A. Notation

Symbol	Expression	Meaning
Θ	-	Set of parameters
k	$\lvert\Theta\rvert$	Number of possible parameters/tasks
n	-	Sample budget for querying the generative model (Section 3)
m	-	Maximum number of tasks in the sequential setting (Section 4)
$\mathcal{S}$	-	Set of S states
$\mathcal{A}$	-	Set of A actions
$\mathcal{U}$	-	Set of reward values (finite with cardinality U for Section 4 only)
T	-	Task-transition matrix
$p_\theta(s'\lvert s,a)$	-	Transition probabilities of MDP $\mathcal{M}_\theta$
$q_\theta(u\lvert s,a)$	-	Reward distribution of MDP $\mathcal{M}_\theta$
$r_\theta(s,a)$	-	Mean reward of MDP $\mathcal{M}_\theta$
γ	-	Discount factor
$V_\theta^\pi(s)$	$\mathbb{E}_\theta^\pi\left[\sum_{t=0}^\infty \gamma^t U_t \lvert S_0 = s\right]$	Value function of policy π in MDP $\mathcal{M}_\theta$
$V_\theta^*(s)$	$\max_\pi V_\theta^\pi(s)$	Optimal value function for MDP $\mathcal{M}_\theta$
$\sigma_\theta^r(s,a)^2$	$\mathbb{V}\mathrm{ar}_{q_\theta(\cdot\lvert s,a)}[U]$	Reward variance in task $\mathcal{M}_\theta$
$\sigma_\theta^p(s,a;\theta')^2$	$\mathbb{V}\mathrm{ar}_{p_\theta(\cdot\lvert s,a)}[V_{\theta'}^*(S')]$	Transition/value-function variance in task $\mathcal{M}_\theta$
$\Delta_{s,a}^r(\theta,\theta')$	$\lvert r_\theta(s,a) - r_{\theta'}(s,a)\rvert$	Reward-gaps between tasks θ and θ'
$\Delta_{s,a}^p(\theta,\theta')$	$\lvert(p_\theta(s,a) - p_{\theta'}(s,a))^T V_\theta^*\rvert$	Transition-gaps between tasks θ and θ'
Δ	-	Estimation error of the approximate models (Assumption 1)
ϵ, δ	-	Accuracy and confidence level for Algorithm 1
$\widehat{r}_t, \widehat{p}_t, \widehat{\sigma}_t^r, \widehat{\sigma}_t^p$	See Algorithm 1	Empirical models after t steps
$C_{t,\delta}^x(s,a)$	See Lemma 3	Bernstein confidence intervals for $x \in \{r, p, \sigma_r, \sigma_p\}$
$\bar{\Theta}_t$	See Algorithm 1	Confidence set a t time t
$\mathcal{I}_{s,a}^r(\theta,\theta')$	See Definition 1	Reward information for discriminating θ, θ'
$\mathcal{I}_{s,a}^p(\theta,\theta')$	See Definition 1	Transition information for discriminating θ, θ'
$\mathcal{I}_t(s,a)$	$\max\{\mathcal{I}_t^r(s,a), \mathcal{I}_t^p(s,a)\}$	Index of s, a at time t (see Algorithm 1)
O	-	Mean-observation matrix containing the flattened MDP models
$\widehat{O}_h, \widehat{T}_h$	-	Estimated observation and task-transition matrices after h tasks
$\widetilde{r}_{h,\theta}, \widetilde{p}_{h,\theta}, \widetilde{\sigma}_{h,\theta}^r, \widetilde{\sigma}_{h,\theta}^p$	-	Estimated models after h tasks
$\widetilde{\Theta}_h$	-	Initial set of models for running PTUM on the h-th task
ρ_x or $\rho_x(\Theta, T)$	See Appendix D	Constants in the analysis of the spectral learning algorithm

Table 1. The notation adopted in this paper.

B. Analysis of the PTUM Algorithm

B.1. Definitions and Assumptions

The analysis is carried out under the following two assumptions.

Assumption 3. *Algorithm 1 always enters the transfer mode. That is, the model uncertainty is such that* $\Delta < \frac{\epsilon(1-\gamma)}{4(1+\gamma)}$.

Assumption 4. *The sample budget* n *is large enough to allow Algorithm 1 to identify an* ϵ*-optimal policy.*

These two assumptions allow us to analyze only the core part of PTUM (i.e., the transfer mode), thus excluding trivial cases in which the chosen (ϵ, δ)-PAC algorithm is called. In fact, if Assumption 3 does not hold, the sample complexity for computing an ϵ-optimal policy is equivalent to the one of the chosen algorithm. Similarly, if Assumption 4 does not hold, the sample complexity is n (the samples collected by the generative model) plus the sample complexity of the chosen algorithm.

We define the event $E := \{\forall t = 1, ..., n : \theta^* \in \bar{\Theta}_t\}$ under which the true model is never eliminated from the active model set. This event will be used extensively throughout the whole analysis.

B.2. Concentration Inequalities

Lemma 1 (Bernstein's inequality (Boucheron et al., 2003))**.** *Let* X *be a random variable such that* $|X| \leq c$ *almost surely,* $X_1, \dots, X_n$ n *i.i.d. samples of* X*, and* $\delta > 0$*. Then, with probability at least* $1 - \delta$*,*

$$\left| \mathbb{E}\left[X\right] - \frac{1}{n} \sum_{i=1}^{n} X_i \right| \leq \sqrt{\frac{2\mathbb{V}\mathrm{ar}[X] \log \frac{2}{\delta}}{n}} + \frac{c \log \frac{2}{\delta}}{3n}.$$

Lemma 2 (Empirical Bernstein's inequality (Maurer & Pontil, 2009))**.** *Let* X *be a random variable such that* $|X| \leq c$ *almost surely,* $X_1, \dots, X_n$ n *i.i.d. samples of* X*, and* $\delta > 0$*. Then, with probability at least* $1 - \delta$*,*

$$\left| \mathbb{E}\left[X\right] - \frac{1}{n} \sum_{i=1}^{n} X_i \right| \leq \sqrt{\frac{2\widehat{\mathbb{V}}\mathrm{ar}[X] \log \frac{4}{\delta}}{n}} + \frac{7c \log \frac{4}{\delta}}{3(n-1)},$$

where $\widehat{\mathbb{V}}\mathrm{ar}[X]$ *denotes the empirical variance of* X *using* n *samples.*

B.3. Lemmas

We begin by showing that the true model is never eliminated from the confidence sets of Algorithm 1 with high probability.

Lemma 3 (Valid confidence sets)**.** *Let* $\delta > 0$ *and, for* $N_t(s, a) > 1$*,*

$$C^r_{t,\delta}(s,a) = \sqrt{\frac{2\widehat{\sigma}^r_t(s,a)^2 \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{7 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)} + \Delta^r_{\max},$$

$$C^p_{t,\delta}(s,a;\theta') = \sqrt{\frac{2\widehat{\sigma}^p_t(s,a;\theta')^2 \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{7 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)(1-\gamma)} + \Delta^p_{\max},$$

$$C^{\sigma^r}_{t,\delta}(s,a) = \sqrt{\frac{2 \log \frac{4SAn(|\Theta|+1)}{\delta}}{N_t(s,a)-1}} + \Delta^{\sigma_r}_{\max},$$

$$C^{\sigma^p}_{t,\delta}(s,a) = \frac{1}{1-\gamma}\sqrt{\frac{2 \log \frac{4SAn(|\Theta|+1)}{\delta}}{N_t(s,a)-1}} + \Delta^{\sigma_p}_{\max}.$$

Set these confidence intervals to infinity if $N_t(x, a) \leq 1$*. Then, the event* $E := \{\forall t = 1, ..., n : \theta^* \in \bar{\Theta}_t\}$ *holds with probability at least* $1 - \delta$*.*

Proof. Take any step $t \geq 1$, any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, any model $\theta' \in \Theta$, and let $\delta' > 0$. We need to show that the conditions of (1)-(4) hold. First notice that these conditions trivially hold if $N_t(s,a) \leq 1$. Thus, suppose that $N_t(s,a) > 1$ so that the confidence intervals are well-defined. Using the triangle inequality we have that

$$|\widehat{r}_t(s,a) - \widetilde{r}_{\theta^*}(s,a)| \leq |\widehat{r}_t(s,a) - r_{\theta^*}(s,a)| + \Delta^r_{\max}$$

$$|(\widetilde{p}_{\theta^*}(s,a) - \widehat{p}_t(s,a))^T \widetilde{V}^*_{\theta'}| \leq |(p_{\theta^*}(s,a) - \widehat{p}_t(s,a))^T \widetilde{V}^*_{\theta'}| + \Delta^p_{\max}$$

$$|\widehat{\sigma}^r_t(s,a) - \widetilde{\sigma}^r_{\theta^*}(s,a)| \leq |\widehat{\sigma}^r_t(s,a) - \sigma^r_{\theta^*}(s,a)| + \Delta^{\sigma_r}_{\max},$$

$$|\widehat{\sigma}^p_t(s,a;\theta') - \widehat{\sigma}^p_{\theta^*}(s,a;\theta')| \leq |\widehat{\sigma}^p_t(s,a;\theta') - \sigma^p_{\theta^*}(s,a;\theta')| + \Delta^{\sigma_p}_{\max}.$$

Using Lemma 2, we have that, with probability at least $1-\delta'$,

$$|\widehat{r}_t(s,a) - r_{\theta^*}(s,a)| \leq \sqrt{\frac{2\widehat{\sigma}^r_t(s,a)^2 \log\frac{4}{\delta'}}{N_t(s,a)}} + \frac{7\log\frac{4}{\delta'}}{3(N_t(s,a)-1)}.$$

Similarly, for any $\theta' \in \Theta$, we have that, with probability at least $1-\delta'$,

$$|(p_{\theta^*}(s,a) - \widehat{p}_t(s,a))^T \widetilde{V}^*_{\theta'}| \leq \sqrt{\frac{2\widehat{\sigma}^p_t(s,a;\widetilde{V}^*_{\theta'})^2 \log\frac{4}{\delta'}}{N_t(s,a)}} + \frac{7\log\frac{4}{\delta'}}{3(N_t(s,a)-1)(1-\gamma)}.$$

From Theorem 10 of (Maurer & Pontil, 2009),

$$|\widehat{\sigma}^r_t(s,a) - \sigma^r_{\theta^*}(s,a)| \leq \sqrt{\frac{2\log\frac{2}{\delta'}}{N_t(s,a)-1}}$$

and

$$|\widehat{\sigma}^p_t(s,a;\theta') - \sigma^p_{\theta^*}(s,a;\theta')| \leq \frac{1}{1-\gamma}\sqrt{\frac{2\log\frac{2}{\delta'}}{N_t(s,a)-1}}$$

hold with probability at least $1-\delta'$, respectively.

Taking union bounds over all state action pairs and over the maximum number of samples n, these four inequalities hold at the same time with probability at least $1 - 2SA(|\Theta|+1)n\delta'$. The result follows after setting $\delta = 2SAn(|\Theta|+1)\delta'$ and rearranging.

□

Next we bound the number of samples required from some state-action pair in order to eliminate a model from the confidence set.

