# On Robust and Adaptive Fidelity Selection for Human-in-the-loop Queues

Piyush Gupta Vaibhav Srivastava

***Abstract*— We consider a human agent servicing a queue of homogeneous tasks. The agent can service a task with normal or high fidelity level, where fidelity refers to the degree of exactness and precision while servicing the task. We assume the parameters of the human's service time distribution depend on the selected fidelity level and her cognitive state and are assumed to be unknown a priori. These parameters are learned online through Bayesian parameter estimation. We formulate a robust adaptive semi-Markov decision process (SMDP) to solve our optimal fidelity selection problem and extend the results on convergence of robust-adaptive Markov decision processes (MDP) to robust-adaptive SMDPs.**

## I. Introduction

Human-in-the-loop systems are pervasive in many safety-critical systems such as robot-assisted inspection, search and rescue, and flight control [1]. Often times, supervisory control environments such as NASA control room, with lack of personnel, can result in high workload condition for the human agents [2]. The performance of the human agents varies with their cognitive load, and hence, it is critical to efficiently manage their cognitive load. This can be achieved by providing optimal fidelity level recommendations to the agent based on her cognitive state and workload. An important challenge is to learn an accurate model of agent's service time, and accordingly adapt the fidelity selection policy, while ensuring robustness to model uncertainty.

There exist approximate models of the human service-time distribution that are estimated using data pooled across multiple human subject experiments [3], [4]. However, due to individual differences, the agent-specific distribution might be unknown a priori. The pooled estimate needs to be adapted online to estimate for the individual. In this paper, we achieve this adaptation via model-based adaptive SMDPs. While such a model-based approach is computationally demanding compared to model-free methods such as Q-learning [5], it is ideal for human agents due to its less data intensive nature.

We study optimal fidelity selection problem for a human agent servicing a stream of homogeneous tasks. The agent's cognitive state evolves based on the selected fidelity level and therefore, impacts her performance in subsequent tasks. We design our reward structure to encourage high quality servicing of tasks with high fidelity, while penalizing the agent for delays in servicing of awaiting tasks. Our previous work [6] elucidates on this trade-off between high quality servicing and resulting service time delays, and provide an optimal policy by formulating an uncertainty-free SMDP. We extend our work in [6] by considering uncertainty in agent's service time distribution.

An SMDP accounts for the system uncertainty through probabilistic state transitions. However, the obtained policy is sensitive to errors in the stochastic models [7], [8]. The large uncertainty in the service time models, especially in the initial stage with limited observation data may lead to sub-optimal policies. Existing methods for reducing transition model uncertainty include risk-constrained MDPs [9] that optimize the Conditional Value-at-Risk, chance-constrained MDPs [10] that provide a probabilistic framework for handling uncertainty in the transition probabilities, and robust MDPs [11], [12]. In this work we consider a human agent with non-memoryless service time distribution, and thus formulate a robust adaptive SMDP to deal with general service time distributions [13], [14] and learn a robust policy.

In this work, we show that the solution of the synchronous and asynchronous value iteration (VI) methods [5] for robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. While there exists a convergence analysis for the robust [15], [16] and adaptive [17] MDPs, such an analysis is missing for robust adaptive SMDPs to the best of our knowledge. A key challenge that we address in comparison to MDPs is the time dependence of the robust adaptive Bellman operator for SMDPs that requires careful comparison between optimal value functions for intermediate SMDPs at different time steps, and the optimal value function for the uncertainty-free SMDP.

The major contributions of this work are fourfold: (i) we pose the optimal fidelity selection problem with uncertain human service time distribution in a robust adaptive SMDP framework, (ii) we continuously improve the service time distribution estimates using Bayesian parametric estimation [18] and utilize it to obtain a robust policy, (iii) we formally show that the solution of both synchronous and asynchronous VI methods for the robust adaptive SMDP converge to the optimal solution for the uncertainty-free SMDP, and (iv) we provide numerical illustrations that show convergence of the robust adaptive SMDP solution to the uncertainty-free SMDP.

This manuscript is structured as follows: in Section II, we provide our problem setup, and formulate it as a robust adaptive SMDP. In Section III, we provide the convergence analysis of the robust SMDP to the uncertainty free SMDP solution. In Section IV, we numerically show that the policy learned from robust adaptive SMDP converges to uncertainty-free SMDP solution. We conclude in Section V.

This work has been supported by NSF Award IIS-1734272.

Piyush Gupta (guptapi1@msu.edu) and Vaibhav Srivastava (vaibhav@egr.msu.edu) are with Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, 48824, USA.

