A PROOF OF THEOREM 1

For any policy \( \mu \), we have

\[
\mathcal{R}^\mu(x) = E_\mu \left[ \sum_{t=0}^{T} r_t^x(X_t, A_t) \right]
\]

\[
= E_\mu \left[ \sum_{t=0}^{T} r_t^x((S_t, Q_t), A_t) \right]
\]

\[
\begin{align*}
\mathcal{R}^\mu(\mu) &= E_\mu \left[ \sum_{t=0}^{T} r_t^\pi(S_t, A_t) \right] = \mathcal{R}^\pi(\mu),
\end{align*}
\]

where the equality in \((a)\) follows from the definition of \( r_t^x \) in (7). Further, using (8), we have

\[
r_f(X_{T+1}) = r_f((S_{T+1}, Q_{T+1})) = 1_{F}(Q_{T+1}).
\]

Following the acceptance condition of the DFA \( \mathcal{A} \), which is equivalent to the LTL \( f \) specification \( \phi \), a run \( \xi \) of the POMDP satisfies \( \phi \) if and only if the word generated by the run satisfies the acceptance condition of the DFA \( \mathcal{A} \), i.e., its run on \( \mathcal{A} \), \( \xi_{\mathcal{A}} \), ends in the acceptance set \( F \). Hence,

\[
\mathcal{R}^f(\mu) = E_\mu \left[ r_f(X_{T+1}) \right] = \mathbb{P}^\mu(\phi).
\]

B PROOF OF LEMMA 1

We have

\[
\mathcal{R}^* = l^*
\]

\[
\leq l^*_{B}
\]

\[
\leq \inf_{0 \leq \lambda \leq B} L(\bar{\mu}, \lambda) + \epsilon
\]

\[
= \mathcal{R}^{\mu^*}(\bar{\mu}) + \inf_{0 \leq \lambda \leq B} \lambda(\mathcal{R}^f(\bar{\mu}) - 1 + \delta) + \epsilon.
\]

There are two possible cases: (i) \( \mathcal{R}^f(\bar{\mu}) - 1 + \delta \geq 0 \) and (ii) \( \mathcal{R}^f(\bar{\mu}) - 1 + \delta < 0 \).

If case (i) is true, then (10) is trivially satisfied. Further, in this case, we have

\[
\inf_{0 \leq \lambda \leq B} \lambda(\mathcal{R}^f(\bar{\mu}) - 1 + \delta) = 0.
\]

Therefore, \( \mathcal{R}^* \leq \mathcal{R}^{\mu^*}(\bar{\mu}) + \epsilon \), and hence, (15) is satisfied.

If case (ii) is true, we have

\[
\inf_{0 \leq \lambda \leq B} \lambda(\mathcal{R}^f(\bar{\mu}) - 1 + \delta) = B(\mathcal{R}^f(\bar{\mu}) - 1 + \delta)
\]

\[
< 0.
\]

Therefore, \( \mathcal{R}^* \leq \mathcal{R}^{\mu^*}(\bar{\mu}) + \epsilon \), and hence, (15) is satisfied. Further, we have

\[
B(\mathcal{R}^f(\bar{\mu}) - 1 + \delta) \geq \mathcal{R}^* - \mathcal{R}^{\mu^*}(\bar{\mu}) - \epsilon
\]

\[
\geq \mathcal{R}^* - R_m - \epsilon.
\]

The last inequality holds because \( R_m \) is the maximum achievable reward. Hence, (16) is satisfied.

C PROOF OF THEOREM 2

Consider the dual of (P4). Let

\[
u^*_{B} := \inf_{0 \leq \lambda \leq B} \sup_{\mu} L(\mu, \lambda).
\]
We have

\[ l_B^* \leq u_B^* \]

\[ = \inf_{0 \leq \lambda \leq B} \sup_{\mu} L(\mu, \lambda) \]

\[ \leq \sup_{\mu} L(\mu, \bar{\lambda}) \]

\[ \leq \frac{1}{K} \sum_{k=1}^{K} L(\mu_k, \lambda_k) \]

\[ \leq \frac{1}{K} \sum_{k=1}^{K} L(\mu_k, \lambda_k) \]

\[ \leq \frac{1}{K} \inf_{0 \leq \lambda \leq B} \sum_{k=1}^{K} L(\mu_k, \lambda) + 2B \sqrt{2 \log 2/K} \]

\[ \leq \inf_{0 \leq \lambda \leq B} L(\bar{\mu}, \lambda) + 2B \sqrt{2 \log 2/K} . \]

The inequality in (a) holds because of weak duality [Boyd and Vandenberghe 2004]. The equality in (b) holds because of the bilinearity (affine) of \( L(\cdot) \). The inequality in (c) holds because \( \mu_k \) is the maximizer associated with \( \lambda_k \). Inequality (d) follows from Corollary 5.7 in [Hazan et al. 2016]. Equality in (e) is again a consequence of bilinearity of \( L(\cdot) \).