Lemma 4 (Model elimination)**.** *Let $\theta \in \Theta$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\Delta := \max\{\Delta^r_{\max}, \Delta^p_{\max}, \Delta^{\sigma_r}_{\max}, \Delta^{\sigma_p}_{\max}\}$, and define*

$$\bar{n}^r_\theta(s,a) := \min_{\theta'' \in \bar{\Theta}_t} \max\left\{ \frac{\widetilde{\sigma}^r_{\theta''}(s,a)^2}{[\widetilde{\Delta}^r_{s,a}(\theta^*,\theta) - 4\Delta]^2_+}, \frac{1}{[\widetilde{\Delta}^r_{s,a}(\theta^*,\theta) - 4\Delta]_+} \right\},$$

$$\bar{n}^p_\theta(s,a) := \min_{\theta' \in \Theta, \theta'' \in \bar{\Theta}_t} \max\left\{ \frac{\widetilde{\sigma}^p_{\theta''}(s,a;\theta')^2}{[\widetilde{\Delta}^p_{s,a}(\theta^*,\theta) - 4\Delta]^2_+}, \frac{1/(1-\gamma)}{[\widetilde{\Delta}^p_{s,a}(\theta^*,\theta) - 4\Delta]_+} \right\}.$$

Then, under event E, if $N_t(s,a) > \bar{n}_\theta(s,a) := 32\log\frac{8SAn(1+|\Theta|)}{\delta}\min\{\bar{n}^r_\theta(s,a), \bar{n}^p_\theta(s,a)\}$, we have that $\theta \notin \bar{\Theta}_t$.

Proof. We split the proof into two parts, dealing with rewards and transitions separately. We then combine these results to obtain the final statement.

Elimination by rewards Assuming $\theta \in \bar{\Theta}_t$, we must have, for all state-action pairs and all $\theta'' \in \bar{\Theta}_t$,

$$\begin{aligned}
\widetilde{\Delta}^r_{s,a}(\theta^*, \theta) &\leq |\widetilde{r}_\theta(s,a) - \widehat{r}_t(s,a)| + |\widetilde{r}_{\theta^*}(s,a) - \widehat{r}_t(s,a)| \leq 2C^r_{t,\delta}(s,a) \\
&\leq 2\sqrt{\frac{2\widehat{\sigma}^r_t(s,a)^2 \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{14 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)} + 2\Delta^r_{\max} \\
&\leq 2\sqrt{\frac{2\widetilde{\sigma}^r_{\theta''}(s,a)^2 \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{4 \log \frac{8SAn(|\Theta|+1)}{\delta}}{(N_t(s,a)-1)} + \frac{14 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)} + 2\Delta^r_{\max} + 2\Delta^{\sigma_r}_{\max} \\
&\leq 2\sqrt{\frac{2\widetilde{\sigma}^r_{\theta''}(s,a)^2 \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{26 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)} + 4\Delta
\end{aligned}$$

where we applied the triangle inequality and used Lemma 3 to upper bound the empirical variance by the variance of a model θ'' in the confidence set. We note that, if the model uncertainty is too high and the denominators of $\bar{n}_\theta(s,a)$ are zero, it is not possible to eliminate θ from the rewards of this state-action pair. If this is not the case, for $N_t(s,a) \geq \bar{n}_\theta(s,a)$ we have that

$$2\sqrt{\frac{2\widetilde{\sigma}^r_{\theta''}(s,a)^2 \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} < \frac{\widetilde{\Delta}^r_{s,a}(\theta^*, \theta) - 4\Delta}{2}$$

and

$$\frac{26 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)} < \frac{\widetilde{\Delta}^r_{s,a}(\theta^*, \theta) - 4\Delta}{2}.$$

Plugging these two inequalities in the first upper bound leads to the contradiction $\widetilde{\Delta}^r_{s,a}(\theta^*, \theta) < \widetilde{\Delta}^r_{s,a}(\theta^*, \theta)$, hence it must be that $\theta \notin \bar{\Theta}_t$.

Elimination by transition The proof proceeds analogously to the previous case. Let $\theta' \in \Theta$ and $\theta'' \in \bar{\Theta}_t$, then

$$\begin{aligned}
\widetilde{\Delta}^p_{s,a}(\theta^*, \theta) &\leq |(\widehat{p}_t(s,a) - \widetilde{p}_{\theta^*}(s,a))^T \widetilde{V}^*_{\theta'}| + |(\widehat{p}_t(s,a) - \widetilde{p}_\theta(s,a))^T \widetilde{V}^*_{\theta'}| \leq 2C^p_{t,\delta}(s,a;\theta') \\
&\leq 2\sqrt{\frac{2\widehat{\sigma}^p_t(s,a;\theta') \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{14 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)(1-\gamma)} + 2\Delta^p_{\max} \\
&\leq 2\sqrt{\frac{2\widetilde{\sigma}^p_{\theta''}(s,a;\theta') \log \frac{8SAn(|\Theta|+1)}{\delta}}{N_t(s,a)}} + \frac{26 \log \frac{8SAn(|\Theta|+1)}{\delta}}{3(N_t(s,a)-1)(1-\gamma)} + 4\Delta,
\end{aligned}$$

where once again we applied the triangle inequality and Lemma 3 to upper bound the empirical variance. Hence, applying the same reasoning as before we obtain a contradiction, which in turns implies that $\theta \notin \bar{\Theta}_t$.

We finally note that if $\bar{n}_\theta(s,a) = +\infty$, i.e., the approximate models are too inaccurate, it is not possible to eliminate θ using this state-action pair. $\square$

The following is a known result which bounds the deviation in value function between different MDPs.

Lemma 5 (Simulation lemma)**.** *Let $\theta, \theta' \in \Theta$, $s \in \mathcal{S}$, and π be any policy. Denote by $\nu^\pi_\theta(s', a'; s)$ the discounted state-action visitation frequencies (Sutton et al., 2000) of π in MDP θ starting from s. Then, the following hold:*

$$|V^\pi_\theta(s) - V^\pi_{\theta'}(s)| \leq \sum_{s',a'} \nu^\pi_{\theta'}(s', a'; s) \left[|r_\theta(s',a') - r_{\theta'}(s',a')| + \gamma |(p_\theta(s',a') - p_{\theta'}(s',a'))^T V^\pi_\theta|\right].$$

$$|V^*_{\theta'}(s) - V^*_\theta(s)| \leq \max_{\pi \in \{\pi^*_\theta, \pi^*_{\theta'}\}} \sum_{s',a'} \nu^\pi_{\theta'}(s', a'; s) \left[|r_\theta(s',a') - r_{\theta'}(s',a')| + \gamma |(p_\theta(s',a') - p_{\theta'}(s',a'))^T V^*_\theta|\right].$$

Proof. See, e.g., Lemma 3 of (Zanette et al., 2019) for the first inequality and Lemma 2 of (Azar et al., 2013b) or Lemma 2 of (Zanette et al., 2019) for the second one.[5] □

Corollary 2 (Value-function error decomposition). *Let $\theta, \theta' \in \Theta$ and $s \in \mathcal{S}$. Then,*

$$|V^*_{\theta'}(s) - V^{\pi^*_\theta}_{\theta'}(s)| \leq 2 \max_{\pi \in \{\pi^*_\theta, \pi^*_{\theta'}\}} \sum_{s',a'} \nu^\pi_{\theta'}(s',a';s) \left[|r_\theta(s',a') - r_{\theta'}(s',a')| + \gamma|(p_\theta(s',a') - p_{\theta'}(s',a'))^T V^*_\theta|\right].$$

Proof. Using the triangle inequality,

$$|V^*_{\theta'}(s) - V^{\pi^*_\theta}_{\theta'}(s)| \leq \underbrace{|V^*_{\theta'}(s) - V^*_\theta(s)|}_{(a)} + \underbrace{|V^*_\theta(s) - V^{\pi^*_\theta}_{\theta'}(s)|}_{(b)}.$$

We can bound (a) using the second inequality in Lemma 5 as

$$|V^*_{\theta'}(s) - V^*_\theta(s)| \leq \max_{\pi \in \{\pi^*_\theta, \pi^*_{\theta'}\}} \sum_{s',a'} \nu^\pi_{\theta'}(s',a';s) \left[|r_\theta(s',a') - r_{\theta'}(s',a')| + \gamma|(p_\theta(s',a') - p_{\theta'}(s',a'))^T V^*_\theta|\right].$$

Similarly, we can use the first inequality in Lemma 5 to bound (b) by noticing that $V^*_\theta = V^{\pi^*_\theta}_\theta$. We have

$$\begin{aligned}|V^*_\theta(s) - V^{\pi^*_\theta}_{\theta'}(s)| &\leq \sum_{s',a'} \nu^{\pi^*_\theta}_{\theta'}(s',a';s) \left[|r_\theta(s',a') - r_{\theta'}(s',a')| + \gamma|(p_\theta(s',a') - p_{\theta'}(s',a'))^T V^*_\theta|\right]\\ &\leq \max_{\pi \in \{\pi^*_\theta, \pi^*_{\theta'}\}} \sum_{s',a'} \nu^\pi_{\theta'}(s',a';s) \left[|r_\theta(s',a') - r_{\theta'}(s',a')| + \gamma|(p_\theta(s',a') - p_{\theta'}(s',a'))^T V^*_\theta|\right].\end{aligned}$$

Combining the two displays above concludes the proof. □

The following lemma ensures that, if the algorithm did not stop at a certain time t, certain models belong to the confidence set.

Lemma 6 (Stopping condition). *Let τ be the random stopping time of Algorithm 1 and*

$$\Theta_\epsilon := \left\{\theta \in \Theta \;\middle|\; \|\widetilde{r}_\theta - \widetilde{r}_{\theta^*}\| > \kappa_\epsilon \vee \|(\widetilde{p}_\theta - \widetilde{p}_{\theta^*})^T \widetilde{V}^*_{\theta^*}\| > \frac{\kappa_\epsilon}{\gamma}\right\},$$

where $\kappa_\epsilon := \frac{(1-\gamma)\epsilon}{4} - \frac{\Delta(1+\gamma)}{2}$. Then, under event E, for all $t < \tau$, there exists at least one model $\theta \in \Theta_\epsilon$ such that $\theta \in \bar{\Theta}_t$.

Proof. We note that, for all $t < \tau$, under event E, it must be that

$$\exists \theta \in \bar{\Theta}_t, s \in \mathcal{S} : \widetilde{V}^{\widetilde{\pi}^*_{\theta^*}}_\theta(s) < \widetilde{V}^*_\theta(s) - \epsilon + 2\Delta\frac{(1+\gamma)}{1-\gamma},$$

otherwise the algorithm would stop before τ. This implies that $|\widetilde{V}^*_\theta(s) - \widetilde{V}^{\widetilde{\pi}^*_{\theta^*}}_\theta(s)| > \epsilon - 2\Delta\frac{(1+\gamma)}{1-\gamma}$ holds as well and, using Corollary 2,

$$2 \max_{\pi \in \{\widetilde{\pi}^*_{\theta^*}, \widetilde{\pi}^*_\theta\}} \sum_{s',a'} \nu^\pi_\theta(s',a';s) \left[|\widetilde{r}_\theta(s',a') - \widetilde{r}_{\theta^*}(s',a')| + \gamma|(\widetilde{p}_\theta(s',a') - \widetilde{p}_{\theta^*}(s',a'))^T \widetilde{V}^*_{\theta^*}|\right] > \epsilon - 2\Delta\frac{(1+\gamma)}{1-\gamma} \tag{5}$$

holds for some $\theta \in \bar{\Theta}_t$ and $s \in \mathcal{S}$. Assume that all models in Θ_ϵ have been eliminated. Then, using that ν sums up to $1/(1-\gamma)$ and that all models must be sufficiently close to θ^*, the left-hand side of this inequality can be upper bounded by $\epsilon - 2\Delta\frac{(1+\gamma)}{1-\gamma}$. Hence, we obtain a contradiction and it must be that $\theta \in \widetilde{\Theta}_t$ for some $\theta \in \Theta_\epsilon$. □

Lemma 7 (Positive index). *Let τ be the random stopping time of Algorithm 1, then, under event E,*

$$\forall t < \tau : \mathcal{I}_t(S_t, A_t) > 0$$

[5]Note that, although the inequalities of, e.g., Azar et al. (2013b) and Zanette et al. (2019) relate the value functions of a fixed MDP with those of its empirical counterpart, they actually hold for any two MDPs.

