\hat{\theta} - \theta\|_2$ on the RHS can be upper bounded by a constant that does not depend on θ and $\hat{\theta}$. The same argument holds for $\tilde{\gamma}^\theta_{\infty,t}$. □

D.5. Proof of Lemma C.2

For a cleaner notation, we omit the dependency on ω in the following analysis. From the definitions, for all $\theta, \theta' \in \Theta$ and $\mu \in \mathcal{M}$,

$$\begin{aligned} Q^s_{\infty,T}(\theta'|\theta) - Q_{\mu,T}(\theta'|\theta) = \frac{1}{T}\Bigg\{ &\sum_{t=2}^{T} \sum_{o_{t-1},b_t} \left[\tilde{\gamma}^\theta_{\infty,t}(o_{t-1}, b_t) - \tilde{\gamma}^\theta_{\mu,t|T}(o_{t-1}, b_t)\right] [\log \pi_b(b_t|s_t, o_{t-1}; \theta'_b)] \\ &+ \sum_{t=1}^{T} \sum_{o_t,b_t} \left[\gamma^\theta_{\infty,t}(o_t, b_t) - \gamma^\theta_{\mu,t|T}(o_t, b_t)\right] [\log \pi_{lo}(a_t|s_t, o_t; \theta'_{lo})] \\ &+ \sum_{t=1}^{T} \sum_{o_t} \left[\gamma^\theta_{\infty,t}(o_t, b_t = 1) - \gamma^\theta_{\mu,t|T}(o_t, b_t = 1)\right] [\log \pi_{hi}(o_t|s_t; \theta'_{hi})] + err\Bigg\}, \end{aligned}$$

where the last term is a small error term associated with $t = 1$ such that,

$$|err| = \left| \sum_{o_0,b_1} \tilde{\gamma}^\theta_{\infty,1}(o_0, b_1) [\log \pi_b(b_1|s_1, o_0; \theta'_b)] \right| \le \max_{b_1,s_1,o_0} |\log \pi_b(b_1|s_1, o_0; \theta'_b)|.$$

The maximum on the RHS is finite due to the non-degeneracy assumption. Furthermore,

$$\begin{aligned} \left|Q^s_{\infty,T}(\theta'|\theta) - Q_{\mu,T}(\theta'|\theta)\right| \le \frac{1}{T}\Bigg\{ &\sum_{t=2}^{T} \max_{b_t,s_t,o_{t-1}} |\log \pi_b(b_t|s_t, o_{t-1}; \theta'_b)| \sum_{o_{t-1},b_t} \left|\tilde{\gamma}^\theta_{\infty,t}(o_{t-1}, b_t) - \tilde{\gamma}^\theta_{\mu,t|T}(o_{t-1}, b_t)\right| \\ &+ \sum_{t=1}^{T} \max_{a_t,s_t,o_t} |\log \pi_{lo}(a_t|s_t, o_t; \theta'_{lo})| \sum_{o_t,b_t} \left|\gamma^\theta_{\infty,t}(o_t, b_t) - \gamma^\theta_{\mu,t|T}(o_t, b_t)\right| \\ &+ \sum_{t=1}^{T} \max_{s_t,o_t} |\log \pi_{hi}(o_t|s_t; \theta'_{hi})| \sum_{o_t} \left|\gamma^\theta_{\infty,t}(o_t, b_t = 1) - \gamma^\theta_{\mu,t|T}(o_t, b_t = 1)\right| + |err|\Bigg\}. \end{aligned}$$

Since the bounds in Lemma D.6 hold for any $k > 0$, they also hold in the limit as $k \to \infty$. Therefore, for any θ, any μ and any $t \in [1 : T]$,

$$\left\|\gamma^\theta_{\mu,t|T} - \gamma^\theta_{\infty,t}\right\|_{\mathrm{TV}} \le \left(1 - \frac{\varepsilon_b^2 \zeta}{|\mathcal{O}|}\right)^{t-1} + \left(1 - \frac{\varepsilon_b^2 \zeta}{|\mathcal{O}|}\right)^{T-t}.$$

For any θ, any μ and any $t \in [2 : T]$,

$$\left\|\tilde{\gamma}^\theta_{\mu,t|T} - \tilde{\gamma}^\theta_{\infty,t}\right\|_{\mathrm{TV}} \le 2\left(1 - \frac{\varepsilon_b^2 \zeta}{|\mathcal{O}|}\right)^{t-2} + \left(1 - \frac{\varepsilon_b^2 \zeta}{|\mathcal{O}|}\right)^{T-t}.$$