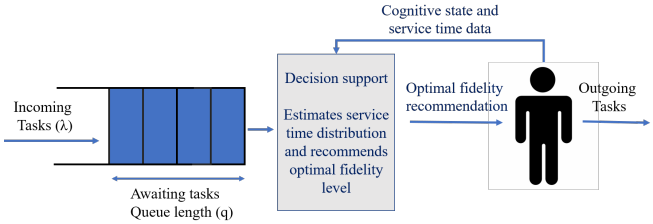


Fig. 1: Schematic of our problem setup. The incoming tasks arrive at a constant arrival rate $\lambda$ and gets stored in a queue. The decision support learns the human service-time distribution based on the online observations and recommends an optimal fidelity level to the human agent based on the system state (queue length and cognitive state).

## II. Background and Problem Formulation

We now present our problem setup, and formulate the optimal fidelity selection problem as a robust adaptive SMDP.

### A. Problem Setup

We consider a human agent with unknown service time distribution servicing a queue of homogeneous tasks. We assume the availability of approximate model for human service time distribution from human subject experiments, which we adapt online to estimate the agent's distribution using Bayesian parametric estimation.

The homogeneous tasks arrive according to a Poisson process with a constant arrival rate $\lambda \in \mathbb{R}_{>0}$, and are stored in a dynamic queue with a maximum capacity $L \in \mathbb{N}$, until serviced by the human agent in a first-come-first-serve discipline. These tasks continuously lose value at a constant rate $c \in \mathbb{R}_{>0}$ while waiting in the queue. The agent can choose to service the tasks with normal or high fidelity level. When the agent services a task with high fidelity, she meticulously looks into the details, which results in higher-quality service but leads to larger service times and increased operator tiredness. We treat the cognitive state of the agent as a lumped parameter that captures psychological factors such as fatigue, stress and situational awareness. We assume that the unknown mean service time of the agent increases with the fidelity level and is a unimodal function of the cognitive state, which is inspired from the experimental psychology literature. For example, according to Yerkes-Dodson law [19], excessive stress overwhelms the operator and too little stress leads to reduction in vigilance. Hence, the human performance is optimal for some intermediate cognitive state.

Fig. 1 shows a schematic of our problem setup. We are interested in design of a decision support system that continuously learns the human service-time distribution and assists the agent by recommending an optimal fidelity level to service each task. The recommendation is made based on the robust adaptive policy learned by the decision support for a given queue length $q \in \mathbb{Z}_{\geq 0}$ and the human cognitive state. We assume to have real time access to the human cognitive state using, e.g., Electroencephalogram (EEG) measurements (see [20] for measures of cognitive load from EEG data).

### B. Robust Adaptive SMDP formulation

We now model our problem as a robust adaptive SMDP $\Gamma^{RA}$. We focus on the unknown service time distributions and refer the interested readers to [6] for more details on modeling of uncertainty-free distributions.

We consider a finite state space $\mathcal{S} := \{(q, \mathrm{cog}) |\ q \in \{0, 1, ..., L\},\ \mathrm{cog} \in \mathcal{C} := \{i/N\}_{i\in\{0,...,N\}}$, for some $N \in \mathbb{N}$, where cog represents the lumped cognitive state. We consider five possible actions for the agent given by: (i) ***Waiting (W)***, when the queue is empty, (ii) ***Resting (R)***, which provides the resting time for the human operator to reach the optimal cognitive state when tired, (iii) ***Skipping (S)***, which allows the operator to skip a task to reduce the queue length and thereby focus on newer tasks, (iv) ***Normal Fidelity (N)*** for servicing the task with normal fidelity, and (v) ***High Fidelity (H)*** for servicing the task more carefully with high precision. Hence, a set of admissible actions $\mathcal{A}_s$ for each state $s \in \mathcal{S}$ is given by: (i) $\mathcal{A}_s := \{W \mid s \in \mathcal{S},\ q = 0\}$ when queue is empty, (ii) $\mathcal{A}_s := \{\{R, S, N, H\} |\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and $\mathrm{cog} > \mathrm{cog}^*$, where $\mathrm{cog}^* \in \mathcal{C}$ is the optimal cognitive state, and (iii) $\mathcal{A}_s := \{\{S, N, H\} |\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and $\mathrm{cog} \leq \mathrm{cog}^*$.