Proof. Recall that the algorithm enters the transfer mode if $\Delta < \frac{\epsilon(1-\gamma)}{4(1+\gamma)}$. Take any time $t < \tau$. Under event E, we have $\theta^* \in \bar{\Theta}_t$ and Lemma 6 implies that $\theta \in \bar{\Theta}_t$ for some $\theta \in \Theta_\epsilon$. The definition of Θ_ϵ implies that either $\|\widetilde{r}_\theta - \widetilde{r}_{\theta^*}\| > \kappa_\epsilon$ or $\|(\widetilde{p}_\theta - \widetilde{p}_{\theta^*})^T \widetilde{V}^*_{\theta^*}\| > \frac{\kappa_\epsilon}{\gamma}$ and both these quantities are strictly greater than zero since $\kappa_\epsilon > \frac{\epsilon(1-\gamma)}{8}$. Since the index contains a maximum over models involving these two, the result follows straightforwardly. □

The following lemma is the key result that allows us to bound the sample complexity of Algorithm 1. It shows that, at any time t, the number of times the chosen state-action action pair (S_t, A_t) has been chosen before is bounded by a quantity proportional to minimum number of samples required from any state-action pair to eliminate any of the active models.

Lemma 8 (Fundamental lemma)**.** *Let (S_t, A_t) be the state-action pair chosen at time t. Then, under event E, the number of queries to such couple prior to time t can be upper bounded by*

$$N_t(S_t, A_t) < \frac{128 \log(8SAn(|\Theta| + 1)/\delta)}{\max_{s,a} \max_{\theta \in \bar{\Theta}_t} \mathcal{I}_{s,a}(\theta^*, \theta)}.$$

Proof. Let $F_t = \mathbb{1}\{\forall s, a \in \mathcal{S} \times \mathcal{A} : \mathcal{I}^r_t(S_t, A_t) \geq \mathcal{I}_t(s, a)\}$ be the event under which, at time t, the maximizer of the index is attained by the reward components. The proof is divided in two parts, based on whether F_t holds or not.

Event F_t holds We start by defining some quantities. Let $\underline{\Theta}_t := \operatorname{argmax}_{\theta, \theta' \in \bar{\Theta}_t} \mathcal{I}^r_t(s, a)$ be the set of active models that attain the maximum in the reward index. Similarly, define

$$\bar{\theta}_t := \operatorname*{argmax}_{\theta \in \bar{\Theta}_t} \widetilde{\Delta}^r_{S_t, A_t}(\theta^*, \theta), \quad \underline{\theta}_t := \operatorname*{argmin}_{\theta \in \bar{\Theta}_t} \widetilde{\Delta}^r_{S_t, A_t}(\theta^*, \theta),$$

as the farthest and closest models from θ^* among the active ones, respectively. Assume, without loss of generality, that the maximums/minimums are attained by single models. If more than one model attains them, the proof follows equivalently by choosing arbitrary ones. Furthermore, let θ^v_t be the (random) model among those in $\underline{\Theta}_t$ whose reward-variance is used to attain the maximum in the index.

We now proceed as follows. First, we prove that an upper bound to the index of the chosen state-action pair directly relates to the sample complexity for eliminating $\bar{\theta}_t$. Then, we use this result to guarantee that (S_t, A_t) cannot be chosen more than the stated quantity prior to time step t, otherwise $\bar{\theta}_t$ could not be an active model.

By assumption we have

$$\begin{aligned}
\mathcal{I}_t(S_t, A_t) = \mathcal{I}^r_t(S_t, A_t) &= \min\left\{ \frac{(\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \underline{\theta}_t) - 8\Delta)^2}{\widetilde{\sigma}^r_{\theta^v_t}(S_t, A_t)^2}, \widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \underline{\theta}_t) - 8\Delta \right\} \\
&\overset{(a)}{\leq} \min\left\{ \frac{\left(\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) + \widetilde{\Delta}^r_{S_t,A_t}(\theta^*, \underline{\theta}_t) - 8\Delta\right)^2}{\widetilde{\sigma}^r_{\theta^v_t}(S_t, A_t)^2}, \widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) + \widetilde{\Delta}^r_{S_t,A_t}(\theta^*, \underline{\theta}_t) - 8\Delta \right\} \\
&\overset{(b)}{\leq} \min\left\{ \frac{(2\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 8\Delta)^2}{\widetilde{\sigma}^r_{\theta^v_t}(S_t, A_t)^2}, 2\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 8\Delta \right\} \\
&\leq 4 \min\left\{ \frac{(\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 4\Delta)^2}{\widetilde{\sigma}^r_{\theta^v_t}(S_t, A_t)^2}, \widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 4\Delta \right\},
\end{aligned}$$

where (a) follows from the triangle inequality and (b) from the definition of $\bar{\theta}_t$ (which was defined as the farthest from the estimate of θ^*). Note that to prove this inequalities we also need that $\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \underline{\theta}_t) - 8\Delta \geq 0$, which is implied by Lemma 7. Since (S_t, A_t) is chosen at time t, it must be that $\mathcal{I}_t(S_t, A_t) \geq \mathcal{I}_t(s, a)$ for all $s, a \in \mathcal{S} \times \mathcal{A}$. This implies that, for all $s, a \in \mathcal{S} \times \mathcal{A}$ and $\theta \in \widetilde{\Theta}_t$,

$$4 \min\left\{ \frac{(\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 4\Delta)^2}{\widetilde{\sigma}^r_{\theta^v_t}(S_t, A_t)^2}, \widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 4\Delta \right\} \geq \mathcal{I}_t(s, a) \geq \mathcal{I}_{s,a}(\theta^*, \theta), \tag{6}$$

where the second inequality holds since the index of (s, a) is by definition larger than the one using the models θ^* and θ. Note that Lemma 4 and 7 ensure that a number of queries to (S_t, A_t) of

$$32 \log \frac{8SAn(|\Theta|+1)}{\delta} \max\left\{ \frac{\widetilde{\sigma}^r_{\theta^v_t}(S_t, A_t)^2}{(\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 4\Delta)^2}, \frac{1}{\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \theta^*) - 4\Delta} \right\}$$

suffices for eliminating $\bar{\theta}_t$. In particular, Lemma 7 implies that $\widetilde{\Delta}^r_{S_t,A_t}(\bar{\theta}_t, \underline{\theta}_t) > 8\Delta$, which, in turn, implies that $\widetilde{\Delta}^r_{S_t,A_t}(\theta^*, \bar{\theta}_t) > 4\Delta$. Therefore, Equation 6 above implies that a number of queries of

$$\frac{128 \log(8SAn(|\Theta|+1)/\delta)}{\max_{s,a} \max_{\theta \in \bar{\Theta}_t} \mathcal{I}_{s,a}(\theta^*, \theta)}.$$

also suffices. We note that the maximums at the denominator can be introduced since (6) holds for all s, a and θ. Hence, it must be that $N_t(S_t, A_t)$ is strictly less than this quantity, otherwise the model $\bar{\theta}_t$ would be eliminated at time step $t-1$ and it could not be active at time t. This concludes the first part of the proof.

Event F_t does not hold In this case, the maximizer of the index must be attained using the transition components, thus $\mathcal{I}_t(S_t, A_t) = \mathcal{I}^p_t(S_t, A_t)$. The proof follows exactly the same steps as before and is therefore not reported. Since the result is the same, combining these two parts proves the main statement.

□

The following lemma ensures that Algorithm 1 returns ϵ-optimal policies with high probability.

Lemma 9 (Correctness). *Let τ be the stopping time of Algorithm 1 and π_τ be the returned policy. Then, under event E, π_τ is ϵ-optimal with respect to θ^*.*

Proof. Recall that π_τ is ϵ-optimal if, for all states, $V^{\pi_\tau}_{\theta^*}(s) \geq V^*_{\theta^*}(s) - \epsilon$. Furthermore, π_τ is optimal for one of the active models at time τ, i.e., $\pi_\tau = \widetilde{\pi}^*_\theta$ for some $\theta \in \bar{\Theta}_\tau$. Since under E we have $\theta^* \in \bar{\Theta}_\tau$, a sufficient condition is that $\|V^*_{\theta'} - V^{\widetilde{\pi}^*_\theta}_{\theta'}\| < \epsilon$ holds for all $\theta' \in \bar{\Theta}_\tau$. Let us upper bound the left-hand side as

$$\|V^*_{\theta'} - V^{\widetilde{\pi}^*_\theta}_{\theta'}\| \leq \|\widetilde{V}^{\widetilde{\pi}^*_\theta}_{\theta'} - V^{\widetilde{\pi}^*_\theta}_{\theta'}\| + \|V^*_{\theta'} - \widetilde{V}^*_{\theta'}\| + \|\widetilde{V}^{\widetilde{\pi}^*_\theta}_{\theta'} - \widetilde{V}^*_{\theta'}\|$$

Using Lemma 5, we can bound the first term by

$$\begin{aligned}
\|\widetilde{V}^{\widetilde{\pi}^*_\theta}_{\theta'} - V^{\widetilde{\pi}^*_\theta}_{\theta'}\| &\leq \sum_{s',a'} \nu^{\widetilde{\pi}^*_\theta}_{\theta'}(s', a'; s)(|r_{\theta'}(s,a) - \widetilde{r}_{\theta'}(s,a)| + \gamma |(p_{\theta'}(s,a) - \widetilde{p}_{\theta'}(s,a))^T \widetilde{V}^{\widetilde{\pi}^*_\theta}_{\theta'}| \\
&\leq \sum_{s',a'} \nu^{\widetilde{\pi}^*_\theta}_{\theta'}(s', a'; s)(\Delta + \gamma\Delta) \leq \frac{\Delta(1+\gamma)}{1-\gamma}
\end{aligned}$$

and the second term by

$$\begin{aligned}
\|V^*_{\theta'} - \widetilde{V}^*_{\theta'}\| &\leq \max_{\pi \in \{\pi^*_{\theta'}, \widetilde{\pi}^*_{\theta'}\}} \sum_{s',a'} \nu^{\pi}_{\theta'}(s', a'; s)(|r_{\theta'}(s,a) - \widetilde{r}_{\theta'}(s,a)| + \gamma |(p_{\theta'}(s,a) - \widetilde{p}_{\theta'}(s,a))^T \widetilde{V}^*_{\theta'}| \\
&\leq \max_{\pi \in \{\pi^*_{\theta'}, \widetilde{\pi}^*_{\theta'}\}} \sum_{s',a'} \nu^{\pi}_{\theta'}(s', a'; s)(\Delta + \gamma\Delta) \leq \frac{\Delta(1+\gamma)}{1-\gamma}.
\end{aligned}$$

Therefore, the stopping condition,

$$\|\widetilde{V}^{\widetilde{\pi}^*_\theta}_{\theta'} - \widetilde{V}^*_{\theta'}\| + 2\Delta \frac{(1+\gamma)}{1-\gamma} \leq \epsilon$$

implies that $\|V^*_{\theta'} - V^{\widetilde{\pi}^*_\theta}_{\theta'}\| \leq \epsilon$, which in turn implies the ϵ-optimality of π_τ. □

B.4. Sample Complexity Bounds

We are now ready to prove the main theorem, which bounds the sample complexity of Algorithm 1.