Combining everything above,

$$\begin{aligned}
&\left|Q^s_{\infty,T}(\theta'|\theta) - Q_{\mu,T}(\theta'|\theta)\right| \\
\leq\ & \frac{1}{T}\Bigg\{ \max_{b_t,s_t,o_{t-1}} |\log \pi_b(b_t|s_t,o_{t-1};\theta'_b)| \left[1 + 2\sum_{t=2}^{T} \left\|\tilde{\gamma}^{\theta}_{\mu,t|T} - \tilde{\gamma}^{\theta}_{\infty,t}\right\|_{\mathrm{TV}}\right] \\
&+ 2\left[\max_{a_t,s_t,o_t} |\log \pi_{lo}(a_t|s_t,o_t;\theta'_{lo})| + \max_{s_t,o_t} |\log \pi_{hi}(o_t|s_t;\theta'_{hi})|\right] \sum_{t=1}^{T} \left\|\gamma^{\theta}_{\mu,t|T} - \gamma^{\theta}_{\infty,t}\right\|_{\mathrm{TV}} \Bigg\} \\
\leq\ & \frac{1}{T}\Bigg\{ \left(1 + \frac{6|O|}{\varepsilon_b^2\zeta}\right) \max_{b_t,s_t,o_{t-1}} |\log \pi_b(b_t|s_t,o_{t-1};\theta'_b)| \\
&\qquad + \frac{4|O|}{\varepsilon_b^2\zeta}\left[\max_{a_t,s_t,o_t} |\log \pi_{lo}(a_t|s_t,o_t;\theta'_{lo})| + \max_{s_t,o_t} |\log \pi_{hi}(o_t|s_t;\theta'_{hi})|\right] \Bigg\} = \frac{C(\theta')}{T},
\end{aligned}$$

where $C(\theta')$ is a positive real number that only depends on θ' and the structural constants $|\mathcal{O}|$, ζ and ε_b. Due to Assumption 2, $C(\theta')$ is continuous with respect to θ'. Since Θ is compact, $\sup_{\theta'\in\Theta} C(\theta') < \infty$. Therefore,

$$\left|Q^s_{\infty,T}(\theta'|\theta) - Q_{\mu,T}(\theta'|\theta)\right| \leq \frac{1}{T}\sup_{\theta'\in\Theta} C(\theta').$$

Taking supremum with respect to θ, θ' and μ completes the proof. □

D.6. Proof of the strong stochastic equicontinuity condition (20)

First, for all $\delta > 0$ and $\omega \in \Omega$,

$$\begin{aligned}
&\limsup_{T\to\infty} \sup_{\theta_1,\theta'_1,\theta_2,\theta'_2\in\Theta;\|\theta_1-\theta_2\|_2+\|\theta'_1-\theta'_2\|_2\leq\delta} \left|Q^s_{\infty,T}(\theta'_1|\theta_1;\omega) - Q^s_{\infty,T}(\theta'_2|\theta_2;\omega)\right| \\
\leq\ & \limsup_{T\to\infty} \frac{1}{T} \sup_{\theta_1,\theta'_1,\theta_2,\theta'_2\in\Theta;\|\theta_1-\theta_2\|_2+\|\theta'_1-\theta'_2\|_2\leq\delta} \left|f_t(\theta'_1|\theta_1;\omega) - f_t(\theta'_2|\theta_2;\omega)\right|.
\end{aligned}$$

Due to the boundedness of $f_t(\theta'|\theta;\omega)$ from Appendix C.2, we can apply the ergodic theorem (Lemma C.3). $\mathbb{P}_{\theta^*}$ almost surely,

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \sup_{\theta_1,\theta'_1,\theta_2,\theta'_2\in\Theta;\|\theta_1-\theta_2\|_2+\|\theta'_1-\theta'_2\|_2\leq\delta} \left|f_t(\theta'_1|\theta_1;\omega) - f_t(\theta'_2|\theta_2;\omega)\right| \\
=\ & \mathbb{E}_{\theta^*}\left[\sup_{\theta_1,\theta'_1,\theta_2,\theta'_2\in\Theta;\|\theta_1-\theta_2\|_2+\|\theta'_1-\theta'_2\|_2\leq\delta} \left|f_1(\theta'_1|\theta_1;\omega) - f_1(\theta'_2|\theta_2;\omega)\right|\right] \\
\leq\ & \mathbb{E}_{\theta^*}\left[\sup_{\theta_1,\theta'_1,\theta'_2\in\Theta;\|\theta'_1-\theta'_2\|_2\leq\delta} \left|f_1(\theta'_1|\theta_1;\omega) - f_1(\theta'_2|\theta_1;\omega)\right|\right] \\
&\qquad + \mathbb{E}_{\theta^*}\left[\sup_{\theta_1,\theta_2,\theta'_2\in\Theta;\|\theta_1-\theta_2\|_2\leq\delta} \left|f_1(\theta'_2|\theta_1;\omega) - f_1(\theta'_2|\theta_2;\omega)\right|\right].
\end{aligned}$$

Notice that for all θ_1, θ'_1, θ'_2 and ω,

$$\begin{aligned}
|f_1(\theta'_1|\theta_1;\omega) - f_1(\theta'_2|\theta_1;\omega)| \leq\ & \max_{o_t} \left|\log \pi_{hi}(o_t|\omega(s_t);\theta'_{1,hi}) - \log \pi_{hi}(o_t|\omega(s_t);\theta'_{2,hi})\right| \\
&+ \max_{o_t} \left|\log \pi_{lo}(\omega(a_t)|\omega(s_t),o_t;\theta'_{1,lo}) - \log \pi_{lo}(\omega(a_t)|\omega(s_t),o_t;\theta'_{2,lo})\right| \\
&+ \max_{o_{t-1},b_t} \left|\log \pi_b(b_t|\omega(s_t),o_{t-1};\theta'_{1,b}) - \log \pi_b(b_t|\omega(s_t),o_{t-1};\theta'_{2,b})\right|.
\end{aligned}$$