Let $\tau$ be the sojourn time spent in state $s$. The sojourn time distribution $\mathbb{P}(\tau |\ s, a)$ represents the service time while servicing the task with normal or high fidelity, resting time, constant time of skip, and time until the next task arrival while waiting. We model service time distributions in Section II-C and refer the readers to [6] for details on resting and waiting time. We define a state transition distribution $\mathbb{P}(s' |\ \tau, s, a)$ from state $s$ to $s'$ conditioned on an action $a \in \mathcal{A}_s$ and sojourn time $\tau$ spent in state $s$. This distribution involves the transition in queue length which is given by Poisson distribution and the cognitive dynamics which we model as a Markov chain such that the cognitive state cog increases with high probability when the operator is busy ($a \in \{N, H\}$), and decreases when the operator is idle ($a \in \{R, W\}$). For a detailed description of the modeling of cognitive dynamics, we refer the interested readers to [6].

For each task, the human agent receives a high (low) immediate reward for servicing the task with high (normal) fidelity, and no reward for not servicing the task. Furthermore, the agent incurs a penalty at a constant rate $c \in \mathbb{R}_{>0}$ for each task waiting in the queue. Hence, it can be shown that the expected net immediate reward received by the agent for selecting an action $a$ in state $s$ is given by:

$$R(s, a) = r(s, a) - \sum_{\tau} \mathbb{P}(\tau | s, a) c \Big(\frac{2q + \lambda\tau}{2}\Big)\tau, \quad (1)$$

where $r : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}_{\geq 0}$ is the reward defined by: (i) $r(s, a) = r_H$, if $a = H$; (ii) $r(s, a) = r_N$, if $a = N$; and $r(s, a) = 0$, if $a \in \{W, R, S\}$, with $r_H, r_N \in \mathbb{R}_{\geq 0}$ and $r_H > r_N$, and $\sum_\tau \mathbb{P}(\tau | s, a) c\tau \Big(\mathbb{E}\Big[\frac{q+q'}{2}\Big|\ \tau, s, a\Big]\Big)$ is the expected penalty due to tasks waiting in the queue.

### C. Modeling human service time and uncertainity set

There are many approximate models used for modeling service time distribution $\mathbb{P}(\tau | s, a)$ for the human agents, most common being lognormal [3] and inverse Gaussian distribution [4]. We assume that the agent's service time distribution follows a log-normal distribution $\mathrm{Lognormal}(\mu, \sigma^2)$

with unknown parameters $\mu$ and $\sigma^2$, that are the functions of the cognitive state and fidelity-level. We utilize the Bayesian parameter estimation with a normal-inverse-chi-squared prior [21] to estimate the distribution parameters using online observations. Using the prior distribution $NI\chi^2(\mu_0, \kappa_0, \nu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma^2/\kappa_0) \times \chi^{-2}(\sigma^2|\nu_0, \sigma_0^2)$ for the parameters $\mu$ and $\sigma^2$, and $n \in \mathbb{N}$ realizations from $\text{Lognormal}(\mu, \sigma^2)$, the posterior distribution of $(\mu, \sigma^2)$ is given by:

$$p(\mu, \sigma^2) = NI\chi^2(\mu_n, \kappa_n, \nu_n, \sigma_n^2), \quad \text{where} \tag{2}$$

$$\mu_n = \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_n},\ \kappa_n = \kappa_0 + n,\ \nu_n = \nu_0 + n,$$
$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0\sigma_0^2 + \sum_i (x_i - \overline{x})^2 + \frac{n\kappa_0}{\kappa_0 + n}(\mu_0 - \overline{x})^2\right),$$

$x_i,\ i \in \{1, \ldots, n\}$ are the service time samples and $\bar{x}$ is the sample mean. Hence, the posterior distribution at any time $t$ can be computed to recursively estimate the model parameters $\mu$ and $\sigma^2$ by using the online observations. Let $\hat{\mathbb{P}}_t(\tau|s, a)$ be the estimate of the service time distribution at time $t$. Note that the estimate $\hat{\mathbb{P}}_t(\tau|s, a)$ can be used to estimate $\mathbb{P}(s', \tau|s, a)$ and $R(s, a)$, resulting in estimates $\hat{\mathbb{P}}_t(s', \tau|s, a)$ and $\hat{R}_t(s, a)$ at time $t$.

