9. Further Details for Section 3

9.1. Proof Sketch of Theorem 1

The proof parallels the proof of Proposition 1 by Strehl et al. (2006) for MDPs, except the horizon (denoted by \( T \) in their paper) needs to be redefined:

\[
H = \frac{1}{1-\gamma} \ln \left( \frac{4}{\epsilon (1-\gamma)} \right) + L + \frac{1}{\sqrt{C}} \ln \frac{2}{\delta}.
\]

This choice of \( H \) ensures that, for any epoch \( t \), the non-stationary policy \( \mathbf{A}_t \) in state \( s_t \) is either \( \Theta(\epsilon) \)-optimal, or will reach an unknown state in \( H \) steps with probability at least \( \epsilon (1-\gamma) \). In either case, the algorithm will reach a next state between step \( \frac{1}{1-\gamma} \ln \left( \frac{4}{\epsilon (1-\gamma)} \right) \) and \( H \), since with probability at least \( 1-\delta' \), the waiting time of taking action \( a_t \) in state \( s_t \) is \( L + \frac{2}{\sqrt{C}} \ln \frac{1}{\delta} \) (Lemma 1). Taking a union bound over all possible non-\( \epsilon \)-optimal steps (which is polynomial in \( \zeta, 1/\epsilon, 1/\delta, \) and \( 1/(1-\gamma) \)), that is, setting \( \delta' \) to \( \delta/\text{poly}(\zeta, 1/\epsilon, 1/\delta, 1/(1-\gamma)) \), we can prove the theorem as done in Strehl et al. (2006). Note that we need not take a union over all epochs, but only those where the decision is potentially non-\( \epsilon \)-optimal; if \( \mathbf{A}_t \) is \( \epsilon \)-optimal in epoch \( t \), it does not count towards the sample complexity anyway.

9.2. Definition of Known-state SMDP

**Definition 3** Let \( M = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \) be an SMDP, \( Q \) is a state–action value function, and \( \mathcal{K} \subseteq \mathcal{S} \times \mathcal{A} \) a set of “known” state–actions. Define the known state–action SMDP (with respect to \( \mathcal{K} \)) as \( M_\mathcal{K} = (\mathcal{S}, \mathcal{A}, P_\mathcal{K}, R_\mathcal{K}, \gamma) \), where

\[
P_\mathcal{K}(s', \tau | s, a) = \begin{cases} 
P(s', \tau | s, a), & \text{if } (s, a) \in \mathcal{K} \\ 
P(s = s', \tau = 1), & \text{otherwise} \end{cases}
\]

\[
R_\mathcal{K}(s, a) = \begin{cases} 
R(s, a), & \text{if } (s, a) \in \mathcal{K} \\ 
(1-\gamma)Q(s, a), & \text{otherwise} \end{cases}
\]

In other words, the known state–action SMDP \( M_\mathcal{K} \) has identical dynamics to \( M \) except in unknown state–actions where (i) the transitions are all self-loops with unit waiting time, and (ii) the \( Q \)-values are exact.

9.3. Proof of Theorem 2

Clearly, the construction leads to optimistic value functions, so the first condition of Theorem 1 holds.

We now consider when a state–action pair \((s, a)\) becomes known. Define the effective transition probabilities by

\[
P_S^\tau(s', a, \tau) = \sum_\tau P(s', \tau | s, a) \gamma^\tau,
\]

and the marginal distribution of waiting time by

\[
P_T^\tau(s, a) = \sum_{s'} P(s', \tau | s, a).
\]

We first generalize the simulation lemma (see, e.g., Kearns & Singh (2002); Strehl et al. (2009)) for MDPs to SMDPs, giving a bound on the value function differences in terms of model estimation errors:

**Lemma 5** Let \( M_i = (\mathcal{S}, \mathcal{A}, P_i, R_i, \gamma) \) \((i = 1, 2)\) be two SMDPs that differ only in reward and transition functions, and \( V_{i}^\gamma \) and \( Q_{i}^\gamma \) the respective optimal value functions. Let \( \bar{\gamma}_{s,a} \) be the effective discount factor for \((s, a)\) under \( M_2 \):

\[
\bar{\gamma}_{s,a} = \sum_\tau \gamma^\tau P_T^\tau(s, a).
\]

and define the discount-adjusted model estimation error by

\[
\varepsilon_{s,a} = \frac{1}{1 - \bar{\gamma}_{s,a}} \left( |R_1(s, a) - R_2(s, a)| + \max s' \| P_1^\tau(s', a) - P_2^\tau(s', a) \|_1 \right).
\]

Then, for any \( s \) and \( a \),

\[
|Q_{1}^\gamma(s, a) - Q_{2}^\gamma(s, a)| \leq \max_{s',a} \varepsilon_{s,a}
\]

\[
|V_{1}^\gamma(s, a) - V_{2}^\gamma(s, a)| \leq \max_{s',a} \varepsilon_{s,a}
\]

