# Supplementary Materials for PAC Continuous State Online Multitask Reinforcement Learning with Identification

Yao Liu
Peking University
Beijing, PR China
liuyao@pku.edu.cn

Zhanhan Guo
Carnegie Mellon University
Pittsburgh, PA, United States
zguo@cs.cmu.edu

Emma Brunskill
Carnegie Mellon University
Pittsburgh, PA, United States
ebrun@cs.cmu.edu

# APPENDIX

## A. ALGORITHM PSEUDO-CODE

---
**Algorithm 1** Continuous State Online Multitask RL with Identification (**COMRLI**)

---
**Require:** $T_1, \bar{C}, L_Q, \epsilon, \Gamma, H$
  **for** $t = 1, 2, \ldots T_1$ **do**
    Receive an unknown MDP $M_t \in \mathcal{M}$
    Run algorithm 2 on $M_t$ with $(\Gamma, D)$-*known*.
    For all remaining steps until H steps, execute C-PACE algorithm on $M_t$
    Store all samples as a set $Sample_{M_t}$
  **end for**
  Cluster all tasks into $\widehat{C} \leq \bar{C}$ groups and combine their sample sets.
  **for** $t = T_1 + 1, \ldots T$ **do**
    Receive unknown MDP $M_t \in \mathcal{M}$
    Run algorithm 3 on $M_t$
    **if** $M_t$ is identified **then**
      Combine samples from $M_t$ to the group
    **end if**
  **end for**

---

## B. PROOFS OF LEMMAS

*Definition 7.* A state-action pair is $(\Gamma, D)$-known when it has at least

$$max \left\{ D^2 \ln \left( \frac{\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)}{\delta} \right), \ \frac{Q_{max}^2}{\Gamma^2 (1-\gamma)^2} \ln \left( \frac{T_1 \mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)}{\delta} \right) \right\}$$

visited neighbors within a distance of $\frac{\Gamma(1-\gamma)}{8L_Q}$.

### B.1 Proposition 1

PROPOSITION 1. *(lemma 4.5 in [2]) There are at most* $k \cdot \mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)$ *number of visits to state-action pairs that are unknown in Algorithm 2, where a known state-action pair means it has k visited neighbors within a distance of* $\frac{\Gamma(1-\gamma)}{8L_Q}$.

---
**Algorithm 2** Phase 1: Continuous PAC Explore

---
**Require:** $T_e, L_Q, \Gamma$
  Set the neighborhood radius to $\frac{min\{\Gamma/4, 1/24\}}{L_Q}$
  **while** some $(s, a)$ is unknown (see Def.7) **do**
    This is a start of new $T_e$-step episode.
    Find a $T_e$-step undiscounted optimistic Q-function $Q_{0,T_e}$ by:
    Initialize: $Q_{T_e,T_e}(s, a) = 0$
    **for** $t = T_e - 1, \ldots, 0$ **do**
      **if** $(s, a)$ is known **then**
        Find k nearest neighbors of $(s, a)$: $(s_j, a_j, r_j, s_j')_{j=1}^k$
        $Q_{t,T_e} = \frac{1}{k} \sum_{j=1}^{k} \left( r_j + \max_a Q_{t+1,T_e} \left( s_j', a \right) + L_Q d_{ij} \right)$
      **else**
        $Q_{t,T_e} = (T_e - t)$
      **end if**
    **end for**
    Take greedy policy of $Q_{0,T_e}(s, a)$ for next $T_e$ steps.
    **if** $(s, a)$ is unknown **then**
      Add $(s, a, r, s')$ to the sample set
    **end if**
  **end while**

---

### B.2 Proof of Proposition 2

PROPOSITION 2. *In Algorithm 2, denoting the exact Bellman operator by B and the upper bound of Q value by $Q_max$, if*

$$\frac{4Q_{max}^2}{\epsilon^2} \ln \left( \frac{4\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)}{\delta} \right) \leq k \leq \frac{4\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)}{\delta}$$