The uncertainty in the human service time models could be large, especially in the initial stage with limited observation data, which may lead to suboptimal policies. This can be mitigated through the use of robust SMDP. The robust SMDP optimizes the worst-case performance to obtain robust policy when the joint distribution $\mathbb{P}(s', \tau|s, a)$ lies in an uncertainty set $\mathcal{P}^a$, i.e, $\mathbb{P}(s', \tau|s, a) \in \mathcal{P}^a$. Note that the robust SMDP formulation is not inherently adaptive in nature and does not explicitly use improved transition models that can be learned using online observations to obtain less conservative policy. The robust adaptive SMDP utilizes the latest improved estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ for the joint probability $\mathbb{P}(s', \tau|s, a)$ and reward $R(s, a)$ at time $t$, respectively.

The choice of uncertainty set $\mathcal{P}^a$ is critical for the performance of the robust algorithm. A poor modeling of the uncertainty set increases the computational complexity and could lead to highly conservative robust policy. Hence, a choice of uncertainty set $\mathcal{P}^a$ is typically made such that the robust policy is not overly conservative and the optimization can be performed in a computationally tractable manner. Let $\mathcal{D}$ be the observation data up to time $t$. To construct an uncertainty set $\mathcal{P}_t^a$ at time $t$, random samples are generated from the posterior distribution of $(\mu, \sigma^2)$ to construct a set $\Delta_t$ comprised of matrices $\hat{\mathbb{P}}_t(s', \tau|s, a)$ for each $s \in \mathcal{S}$ and $a \in \mathcal{A}_S$. Finally, a $\Psi_t$-confidence level subset of the transition probabilities $\Delta_t$ defined by

$$\mathcal{P}_t^a(\Psi_t) = \{p_t^{sa} \in \Delta_t : \|p_t^{sa} - \overline{p}_t^{sa}\|_1 \leq \Psi_t,\ s \in \mathcal{S}\}, \tag{3}$$

where $\overline{p}_t^{sa}$ is the nominal transition given by $\overline{p}_t^{sa} = \mathbb{E}[p_t^{sa}|\mathcal{D}]$, is used as a choice for the uncertainty set. We seek to find $\Psi_t$-confidence sets for state transition probability vector for every state-action pair at time $t$. We choose $\Psi_t = \frac{6\alpha}{|\mathcal{S}\|\mathcal{A}_S|\pi^2 t^2}$ such that union bounds applied over each state-action pair and time yield that all state transition probabilities belong to respective confidence sets with at least probability $\alpha$. In the following we choose $\alpha = 0.95$.

Using the uncertainty set $\mathcal{P}_t^a$ constructed at time $t$ based on the latest improved estimates $\hat{\mathbb{P}}_t$, the robust adaptive SMDP solves the following robust Bellman equation (4),

$$V^*(s) = \max_{a \in \mathcal{A}_S} \min_{\hat{\mathbb{P}}_t \in \mathcal{P}_t^a} \left\{\hat{R}_t(s, a) + \sum_\tau \sum_{s'} \gamma^\tau \hat{\mathbb{P}}_t(s', \tau|s, a) V^*(s')\right\}, \tag{4}$$

where $0 < \gamma < 1$ is the discount factor, to obtain a robust policy $\pi^* = \text{argmax}_{a \in \mathcal{A}_S} V(s)$, which optimizes the worst-case performance through minimization with respect to the uncertainty set $\mathcal{P}_t^a$. In the next section, we show that the learned policy through robust adaptive SMDP converges to an optimal policy for the uncertainty-free formulation.

## III. Convergence of Robust Adaptive SMDP

We study the convergence properties of the robust adaptive SMDP under the following assumptions.

(A1) State space $\mathcal{S}$ and actiona space $\mathcal{A}_S$ are finite.
(A2) $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ remains bounded for any $t$.
(A3) $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ converges to their true values $\mathbb{P}$ and $R$, respectively, with probability 1.
(A4) The uncertainty set $\mathcal{P}_t^a$ converges to a singleton estimate $\mathbb{P}$ with probability 1.
(A5) Each admissible action is executed from every state infinitely often.

Since agent's service time distribution is independent of the queue length, assumption (A5) can be relaxed to executing each action at every cognitive state. Furthermore, assumptions (A3) and (A5) can be satisfied by adopting exploration strategies such as Gibbs/Boltzman distribution method for action selection [5], wherein an action is selected with probability proportional to the current estimate of the state-action value function divided by a temperature parameter. The temperature parameter is annealed to ensure that the action selection rule converges over time to a greedy policy with respect to the value function estimate, while ensuring each cognitive state-action pair is selected infinitely often.