**Proof** Let \((s, a)\) be the state–action pair that achieves maximum difference of \( |Q_{1}^\gamma(s', a) - Q_{2}^\gamma(s', a)| \). To simplify notation, define

\[
\varepsilon_R = |R_1(s, a) - R_2(s, a)|
\]

\[
\varepsilon_P = \| P_1^\tau(s', a) - P_2^\tau(s', a) \|_1
\]

\[
\Delta = |Q_{1}^\gamma(s, a) - Q_{2}^\gamma(s, a)|
\]
Then,

\[ \Delta = |Q_1^*(s, a) - Q_2^*(s, a)| \]

\[ = \left| R_1(s, a) + \sum_{s', \tau} \gamma^\tau P_1(s', \tau)|s, a|V_1^*(s') \right| \\
- \left| R_2(s, a) + \sum_{s', \tau} \gamma^\tau P_2(s', \tau)|s, a|V_2^*(s') \right| \\
\leq |R_1(s, a) - R_2(s, a)| \\
+ \left| \sum_{s', \tau} \gamma^\tau (P_1(s', \tau)|s, a| - P_2(s', \tau)|s, a|) V_1^*(s') \right| \\
+ \left| \sum_{s', \tau} \gamma^\tau P_2(s', \tau)|s, a|(V_1^*(s') - V_2^*(s')) \right| \\
\leq \varepsilon_R + V_{\max|s, a, \tau} - \Delta \left| \sum_{s', \tau} \gamma^\tau P_2(s', \tau)|s, a| \right| \\
= (\varepsilon_R + V_{\max|s, a, \tau}) + \gamma_{s,a,\Delta} \\
= (1 - \gamma_{s,a})\varepsilon_{s,a} + \gamma_{s,a,\Delta}. \\
\]

Rearranging terms, we have

\[ \Delta \leq \varepsilon_{s,a} \leq \max_{s',a'} \varepsilon_{s',a}. \]

The case for \( V^*_M - V^*_M \) follows immediately from the following observation: for any state \( s \),

\[ |V_1^*(s) - V_2^*(s)| = \max_a Q_1^*(s, a) - \max_a Q_2^*(s, a) \]

\[ \leq \max_a |Q_1(s, a) - Q_2(s, a)| \leq \Delta. \]

\[ \square \]

Clearly, \( R(s, a) \in [0, \frac{1}{1-\gamma}] \). Using a concentration argument based on Hoeffding’s inequality, one can establish that \( n/(\varepsilon^2(1-\gamma)^2) \) samples suffice to ensure \( \varepsilon \) accuracy in the reward estimate. Similarly, the effective transition probabilities \( P(s'|s, a) \) can also be within \( \varepsilon \) total variation with \( n/(Na_a/\varepsilon^2) \) samples. Therefore, by setting \( \varepsilon \) appropriately, the accuracy condition in Theorem 1 can be satisfied.

Finally, there are at most \( SA \) many state-actions, each known when it is visited sufficiently often. The bounded-surprises condition in Theorem 1 thus holds.

Therefore all three conditions of Theorem 1 hold, and the result follows.

10. Further Details for Section 4

10.1. Proof Sketch of Lemma 3

Fix a non-\( \varepsilon \)-optimal option set \( O' \subset O^* \) with \( |O'| \leq \bar{O} \).

By assumption, it fails to represent a near-optimal policy for MDPs drawn i.i.d. from \( \nu \) over \( \mathcal{M} \). Following the same argument for Lemma 1 of Brunskill & Li (2013), \( p_{\min}\ln \xi \) many tasks suffices to reveal the non-\( \varepsilon \)-optimality of \( O' \), with probability at least \( 1 - \delta/C \). Taking a union bound over all \( C \) subsets of \( O^* \) up to size \( \bar{O} \), one finishes the proof of the lemma.

10.2. Proof Sketch of Lemma 4

For convenience, define \( \epsilon_1 = (\varepsilon - \varepsilon)/4 \). The proof relies on three major steps, each holding with probability at least \( 1 - \delta \).

- **The MDP models are all estimated to sufficient accuracy:** The condition together with Lemma 2 implies every state-action will be visited at least \( \Omega(NV_{\max|s, a, \tau}^2(1-\gamma)^2 \ln 1/\delta) \) times. Applying Hoeffding’s inequality together with Lemma 8.5.5 of Kakade (2003), the reward and transition probabilities of every state-action pair are estimated with \( \epsilon_1(1-\gamma)/V_{\max|s, a, \tau} \) accuracy. By the simulation lemma (c.f., Kearns & Singh (2002); Strehl et al. (2009)), \( |V^*_M(s) - V^*_M(s)| < \epsilon_1 \), and similarly, \( |V^*_M(s) - V^*_M(s)| < \epsilon_1 \), where \( M \) and \( \hat{M} \) are the underlying/estimated MDPs, and \( M' \) and \( \hat{M}' \) the corresponding SMDPs induced by the discovered option set \( \hat{O} \).