The RHS does not depend on θ_1. Due to Assumption 2, π_{hi}, π_{lo} and π_b as functions of the parameter θ are uniformly continuous on Θ, with any other input arguments. Therefore it is straightforward to verify that, for any ω,

$$\lim_{\delta\to 0} \sup_{\theta_1,\theta'_1,\theta'_2\in\Theta;\|\theta'_1-\theta'_2\|_2\leq\delta} |f_1(\theta'_1|\theta_1;\omega) - f_1(\theta'_2|\theta_1;\omega)| = 0.$$

Applying the dominated convergence theorem,

$$\lim_{\delta\to 0}\mathbb{E}_{\theta^*}\left[\sup_{\theta_1,\theta_1',\theta_2'\in\Theta;\|\theta_1'-\theta_2'\|_2\le\delta}|f_1(\theta_1'|\theta_1;\omega)-f_1(\theta_2'|\theta_1;\omega)|\right]=0.$$

Similarly, using Lemma C.1 we can show that for any ω,

$$\lim_{\delta\to 0}\sup_{\theta_1,\theta_2,\theta_2'\in\Theta;\|\theta_1-\theta_2\|_2\le\delta}|f_1(\theta_2'|\theta_1;\omega)-f_1(\theta_2'|\theta_2;\omega)|=0.$$

Using the dominated convergence theorem gives the convergence of the expectation as well. Combining the above gives the strong stochastic equicontinuity condition (20). □

D.7. Proof of Lemma C.5

Consider the following joint distribution on the graphical model shown in Figure 1: the prior distribution of (O_0, S_1) is ν^*, and the joint distribution of the rest of the graphical model is determined by an options with failure policy with parameters ζ and θ. Notice that this is the *correct* graphical model for the inference of the true parameter θ^*, since the assumed prior distribution of (O_0, S_1) coincides with the correct one.

For clarity, we use the same notations as in Appendix B.3 for the full likelihood function, the marginal likelihood function and the (unnormalized) Q-function. Specifically, such quantities used in this proof have the same symbols as those defined in Appendix B.3, but mathematically they are not the same.

Parallel to Appendix B.3, the full likelihood function is

$$L(s_{1:T},a_{1:T},o_{0:T},b_{1:T};\theta)=\nu^*(o_0,s_1)\mathbb{P}_{\theta,o_0,s_1}(S_{2:T}=s_{1:T},A_{1:T}=a_{1:T},O_{1:T}=o_{1:T},B_{1:T}=b_{1:T}).$$

The marginal likelihood function is

$$L^m(s_{1:T},a_{1:T};\theta)=\sum_{o_0}\nu^*(o_0,s_1)\mathbb{P}_{\theta,o_0,s_1}(S_{2:T}=s_{1:T},A_{1:T}=a_{1:T}).$$

Let μ^* be the conditional distribution of O_0 given s_1. For any $o_0\in\mathcal{O}$,

$$\mu^*(o_0|s_1)=\frac{\nu^*(o_0,s_1)}{\sum_{o_0'\in\mathcal{O}}\nu^*(o_0',s_1)}.$$

Therefore, for the inference of θ^* considered in this proof, the (unnormalized) Q-function can be expressed as

$$\begin{aligned}\tilde{Q}_{\mu^*,T}(\theta'|\theta)=&\sum_{o_{0:T},b_{1:T}}\frac{L(s_{1:T},a_{1:T},o_{0:T},b_{1:T};\theta)}{L^m(s_{1:T},a_{1:T};\theta)}\log L(s_{1:T},a_{1:T},o_{0:T},b_{1:T};\theta')\\=&\sum_{o_{0:T},b_{1:T}}\mu^*(o_0|s_1)\mathbb{P}_{\theta,o_0,s_1}(S_{2:T}=s_{2:T},A_{1:T}=a_{1:T},O_{1:T}=o_{1:T},B_{1:T}=b_{1:T})\\&\times z^{\theta}_{\gamma,\mu^*}\log[\nu^*(o_0,s_1)\mathbb{P}_{\theta',o_0',s_1}(S_{2:T}=s_{1:T},A_{1:T}=a_{1:T},O_{1:T}=o_{1:T},B_{1:T}=b_{1:T})].\end{aligned}$$