Let $T : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ be the Bellman operator for an uncertainty-free SMDP $\Gamma$ defined by:

$$T(V(s)) = \max_{a \in \mathcal{A}_S} \left\{R(s, a) + \sum_\tau \sum_{s'} \gamma^\tau \mathbb{P}(s', \tau|s, a) V(s')\right\}. \tag{5}$$

We first show the convergence of the adaptive asynchronous VI method followed by the convergence of the robust approach. In asynchronous VI, an effective approach to deal with large state spaces, the value of only a subset of states $s \in B_t \subseteq \mathcal{S}$ are updated at any time $t$. The adaptive VI method adapts to the latest estimates of the model to compute the control policy, which is continuously improved through online observations. Hence, the adaptive asynchronous VI update at time $t$ is given by:

$$V_{t+1}(s) = \begin{cases} \hat{T}_t(V_t), & \text{if } s \in B_t, \\ V_t(s), & \text{otherwise}, \end{cases} \tag{6}$$

where $B_t \subseteq \mathcal{S}$ is the subset of states that are updated at time $t$, and $\hat{T}_t$ is the Bellman operator for the SMDP estimate $\hat{\Gamma}_t$ at time $t$, that utilizes the estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ in (5). The set $B_t$ can be chosen using prioritized sweeping [22] for improved computational performance.

*Theorem 1:* Under Assumptions A1-A5, the adaptive asynchronous VI converges to the optimal value function $V^*$ for the uncertainty-free SMDP $\Gamma$ with probability 1.

*Proof:* See Appendix A for the proof. ∎

Let $\mathcal{P}^a$ be the uncertainty set for the probability $\mathbb{P}(s', \tau | s, a)$ for a given action $a \in \mathcal{A}_S$. Then, the robust adaptive asynchronous VI update at time $t$ is given by:

$$V_{t+1}(s) = \begin{cases} \max_{a \in \mathcal{A}_S} \min_{\mathbb{P} \in \mathcal{P}^a} \left\{ \mathcal{J}^a(V_t) \right\}, & \text{if } s \in B_t, \\ V_t(s), & \text{otherwise,} \end{cases} \tag{7}$$

where for a given $a \in \mathcal{A}_S$, $\mathcal{J}^a : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ is given by

$$\mathcal{J}^a(V(s)) = R(s,a) + \sum_{\tau} \sum_{s'} \gamma^{\tau} \mathbb{P}(s', \tau | s, a) V(s'). \tag{8}$$

Let $T_r : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ be the robust Bellman operator for SMDP $\Gamma$ defined by:

$$T_r(V(s)) = \max_{a \in \mathcal{A}_S} \min_{\mathbb{P} \in \mathcal{P}^a} \left\{ \mathcal{J}^a(V(s)) \right\}. \tag{9}$$

*Theorem 2:* Under Assumptions A1-A5, the robust adaptive asynchronous VI converges to the optimal value function $V^*$ for the uncertainty-free SMDP $\Gamma$ with probability 1. In addition, for $\Psi_t = \frac{6\alpha}{|\mathcal{S}||\mathcal{A}_S|\pi^2 t^2}$, the union bounds applied over each state-action pair and time yield that at any time, the obtained policy is robust with respect to uncertainty in service time distributions with at least probability $\alpha$.

*Proof:* See Appendix B for the proof. ∎

## IV. Numerical Illustrations

Fig. 2a and 2b shows an optimal policy and the optimal value function obtained using VI algorithm for the uncertainty-free SMDP $\Gamma$ with $\text{cog}^* = 0.6$ as the optimal cognitive state. The optimal policy selects high fidelity around the $\text{cog}^*$ for small queue lengths, and then transitions to normal fidelity as the queue length increases in sub-optimal cognitive states. The optimal action is to rest when the queue length is large and cognitive state is high. Similarly, skipping of tasks is an optimal action for large queue lengths in sub-optimal cognitive states. The corresponding optimal value function is a decreasing function of $q$ and is a uni-modal function of the cognitive state, the maximum for which occurs at $\text{cog}^*$ for each $q$; see [6] for more details.

Fig. 2c and 2d shows a robust adaptive policy, and the corresponding optimal value function for the uncertain SMDP $\Gamma^{RA}$ obtained from asynchronous VI algorithm. In our numerical illustrations, we only estimate the parameter $\mu$ of the $\text{Lognormal}(\mu, \sigma^2)$ distribution using Bayesian estimation. In case when $\sigma^2$ is unknown, Markov chain Monte Carlo (MCMC) methods [23] can be used to sample from the posterior inverse-chi-squared-distribution to create the uncertainty set $\mathcal{P}_t^a$ at time $t$. An identical policy and value function is obtained in case of synchronous VI algorithm. Hence, the solution of the synchronous and asynchronous VI converges to the optimal solution of uncertainty-free SMDP.