*and the radius of the neighborhood is no more than* $\frac{\Gamma}{2L_Q}$, *then w.p.* $1 - \delta/2$, *for all known* $(s, a)$ *(Known is defined in proposition 1), we have:*

$$|Q_{t,T_e}(s, a) - BQ_{t+1,T_e}(s, a)| \leq \Gamma$$

*where $T_e$ is the finite horizon length in algorithm 2, and $t < T_e$.*

PROOF. By proposition 1, there should be at most $k\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)$ unknown samples in algorithm 2. By lemma 3.13 in [4], we have that if $\frac{Q_{max}^2}{\Gamma^2} \ln \left( \frac{4\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)}{\delta} \right) \leq k \leq \frac{4\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)}{\delta}$, then w.p. $1 - \delta/2$ for any known $(s, a)$ and t,

$$-\Gamma/2 \leq \widehat{B}Q_{t,T_e}(s, a) - BQ_{t,T_e}(s, a) \leq \Gamma/2$$

**Algorithm 3** Phase 2: Continuous Identify

---

**Require:** $\Gamma, \epsilon, \bar{C}, H, T_i, n$

  Initialize version space: $\mathcal{C} \leftarrow \{1, \ldots, \bar{C}\}$
  **while** $h < H$ **do**
    **for** $c \in \mathcal{C}$ **do**
      Use algorithm 2 with samples from $M_c$ to find a
  informative pair $(s, a), s.t. \left| Q_{M_i}^{\pi_i}(s,a) - Q_{M_j}^{\pi_j}(s,a) \right| \geq \frac{7\Gamma}{8}$
      **if** informative pair $(s, a)$ is reached **then**
        Break the loop.
      **end if**
    **end for**
    **for** $g = \{i, j\}$ **do**
      **for** $t = 1 \ldots T_i$ **do** (Monte Carlo estimate)
        Run the greedy policy of $Q_{M_g}$, $\pi_g$, for n steps
        $R_{gt} \leftarrow r(s,a) + \sum_{l=1}^{n} \gamma^l r_l, h \leftarrow h + n$
        $\widetilde{D} \leftarrow 1$
        **while** Haven't returned to $(s, a)$ **do**
          **for** $c \in \mathcal{C}$ **do**
            Use $M_c$ to create an informative MDP,
  and try to go back to $(s, a)$ within $12\widetilde{D} \ln \frac{\bar{C}}{\delta}$ steps.
            **if** get back to $(s, a)$ **then**
              Break the loop.
            **end if**
          **end for**
          $\widetilde{D} \leftarrow 2\widetilde{D}$
        **end while**
      **end for**
      $R_g \leftarrow \frac{1}{T_i} \sum_{t=1}^{T_i} R_{gt}$
      **if** $\left| R_g - Q_{M_g}(s,a) \right| \geq \frac{\Gamma}{8}$ **then**
        $\mathcal{C} \leftarrow \mathcal{C} \backslash \{g\}$
      **end if**
    **end for**
    **if** Both i,j haven't been eliminated **then**
      Eliminate the model k with a smaller $R_g$.
    **end if**
    **if** Only one group left **then**
      Combine the sample sets and run C-PACE
    **end if**
  **end while**

---

We also have

$$0 \leq \widetilde{B}Q_{t,T_e}(s,a) - \widehat{B}Q_{t,T_e}(s,a) \leq \Gamma/2$$

and $Q_{t,T_e} = \widetilde{B}Q_{t+1,T_e}$ by definition. Then combining them together we get

$$-\Gamma \leq Q_{t,T_e}(s,a) - BQ_{t+1,T_e}(s,a) \leq \Gamma$$

$\square$

## B.3   Proof of Proposition 3

PROPOSITION 3. *Assume $R \in [0, 1]$. Suppose in Algorithm 2, $|Q_{t,T_e}(s,a) - BQ_{t+1,T_e}(s,a)| \leq \Gamma$, $\pi$ is a $T_e$-step greedy policy introduced by $Q_{t,T_e}$, and $Q_{t,T_e}^{\pi}$ is the Q value of this policy. Then $\forall$ known $(s,a)$ (Known is defined in proposition 1), $t < T$, $\left| Q_{t,T_e}^*(s,a) - Q_{t,T_e}^{\pi}(s,a) \right| \leq 2(T_e - t)\Gamma$, and*

$$\left| V_{t,T_e}^*(s) \leq V_{t,T_e}^{\pi}(s) \right| \leq 2(T_e - t)\Gamma$$

PROOF. Firstly we need to clarify some notations: $B$ is the exact Bellman operator. $\widehat{B}$ is the approximate Bellman