## E ADDITIONAL DETAILS ON EXPERIMENTS

### E.1 MODEL DESCRIPTION

In this subsection, we provide further details on the grid world POMDP models used in our experiments. The images corresponding to the various models indicate the state space and the labeling function, e.g., in Fig. 1a we have \( L((1,2)) = \{b\}, L((3,3)) = \{a\}, \) and \( L((i,j)) = \{} \) for all other grid locations \((i,j)\). In all models, the agent starts from the grid location \((0,0)\). Further, the reward for all actions is 0 in all grid locations, unless specified otherwise. In the supplementary material, we also provide videos that capture some representative behaviors of the policies generated by Algorithm 1. We will discuss them in greater detail below.

#### E.1.1 Location Uncertainty

**Reach-Avoid Tasks.** In model \( \mathcal{M}_1 \), reward \( r((0,3), a) = 2 \) and \( r((3,3), a) = 1 \) for all actions \( a \). We observe that the agent satisfies the reach-avoid constraint with high probability and ends up in the top-right corner where the reward is highest. A representative trajectory for this model can be found in the video mu1_1.mp4.

In model \( \mathcal{M}_2 \), reward \( r((1,6)) = 3, r((4,3), a) = 3, \) and \( r((7,7), a) = 1 \) for all actions \( a \). In this model, we observe two characteristic behaviors. The agent reaches the goal state \( a \) and remains there (see video mu2_1.mp4). This behavior ensures that the specification is met but the reward is relatively lower. The other behavior is that the agent goes towards the location \((4,3)\) and tries to remain there to obtain higher reward (see video mu2_2.mp4). However, since the obstacle is very close and the transitions are stochastic, it is prone to violating the constraint. Nonetheless, this violation is rare enough such that the overall satisfaction probability exceeds the desired threshold.

**Ordered Tasks.** For models \( \mathcal{M}_3, \mathcal{M}_4, \) and \( \mathcal{M}_5 \), reward \( r((3,3), a) = 1 \) for all actions \( a \). In model \( \mathcal{M}_3 \), the agent visits \( a \) and then \( b \) in that order most of the time (see video mu3_1.mp4). Very rarely, the agent narrowly misses one
of the goals due to the stochasticity in transitions and partial observability (see video mu3_2.mp4). In model $M_4$, the agent is almost always successful in satisfying the constraint and maximizing the reward (see video mu4_1.mp4). In model $M_5$, we see both successes (see video mu5_1.mp4) and failures (see video mu5_2.mp4). However, the failure probability is within the threshold, as suggested by Table 1.

**E.1.2 Predicate Uncertainty**

In the experiments of this section, there are two possible locations for object $b$: $(3,0)$ and $(0,3)$. In both cases, whenever the agent is 'far' away (Manhattan distance greater than 1) from the object $b$, it gets an observation 'F' indicating that it is far with probability 1. When the object is at the bottom left and the agent is adjacent to it, the agent gets an observation 'C' with probability 0.9 indicating that the object is close. However, if object $b$ is at the top right and the agent is adjacent to it, the agent gets an observation 'C' only with probability 0.1. Therefore, the detection capability of the agent is stronger when the object is in the bottom-left location as opposed to when it is in the top-right location.

**Reach-Avoid Tasks.** In model $M_8$, reward $r((3,0), a) = 2$ and $r((0,3), a) = 4$ for all actions $a$. In this model, generally, the agent first collects some information from the bottom-left, reaches $a$, and goes to the rewarding location that is not an obstacle (see videos mu8_1.mp4, mu8_2.mp4, mu8_3.mp4). We see rare instances where the agent completely ignores the constraint and maximizes the reward (see video mu8_4.mp4).