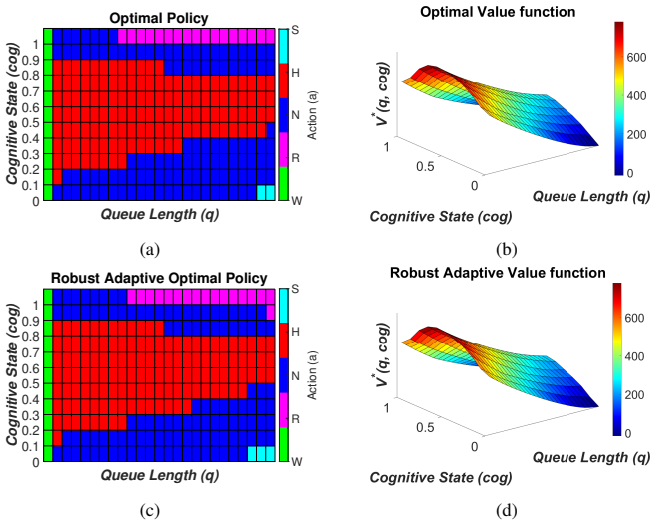


Fig. 2: (a) Optimal policy and (b) optimal value function for the uncertainty-free SMDP $\Gamma$. (c) Robust adaptive optimal policy, and corresponding (d) optimal value function for the uncertain SMDP $\Gamma^{RA}$.

Fig. 3 shows the policy updates for the synchronous and asynchronous robust adaptive algorithm after 4, 8, 32 and 76 iterations, respectively. The synchronous robust adaptive algorithm performs a VI update at each time step in (6) for all states, i.e. $B_t = \mathcal{S}$, while in the asynchronous robust adaptive algorithm, the VI update is performed on a randomly chosen subset $B_t \in S$ at each time step. The asynchronous robust adaptive algorithm converges (80 iterations) much faster than the synchronous robust adaptive algorithm (404 iterations), while both eventually converge to the optimal policy for the uncertainty-free SMDP $\Gamma$ as shown in Fig. 2.

## V. Conclusions

We studied optimal fidelity selection problem for a human agent servicing a stream of homogeneous tasks. The parameters of the service time distribution for the agent are unknown a priori which are learned online through Bayesian parametric estimation. We utilize the robust-adaptive SMDP approach which adapts to the latest estimates of the distribution model, while obtaining robust policy towards the worst-case performance. We formally extend the convergence results of the robust adaptive MDP to robust adaptive SMDP, and show that the solution of the robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. Furthermore, we numerically illustrate the convergence of the synchronous and asynchronous robust adaptive policy to the uncertainty-free optimal policy.

## Appendix

### A. Proof of Theorem 1

A key challenge we address is the time-dependence of the asynchronous adaptive Bellman updates that adapt to the latest estimates of the service-time distribution. Using Lemmas 1 and 2, we upper-bound the difference between optimal value functions for intermediate SMDPs at subsequent time steps, and the optimal value function for the uncertainty-free SMDP, which are used to establish the convergence result.