operator which is defined in [4]:

$$\widehat{B}Q(s,a) = \frac{1}{k} \sum_{i=1}^{k} \left( r_i + \gamma V(s_i^{'}) \right)$$

where $(s_i, a_i, r_i, s_i^{'})$ ranges over the k nearest neighbor tuples. $\widetilde{B}$ denotes the approximate Bellman operator defined by the right side of equation 1 in [4]:

$$\widetilde{B}Q(s,a) = \frac{1}{k} \sum_{i=1}^{k} \left( r_i + \gamma V(s_i^{'}) + L_Q d_i \right)$$

where $L_Q$ is the Lipschitz constant and $d_i$ is the distance between $(s,a)$ and $(s_i, a_i)$. Then, we will prove

$$\left| Q_{t,T_e}^*(s,a) - Q_{t,T_e}(s,a) \right| \leq (T_e - t)\Gamma$$

for all $0 \leq t \leq T_e$ by induction. For $t = T_e$, $Q_{t,T_e}^*(s,a) = Q_{t,T_e}(s,a) = 0$. Then assuming the inequality holds for $t+1$, we want to prove the result also holds for t:

$$
\begin{aligned}
\left| Q_{t,T_e}^* - Q_{t,T_e} \right| &\leq \left| BQ_{t+1,T_e}^* - BQ_{t+1,T_e} \right| \\
&\quad + \left| BQ_{t+1,T_e} - Q_{t,T_e} \right| \\
&\leq \left| \sum_{s'} P(s'|s,a) \left( max_{a'} Q_{t+1,T_e}^*(s',a') \right. \right. \\
&\quad \left. \left. - max_{a'} Q_{t+1,T_e}(s',a') \right) \right| + \Gamma \\
&\leq (T_e - t - 1)\Gamma + \Gamma \\
&= (T_e - t)\Gamma
\end{aligned}
$$

The first step follows from triangle inequality and the fact $Q_{t,T_e}^* = BQ_{t+1,T_e}^*$. The second line follows from the assumption in the proposition. The third line follows from the assumption of induction hypothesis. Now we have

$$\left| Q_{t,T_e}^*(s,a) - Q_{t,T_e}(s,a) \right| \leq (T_e - t)\Gamma$$

Then we will bound the difference between $Q_{t,T_e}^{\pi}$ and $Q_{t,T_e}$ in a similar way. Here we define a new Bellman operator $B^{\pi}$:

$$B^{\pi} Q_{t,T_e}(s,a) = R(s,a) + \sum_{s' \in \mathcal{S}} P(s'|s,a) Q_{t+1,T_e}(s', \pi(s'))$$

From the definition, we know $Q_{t,T_e}^{\pi} = B^{\pi} Q_{t+1,T_e}^{\pi}$. Since $\pi$ is the greedy policy over $Q$, we have $BQ = B^{\pi} Q$. Replace the $B$ operator above with $B^{\pi}$, then following similar inequalities we have $\left| Q_{t,T_e}^{\pi}(s,a) - Q_{t,T_e}(s,a) \right| \leq 2(T_e - t)\Gamma$. Then by triangle inequality we get the result. After we have the bound on Q, the the bound on V immediately follows. $\square$

## B.4   Proof of Lemma 1

LEMMA 1. *After no more than $O\left(\left(k\mathcal{N}_{\mathcal{S}\mathcal{A}}, -(L_Q) + \ln \frac{1}{\delta}\right) D\right)$ steps of algorithm 2, every state-action pair will have at least k visited neighbors, with probability of $1 - \delta$, where $k \geq k_{min} = O\left(D^2 \ln \left(\frac{\mathcal{N}_{\mathcal{S}\mathcal{A}}(L_Q, \Gamma)}{\delta}\right)\right)$, $k \leq k_{max} = O\left(\frac{\mathcal{N}_{\mathcal{S}\mathcal{A}}(L_Q, \Gamma)}{\delta}\right)$.*