- **The discovered option set \( \hat{O} \) is \( \varepsilon \)-optimal for all MDPs in \( \mathcal{M} \):** Using the triangle inequality together with the two inequalities established in the previous step, we have

\[ V^*_M(s) - V^*_M(s) \]

\[ \leq |V^*_M(s) - V^*_M(s)| + |V^*_M(s) - V^*_M(s)| \]

\[ + |V^*_M(s) - V^*_M(s)| \]

\[ \leq 2\epsilon_1 + |V^*_M(s) - V^*_M(s)|. \]

In the option-discovery step, \( \hat{O} \) must satisfy \( V^*_M(s) - V^*_M(s) \leq (\varepsilon + \varepsilon)/2 \). Therefore, \( V^*_M(s) - V^*_M(s) \leq 2\epsilon_1 + (\varepsilon + \varepsilon)/2 = \varepsilon \); that is, the option set \( \hat{O} \) is \( \varepsilon \)-optimal for all MDPs encountered in phase 1. According to Lemma 3, \( \hat{O} \) must also be \( \varepsilon \)-optimal for all MDPs in \( \mathcal{M} \); otherwise, it will fail to represent \( \varepsilon \)-optimal policies in at least one MDP in phase 1.

- **There exists at least one option set that satisfies the criterion of Equation 2:** According to the assumption, there exists some option set \( \hat{O} \) that is \( \varepsilon \)-optimal for \( \mathcal{M} \): for any \( M \) and any \( s, V^*_M(s) - V^*_M(s) < \varepsilon \), where
M′ is the SMDP induced by M and $\bar{O}$. Using the triangle inequality as well as the accuracy guarantee established in step 1, one gets

$$V^*_{M'}(s) - V^*_{M}(s) < V^*_{M'}(s) + \epsilon_1 - V^*_{M'}(s) + \epsilon_1$$

$$< \epsilon + 2\epsilon_1$$

$$= (\epsilon + \epsilon)/2.$$

In other words, $\bar{O}$ will satisfy Equation 2.

The overall failure probability is at most $\delta$: All three steps above hold with high probability. The first two steps require a union bound over all possible subsets of $O^*$ with size up to $\bar{O}$. There are $C = \bar{O} \left( (O^*)^{\bar{O}} \right)$ many such subsets. It suffices to set $\delta \leftarrow \delta/C$ for the union bound to complete the whole proof.

10.3. Proof of Theorem 3

The sample complexity can be divided into two terms, corresponding to tasks in phase 1 and in phase 2, respectively. The sample complexity of the MDP tasks in phase 1 is simply the number of tasks in phase 1, $T_1$, multiplied by the sample complexity of the $E^3$ algorithm.

11. Further Details for Section 5

We now illustrate the process of evaluating the bound on the sample complexity benefit with the small example shown in Figure 1. In this example there are 2 states and 4 MDPs, and each MDP has a single $\epsilon$-optimal action in each state, shown in the Figure’s table. Assume that state $s_1$ deterministically transitions to $s_2$. Before introducing an option, there were 4 state-action combinations $(s_1, a_1), (s_1, a_2), (s_2, a_4), (s_2, a_5)$ needed to cover the $\epsilon$-optimal policies of each MDP, resulting in a sample complexity bound of $\bar{O} \left( \frac{4}{1 - \gamma^2} \right)$. Now consider adding the option whose initiation state is $s_1$ and that takes action $a_2$ in state $s_1$ and action $a_5$ in state $s_2$. The length of this option is always 2, so from the prior section the option’s contribution to the sample complexity is $\bar{O} \left( \frac{4}{(1 - \gamma^2)(1 - \gamma)^2} \right) \left( 2 + \frac{1}{1 - \gamma} \right)$. This option covers MDPs $m_3$ and $m_4$. To cover $s_1$ and $s_2$ for the remaining uncov-ered MDPs requires just 2 primitive state-action pairs, with a resulting $\bar{O} \left( \frac{2}{(1 - \gamma)^3} \right)$ contribution to the sample complexity bound. Therefore, introducing the option will reduce this upper bound on the sample complexity if

$$\frac{1}{(1 - \gamma^2)(1 - \gamma)^3} (2 + \frac{1}{1 - \gamma}) + \frac{2}{(1 - \gamma)^6} < \frac{4}{(1 - \gamma)^6}$$

$$\Leftrightarrow 5 < 6\gamma + \gamma^2$$

which holds for large $\gamma$, such as $\gamma = 0.9$. The algorithm evaluates this expression for the input $\gamma$, and keeps the option if the expression holds.