Theorem 1. *Assume Δ is such that Algorithm 1 enters the transfer mode. Let τ be the random stopping time and π_τ be the returned policy. Then, with probability at least $1-\delta$, π_τ is ϵ-optimal for θ^* and the total number of queries to the generative model can be bounded by*

$$\tau \leq \frac{128 \min\{SA, |\Theta|\} \log(8SAn(|\Theta|+1)/\delta)}{\max_{s,a} \min_{\theta \in \Theta_\epsilon} \mathcal{I}_{s,a}(\theta^*, \theta)},$$

where, for $\kappa_\epsilon := \frac{(1-\gamma)\epsilon}{4} - \frac{\Delta(1+\gamma)}{2}$, the set $\Theta_\epsilon \subseteq \Theta$ is

$$\Theta_\epsilon := \left\{ \theta \;\middle|\; \|\widetilde{r}_\theta - \widetilde{r}_{\theta^*}\| > \kappa_\epsilon \vee \|(\widetilde{p}_\theta - \widetilde{p}_{\theta^*})^T \widetilde{V}^*_{\theta^*}\| > \frac{\kappa_\epsilon}{\gamma} \right\}.$$

Proof. Lemma 3 ensures that event E holds with probability at least $1-\delta$. Therefore, we shall carry out the proof conditioned on E.
We split the proof into two parts. In the first one, we bound the number of times each state-action pair can be visited before the algorithm stops. In the second part, we directly bound the number of steps in which each model can be active.

Bound over $\mathcal{S} \times \mathcal{A}$ Take any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any sequence $\{n_t\}_{t \geq 1}$, its number of visits can be written as

$$\begin{aligned} N_\tau(s,a) &= \sum_{t=1}^{\tau} \mathbb{1}\{S_t = s \wedge A_t = a | E\} \\ &= \underbrace{\sum_{t=1}^{\tau} \mathbb{1}\{S_t = s \wedge A_t = a \wedge N_t(s,a) < n_t | E\}}_{(a)} + \underbrace{\sum_{t=1}^{\tau} \mathbb{1}\{S_t = s \wedge A_t = a \wedge N_t(s,a) \geq n_t | E\}}_{(b)}. \end{aligned}$$

For

$$n_t := \frac{128 \log(8SAn(|\Theta|+1)/\delta)}{\max_{s,a} \max_{\theta \in \bar{\Theta}_t} \mathcal{I}_{s,a}(\theta^*, \theta)},$$

Lemma 8 ensures that, under event E, $(b) = 0$. Thus, we only need to bound (a). For all $t < \tau$, Lemma 6 implies that there exists a model $\theta \in \Theta_\epsilon$ which also belongs to the confidence set at time t, $\theta \in \bar{\Theta}_t$. Therefore, for all $t < \tau$ and $(s', a') \in \mathcal{S} \times \mathcal{A}$, we have $\max_{\theta \in \bar{\Theta}_t} \mathcal{I}_{s',a'}(\theta^*, \theta) \geq \min_{\theta \in \Theta_\epsilon} \mathcal{I}_{s',a'}(\theta^*, \theta)$. Since we removed all random quantities from n_t, we can now bound (a) as

$$(a) < \frac{128 \log(8SAn(|\Theta|+1)/\delta)}{\max_{s,a} \min_{\theta \in \Theta_\epsilon} \mathcal{I}_{s,a}(\theta^*, \theta)}. \tag{7}$$

This immediately yields a bound on the stopping time,

$$\tau = \sum_{s,a} N_\tau(s,a) < \frac{128 SA \log(8SAn(|\Theta|+1)/\delta)}{\max_{s,a} \min_{\theta \in \Theta_\epsilon} \mathcal{I}_{s,a}(\theta^*, \theta)}.$$

Bound over Θ From the first part of the proof, we know that, for $t < \tau$, the confidence set $\bar{\Theta}_t$ must contain a model that is also in Θ_ϵ, otherwise the algorithm would stop. Therefore, the stopping time can be bounded by

$$\tau \leq \sum_{t=1}^{n} \mathbb{1}\left\{\exists \theta \in \bar{\Theta}_t : \theta \in \Theta_\epsilon | E\right\}.$$

By definition of the algorithm, the state-action pair chosen at each time step does not change until the set of active models $\underline{\Theta}_t$ (those that control the maximizer of the index as in the proof of Lemma 8) does not change. Furthermore, once a model has been eliminated, it cannot become active again. Consider a sequence $\{\tau_h\}_{h\geq 1}$ with $\tau_1 = 1$. We can partition the time line into different contiguous intervals (from now on called phases) $\mathcal{T}_h := [\tau_h, \tau_{h+1} - 1]$ such that the set of active models does not change within $\mathcal{T}_h$ and a change of phase occurs only when a model is eliminated. Let $\underline{\Theta}_h$ be the set of active models in phase h. We have $\tau_{h+1} = \inf_{t\geq 1}\{t \mid \exists\theta \in \underline{\Theta}_h : \theta \notin \bar{\Theta}_t\}$. That is, the beginning of the new phase $h+1$ is the step where one of the previously-active models is eliminated. Let $\bar{\theta}_h$ be any such model and $h(t)$ be the (unique) phase containing time t. Note that, for each $\theta \in \Theta \setminus \{\theta^*\}$, there exists at most one phase $h(\theta)$ where $\bar{\theta}_{h(\theta)} = \theta$. Then,

$$\begin{aligned}
\tau &\leq \sum_{\theta\in\Theta\setminus\{\theta^*\}} \sum_{t=1}^{n} \mathbb{1}\left\{\bar{\theta}_{h(t)} = \theta \wedge \exists\theta' \in \bar{\Theta}_t : \theta' \in \Theta_\epsilon | E\right\} \\
&\leq \sum_{\theta\in\Theta\setminus\{\theta^*\}} \sum_{t=\tau_{h(\theta)}}^{\tau_{h(\theta)+1}-1} \mathbb{1}\left\{\bar{\theta}_{h(t)} = \theta \wedge \exists\theta' \in \bar{\Theta}_t : \theta' \in \Theta_\epsilon | E\right\} \leq \frac{128(|\Theta|-1)\log(8SAn(|\Theta|+1)/\delta)}{\max_{s,a}\min_{\theta\in\Theta_\epsilon}\mathcal{I}_{s,a}(\theta^*,\theta)},
\end{aligned}$$

where in the last inequality we applied Lemma 8 by noticing that, within the same phase, the chosen state-action pair does not change and used the fact that a model in Θ_ϵ still survives to upper bound the minimum over models in the confidence set. The proof follows by taking the minimum of the two bounds.

□

Corollary 1. *Let Γ be the minimum gap between θ^* and any other model in Θ,*

$$\Gamma := \min_{\theta\neq\theta^*} \max\left\{\|\widetilde{r}_\theta - \widetilde{r}_{\theta^*}\|, \|(\widetilde{p}_\theta - \widetilde{p}_{\theta^*})^T \widetilde{V}^*_{\theta^*}\|\right\}.$$

Then, with probability at least $1-\delta$,

$$\tau \leq \widetilde{\mathcal{O}}\left(\frac{\min\{SA, |\Theta|\}\log(1/\delta)}{\max\{\Gamma^2, \epsilon^2\}(1-\gamma)^4}\right).$$

Proof. We notice that each model $\theta \in \Theta_\epsilon$ is, by definition, such that either $\|\widetilde{r}_\theta - \widetilde{r}_{\theta^*}\| \geq \max\{\Gamma, \kappa_\epsilon\}$ or $\|(\widetilde{p}_\theta - \widetilde{p}_{\theta^*})^T \widetilde{V}^*_{\theta^*}\| \geq \max\{\Gamma, \kappa_\epsilon\}$. By the transfer condition, we also have that $\kappa_\epsilon \geq \frac{(1-\gamma)\epsilon}{8}$. Then, it is easy to see that

$$\max_{s,a}\min_{\theta\in\Theta_\epsilon} \mathcal{I}_{s,a}(\theta^*,\theta) \geq \max\{\Gamma^2, \kappa_\epsilon^2\}(1-\gamma)^2 \geq \frac{1}{8}\max\{\Gamma^2, \epsilon^2\}(1-\gamma)^4,$$

where we use the previous lower bounds and upper bounded the value-function variance by $1/(1-\gamma)^2$. Then, the result follows by rewriting in $\widetilde{\mathcal{O}}$ notation.

□

C. Learning HMMs by Tensor Decomposition

After reducing our setting to a HMM learning problem, we can almost immediately plug the agent's observations into the tensor decomposition approach of Anandkumar et al. (2014) and obtain estimates $\widehat{O}$ and $\widehat{T}$ of the desired matrices. We now briefly describe how this method works as some of its features are needed for our analysis later. The detailed steps are reported in Algorithm 3. The key intuition behind the method of Anandkumar et al. (2014) is that the second and third moments of the HMM observations posses a low-rank tensor structure. More precisely, it is possible to find a transformation of these observations such that the resulting third moment is a symmetric and orthogonal tensor whose spectral decomposition directly yields (transformations of) the HMM parameters. To see this, we first formulate the problem as an instance of a multi-view model (Sun, 2013). Take three consecutive observations (our "views"), say o_1, o_2, o_3, and let $\Sigma_{i,j} := \mathbb{E}[o_i \otimes o_j]$, for $i,j \in \{1,2,3\}$, be their covariance matrices, where $\otimes$ denotes the tensor product. Define the transformed views as $\widetilde{o}_1 = \Sigma_{3,2}\Sigma_{1,2}^\dagger o_1$ and $\widetilde{o}_2 = \Sigma_{3,1}\Sigma_{2,1}^\dagger o_2$, and let the second and third cross-view moments be $M_2 = \mathbb{E}[\widetilde{o}_1 \otimes \widetilde{o}_2]$ and $M_3 = \mathbb{E}[\widetilde{o}_1 \otimes \widetilde{o}_2 \otimes o_3]$, respectively. Then, Theorem 3.6 of Anandkumar et al. (2014) shows that $M_2 = \sum_{j=1}^k \omega_j \mu_{3,j} \otimes \mu_{3,j}$ and $M_3 = \sum_{j=1}^k \omega_j \mu_{3,j} \otimes \mu_{3,j} \otimes \mu_{3,j}$, where $\mu_{3,j} = \mathbb{E}[o_3 | \theta_3^* = \theta_j]$ and $\omega_j = \omega(\theta_j)$. Hence, these moments posses a low-rank tensor structure, as they can be decomposed into tensor products of vectors, and it is

Algorithm 3 Learning HMMs by Tensor Decomposition

Require: Observations $\{o_l\}_{l=1}^h$ with $o_l \in \mathbb{R}^d$, number of tasks k

Ensure: Estimated observation matrix $\widehat{O}$ and transition matrix $\widehat{T}$

1: Split observations into $m = \lfloor h/3 \rfloor$ triples: $\{(o_{1,l}, o_{2,l}, o_{3,l})\}_{l=1}^m$

2: Estimate covariance matrices: $\widehat{\Sigma}_{i,j} = \frac{1}{m}\sum_{l=1}^m o_{i,l} \otimes o_{j,l}$, for $i, j \in \{1, 2, 3\}$

3: Get transformed observations: $\widetilde{o}_{1,l} = \widehat{\Sigma}_{3,2}\widehat{\Sigma}_{1,2}^{\dagger} o_{1,l}$, $\widetilde{o}_{2,l} = \widehat{\Sigma}_{3,1}\widehat{\Sigma}_{2,1}^{\dagger} o_{2,l}$, $l \in [m]$

4: Estimate 2nd and 3rd moments: $\widehat{M}_2 = \frac{1}{m}\sum_{l=1}^m \widetilde{o}_{1,l} \otimes \widetilde{o}_{2,l}$, $\widehat{M}_3 = \frac{1}{m}\sum_{l=1}^m \widetilde{o}_{1,l} \otimes \widetilde{o}_{2,l} \otimes o_{3,l}$