PROOF. For convenience, we denote a visit to an unknown state-action pair in algorithm 2 as an escape. By the diameter assumption and Markov's inequality, we have that there exists a policy $\pi$ that will escape within $2D$ steps with a probability of at least $1/2$. So such a policy would get a reward of $D$ in $T = 3D$ steps in $M_K$. So the optimal policy will get at least $D$ reward if we set $T_e = 3D$ steps, which means $V_{0,T_e}^*(s) \geq D$. By applying proposition 3, we get

*w.p.* $1 - \delta/2$, $V_{0,T_e}^\pi (s)$ is at least $D - 2T_e\Gamma$, which could also be expressed as

$$\sum_{t=1}^{T_e} Pr\,(escape\ at\ t)\,(T_e - t)$$

So with probability of $1 - \delta/2$ we get the probability of escape in T steps, $p_e$, could be at least a constant such as $\frac{1}{4}$ by using a $\Gamma$ smaller than $1/24$:

$$
\begin{aligned}
p_e \quad &= \sum_{t=1}^{T_e} Pr\,(escape\ at\ t) \geq \sum_{t=1}^{T_e} Pr\,(escape\ at\ t)\,\frac{T_e - t}{T_e} \\
&\geq \frac{D}{T_e} - 2\Gamma = \frac{1}{3} - 2\Gamma \geq \frac{1}{4}
\end{aligned}
$$

Every $T_e$-step episode has at least probability $p_e$ of escaping. Since there are at most $k\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma)$ number of escapes before everything is known, we can bound how many episodes there are until everything is known with high probability $(1 - \delta/2)$. Lemma 56 from [3] yields $O\left(k\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma) + \ln\frac{1}{\delta}\right)$ for the number of episodes. Then the total number of time-steps required is $O\left(\left(k\mathcal{N}_{\mathcal{SA}}(L_Q, \Gamma) + \ln\frac{1}{\delta}\right)D\right)$. The lower bound of $p_e$ holds with probability $1 - \delta/2$, so the whole theorem holds by a union bound with probability $1 - \delta$. Note that the $Q_{max}$ in the constraint of k is no more that $T = O(D)$. $\square$

## B.5 Lemma 2

LEMMA 2. *(lemma 1 in [1]) If we set* $T_1 = \frac{\ln\frac{\bar{C}}{\delta}}{p_{min}}$ *then w.p.* $1 - \delta$, *all distinct MDPs will be encountered in phase 1.*

## B.6 Proof of Lemma 3

LEMMA 3. *If $M_1$ and $M_2$ are 2 MDPs such that for any (s,a) pair,*

$$|r_{M_1}(s,a) - r_{M_2}(s,a)| < \frac{\epsilon(1-\gamma)}{2}$$

$$\int_{\mathcal{S}} |T_{M_1}(s'|s,a) - T_{M_2}(s'|s,a)|ds' < \frac{\epsilon(1-\gamma)^2}{2}$$

*then the optimal Q-functions for $M_1$ and $M_2$, $Q_{M_1}$ and $Q_{M_2}$, satisfy that for any (s,a) pair*

$$|Q_{M_1}(s,a) - Q_{M_2}(s,a)| < \epsilon$$

PROOF. Let $B_1$ and $B_2$ denote the Bellman operators of $M_1$ and $M_2$, and $Q_0$ denotes an arbitrary function over $\mathcal{S} \times \mathcal{A}$. To show the result, it is sufficient to prove that $|B_1^i Q_0(s,a) - B_2^i Q_0(s,a)| \leq \sum_{j=0}^{i} \gamma^j \epsilon(1-\gamma)$, and then take the limit as $i \to \infty$. We prove this by induction. The base case is trivial since $Q_0(s,a) = Q_0(s,a)$. Assuming the state-

ment holds for i, we consider the case of i+1:

$$
\begin{aligned}
&|B_1^{i+1}Q_0(s,a) - B_2^{i+1}Q_0(s,a)| \\
\leq\quad &|r_{M_1}(s,a) - r_{M_2}(s,a)| \\
+\quad &\gamma\left|\int_{\mathcal{S}} T_{M_1}(s'|s,a)max_{a'}B_1^i Q_0(s,a)ds'\right. \\
&\left. - \int_{\mathcal{S}} T_{M_2}(s'|s,a)max_{a'}B_2^i Q_0(s,a)ds'\right| \\
\leq\quad &|r_{M_1}(s,a) - r_{M_2}(s,a)| \\
+\quad &\gamma\left|\int_{\mathcal{S}} (T_{M_1}(s'|s,a) - T_{M_2}(s'|s,a))max_{a'}B_1^i Q_0(s,a)ds'\right. \\
&+ \int_{\mathcal{S}} T_{M_2}(s'|s,a)max_{a'}B_1^i Q_0(s,a)ds' \\
&\left. - \int_{\mathcal{S}} T_{M_2}(s'|s,a)max_{a'}B_2^i Q_0(s,a)ds'\right| \\
\leq\quad &\frac{\epsilon(1-\gamma)}{2} + \gamma\left|Q_{max}\int_{\mathcal{S}} |T_{M_1}(s'|s,a) - T_{M_2}(s'|s,a)|ds'\right| \\
+\quad &\gamma\left|\int_{\mathcal{S}} T_{M_2}(s'|s,a)|max_{a'}B_1^i Q_0(s,a) - max_{a'}B_2^i Q_0(s,a)|ds'\right| \\
\leq\quad &\frac{\epsilon(1-\gamma)}{2} + \gamma\frac{1}{1-\gamma} \times \frac{\epsilon(1-\gamma)^2}{2} + \gamma\sum_{j=0}^{i}\gamma^j\epsilon(1-\gamma) \\
\leq\quad &\epsilon(1-\gamma) + \gamma\sum_{j=0}^{i}\gamma^j\epsilon(1-\gamma) \\
=\quad &\sum_{j=0}^{i+1}\gamma^j\epsilon(1-\gamma)
\end{aligned}
$$

The first inequality follows from the definition of Bellman operator. The second inequality follows by adding and subtracting the same thing. The third inequality follows by the triangle inequality. The fourth inequality follows from the condition of the lemma and the inductive assumption. $\square$

## B.7 Proof of Lemma 4

LEMMA 4. *If all tasks in phase 1 are run for at least $H_{min}$ steps, with probability $1 - \delta$, the following holds:*

1. *For all tasks, any state-action pair $(s,a)$ will receive at least $O\left(\frac{Q_{max}^2}{\Gamma^2(1-\gamma)^2}\ln\left(\frac{T_1\mathcal{N}_{\mathcal{SA}}(L_Q,\Gamma)}{\delta}\right)\right)$ visited neighbors whose distance with $(s,a)$ is no more than $\frac{\Gamma}{8L_Q}$ $((\Gamma, D)$-known).*

2. *Tasks in phase 1 will be clustered correctly with high probability.*

3. *For any cluster, the max-norm distance between the approximate Q-function and the true optimal Q-function for any task in this cluster is at most $\frac{5*\Gamma}{16}$.*

PROOF. The first statement could be proved immediately by lemma 1.

Before we prove the second statement, we firstly clarify how to cluster tasks at the end of phase 1. For each task, we find a fixed-point solution $Q_{M_i}$ of $Q = \widetilde{B}Q$, where $\widetilde{B}$ is defined in C-PACE. Then we check all the state-action pairs in a covering set and put two tasks that have $\frac{\Gamma}{2}$-close value on all the pairs into one cluster.

We are going to prove that after clustering like this, different tasks would be clustered into different groups and the

same tasks would be put into the same group. We will first prove there must exist an $\frac{3\Gamma}{4}$ difference between $Q^*$ of different tasks on the covering set, and the $Q^*$ of tasks within one underlying cluster is $\frac{\Gamma}{4}$ close. Then we prove the fixed solution $Q_{M_i}$ is a $\frac{\Gamma}{16}$-close approximation of $Q^*$. These will prove that the distance of approximate Q-functions from a same underlying cluster should be at most $\frac{3\Gamma}{8} = \frac{\Gamma}{4} + 2 * \frac{\Gamma}{16}$. Thus a threshold of $\frac{\Gamma}{2}$ would ensure the correctness of clustering.

By the assumption of a Q-gap $\Gamma$ and the Lipschitz smoothness, we could say for any two distinct tasks $i, j$, there exists at least a state-action pair such that the optimal Q-function is different in its neighborhood. By the definition of a covering set, there must exists one point in the covering set in such a neighborhood. Therefore for such a point, the optimal Q function has a gap of at least $\frac{3\Gamma}{4}$.