We can rewrite $\tilde{Q}_{\mu^*,T}(\theta'|\theta)$ using the structure of the options with failure framework, drop the terms irrelevant to θ' and normalize using T. The result is the following definition of the (normalized) Q-function:

$$\begin{aligned}Q^*_T(\theta'|\theta):=&\frac{\sum_{o_0,b_1}\nu^*(o_0|s_1)\mathbb{P}_{\theta,o_0,s_1}(S_{2:T}=s_{2:T},A_{1:T}=a_{1:T},B_1=b_1)[\log\pi_b(b_1|s_1,o_0;\theta_b')]}{T\sum_{o_0}\nu^*(o_0,s_1)\mathbb{P}_{\theta,o_0,s_1}(S_{2:T}=s_{1:T},A_{1:T}=a_{1:T})}\\&+\frac{1}{T}\sum_{t=1}^{T}\sum_{o_t,b_t}\gamma^{\theta}_{\mu^*,t|T}(o_t,b_t)[\log\pi_{lo}(a_t|s_t,o_t;\theta_{lo}')]+\frac{1}{T}\sum_{t=1}^{T}\sum_{o_t}\gamma^{\theta}_{\mu^*,t|T}(o_t,b_t=1)[\log\pi_{hi}(o_t|s_t;\theta_{hi}')]\\&+\frac{1}{T}\sum_{t=2}^{T}\sum_{o_{t-1},b_t}\tilde{\gamma}^{\theta}_{\mu^*,t|T}(o_{t-1},b_t)[\log\pi_b(b_t|s_t,o_{t-1};\theta_b')].\end{aligned}$$

We draw a comparison between $Q_T^*(\theta'|\theta)$ and $Q_{\mu^*,T}(\theta'|\theta)$ defined in (3): their difference is in the first term of $Q_T^*(\theta'|\theta)$. Maximizing $Q_T^*(\theta'|\theta)$ with respect to θ' is equivalent to maximizing the (unnormalized) Q-function $\tilde{Q}_{\mu^*,T}(\theta'|\theta)$. In Algorithm 1, since $Q_T^*(\theta'|\theta)$ is unavailable, we use $Q_{\mu^*,T}(\theta'|\theta)$ as its approximation.

$Q_T^*(\theta'|\theta)$ depends on the observation sequence, therefore it is a function of a sample path $\omega \in \Omega$. In the following we explicitly show this dependency by writing $Q_T^*(\theta'|\theta;\omega)$. Clearly, for any $\theta, \theta' \in \Theta$, $\omega \in \Omega$ and $T \geq 2$,

$$|Q_T^*(\theta'|\theta;\omega) - Q_{\mu^*,T}(\theta'|\theta;\omega)| \leq \frac{1}{T} \sup_{\theta' \in \Theta} \max_{b_1,s_1,o_0} |\log \pi_b(b_1|s_1,o_0;\theta_b')| \,.$$

Combining this with the stochastic convergence of $Q_{\mu^*,T}$ as shown in Theorem 1, we have, that for any $\theta \in \Theta$, as $T \to \infty$,

$$\left|Q_T^*(\theta|\theta^*;\omega) - \bar{Q}(\theta|\theta^*)\right| \to 0, \ P_{\theta^*}\text{-a.s.}$$

Using the dominated convergence theorem, such a convergence holds in expectation as well. For any $\theta \in \Theta$,

$$\lim_{T\to\infty} \mathbb{E}_{\theta^*}\left[Q_T^*(\theta|\theta^*;\omega)\right] = \bar{Q}(\theta|\theta^*).$$

Since maximizing $Q_T^*(\theta|\theta^*)$ with respect to θ is equivalent to maximizing the (unnormalized) Q-function $\tilde{Q}_{\mu^*,T}(\theta|\theta^*)$, the standard monotonicity property of the EM update holds as well. For any $\theta \in \Theta$, $\omega \in \Omega$ and $T \geq 2$,

$$\log L^m[\omega(s_{1:T}), \omega(a_{1:T}); \theta] - \log L^m[\omega(s_{1:T}), \omega(a_{1:T}); \theta^*] \geq T\left[Q_T^*(\theta|\theta^*;\omega) - Q_T^*(\theta^*|\theta^*;\omega)\right].$$

Taking expectation on both sides, we have

$$\mathbb{E}_{\theta^*}[\text{LHS}] = \sum_{s_{1:T},a_{1:T}} L^m(s_{1:T},a_{1:T};\theta^*) \log \frac{L^m(s_{1:T},a_{1:T};\theta)}{L^m(s_{1:T},a_{1:T};\theta^*)} \leq 0,$$

due to the non-negativity of the Kullback-Leibler divergence. Therefore, $\mathbb{E}_{\theta^*}[Q_T^*(\theta|\theta^*;\omega)] \leq \mathbb{E}_{\theta^*}[Q_T^*(\theta^*|\theta^*;\omega)]$, and in the limit we have $\bar{Q}(\theta|\theta^*) \leq \bar{Q}(\theta^*|\theta^*)$ for any $\theta \in \Theta$. Applying the identifiability assumption for the uniqueness of $\bar{M}(\theta^*)$ completes the proof. □

E. Empirical results

In this section, we demonstrate the empirical performance of Algorithm 1 using a simple numerical example. Consider the Markov Decision Process (MDP) illustrated in Figure 3. 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 $a_t =$ RIGHT, then the next state is sampled uniformly from the states on the right of state s_t (including s_t itself). Symmetrically, if $a_t =$ LEFT, then the next state is sampled uniformly from the states on the left of state s_t (including s_t).

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.

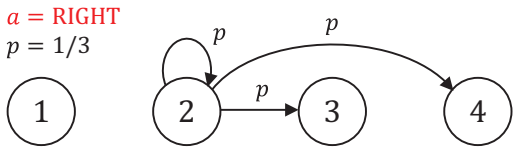

Figure 3. The state transition structure.