5: Find the k largest eigenvectors $\widehat{D} \in \mathbb{R}^{d\times k}$ and eigenvalues $\widehat{\Lambda} \in \mathbb{R}^{k\times k}$ of $\widehat{M}_2$

6: Compute the whitening matrix: $\widehat{W} = \widehat{D}\widehat{\Lambda}^{-\frac{1}{2}}$

7: Run Algorithm 1 of Anandkumar et al. (2014) on $\widehat{M}_3(\widehat{W}, \widehat{W}, \widehat{W})$, obtain estimated eigenvalues $\{\widehat{\lambda}_j\}_{j=1}^k$ and eigenvectors $\{\widehat{v}_j\}_{j=1}^k$

8: Estimate conditional means of the third view: $\widehat{\mu}_{3,j} = \widehat{\lambda}_j(\widehat{W}^T)^{\dagger}\widehat{v}_j$

9: Estimate mean observations: $[\widehat{O}]_{:,j} = \widehat{\Sigma}_{2,1}\widehat{\Sigma}_{3,1}^{\dagger}\widehat{\mu}_{3,j}$

10: Estimate transition matrix: $[\widehat{T}]_{:,j} = \widehat{O}^{\dagger}\widehat{\mu}_{3,j}$

possible to recover the conditional means $\mu_{3,j}$ using the robust tensor power (RTP) method of Anandkumar et al. (2014). Given these, from Proposition 4.2 of Anandkumar et al. (2012), we can recover the HMM parameters as follows:

$$[O]_{:,j} = \Sigma_{2,1}\Sigma_{3,1}^{\dagger}\mu_{3,j},$$
$$[T]_{:,j} = O^{\dagger}\mu_{3,j}.$$

In practice, all the required moments can be estimated from samples. Suppose the agent has observed $3m$ tasks and split these in m triples of contiguous tasks. Then, we can estimate the covariance matrices between views as $\widehat{\Sigma}_{i,j} = \frac{1}{m}\sum_{h=1}^m o_{i,h} \otimes o_{j,h}$, and similarly for the second and third cross moments of the transformed observations.

D. Analysis of the Sequential Transfer Algorithm

D.1. Definitions and Assumptions

In this section, we analyze the approximate models computed by the sequential transfer algorithm (Algorithm 2) using the RTP method (Algorithm 3). To simplify notation, we drop the task index h whenever clear from the context.

We introduce some additional vector/matrix notation. We use $\|\cdot\|_2$ to denote the spectral norm and $\|\cdot\|_F$ to denote the Frobenius norm. For a matrix A, we denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ its minimum and maximum singular value, respectively.

We make use of the following assumption in order to bound the estimation error of the reward and transition variances.

Assumption 5. *There exists a positive constant $\underline{\sigma} > 0$ which lower-bounds both the true reward and value variances, $\sigma_\theta^r(s, a)$ and $\sigma_\theta^p(s, a; \theta')$, and their estimates $\widetilde{\sigma}_{h,\theta}^r(s, a)$ and $\widetilde{\sigma}_{h,\theta}^p(s, a; \theta')$ at any h.*

We believe this assumption could be removed, at least for the approximate models, but at the cost of more complicated proofs.

D.2. Supporting Lemmas

Lemma 10 (Lemma 5 of Azizzadenesheli et al. (2016))**.** *Let $\{\widehat{\mu}_{3,j}\}_{j=1}^k$ be the columns of the third-view matrix estimated by the RTP method (Algorithm 3 in Appendix C) after observing h tasks. Then, there exists two constants $\rho_1(\Theta, T)$, $\rho_2(\Theta, T)$ such that, for any $\delta' \in (0, 1)$, if*

$$h > \rho_1(\Theta, T) \log \frac{2SA(S + U)}{\delta'},$$

then, under Assumption 2, with probability at least $1-\delta'$ and up to some permutation of the columns of the third view,

$$\|\widehat{\mu}_{3,j} - \mu_{3,j}\|_2 \le \rho_2(\Theta, T)\sqrt{\frac{\log\left(2SA(S+U)/\delta'\right)}{h}}.$$

In Lemma 10, compared to the original result, we collapsed all terms of minor relevance for our purpose into two constants ρ_1, ρ_2. These are functions of the given family of tasks (through maximum/minimum eigenvalues of the covariance matrices introduced before) and of the underlying Markov chain. We refer the reader to Appendix C of Azizzadenesheli et al. (2016) for their full expression.

We now bound the estimation error of the observation matrix O, which will be directly used to bound the errors of the approximate MDP models.

Lemma 11 (Estimation Error of O)**.** *Let $\widehat{O}$ be the observation matrix estimated by Algorithm 3 using h tasks, $\delta' \in (0,1)$, and*

$$\rho_3(\Theta, T) := \max\left\{\rho_1(\Theta, T), \frac{1}{\lambda_{\min}(\Sigma_{2,1})^2}, \frac{4}{\lambda_{\min}(\Sigma_{3,1})^2}\right\}.$$

Then, if $h \ge \rho_3(\Theta, T)\log(6SA(S+U)/\delta')$, we have that, with probability at least $1-\delta'$,

$$\left\|[O]_{:,j} - [\widehat{O}]_{:,j}\right\|_2 \le \rho_4(\Theta, T)\sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

where

$$\rho_4(\Theta, T) := \frac{4\lambda_{\max}(\Sigma_{2,1})\rho_2(\Theta, T) + 4SA + 8SA\lambda_{\max}(\Sigma_{2,1})}{\lambda_{\min}(\Sigma_{3,1})}.$$

Proof. Recall that $[O]_{:,j} = \Sigma_{2,1}\Sigma_{3,1}^{\dagger}\mu_{3,j}$ and similarly for its estimate $[\widehat{O}]_{:,j} = \widehat{\Sigma}_{2,1}\widehat{\Sigma}_{3,1}^{\dagger}\widehat{\mu}_{3,j}$. Let us decompose the total error into the deviations of each single component,

$$\begin{aligned}
\left\|[O]_{:,j} - [\widehat{O}]_{:,j}\right\|_2 &= \left\|\Sigma_{2,1}\Sigma_{3,1}^{\dagger}\mu_{3,j} - \widehat{\Sigma}_{2,1}\widehat{\Sigma}_{3,1}^{\dagger}\widehat{\mu}_{3,j} \pm \widehat{\Sigma}_{2,1}\widehat{\Sigma}_{3,1}^{\dagger}\mu_{3,j}\right\|_2 \\
&\le \left\|\widehat{\Sigma}_{2,1}\right\|_2\left\|\widehat{\Sigma}_{3,1}^{\dagger}\right\|_2\|\mu_{3,j} - \widehat{\mu}_{3,j}\|_2 + \left\|\Sigma_{2,1}\Sigma_{3,1}^{\dagger}\mu_{3,j} - \widehat{\Sigma}_{2,1}\widehat{\Sigma}_{3,1}^{\dagger}\mu_{3,j} \pm \widehat{\Sigma}_{2,1}\Sigma_{3,1}^{\dagger}\mu_{3,j}\right\|_2 \\
&\le \left\|\widehat{\Sigma}_{2,1}\right\|_2\left\|\widehat{\Sigma}_{3,1}^{\dagger}\right\|_2\|\mu_{3,j} - \widehat{\mu}_{3,j}\|_2 + \left\|\Sigma_{3,1}^{\dagger}\right\|_2\|\mu_{3,j}\|_2\left\|\Sigma_{2,1} - \widehat{\Sigma}_{2,1}\right\|_2 \\
&\qquad + \left\|\widehat{\Sigma}_{2,1}\right\|_2\|\mu_{3,j}\|_2\left\|\Sigma_{3,1}^{\dagger} - \widehat{\Sigma}_{3,1}^{\dagger}\right\|_2.
\end{aligned}$$

We now bound all these components separately. We first notice that, from Proposition 6 of Azizzadenesheli et al. (2016),

$$\left\|\Sigma_{2,1} - \widehat{\Sigma}_{2,1}\right\|_2 \le \sqrt{\frac{\log 1/\delta'}{h}}, \quad \left\|\Sigma_{3,1} - \widehat{\Sigma}_{3,1}\right\|_2 \le \sqrt{\frac{\log 1/\delta'}{h}}, \tag{8}$$

Using this result,

$$\begin{aligned}
\left\|\widehat{\Sigma}_{2,1}\right\|_2 &\le \|\Sigma_{2,1}\|_2 + \left\|\Sigma_{2,1} - \widehat{\Sigma}_{2,1}\right\|_2 \le \lambda_{\max}(\Sigma_{2,1}) + \left\|\Sigma_{2,1} - \widehat{\Sigma}_{2,1}\right\|_2 \\
&\le \lambda_{\max}(\Sigma_{2,1}) + \sqrt{\frac{\log 1/\delta'}{h}} \le 2\lambda_{\max}(\Sigma_{2,1}),
\end{aligned}$$

which holds for

$$h \ge \frac{\log 1/\delta'}{\lambda_{\max}(\Sigma_{2,1})^2}. \tag{9}$$

Using Lemma E.1 of Anandkumar et al. (2012),

$$\left\|\widehat{\Sigma}_{3,1}^{\dagger}\right\|_2 \le \frac{1}{\lambda_{\min}(\widehat{\Sigma}_{3,1})} \le \frac{2}{\lambda_{\min}(\Sigma_{3,1})},$$

which holds when

$$\frac{\|\Sigma_{3,1} - \widehat{\Sigma}_{3,1}\|}{\lambda_{\min}(\Sigma_{3,1})} \leq \frac{1}{2}.$$

From (8), a sufficient condition for this to hold is

$$h \geq \frac{4 \log 1/\delta'}{\lambda_{\min}(\Sigma_{3,1})^2}. \tag{10}$$

Under this same condition, we can apply Proposition 7 of Azizzadenesheli et al. (2016) to bound

$$\left\|\Sigma_{3,1}^{\dagger} - \widehat{\Sigma}_{3,1}^{\dagger}\right\|_2 \leq \frac{2}{\lambda_{\min}(\Sigma_{3,1})}\sqrt{\frac{\log 1/\delta'}{h}}.$$

Since the columns $\mu_{3,j}$ of the third-view sum up to one every S components (for the transition model) and every U components (for the reward model), we have $\|\mu_{3,j}\|_2 \leq \|\mu_{3,j}\|_1 \leq 2SA$. Finally, we can bound the error in estimating such columns using Lemma 10,

$$\|\widehat{\mu}_{3,j} - \mu_{3,j}\|_2 \leq \rho_2(\Theta, T)\sqrt{\frac{\log\left(2SA(S+U)/\delta'\right)}{h}}.$$

Plugging everything into the initial error decomposition and rearranging,

$$\left\|[O]_{:,j} - [\widehat{O}]_{:,j}\right\|_2 \leq \frac{4\lambda_{\max}(\Sigma_{2,1})\rho_2(\Theta,T) + 2SA + 4SA\lambda_{\max}(\Sigma_{2,1})}{\lambda_{\min}(\Sigma_{3,1})}\sqrt{\frac{\log\left(2SA(S+U)/\delta'\right)}{h}},$$

which holds for when h satisfies the conditions of Lemma 10, Equation 9, and Equation 10. The two bounds of (8) and that of Lemma 10 hold each with probability at least $1-\delta'$, hence the final bound holds with probability at least $1-3\delta'$. Renaming δ' into $3\delta'$ concludes the proof. □

Similarly to O, we also bound the estimation error of T. The proof follows Theorem 3 of Azizzadenesheli et al. (2016).