Note that when we define the underlying MDP cluster, we guarantee that their parameters are within a distance of $\frac{\epsilon(1-\gamma)^2}{16}$, and $\Gamma$ is at least $\frac{\epsilon}{2}$. Thus by lemma 3, the max-norm distance of optimal Q functions within one true underlying cluster are at most $\frac{\Gamma}{4}$.

Then we prove that $\left| Q_{M_i}(s,a) - Q^*_{M_i} \right| \le \frac{\Gamma}{16}$ for all $(s,a)$ in the covering set and $M_i$ in the $T_1$ tasks. We have at least $k = \frac{16^2 Q_{max}^2}{\Gamma^2 (1-\gamma)^2} \ln \left( \frac{\mathcal{N}_{\mathcal{S}\mathcal{A}} T_1}{\delta} \right)$ neighbors for any point in the covering set $S_c$. Note the size of the covering set is $O(\mathcal{N}_{\mathcal{S}\mathcal{A}}(L_Q, \Gamma))$ and we certainly have $T_1$ tasks. By Lemma 3.14 in [4], the fixed point solution Q satisfies that

$$|Q_{M_i} - BQ_{M_i}| \le \frac{\Gamma}{16(1-\gamma)}$$

for any $M_i$. Then using Proposition 4.1 in [5], we have $|Q^* - Q| \le \frac{\Gamma}{16}$ for all the $T_1$ tasks.

Now we have already proved the third statement: The distance of approximate Q functions of any MDP with its optimal Q function is at most $\frac{\Gamma}{16}$, and the distance between optimal Q of MDPs within a cluster is $\frac{\Gamma}{4}$. So the distance between approximate Q-function with the optimal Q-functions for any MDP in the same cluster is $\frac{\Gamma}{4} + \frac{\Gamma}{16} \le \frac{5\Gamma}{16}$. $\square$

## B.8 Proof of Lemma 5 (in the paper)

LEMMA 5. *If every state-action pair is $(\Gamma, D)$-known, then given any start state and desired state-action pair, it is possible to visit the desired state-action pair's neighborhood in no more than $\widetilde{O}(D)$ steps with high probability.*

PROOF. Firstly, we construct an MDP $M_{inform}$ such that the desired state-action pairs have unit reward and all others have 0 reward. The desired pair is a self-loop and other transition probabilities are inherited from the true MDP dynamics. We know which point is desired, so we can modify the original samples to become samples in the new MDP $M_{inform}$. Because now every pair is $(\Gamma, D)$-known, so we have $O(D^2)$ samples in every state-action's neighborhood in this $M_{inform}$. Following a similar analysis of lemma 1, we could find a policy whose probability of reaching the desired region within $3D$ steps is at least $\frac{1}{4}$. Then after $3D \log_{\frac{3}{4}} \delta$ steps the policy could reach the desired region with probability of $1 - \delta$. $\square$

## B.9 Proof of Lemma 6 (in the paper)

LEMMA 6. *When we face an unknown task in phase 2, we could reach any desired state-action pair within $O(\bar{C}D \ln \frac{\bar{C}}{\delta})$ steps with probability $1 - \frac{\delta}{C}$.*

PROOF. First, consider the case where we know the diameter D. The unknown task must be one of the $\bar{C}$ tasks from phase 1. Our algorithm tries to run each policy in $3D \ln \frac{\bar{C}}{\delta}$ steps to the desired state-action pair using the samples from one of the $\bar{C}$ tasks, thus lemma 6 holds with probability $1 - \frac{\bar{C}}{\delta}$. By trying all policies from the $\bar{C}$ tasks, we will encounter the one policy that corresponds to the same task and reach the desired region with high probability.