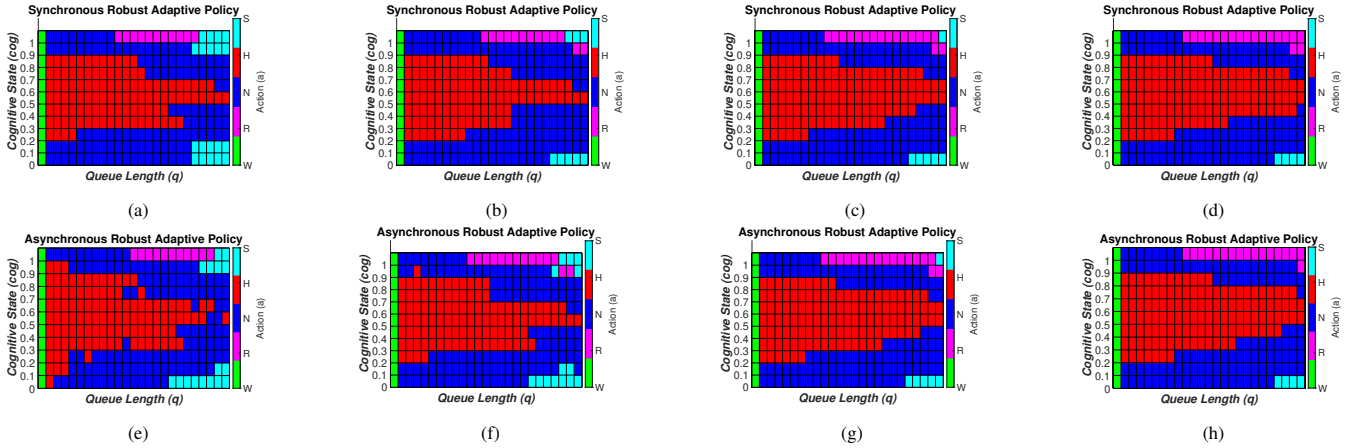


Fig. 3: Policy updates for the synchronous ((a)-(d)) and asynchronous ((e)-(h)) robust adaptive algorithm after 4, 8, 32 and 76 iterations, respectively.

*Theorem 3 (adapted from [24, Chapter 10]):* $T$ is a contraction mapping and therefore, there exists a unique fixed point satisfying $T(V^*) = V^*$, where $V^*$ is the optimal value function for the uncertainty-free SMDP $\Gamma$.

Let $V_t^*$ be the optimal value function for SMDP $\hat{\Gamma}_t$ defined by the estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$. Therefore, $V_t^* = \hat{T}_t(V_t^*)$. Let $\|\cdot\|$ be the max-norm given by $\|v\| = \max\{|v_1|, |v_2| \ldots, |v_n|\}$, for any vector $v = (v_1, v_2, \ldots, v_n)$. Let $\tau_{\min} \geq 1$ be the minimum number of time steps spent in any state $s \in \mathcal{S}$, for any action $a \in \mathcal{A}_{\mathcal{S}}$.

*Lemma 1:* For any state $s \in \mathcal{S}$, following statements hold:

(i) $|\hat{T}_t(V_1(s)) - \hat{T}_t(V_2(s))| \leq \gamma^{\tau_{\min}} \|V_1 - V_2\|$ for any $s \in \mathcal{S}$, where the Bellman operator $\hat{T}_t$ at any time $t$ is applied on value function estimates $V_1$ and $V_2$.

(ii) $|V_{t+1}(s) - V_t^*(s)| \leq \gamma^{\tau_{\min}} \|V_t - V_t^*\|$, if $s \in B_t$.

*Proof:* We prove the first statement. For any $s \in \mathcal{S}$, let $\mathcal{V} := |\hat{T}_t(V_1(s)) - \hat{T}_t(V_2(s))|$. Therefore,

$$\mathcal{V} \leq \max_{a \in \mathcal{A}_{\mathcal{S}}} \left| \sum_{\tau} \sum_{s'} \gamma^{\tau} \hat{\mathbb{P}}_t(s', \tau | s, a)(V_1(s') - V_2(s')) \right|$$
$$\leq \|V_1 - V_2\| \sum_{\tau} \gamma^{\tau} \hat{\mathbb{P}}_t(\tau | s, a) \leq \gamma^{\tau_{\min}} \|V_1 - V_2\|.$$

The second statement follows from the first by noting that $|V_{t+1}(s) - V_t^*(s)| = |\hat{T}_t(V_t) - \hat{T}_t(V_t^*)|$ if $s \in B_t$. ∎

*Lemma 2:* Under Assumptions A1-A5, for any given $\epsilon > 0$, there exists a time $\tilde{t}$, such that for any $t \geq \tilde{t}$, (i) $\|V_t^* - V^*\| \leq \epsilon$, and (ii) $\|V_{t+1}^* - V_t^*\| \leq 2\epsilon$ with probability 1.

*Proof:* Since the estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ are assumed to be bounded at any time $t$ (assumption (A2)), the value function estimate $V_t$ at any time $t$ also remains bounded. Furthermore, using assumption (A3), $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ converge to their true values $\mathbb{P}$ and $R$, respectively, with probability 1, i.e, for any $\epsilon_1, \epsilon_2 > 0$, there exists a time $\tilde{t}_0$ such that, for any $t \geq \tilde{t}_0$, $|\hat{R}_t - R| \leq \epsilon_1$ and $|\hat{p}_t^{ij} - p^{ij}| \leq \epsilon_2$, where $\hat{p}_t^{ij}$ and $p^{ij}$ are the elements of $\hat{\mathbb{P}}_t$ and $\mathbb{P}$, respectively. In asynchronous VI, the value of only the states $s \in B_t \subseteq \mathcal{S}$ are updated at any time $t$, however, each state is assumed to be updated infinitely often. Therefore, the sequence (6) converges with probability 1, i.e., for any $\epsilon_3 > 0$, there exists a time $\tilde{t}_1 \geq \tilde{t}_0$ such that $\|V_{t+1} - V_t\| \leq \epsilon_3$, for $t \geq \tilde{t}_1$.