In our experiment, we investigate the behavior of $\|\theta^{(n)} - \theta^*\|_2$ as a random variable dependent on n and T. 50 sample paths of length 10000 are sampled from (approximately) the stationary Markov chain induced by the expert policy. Then, the first T state-action pairs are used as the observation sequence $\{s_{1:T}, a_{1:T}\}$, with $T \in \{5000, 8000, 10000\}$. For all s_1, $\mu(o_0 = \text{RIGHTEND}|s_1) = 1$. The initial parameter estimate $(\theta_{hi}^{(0)}, \theta_{lo}^{(0)}, \theta_b^{(0)}) = (0.5, 0.6, 0.7)$. The parameter spaces Θ_{hi},

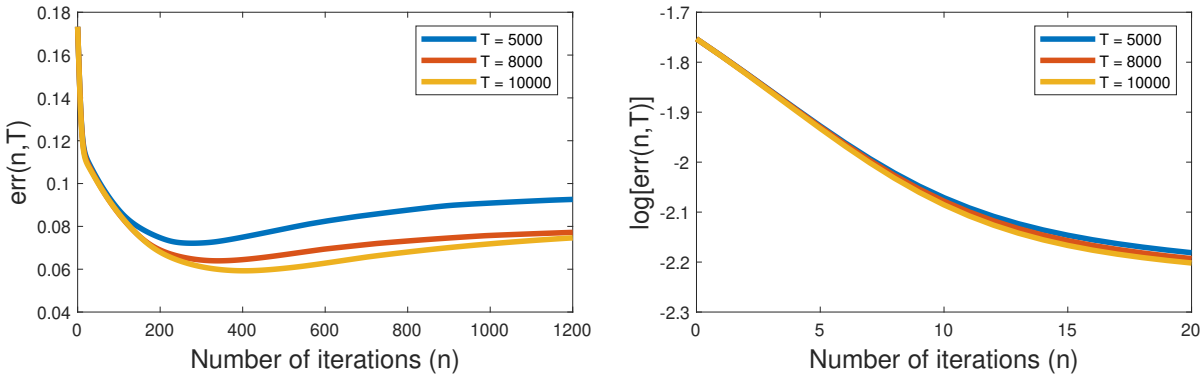

Figure 4. Plots of $err(n, T)$ and $\log err(n, T)$ with varying n and T.

Θ_{lo} and Θ_b are all equal to the interval $[0.1, 0.9]$. 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 4.

Although the regularity conditions in Theorem 3 are hard to validate even for this simple example, we observe that the empirical result qualitatively matches the performance guarantee. From Figure 4, $err(n, T)$ decreases exponentially in the early phase of the algorithm (roughly the first 10 iterations). Then, as n further increases, $err(n, T)$ remains at low values but does not decrease to 0. One caveat is that when $n > 300$, $err(n, T)$ slightly increases. A possible explanation of this behavior consistent with Theorem 3 is the following: there exists a low probability event that the upper bound fails, and the algorithm converges to a stationary point of the likelihood function outside the vicinity of the true parameter. After all the sample paths converge, the slope of $err(n, T)$ becomes roughly flat again at $n = 1200$. Finally, consistent with Theorem 3, the algorithm achieves better performance as T increases.

In the following, we present details on the experiment and a few auxiliary results.

E.1. Generation of the observation sequences

We first introduce the method we use to sample observation sequences from the stationary Markov chain induced by the expert policy. Using the expert policy and a fixed (o_0, s_1) pair, we generate 50 observation sequences of length 20,000. Then, the first 10,000 samples in each observation sequence are discarded, and the rest is saved as the observation sequence used in the algorithm. Such a procedure is motivated by Lemma A.1: it can be easily verified that our first three assumptions are satisfied in our numerical example. Therefore, from Lemma A.1, the distribution of X_t approaches the stationary distribution ν_{θ^*} regardless of the initial (o_0, s_1) pair.

E.2. Analytical expression of the parameter update

For our numerical example, the parameter update step in Algorithm 1 has a unique analytical solution. For all $\theta \in \Theta$, $\omega \in \mathcal{X}^{\mathbb{Z}}$, $T \geq 2$ and $\mu \in \mathcal{M}$, we first derive the analytical expression of $M_{\mu,T}(\theta; \omega)_{hi}$ which is the updated parameter for π_{hi} based on the previous parameter θ. Such a notation is borrowed from Assumption 5. Using the expression of the Q-function (3), we have

$$M_{\mu,T}(\theta;\omega)_{hi} \in \arg\max_{\theta'_{hi}\in\Theta_{hi}} \sum_{t=1}^{T} \sum_{o_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t = 1)[\log \pi_{hi}(o_t|s_t; \theta'_{hi})],$$

where s_t on the RHS is the state value $\omega(s_t)$ from the sample path ω. We omit ω on the RHS for a cleaner notation. Let $f(\theta'_{hi})$ denote the sum inside the argmax. Then,