If we don't know the diameter, we could use the doubling trick to find an upper bound on D without an increase in the sample complexity. First we try the whole process with $\widetilde{D} = 1$. If we fail, we double the $\widetilde{D}$ and begin a new trial. When $\widetilde{D}$ is bigger than the true value of D, the rest of the analysis is the same as when we know the true diameter. $\square$

## B.10 Proof of Lemma 7

LEMMA 7. *If $T_i$ in algorithm 3 is at least $O\left( \frac{Q_{max}^2}{\Gamma^2} \ln \frac{\bar{C}}{\delta} \right)$ and $n$ is at least $O(\log_\gamma \Gamma)$, we could compute an approximate Q value of policy $\pi$ over the current task $M'$: $\widehat{Q}^{\pi_i}_{M'}(s,a) = R_i$ such that for any $(s,a)$, $\left| \widehat{Q}^{\pi_i}_{M'}(s,a) - Q^{\pi_i}_{M'}(s,a) \right| \le \frac{\Gamma}{16}$ with probability $1 - \frac{\delta}{C}$.*

PROOF. By setting $n$ to at least $\log_\gamma \frac{\Gamma(1-\gamma)}{32}$ we have that for each episode t, $\|R_{it} - \sum_{l=0}^{\infty} \gamma^l r_l\| \le \frac{\Gamma}{32}$. Note that the expectation of $\sum_{l=0}^{\infty} \gamma^l r_l$ is $Q^{\pi_i}_{M'}(s,a)$. We bound the error as follows:

$$
\begin{aligned}
&P\left( \left| \widehat{Q}^{\pi_i}_{M'}(s,a) - Q^{\pi_i}_{M'}(s,a) \right| \ge \frac{\Gamma}{16} \right) \\
=\ & P\left( \left| \frac{1}{T_i} \sum_{t=1}^{T_i} R_i t - Q^{\pi_i}_{M'}(s,a) \right| \ge \frac{\Gamma}{16} \right) \\
\le\ & P\left( \left| \frac{1}{T_i} \sum_{t=1}^{T_i} \left( \sum_{l=0}^{\infty} \gamma^l r_l \right) - Q^{\pi_i}_{M'}(s,a) \right| \ge \frac{\Gamma}{32} \right) \\
\le\ & 2 exp\left\{ -\frac{2 T_i \Gamma^2}{32^2 Q_{max}^2} \right\}
\end{aligned}
$$

The first step follows from the definition of $\widehat{Q}^{\pi_i}_{M_i}(s,a)$. The second step follows from the fact $\|R_{it} - \sum_{l=0}^{\infty} \gamma^l r_l\| \le \frac{\Gamma}{32}$. The third step follows from the Hoeffding inequality. Setting the probability above to $\frac{\delta}{C}$ and solving for $T_i$, we have that

$$T_i = O\left( \frac{Q_{max}^2}{\Gamma^2} \ln \frac{\bar{C}}{\delta} \right)$$

is sufficient to guarantee our desired result. $\square$

## B.11 Proof of Lemma 8

LEMMA 8. *If both model $M_i$ and $M_j$ haven't been eliminated by $\left| R_g - Q_{M_g}(s,a) \right| \ge \frac{\Gamma}{8}$ in algorithm 3, then the true model should have a greater $R_g$ with high probability.*

PROOF. Without loss of generality, we assume $M_i$ is in

the same underlying cluster with current task $M'$. Then:

$$\left| \widehat{Q}_{M_i}^{\pi_i}(s,a) - Q_{M'}^*(s,a) \right|$$
$$\leq \quad \left| \widehat{Q}_{M_i}^{\pi_i}(s,a) - Q_{M_i}^{\pi_i}(s,a) \right| + \left| Q_{M_i}^{\pi_i}(s,a) - Q_{M_i}(s,a) \right|$$
$$+ \left| Q_{M_i}(s,a) - Q_{M'}^*(s,a) \right|$$
$$\leq \quad \frac{\Gamma}{16} + \frac{\Gamma}{16} + \frac{5\Gamma}{16}$$
$$= \quad \frac{7\Gamma}{16}$$

The first step follows from triangle inequality. In the second step, the first replacement follows from Lemma 7, the second replacement follows from proposition 4.1 in [5], and the third replacement follows from Lemma 4. Because $M_j$ also hasn't been eliminated:

$$\left| Q_{M'}^{\pi_j}(s,a) - Q_{M_j}(s,a) \right|$$
$$\leq \quad \left| Q_{M'}^{\pi_j}(s,a) - \widehat{Q}_{M'}^{\pi_j}(s,a) \right| + \left| \widehat{Q}_{M'}^{\pi_j}(s,a) - Q_{M_j}(s,a) \right|$$
$$\leq \quad \frac{\Gamma}{16} + \frac{\Gamma}{8}$$
$$= \quad \frac{3\Gamma}{16}$$