Lemma 12 (Estimation error of T)**.** *Let $\widehat{T}$ be the task-transition matrix estimated by Algorithm 3 using h samples and*

$$\rho_5(\Theta, T) := \max\left\{\rho_3(\Theta, T), \frac{4k\rho_4(\Theta,T)^2}{\lambda_k(O)^2}\right\}.$$

Then, for any $\delta' \in (0,1)$, if $h \geq \rho_5(\Theta,T)\log(6SA(S+U)/\delta')$, we have that, with probability at least $1-\delta'$,

$$\left\|[T]_{:,j} - [\widehat{T}]_{:,j}\right\|_2 \leq \rho_6(\Theta,T)\sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

where

$$\rho_6(\Theta,T) := 8SA\sqrt{k}\frac{1+\sqrt{5}}{\lambda_k(O)^2}\rho_4(\Theta,T).$$

Proof. Recall that $[T]_{:,j} = O^{\dagger}\mu_{3,j}$ and $[\widehat{T}]_{:,j} = \widehat{O}^{\dagger}\widehat{\mu}_{3,j}$. Similarly to Lemma 11, we decompose the error into

$$\left\|[T]_{:,j} - [\widehat{T}]_{:,j}\right\|_2 = \left\|O^{\dagger}\mu_{3,j} - \widehat{O}^{\dagger}\widehat{\mu}_{3,j}\right\|_2 \leq \|\mu_{3,j}\|_2\left\|O^{\dagger} - \widehat{O}^{\dagger}\right\|_2 + \left\|\widehat{O}^{\dagger}\right\|_2\|\mu_{3,j} - \widehat{\mu}_{3,j}\|_2.$$

In the proof of Lemma 11 we already bounded $\|\mu_{3,j}\|_2 \leq 2SA$, while the term $\|\mu_{3,j} - \widehat{\mu}_{3,j}\|_2$ was bounded in Lemma 10. Let us bound the remaining two terms. Take $\lambda_k(\widehat{O})$ as the k-th singular value of $\widehat{O}$. Then, following Azizzadenesheli et al. (2016),

$$\left\|\widehat{O}^{\dagger}\right\|_2 \leq \frac{1}{\sigma_k(\widehat{O})} \leq \frac{2}{\sigma_k(O)},$$

where the second inequality follows from Lemma E.1 of Anandkumar et al. (2012) under the assumption that

$$\frac{\|O - \widehat{O}\|_2}{\sigma_k(O)} \leq \frac{1}{2}. \tag{11}$$

Lemma 11 already bounds the l_2 error in the columns of O. Therefore,

$$\|O - \widehat{O}\|_2 \leq \|O - \widehat{O}\|_F \leq \sqrt{k} \max_{j \in [k]} \|O_j - \widehat{O}_j\|_2 \leq \rho_4(\Theta, T) \sqrt{\frac{k \log(6SA(S+U)/\delta')}{h}}.$$

Therefore, a sufficient condition for (11) is

$$m \geq \frac{4k\rho_4(\Theta, T)^2 \log(6SA(S+U)/\delta')}{\lambda_k(O)^2}.$$

In order to bound the deviation of the pseudo-inverse of O, we apply Theorem 1.1 of Meng & Zheng (2010),

$$\left\|O^\dagger - \widehat{O}^\dagger\right\|_2 \leq \frac{1+\sqrt{5}}{2} \max\left\{\left\|O^\dagger\right\|_2^2, \left\|\widehat{O}^\dagger\right\|_2^2\right\} \|O - \widehat{O}\|_2 \leq \frac{2+2\sqrt{5}}{\lambda_k(O)^2} \rho_4(\Theta, T) \sqrt{\frac{k \log(6SA(S+U)/\delta')}{h}}.$$

Finally, plugging everything back into the first error decomposition,

$$\left\|[T]_{:,j} - [\widehat{T}]_{:,j}\right\|_2 \leq 8SA \frac{1+\sqrt{5}}{\lambda_k(O)^2} \rho_4(\Theta, T) \sqrt{\frac{k \log(6SA(S+U)/\delta')}{h}}.$$

□

We also need the following technical lemma which bounds the difference in standard deviations between random variables under different distributions.

Lemma 13. *Let f be a function which takes values in $[0, b]$, for some $b > 0$, and $p, q \in \mathcal{X}$ two probability distributions on a finite set $\mathcal{X}$. Denote by $\mathbb{V}\mathrm{ar}_p[f]$ and $\mathbb{V}\mathrm{ar}_q[f]$ the variance of f under p and q, respectively, and assume both to be larger than some constant $c > 0$. Then,*

$$\left|\sqrt{\mathbb{V}\mathrm{ar}_p[f]} - \sqrt{\mathbb{V}\mathrm{ar}_q[f]}\right| \leq b\left(\frac{b}{2\sqrt{c}} + 1\right) \|p - q\|_1,$$

Proof. Let $\mu_p := \mathbb{E}_p[f]$ and $\mu_q := \mathbb{E}_q[f]$. For clarity, rewrite the standard deviations as

$$\sqrt{\mathbb{V}\mathrm{ar}_p[f]} = \sqrt{\sum_{x \in \mathcal{X}} p(x)(f(x) - \mu_p)^2} = \|f - \mu_p\|_{2,p},$$

and similarly for q. Here $\|\cdot\|_{2,p}$ denotes the l_2-norm weighted by p. Then,

$$|\|f - \mu_p\|_{2,p} - \|f - \mu_q\|_{2,q}| \leq |\|f - \mu_p\|_{2,p} - \|f - \mu_p\|_{2,q}| + |\|f - \mu_p\|_{2,q} - \|f - \mu_q\|_{2,q}|. \tag{12}$$

Let us bound these two terms separately. For the second one, a direct application of Minskowsky's inequality yields,

$$\|f - \mu_p\|_{2,q} \leq \|\mu_p - \mu_q\|_{2,q} + \|f - \mu_q\|_{2,q} = |\mu_p - \mu_q| + \|f - \mu_q\|_{2,q}.$$

Therefore, applying the same reasoning to the other side, we obtain

$$|\|f - \mu_p\|_{2,q} - \|f - \mu_q\|_{2,q}| \leq |\mu_p - \mu_q| \leq b\|p - q\|_1.$$

We now take care of the first term. Since we have a term of the form $|\sqrt{x} - \sqrt{y}|$ and the concavity of the square root implies $|\sqrt{x} - \sqrt{y}| \leq \frac{1}{2}\max\{\frac{1}{\sqrt{x}}, \frac{1}{\sqrt{y}}\}|x - y|$, we can reduce the problem to bounding the difference of variances,

$$\left|\|f - \mu_p\|_{2,p}^2 - \|f - \mu_p\|_{2,q}^2\right| \leq b^2\|p - q\|_1,$$

where we have used the fact that the term $(f(x) - \mu_p)^2$ is bounded by b^2. By assumption, $\|f - \mu_p\|_{2,p} \geq \sqrt{c}$ and $\|f - \mu_p\|_{2,q} \geq \|f - \mu_q\|_{2,q} \geq \sqrt{c}$. Therefore, plugging these two bounds back into (12),

$$|\|f - \mu_p\|_{2,p} - \|f - \mu_q\|_{2,q}| \leq b\left(\frac{b}{2\sqrt{c}} + 1\right) \|p - q\|_1,$$

which concludes the proof. □

D.3. Main Results

We are now ready to bound the estimation error of the different MDP components. The following Lemma does exactly this for a fixed number of tasks. Theorem 2 extends this to the sequential setting.

Lemma 14. *Let $\widetilde{p}_\theta(s'|s,a)$ and $\widetilde{q}_\theta(u|s,a)$ be the transition and reward distributions estimated by Algorithm 3 after h tasks, $\widetilde{V}^*_\theta$ be the optimal value functions of these models, and $\widetilde{\sigma}^r_\theta(s,a)$, $\widetilde{\sigma}^p_\theta(s,a;\theta')$ be the corresponding variances. Assume that the latter are both bounded below by some positive constant $\bar{\sigma} > 0$ for all s, a, θ, θ'. Then, with probability at least $1-\delta'$, the following hold simultaneously:*

$$|r_\theta(s,a) - \widetilde{r}_\theta(s,a)| \leq \rho_4(\Theta, T) U \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

$$|(p_\theta(s,a) - \widetilde{p}_\theta(s,a))^T \widetilde{V}^*_{\theta'}| \leq \frac{\rho_4(\Theta,T)S}{1-\gamma} \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

$$|\sigma^r_\theta(s,a) - \widetilde{\sigma}_\theta(s,a)| \leq \left(\frac{1}{2\bar{\sigma}} + 1\right) \rho_4(\Theta, T) U \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

$$|\sigma^r_\theta(s,a;\widetilde{V}^*_{\theta'}) - \widetilde{\sigma}_\theta(s,a;\widetilde{V}^*_{\theta'})| \leq \left(\frac{1}{2\bar{\sigma}(1-\gamma)} + 1\right) \frac{\rho_4(\Theta,T)S}{1-\gamma} \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}}.$$

Proof. Recall that the estimated transition and reward probabilities are extracted from the columns of the observation matrix $\widehat{O}$. Therefore, we can directly use Lemma 11 to bound their error. For each state s, action a, next state s', and reward u, we have

$$|q_\theta(u|s,a) - \widetilde{q}_\theta(u|s,a)| \leq \rho_4(\Theta, T) \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

$$|p_\theta(s'|s,a) - \widetilde{p}_\theta(s|s,a)| \leq \rho_4(\Theta, T) \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}}.$$

These inequalities hold simultaneously with probability at least $1-\delta'$. Therefore, since rewards are bounded in $[0,1]$,

$$|r_\theta(s,a) - \widetilde{r}_\theta(s,a)| \leq \|q_\theta(\cdot|s,a) - \widetilde{q}_\theta(\cdot|s,a)\|_1 \leq \rho_4(\Theta, T) U \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}}.$$

Similarly, for any function taking values in $[0, 1/(1-\gamma)]$,

$$|(p_\theta(s,a) - \widetilde{p}_\theta(s,a))^T V| \leq \frac{1}{1-\gamma} \|p_\theta(\cdot|s,a) - \widetilde{p}_\theta(\cdot|s,a)\|_1 \leq \frac{\rho_4(\Theta,T)S}{1-\gamma} \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}}.$$

Finally, a direct application of Lemma 13 yields the desired bounds on the variances,

$$|\sigma^r_\theta(s,a) - \widetilde{\sigma}^r_\theta(s,a)| \leq \left(\frac{1}{2\bar{\sigma}} + 1\right) \rho_4(\Theta, T) U \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}},$$

$$|\sigma^r_\theta(s,a;\theta') - \widetilde{\sigma}^p_\theta(s,a;\theta')| \leq \left(\frac{1}{2\bar{\sigma}(1-\gamma)} + 1\right) \frac{\rho_4(\Theta,T)S}{1-\gamma} \sqrt{\frac{\log\left(6SA(S+U)/\delta'\right)}{h}}.$$

□

Theorem 2. *Let $\{\widetilde{\mathcal{M}}^h_\theta\}_{\theta\in\Theta, h\geq 1}$ be the sequence of MDP models estimated by Algorithm 2, with $\widetilde{p}_{h,\theta}(s'|s,a)$ and $\widetilde{q}_{h,\theta}(u|s,a)$ the transition and reward distributions, $\widetilde{V}^*_{h,\theta}$ the optimal value functions, and $\widetilde{\sigma}^r_{h,\theta}(s,a)$, $\widetilde{\sigma}^p_{h,\theta}(s,a;\theta')$ the corresponding variances. There exist constants $\rho, \rho_r, \rho_p, \rho_{\sigma_r}, \rho_{\sigma_p}$ such that, for $\delta' \in (0,1)$, if*

$$h > \rho \log \frac{2h^2 SA(S+U)}{\delta'},$$

then, with probability at least $1-\delta'$, the following hold simultaneously for all $h \geq 1$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $\theta, \theta' \in \Theta$:

$$|r_\theta(s,a) - \widetilde{r}_{h,\theta}(s,a)| \leq \rho_r \sqrt{\frac{\log c_{h,\delta'}}{h}},$$
$$|(p_\theta(s,a) - \widetilde{p}_{h,\theta}(s,a))^T \widetilde{V}^*_{h,\theta'}| \leq \rho_p \sqrt{\frac{\log c_{h,\delta'}}{h}},$$
$$|\sigma^r_\theta(s,a) - \widetilde{\sigma}^r_{h,\theta}(s,a)| \leq \rho_{\sigma_r} \sqrt{\frac{\log c_{h,\delta'}}{h}},$$
$$|\sigma^r_\theta(s,a;\theta') - \widetilde{\sigma}^p_{h,\theta}(s,a;\theta')| \leq \rho_{\sigma_p} \sqrt{\frac{\log c_{h,\delta'}}{h}},$$

where $c_{h,\delta'} := \pi^2 h^2 SA(S+U)/\delta'$.