**Ordered Tasks.** In model $M_9$, reward $r((0,0), a) = 2$ for all actions $a$. In this model, we observe that the agent mostly succeeds in satisfying the constraint and maximizing the reward (see videos mu9_1.mp4 and mu9_2.mp4).
E.2 HYPER-PARAMETERS AND RUNTIMES

The parameter $\delta$ in all the experiments is chosen in the following manner: i) We first solve a POMDP problem in which we are only interested in maximizing the probability of satisfaction of the LTL$_f$ constraint. Let this probability be denoted by $p_{\text{max}}$. The SARSOP solver gives concrete approximation bounds on its solution, and therefore, on our estimate of $p_{\text{max}}$. ii) Since any threshold $1 - \delta$ larger than $p_{\text{max}}$ is infeasible, we choose a $\delta$ such that $1 - \delta$ is around $0.9p_{\text{max}}$. The values $\eta$ and $B$ are hyperparameters in our experiments. The $\eta$ suggested by Theorem 2 in our paper is guaranteed to result in convergence, but in practice, slightly larger values of $\eta$ can lead to faster convergence.

In Table 1 we provide additional hyper-parameters that were used in our experiments. The parameter $\text{simu}$ denotes the number of Monte-Carlo simulations that were used to estimate the constraint in each iteration. $T_{\text{solve}}$ is the total time (over $K$ iterations) spent in solving the unconstrained POMDP using the SARSOP solver [Kurniawati et al. 2008]. $T_{\text{simu}}$ is the total time spent in simulating policies generated by the SARSOP solver. $T_{\text{total}}$ is the overall computation time for that model.

Most of our models have a state size of 16 ($4 \times 4$). However, the runtime (see Table 1) for these models is drastically different. This is because of two factors: (i) the DFA size and (ii) the complexity of the POMDP problem. The size of the DFA can be large for a complex task. This naturally scales up the state space of the product POMDP. SARSOP returns an alpha-vector policy [Kurniawati et al. 2008]. When the POMDP is complex, the alpha-vector policy returned by SARSOP may have many alpha vectors. This would imply that whenever the agent has to make a decision, it needs to solve a fairly large maximization problem. This makes the simulations time-consuming.

| Model | Spec | $|S|$ | $|Q|$ | $R^{\omega}(\bar{\mu})$ | $R^{f}(\bar{\mu})$ | $1 - \delta$ | $B$ | $\eta$ | $K$ | $\text{simu}$ | $T_{\text{solve}}$ | $T_{\text{simu}}$ | $T_{\text{total}}$ |
|-------|------|-----|-----|----------------|----------------|-------------|---|------|-----|----------|---------------|-------------|--------------|
| $\mathcal{M}_1$ | $\varphi_1$ | 16 | 3 | 1.72 | 0.75 | 0.75 | 5 | 2 | 100 | 200 | 142 | 3518 | 3661 |
| $\mathcal{M}_2$ | $\varphi_1$ | 64 | 3 | 0.95 | 0.70 | 0.70 | 8 | 2 | 50 | 100 | 17299 | 7825 | 25125 |
| $\mathcal{M}_3$ | $\varphi_2$ | 16 | 3 | 0.83 | 0.76 | 0.75 | 5 | 2 | 100 | 200 | 158 | 3614 | 3773 |
| $\mathcal{M}_4$ | $\varphi_1$ | 16 | 4 | 0.80 | 0.71 | 0.70 | 6 | 2 | 100 | 200 | 1893 | 14534 | 16428 |
| $\mathcal{M}_5$ | $\varphi_2$ | 16 | 4 | 0.83 | 0.71 | 0.70 | 6 | 2 | 100 | 200 | 368 | 8440 | 8809 |
| $\mathcal{M}_6$ | $\varphi_1$ | 16 | 4 | 1.01 | 0.79 | 0.80 | 10 | 2 | 100 | 200 | 109 | 718 | 828 |
| $\mathcal{M}_7$ | $\varphi_1$ | 32 | 10 | 4.28 | 0.82 | 0.80 | 25 | 2 | 50 | 100 | 5865 | 57833 | 63999 |
| $\mathcal{M}_8$ | $\varphi_2$ | 32 | 3 | 2.73 | 0.81 | 0.85 | 20 | 0.02 | 100 | 200 | 370 | 21676 | 22046 |
| $\mathcal{M}_9$ | $\varphi_4$ | 32 | 4 | 1.68 | 0.81 | 0.75 | 10 | 0.2 | 100 | 200 | 973 | 25618 | 26591 |