$$\begin{aligned}
f(\theta'_{hi}) = \sum_{t=1}^{T} \Big\{ & \gamma^{\theta}_{\mu,t|T}(o_t = \text{LEFTEND}, b_t = 1)[\log \pi_{hi}(o_t = \text{LEFTEND}|s_t; \theta'_{hi})] \\
& + \gamma^{\theta}_{\mu,t|T}(o_t = \text{RIGHTEND}, b_t = 1)[\log \pi_{hi}(o_t = \text{RIGHTEND}|s_t; \theta'_{hi})] \Big\} \\
= \sum_{t=1}^{T} \Big\{ & \gamma^{\theta}_{\mu,t|T}(o_t = \text{LEFTEND}, b_t = 1) \Big[\mathbb{1}[s_t = 1, 2] \log \theta'_{hi} + \mathbb{1}[s_t = 3, 4] \log(1 - \theta'_{hi}) \Big] \\
& + \gamma^{\theta}_{\mu,t|T}(o_t = \text{RIGHTEND}, b_t = 1) \Big[\mathbb{1}[s_t = 3, 4] \log \theta'_{hi} + \mathbb{1}[s_t = 1, 2] \log(1 - \theta'_{hi}) \Big] \Big\}.
\end{aligned}$$

Taking the derivative of $f(\theta'_{hi})$, we can verify that $f(\theta'_{hi})$ is strongly concave on $[0.1, 0.9]$. Therefore, the parameter update for π_{hi} is unique.

$$M_{\mu,T}(\theta;\omega)_{hi} = \begin{cases} 0.1, & \text{if } \tilde{M}_{\mu,T}(\theta;\omega)_{hi} < 0.1, \\ \tilde{M}_{\mu,T}(\theta;\omega)_{hi}, & \text{if } 0.1 \le \tilde{M}_{\mu,T}(\theta;\omega)_{hi} \le 0.9, \\ 0.9, & \text{if } \tilde{M}_{\mu,T}(\theta;\omega)_{hi} > 0.9, \end{cases}$$

where $\tilde{M}_{\mu,T}(\theta;\omega)_{hi}$ is the unconstrained parameter update given as

$$\begin{aligned}
\tilde{M}_{\mu,T}(\theta;\omega)_{hi} = & \frac{\sum_{t=1}^{T} \gamma^{\theta}_{\mu,t|T}(o_t = \text{LEFTEND}, b_t = 1) \mathbb{1}[s_t = 1, 2]}{\sum_{t=1}^{T} \sum_{o_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t = 1)} \\
& + \frac{\sum_{t=1}^{T} \gamma^{\theta}_{\mu,t|T}(o_t = \text{RIGHTEND}, b_t = 1) \mathbb{1}[s_t = 3, 4]}{\sum_{t=1}^{T} \sum_{o_t} \gamma^{\theta}_{\mu,t|T}(o_t, b_t = 1)}.
\end{aligned}$$

Similarly, the unconstrained parameter updates for π_{lo} and π_b are the following:

$$\tilde{M}_{\mu,T}(\theta;\omega)_{lo} = \frac{1}{T} \sum_{t=1}^{T} \sum_{b_t} \Big\{ \gamma^{\theta}_{\mu,t|T}(o_t = \text{LEFTEND}, b_t) \mathbb{1}[a_t = \text{LEFT}] + \gamma^{\theta}_{\mu,t|T}(o_t = \text{RIGHTEND}, b_t) \mathbb{1}[a_t = \text{RIGHT}] \Big\}.$$

$$\tilde{M}_{\mu,T}(\theta;\omega)_{b} = \frac{1}{T-1} \sum_{t=2}^{T} \sum_{o_{t-1}} \Big\{ \tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t = 1) \mathbb{1}[\text{event}] + \tilde{\gamma}^{\theta}_{\mu,t|T}(o_{t-1}, b_t = 0) \mathbb{1}[\neg\text{event}] \Big\},$$

where the event $= \{(s_t = 1, o_{t-1} = \text{LEFTEND}) \vee (s_t = 4, o_{t-1} = \text{RIGHTEND})\}$. The parameter updates $M_{\mu,T}(\theta;\omega)_{lo}$ and $M_{\mu,T}(\theta;\omega)_b$ are the projections of $\tilde{M}_{\mu,T}(\theta;\omega)_{lo}$ and $\tilde{M}_{\mu,T}(\theta;\omega)_b$ onto $[0.1, 0.9]$, respectively.

E.3. Supplementary results to Figure 4

In this subsection we present supplementary results to Figure 4. In Figure 4, $err(n, T)$ is defined as the average of $\|\theta^{(n)} - \theta^*\|_2$ over all the 50 sample paths. Here, we divide the set of sample paths into smaller sets (by percentiles) and evaluate the average of $\|\theta^{(n)} - \theta^*\|_2$ over these smaller sets separately. The settings for the computation of parameter estimates are the same as before. The following procedure serves as the post-processing step of the obtained parameter estimates.