Proof. Let E_h be the event under which the bounds of Lemma 14 all hold after observing h tasks. We need to prove that, with high probability, there is no h in which the event does not hold. We know that E_h hold with probability at least $1-\delta''$ from Lemma 14, where $\delta'' = \frac{6\delta'}{\pi^2 h^2}$. Too see this, notice that we introduced an extra $\pi^2 h^2/6$ term in the confidence level $c_{h,\delta'}$. Then,

$$\mathbb{P}\left\{\exists h \geq 1 : E_h = 0\right\} \leq \sum_{h=1}^{\infty} \mathbb{P}\left\{E_m = 0\right\} \leq \frac{6\delta'}{\pi^2} \sum_{h=1}^{\infty} \frac{1}{h^2} = \delta',$$

where the first inequality is the union bound, the second is from Lemma 14, and the last equality is from the value of the p-series for $p = 2$. Then, the main theorem follows after renaming the constants in Lemma 14. $\square$

Lemma 15. *Let $\bar{\Theta}_h$ be a set that contains θ^*_h with probability at least $1-\delta$. Then, for any $\delta' \in (0,1)$ and $\theta \in \Theta$, with probability at least $1-\delta'$,*

$$\mathbb{P}\left\{\theta^*_{h+1} = \theta\right\} \leq \sum_{\theta' \in \bar{\Theta}_h} \widehat{T}(\theta,\theta') + \delta k + \rho_T(\Theta, T) k \sqrt{\frac{\log 6SA(S+U)/\delta'}{h}}. \tag{13}$$

Proof. We start by bounding the probability that the next task is θ as

$$\begin{aligned}\mathbb{P}\left\{\theta^*_{h+1} = \theta\right\} &= \sum_{\theta' \in \Theta} \mathbb{P}\left\{\theta^*_{h+1} = \theta | \theta^*_h = \theta'\right\} \mathbb{P}\left\{\theta^*_h = \theta'\right\} \leq \sum_{\theta' \in \bar{\Theta}_h} \mathbb{P}\left\{\theta^*_{h+1} = \theta | \theta^*_h = \theta'\right\} \mathbb{P}\left\{\theta^*_h = \theta'\right\} + \delta k \\ &\leq \sum_{\theta' \in \bar{\Theta}_h} \mathbb{P}\left\{\theta^*_{h+1} = \theta | \theta^*_h = \theta'\right\} + \delta k = \sum_{\theta' \in \bar{\Theta}_h} T(\theta,\theta') + \delta k,\end{aligned}$$

where the first inequality follows from the condition on $\bar{\Theta}_h$, the second by bounding the probability by one, and the last equality by definition of T. The result follows by applying Lemma 12 and renaming the constants. $\square$

Theorem 3. *Let $\delta' \in (0,1)$ and $\delta \leq \frac{\delta'}{3m^2}$ be the confidence value for Algorithm 1. Suppose that, before each task h, a model θ is eliminated from the initial active set if:*

$$\sum_{\theta' \in \bar{\Theta}_h} \widehat{T}_h(\theta,\theta') + \delta k + \rho_T k \sqrt{\frac{\log(9kdm^2/\delta')}{h}} \leq \eta.$$

Then, for $\eta = \frac{\delta'}{3km^2}$, with probability at least $1-\delta'$, at any time the true model is never eliminated from the initial set.

Proof. We have

$$\mathbb{P}\left\{\theta^*_h \notin \widetilde{\Theta}_h\right\} \leq \mathbb{P}\left\{\theta^*_h \notin \widetilde{\Theta}_h \wedge \theta^*_{h-1} \in \widetilde{\Theta}_{h-1}\right\} + \mathbb{P}\left\{\theta^*_{h-1} \notin \widetilde{\Theta}_{h-1}\right\} \leq \sum_{l=2}^{h} \mathbb{P}\left\{\theta^*_l \notin \widetilde{\Theta}_l \wedge \theta^*_{l-1} \in \widetilde{\Theta}_{l-1}\right\},$$

where we applied the first inequality recursively to obtain the second one. We can bound each term in the second sum by

$$\begin{aligned}\mathbb{P}\left\{\theta_l^* \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1}\right\} &\leq \mathbb{P}\left\{\theta_l^* \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1} \wedge \theta_{l-1}^* \in \bar{\Theta}_{l-1}\right\} + \mathbb{P}\left\{\theta_{l-1}^* \notin \bar{\Theta}_{l-1}\right\} \\ &\leq \mathbb{P}\left\{\theta_l^* \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1} \wedge \theta_{l-1}^* \in \bar{\Theta}_{l-1}\right\} + \delta,\end{aligned}$$

where $\bar{\Theta}_{l-1}$ is the set of active models after the application of Algorithm 1 to $\widetilde{\Theta}_{l-1}$ and the second inequality follows from the fact this algorithm discards θ_{l-1}^* with probability at most $1-\delta$. In order to bound the first term,

$$\mathbb{P}\left\{\theta_l^* \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1} \wedge \theta_{l-1}^* \in \bar{\Theta}_{l-1}\right\} \leq \sum_{\theta \in \Theta} \mathbb{P}\left\{\theta_l^* = \theta \wedge \theta \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1} \wedge \theta_{l-1}^* \in \bar{\Theta}_{l-1}\right\}.$$

Let us split this probability based on whether the concentration event on T of Lemma 12 holds or not. Using $\frac{\delta'}{3km^2}$ as confidence value, this probability is trivially bounded by $k\frac{\delta'}{3km^2}$ when the event of Lemma 12 does not hold. Assume this event holds. Then, Lemma 15 implies that the probability of the next task being θ is bounded by $\frac{\delta'}{3km^2}$, and the overall sum is bounded by $k\frac{\delta'}{3km^2}$. Notice that we divided δ' by $3km^2$ w.r.t. the value used in Lemma 12 and Lemma 15. Putting these together,

$$\mathbb{P}\left\{\theta_l^* \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1}\right\} \leq \delta + \frac{2\delta'}{3m^2}.$$

Using the union bound,

$$\mathbb{P}\left\{\exists h \in [m] : \theta_h^* \notin \widetilde{\Theta}_h\right\} \leq \sum_{h=2}^{m} \mathbb{P}\left\{\theta_h^* \notin \widetilde{\Theta}_h\right\} \leq \sum_{h=2}^{m}\sum_{l=2}^{h} \mathbb{P}\left\{\theta_l^* \notin \widetilde{\Theta}_l \wedge \theta_{l-1}^* \in \widetilde{\Theta}_{l-1}\right\} \leq \left(\delta + \frac{2\delta'}{3m^2}\right) m^2.$$

Therefore, the result holds with probability at least $1-\delta'$ by taking $\delta \leq \frac{\delta'}{3m^2}$. $\square$

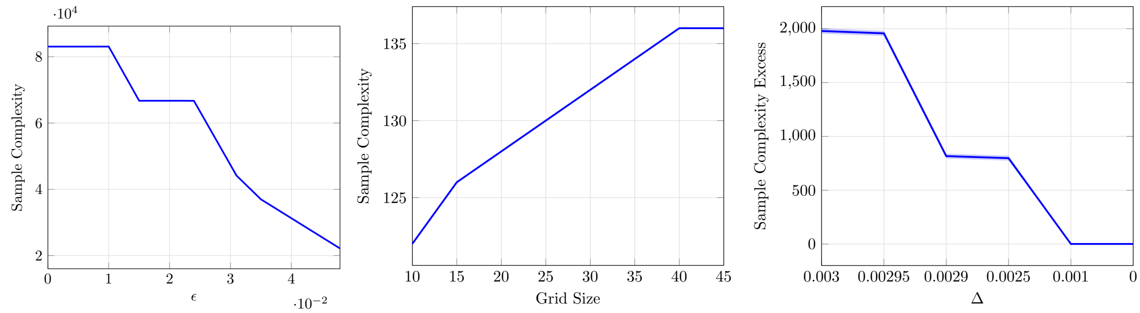

Figure 3. Ablation studies for the main parameters of PTUM. (left) The sample complexity is a piece-wise constant and bounded function of ϵ. (middle) The sample complexity grows logarithmically with the grid size (i.e., the number of states). (right) The excess in sample complexity (w.r.t. the one with perfect models) decreases with the approximation error Δ.

E. Additional Details on the Experiments

In this section, we provide all the necessary details to allow reproducibility of our experiments. We also provide further results and ablation studies to better understand how the proposed algorithms behave in different settings and with different parameters.

E.1. Experiments with PTUM

For these experiments we used a variant of the 4-rooms domain of Sutton et al. (1999). Here the agent navigates a grid with two rooms divided by a vertical wall in the middle and connected by a door. There are four actions (up, down, right, left) and there is a probability of 0.1 that the action chosen by the agent fails and a random one is taken instead. The agent always starts in the bottom-left corner and must reach certain goal states. The reward function (e.g., the goal locations) and the size of the grid vary in each experiment and will be specified later.

Experiments of Figure 1 *left* The agent acts in a 12x12 grid and it receives reward of 1 when reaching a goal state, and 0 otherwise. There are twelve possible tasks with goals and doors in different positions and whose models are fully known to the agent. The parameters for running PTUM are: $\gamma = 0.99$, $\delta = 0.01$, $\epsilon = 0.1$, and $n = 100000$. Both Rmax and MaxQInit switch state-action pairs from unknown to known after 10 samples. Online algorithms without a generative model are run for 100 episodes of 100 steps each. Since goal states are modelled as absorbing states with reward 1, these algorithms are able to retrieve significant information in each episode to make the comparison with PTUM (which uses a generative model) fair. In fact, it suffices to reach a goal to get information about it. For this reason, each sample retrieved by PTUM from the generative model is considered as a single step in an episode.

Experiments of Figure 1 *right* Here the agent acts in a 12x12 grid without any wall, and has to discriminate among 7 possible tasks. Each task has 7 goals whose position is shared. However, given a goal state, reward values can differ between tasks. To be more precise, goals are placed in the corners or nearby them. The true task has the optimal goal giving a reward of 0.8, while all the other tasks have a different best-goal with a reward of 0.81. We set $\epsilon = 1$, $\gamma = 0.9999$, $\delta = 0.01$, and $n = 100000$. Both RMax and MaxQInit switch state-action pairs from unknown to known after 240 samples. The online algorithms without a generative model are run for 1000 episodes of 100 steps each. Due to the fact that goal states are modeled as absorbing state with reward 0 and, since all tasks have goals in the same position, online algorithms are able to get at most one informative sample in each episode (i.e., one with non-zero reward). For this reason, each sample retrieved by PTUM from the generative model is reported as an entire episode in the plots, which makes comparison fair.