The first inequality is from the triangle inequality. The second inequality follows from Lemma 7 (for the first erm) and the elimination condition at line 16 in Algorithm 3 (for the second term). We know there is a gap in the $Q$-function between $M_i$ and $M_j$ because (s,a) is an informative state-action pair: $\left| Q_{M_i}(s,a) - Q_{M_j}(s,a) \right| \geq \frac{7\Gamma}{8}$. Then $Q_{M_j}(s,a)$ must be smaller than $Q_{M_i}(s,a)$. Otherwise it implies that $Q_{M_i}^{\pi_j}(s,a)$ is larger than $Q_{M'}^*(s,a)$. However that is impossible because $Q_{M'}^*$ is the optimal policy's Q value. Therefore,

$$\widehat{Q}_{M'}^{\pi_i}(s,a) - \widehat{Q}_{M'}^{\pi_j}(s,a)$$
$$= \quad (\widehat{Q}_{M'}^{\pi_i}(s,a) - Q_{M_i}(s,a)) + (Q_{M_j}(s,a) - \widehat{Q}_{M'}^{\pi_j}(s,a))$$
$$+ \quad (Q_{M_i}(s,a) - Q_{M_j}(s,a))$$
$$\geq \quad -\frac{\Gamma}{8} - \frac{\Gamma}{8} + \frac{7\Gamma}{8}$$
$$\geq \quad \frac{5\Gamma}{8}$$

The first inequality's first two terms follows from the elimination condition at line 16 in Algorithm 3. The third term follows because $(s,a)$ is an informative pair and the fact that we just showed that $Q_{M_i}(s,a) > Q_{M_j}(s,a)$, so $Q_{M_i}(s,a) - Q_{M_j}(s,a) \geq \frac{7\Gamma}{8}$. Now we have shown that the approximate Q value of the true model's policy must be larger than other. Thus when we eliminate a candidate with the smaller Q-value, we will not eliminate the true candidate of the current task. $\square$

## B.12 Proof of Lemma 9

LEMMA 9. *In phase 2, After*

$$O\left( \frac{Q_{max}^2}{\Gamma^2} \bar{C} \ln \frac{\bar{C}}{\delta} \left( \bar{C} D \ln \frac{\bar{C}}{\delta} + \log_\gamma \Gamma \right) \right)$$

*steps, we correctly identify the new task with probability at least $1 - \delta$.*

PROOF. From Lemma 7, $O\left( \frac{Q_{max}^2}{\Gamma^2} \bar{C} \ln \frac{\bar{C}}{\delta} \log_\gamma \Gamma \right)$ total steps (across multiple trajectories) is sufficient to closely estimate

$Q_{M'}^{\pi_i}(s,a)$ for each $i \in \mathcal{C}$ with probability at least $1 - \frac{\delta}{2\bar{C}}$. Then by Lemma 8, we eliminate at least one candidate model per one informative state-action pair. However, each possible trajectory must start at the same informative state–action pair.

From Lemma 6 $O(\bar{C}D)$ steps are sufficient to return to a desired informative state–action pair with high probability. Therefore the total number of steps required to identify the current task is bounded by

$$O\left( \frac{Q_{max}^2}{\Gamma^2} \bar{C} \ln \frac{\bar{C}}{\delta} \left( D \ln \frac{\bar{C}}{\delta} + \log_\gamma \Gamma \right) \right)$$

where we have applied the union bound to ensure the final bound holds with probability at least is $1 - \delta$. $\square$

## REFERENCES

[1] E. Brunskill and L. Li. Sample complexity of multi-task reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

[2] S. Kakade, M. Kearns, and J. Langford. Exploration in metric state spaces. In *ICML*, volume 3, pages 306–312, 2003.

[3] L. Li. *A unifying framework for computational reinforcement learning theory*. PhD thesis, Rutgers, The State University of New Jersey, 2009.

[4] J. Pazis and R. Parr. Pac optimal exploration in continuous space markov decision processes. In *AAAI Conference on Artificial Intelligence*. Citeseer, 2013.

[5] R. J. Williams and L. C. Baird. Tight performance bounds on greedy policies based on imperfect value functions. Technical report, Citeseer, 1993.