Concretely, as defined before, we obtain a sequence $\{\|\theta^{(n)} - \theta^*\|_2; \omega, T\}_{n \in [0:N]}$ after running Algorithm 1 with any sample path ω and any T. After fixing T and letting $n = N$, $\|\theta^{(N)} - \theta^*\|_2$ is a function of ω only. With a given threshold interval $I = [I_1, I_2]$, we define a smaller set of sample paths as the set of ω with $\|\theta^{(N)} - \theta^*\|_2$ greater than the I_1-th percentile and less than the I_2-th percentile. Let $err(n, T, I)$ be the average of $\|\theta^{(n)} - \theta^*\|_2$ over this smaller set of sample path specified by interval I. For $T = 8000$, the values of $err(n, T, I)$ with specific choices of I are plotted below. If $I = [0, 100]$, $err(n, T, I)$ is equivalent to $err(n, T)$ investigated in Figure 4.

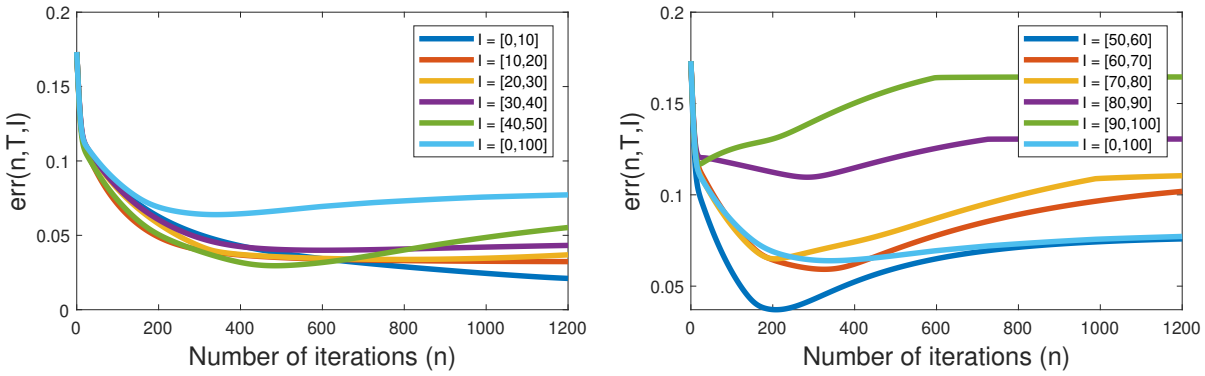

Figure 5. Plots of $err(n, T, I)$ with varying n and I; T is fixed as 8000.

Figure 5 suggests that with probability around 0.6, our algorithm with the particular choice of T and $\theta^{(0)}$ achieves decent performance, decreasing the estimation error by at least a half. A worth-noting observation is that, for all the choices of I (including $I = [90, 100]$ representing the *failed* sample paths), $err(n, T, I)$ roughly follows the same exponential decay in the early stage of the algorithm (roughly the first 10 iterations). The same behavior can be observed for $T = 5000$ and $T = 10000$ as well. It is not clear whether this behavior is general or specific to our numerical example. Detailed investigation is required in future work.

E.4. Varying μ

In this subsection we investigate the effect of μ on the performance of Algorithm 1. Intuitively, from the uniform forgetting analysis throughout this paper, it is reasonable to expect that at each iteration, the effect of μ on the parameter update is negligible if T is large. However, such a negligible error could accumulate if N is large. The effect of μ on the final parameter estimate is not clear without experiments.

We use the same observation sequences as before. T is fixed as 5000. $\theta^{(0)} = (0.5, 0.6, 0.7)$, and the parameter space for all the three parameters remains the same as $[0.1, 0.9]$. For all s_1, $\mu(o_0 = \text{RIGHTEND}|s_1) \in \{0.2, 0.5, 0.8\}$. The performance of the algorithm is evaluated by $err(n, T)$. The result is presented in Figure 6, which shows that indeed, the effect of μ on the final performance of the algorithm is negligible. For $n = 1000$, $\max_\mu err(n, T)$ is 0.7% higher than $\min_\mu err(n, T)$.

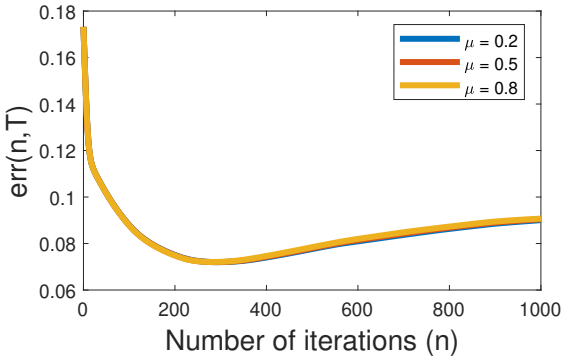

Figure 6. Plots of $err(n, T)$ with varying n and μ; T is fixed as 5000.

E.5. Random initialization

Up to this point, all the empirical results use the same initial parameter estimate $\theta^{(0)} = (0.5, 0.6, 0.7)$ on all the 50 sample paths. In this subsection, we evaluate the effect of the initial estimation error $\{\theta^{(0)} - \theta^*\}_2$ on the performance of the algorithm, by applying random $\theta^{(0)}$.