It is important to note that in both these experiments the parameter needed by Rmax to switch between unknown and known state-action pairs (which directly affects the learning speed) is set way below the one recommended by the theory (see Brafman & Tennenholtz (2002)).

Additional results In order to verify the main theoretical results from the sample-complexity analysis of PTUM, we report ablation studies for its main parameters.

Sample complexity vs accuracy The setting in this case is a grid of dimension 12x12 where a reward of 1 is given when reaching a goal state and 0 otherwise. The agent has knowledge of 12 possible perfect models, each of which has a single goal state located in the opposite corner from the starting one. The door position is different in each task. We set γ to 0.99, δ to 0.01, n to 1000000, and test our algorithm for different values of ϵ.

Figure 3(*left*) shows how the sample complexity changes as a function of ϵ. First, we notice that the function is piece-wise constant, which is a direct consequence of the finite set of models. In fact, varying ϵ changes the set of models Θ_ϵ that must be discarded by the algorithm before stopping. Second, we notice that the function is bounded in ϵ, i.e., we are allowed to set $\epsilon = 0$ and the algorithm returns an optimal policy after discarding all the models different from the true one.

Sample complexity vs number of states In this experiment, the agent faces grids of increasing sizes. The agent obtains reward 1 when ending up in a goal state and 0 otherwise. We consider three known models, each with the goal state in a corner different from the starting one and a different door position. We set ϵ to 0.0001, γ to 0.99, δ to 0.01, and n to 100000.

Figure 3(*middle*) shows how the sample complexity changes when the grid size (i.e., the number of states) increases. As expected, we obtain a logarithmic growth due to the union bounds used to form the confidence sets. As before, the function is piece-wise due to the finite number of models.

Sample complexity vs model error A grid of fixed size 6x6 is considered. The agent has knowledge of 6 models, each of which has the goal state placed in the opposite corner w.r.t. the starting one. All models have reward 1 when reaching the goal state and 0 otherwise. Each task differs from the others in the door position. We set ϵ to 0.13, γ to 0.9, δ to 0.1, and n to 1000000. Here we study the sample complexity of PTUM with the exact set of models when varying the maximum uncertainty Δ.

Figure 3(*right*) shows the excess in sample complexity w.r.t. the case with perfect models when the bound Δ on the approximation error decreases. Once again, we obtain a piece-wise constant function since, similarly to ϵ, a higher error bound Δ changes the set of models that must be discarded by the algorithm and hence its sample complexity. Notably, we do not require $\Delta = 0$ to recover the "oracle" sample complexity.

E.2. Sequential Transfer Experiments

For these experiments we consider a grid-world similar to the objectworld domain of Levine et al. (2011). Here we have an agent navigating a 5×5 grid where each cell is either empty or contains one item (among a set of possible items). Each item is associated to a different reward value and can be picked up by the agent when performing a specific action (among up, down, left, right) in the state containing the item. To keep the problem simple and Markovian, we suppose items immediately re-spawn after being picked up. That is, the agent can pick up the same item indefinitely as far as it manages to return to the state containing it. The transition dynamics include a certain probability of failing the action as in the previous experiments. To be more precise, in this setting failing actions regards exclusively the transition to other states, while the intended item is picked up with another fixed probability. The agent sequentially faces 8 different tasks as follows. Suppose the tasks are ordered in a list. Then, given the current task, with high probability the next task to be faced is the successor in the list, while there is a small probability of skipping one task and going two steps ahead or staying in the same task. The values of these probabilities (rewards, transitions, and task-transitions) will be specified in each experiment. For what concerns the estimation of the models, RTP is run with 100 restarts for 100 iterations. We use low-rank singular value decomposition to pre-process the empirical moments estimated by Algorithm 3 as explained by Anandkumar et al. (2014). This, empirically, seems to be somehow critical for the stability of the algorithm and for its computational efficiency. Regarding this last point, we note that after SVD is applied, Algorithm 3 only needs to decompose a tensor of size $k \times k \times k$, where k is the number of models ($k = 8$ in this case).

Experiments of Figure 2 As mentioned in the main paper, in this experiment only the reward changes between tasks, while the transition dynamics are fixed and have a failure probability of 0.1. The failure probability of rewards is instead fixed to 0.012 between tasks. The possible objects that the agent can get have the following rewards: $\{0, 0.02, 0.04, 0.2, 0.22, 0.24, 0.5, 0.52, 0.54, 0.96, 0.98, 1\}$. Tasks are generated in the following way: first, task with indexes in $\{0, 2, 3, 5\}$ are randomly generated using only objects with values $\{0, 0.2, 0.5, 0.96\}$. Then tasks with index 1 and 7 are generated starting from model 0, in order for them to be ϵ-optimal with it. The same is done for task with index 4 and 5 w.r.t. task with index 5. In particular, these modifications are carried out randomly, using objects with a slightly higher reward w.h.p. in each state-action pair. A small probability of using objects with smaller rewards is also present. We

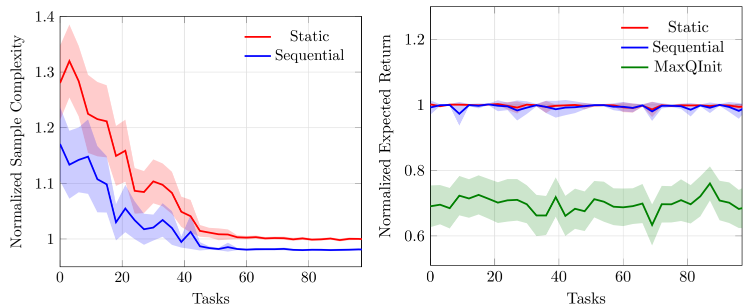

Figure 4. Sequential transfer experiment when both rewards and transition probabilities change across tasks. (left) The sample complexity normalized by the one of PTUM with known models. (right) The expected return normalized by the optimal one for each specific task.

used this task-generation process in order to guarantee that there exist tasks which are hard to distinguish by PTUM, which makes the problem more challenging. If we simply generated tasks randomly as described before (without creating any "similar" copies), the identification problem would become almost trivial even in the presence of estimated models.

The task-transitions succeed to the natural next task in the list with probability 0.97, and fail in favor of the two adjacent tasks with probability 0.015 each. A no-transfer uniform-sampling strategy is blindly run for 300 tasks, querying 50 samples per each state-action pair. This is done to make sure that the estimated models eventually become accurate enough to make PTUM enter the transfer mode. Once this 300 tasks are over, the transfer begins and goes on for 100 tasks. Once the model has been identified, a post-sample of 30 queries is run in each state-action pair in order to obtain a minimum accuracy level in the empirical models. When the identification of PTUM fails (i.e., the algorithm exceeds the budget n), no-transfer uniform sampling is run, and its complexity is reported in the plots.

For the computation of the model inaccuracy bounds (as prescribed by Theorem 2) we set $\rho = 0.135$ for all models. We chose this value so that the inaccuracies after the start-up phase are small enough to make the algorithm enter the transfer mode. Since we noticed that the estimated models become accurate rather quickly, we further decided to decay ρ for the first 100 tasks when PTUM enters the transfer mode from its initial value to 0.006. This allows us to plot a faster and more clear transitory in the sample complexity, but in principle this step could be ignored as the model inaccuracies naturally decay by Theorem 2. For what concerns the computation of the initial sets of models for PTUM from the estimated task-transition matrix $\widehat{T}$, we use the technique described in Theorem 3 with $\eta = 0.087$ and $\rho_T = 0.001$. These values were chosen so that the algorithm is likely not to discard models whose predicted probability is above 0.005. To add further robustness, we always keep the top-3 most probable next models (according to $\widehat{T}$), even if they would be eliminated by the previous condition.

At each step, the agent is required to find a ϵ-optimal policy with $\epsilon = 0.5$ and $\delta = 0.01$. The discount factor γ is set to 0.9. The expected returns of Figure 2 *right* are estimated by running the obtained policies for 30 episodes of 10 steps each and averaging the results. Finally, the update constant for MaxQInit is fixed to 100.

Changing both rewards and transition probabilities Since in our original domain only the positions of the items (i.e., the rewards) change across tasks, here we consider a variant where the transition probabilities change as well. We now generate tasks as follows. The failure probability of the actions changes as well (ranging from 0.1 up to 0.45), and we randomize the rewards as explained before. Again, tasks are generated to be ϵ-optimal as described above, and all the other parameters are set in the same way, except for the uniform sampling costant: no-transfer iterations query the generative model for 150 samples in each state-action pair, while the post-sampling (after identification) takes 100 samples.

The results, shown in Figure 4, are coherent with the ones of Figure 2 for the case where only rewards change. This is an interesting result since the number of parameters that must be estimated by the spectral learning algorithm doubled.

Increasing the MDP stochasticity In the experiments reported above, both the rewards and the transition probabilities are almost deterministic. We now repeat these experiments by increasing the stochasticity in both components. For what concerns the rewards, the possible objects are $\{0, 0.02, 0.2, 0.22, 0.5, 0.52, 0.96, 1\}$, and they are all used when generating the tasks. Moreover, the fact of generating ϵ-optimal tasks is dropped here. The failure probability of the reward is set to 0.3. The transition failure probability is instead the same of the previous experiment, ranging from 0.1 to 0.45, depending on the

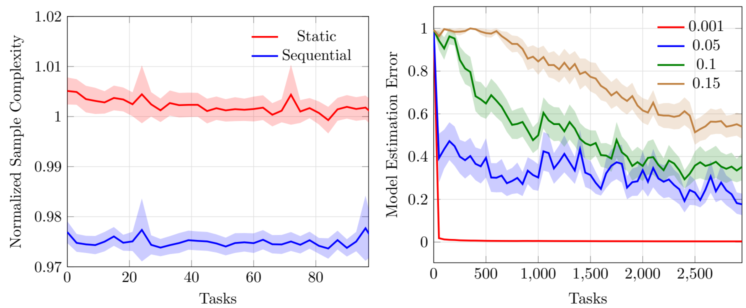

Figure 5. (left) Sequential transfer experiment with higher reward and transition stochasticity. (right) Sequential transfer experiment with higher stochasticy in the transition between tasks. Each curve corresponds to a different value of failure probability in the task-transition matrix.

task. Due to the higher stochasticity, no-transfer uniform sampling requires 250 samples per each state-action pair, while transfer post-sampling is set to 200. All the other parameters are kept the same.

Figure 5(*left*) shows the results. Once again, we have the sequential transfer algorithm outperforming its static counterpart. Compared to the previous experiments, here we notice only a small improvement of the sample complexity as a function of the number of tasks (i.e., rather flat curves). This is mostly due to the fact that now it is harder to have tasks that are very close to each other. This implies that the set of models that need to be discarded by PTUM remains roughly the same after the algorithm starts entering the transfer mode, and thus improvements in the model accuracy only lead to small improvements in the sample complexity.

Increasing the task-transition stochasticity In the previous experiments the transitions between tasks are almost deterministic. The stochasticity in these components seems to be the most critical parameter for what concerns the number of tasks needed by RTP to guarantee accurate estimations. For this purpose, we now show the l_∞ error of the estimated MDP models (with respect to the true ones) for different levels of stochasticity in the task-transition probabilities. Here we fix 10 different tasks, generated by randomizing the items as explained before, and run the RTP algorithm sequentially to obtain the error estimates.

Figure 5(*right*) shows the results. Here we clipped the maximum error to 1 for better visualization. This is anyway reasonable since the algorithm estimates probabilities and we could alternatively normalize the estimates. We note that the estimation error decays as expected for all values of stochasticity. The algorithm requires, however, many more tasks to get accurate estimates when the stochasticity increases.