In this experiment, we use the same observation sequences as before. T is fixed to 8000. For all s_1, $\mu(o_0 = \text{RIGHTEND}|s_1) = 1$. The parameter space for all the three parameters remains the same as $[0.1, 0.9]$. For each observation sequence, we first generate three independent samples x_{hi}, x_{lo} and x_b uniformly from the interval $[0, 1]$. Then, $\theta^{(0)}$ is generated as follows: with a scale factor $w \in \{0.1, 0.2, 0.3\}$, let $\theta^{(0)}_{hi} = \theta^*_{hi} - wx_{hi}$, $\theta^{(0)}_{lo} = \theta^*_{lo} - wx_{lo}$ and $\theta^{(0)}_b = \theta^*_b - wx_b$. As a result, $\theta^{(0)}$ dependent on w is different for different observation sequences. The choices of $\theta^{(0)}$ are not symmetrical with respect to θ^* due to the restriction of the bounded parameter space. The result is shown in Figure 7.

From Figure 7, the curves corresponding to $w = 0.1$ and $w = 0.2$ qualitatively match the performance guarantee in

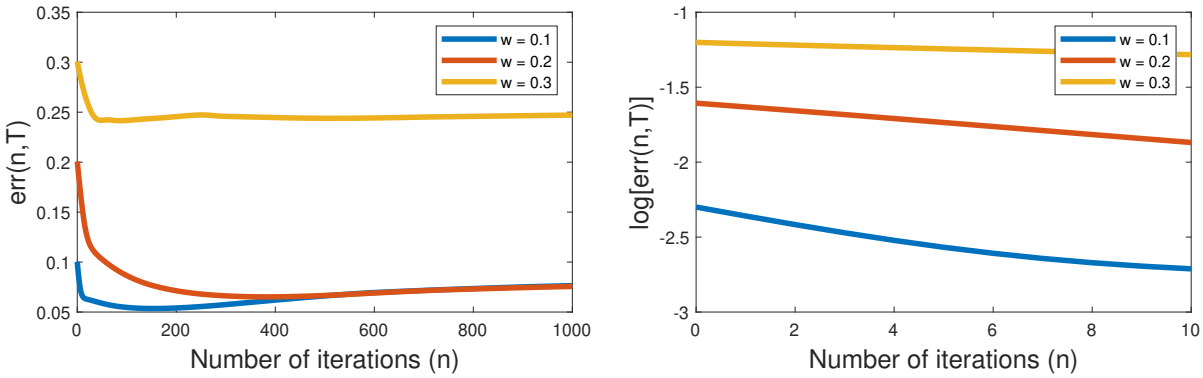

Figure 7. Plots of $err(n, T)$ with varying n and $\theta^{(0)}$; T is fixed to 8000.

Theorem 3. The algorithm achieves decent performance when $\{\theta^{(0)} - \theta^*\}_2$ is intermediate (the case of $w = 0.2$), where the average estimation error $err(n, T)$ is reduced by at least a half. If $\{\theta^{(0)} - \theta^*\}_2$ is small (the case of $w = 0.1$), the parameter estimates cannot improve much from $\theta^{(0)}$. If $\{\theta^{(0)} - \theta^*\}_2$ is large (the case of $w = 0.3$), the algorithm cannot converge to the vicinity of the true parameter, which is consistent with our local convergence analysis.

F. Discussion

In this section we discuss our scope and possible extensions to our work.

First, we assume that the parametric structure of the expert policy is known, which effectively transforms the HIL problem into a well-posed parametric inference problem. In practice, such a parametric structure may not be known a priori. Instead, we can choose an expressive enough parametric model for the learned policy (e.g., a neural network) and infer the optimal parameters. Our algorithm can be applied without modification to this case. Although the performance guarantee becomes invalid, the algorithm still converges to a stationary point of the (approximate) finite sample likelihood function, and decent empirical performance could be obtained assuming good initialization. Two caveats need to be emphasized. The first one is that, regularization is often required since these expressive parametric models often have high dimensional parameter spaces. A common regularization procedure is to penalize the similarity between low level policies corresponding to different options (Daniel et al., 2016a). In this way, the obtained set of options can be more versatile. The second caveat is that, the size of the option space needs to be set when selecting the parametric model. This is similar to the classical order estimation problem in HMMs, and relevant techniques could be borrowed from there.

Second, we restrict the state space $\mathcal{S}$, the action space $\mathcal{A}$ and the option space $\mathcal{O}$ to be finite. For $\mathcal{S}$ and $\mathcal{A}$, such a restriction is not very stringent: if we allow continuous $\mathcal{S}$ and $\mathcal{A}$ assuming the existence of density functions, the algorithm can still be applied. We only need to replace π_{lo} in Theorem 4 by its density. As for the performance guarantee (Theorem 3), the structure of the proof remains the same. The assumptions need to be modified, but such a modification is more technical rather than essential. On the other hand, we emphasize that the restriction of finite $\mathcal{O}$ is important. If $\mathcal{O}$ is continuous, then each step in the forward-backward recursion (Theorem 4) is an integral instead of a sum, and techniques like Sequential Monte Carlo (SMC) need to be applied. Fortunately, it is widely accepted that the options framework should have a finite option space, since the options need to be distinct and separate (Daniel et al., 2016a). Based on this argument, it is sufficient to consider only finite $\mathcal{O}$.