Consider $s \in B_t$ such that $V_{t+1}(s) = \hat{T}_t(V_t(s))$. Let $\mathcal{V}_t^* := \|V_t^* - V^*\|$, where we only consider the states $s \in B_t$ in the vectors $V_t^*$ and $V^*$. Therefore,

$$\mathcal{V}_t^* \leq \|V_t^* - V_{t+1}\| + \|V_{t+1} - V^*\| =: \mathcal{Z}_1 + \mathcal{Z}_2. \quad (10)$$

$\mathcal{Z}_1 = \|V_t^* - V_{t+1}\|$ is upper bounded by:

$$\mathcal{Z}_1 \leq \|V_t^* - \hat{T}_t(V_{t+1})\| + \|\hat{T}_t(V_{t+1}) - \hat{T}_t(V_t)\| =: \mathcal{Z}_1^1 + \mathcal{Z}_1^2. \quad (11)$$

Since $\mathcal{Z}_1^1 = \|V_t^* - \hat{T}_t(V_{t+1})\| = \|\hat{T}_t(V_t^*) - \hat{T}_t(V_{t+1})\|$, from statement (i) of Lemma 1, we get

$$\mathcal{Z}_1^1 \leq \gamma^{\tau_{\min}} \|V_t^* - V_{t+1}\|, \text{ and } \mathcal{Z}_1^2 \leq \gamma^{\tau_{\min}} \|V_{t+1} - V_t\|. \quad (12)$$

Substituting (12) in (11), we get:

$$\mathcal{Z}_1 \leq \frac{\gamma^{\tau_{\min}}}{1 - \gamma^{\tau_{\min}}} \|V_{t+1} - V_t\|. \quad (13)$$

$\mathcal{Z}_2 = \|V^* - V_{t+1}\|$ is upper bounded by:

$$\mathcal{Z}_2 \leq \|V^* - T(V_{t+1})\| + \|T(V_{t+1}) - V_{t+1}\| =: \mathcal{Z}_2^1 + \mathcal{Z}_2^2. \quad (14)$$

Again using statement (i) of Lemma 1, we have

$$\mathcal{Z}_2^1 \leq \gamma^{\tau_{\min}} \|V^* - V_{t+1}\| = \gamma^{\tau_{\min}} \mathcal{Z}_2. \quad (15)$$

Furthermore, $\mathcal{Z}_2^2 = \|T(V_{t+1}) - V_{t+1}\| = \|T(V_{t+1}) - \hat{T}_t(V_t)\|$ is upper bounded by:

$$\mathcal{Z}_2^2 \leq \max_{a \in \mathcal{A}_{\mathcal{S}}} \|R - \hat{R}_t\| + \max_{a \in \mathcal{A}_{\mathcal{S}}} \Big\| \sum_{\tau} \sum_{s'} \gamma^{\tau} \mathbb{P}(s', \tau | s, a) V_{t+1}(s') -$$
$$\sum_{\tau} \sum_{s'} \gamma^{\tau} \hat{\mathbb{P}}_t(s', \tau | s, a) V_t(s') \Big\|. \quad (16)$$

Recall that for any $t \geq \tilde{t}_0$, $|\hat{R}_t - R| \leq \epsilon_1$ and $|\hat{p}_t^{ij} - p^{ij}| \leq \epsilon_2$. Therefore, for $t \geq \tilde{t}_0$, (16) is upper bounded by:

$$\mathcal{Z}_2^2 \leq \epsilon_1 + \max_{a \in \mathcal{A}_{\mathcal{S}}} \Big\| \sum_{\tau} \sum_{s'} \gamma^{\tau} |\mathbb{P}(s', \tau | s, a) - \hat{\mathbb{P}}_t(s', \tau | s, a)| V_{t+1}(s') \Big\|$$
$$+ \Big\| \sum_{\tau} \sum_{s'} \gamma^{\tau} \hat{\mathbb{P}}_t(s', \tau | s, a)(V_{t+1}(s') - V_t(s')) \Big\|$$
$$\overset{(1^*)}{\leq} \epsilon_1 + \|V_{t+1}\| \sum_{\tau} \sum_{s'} \gamma^{\tau} \epsilon_2 + \gamma^{\tau_{\min}} \|V_{t+1} - V_t\|,$$
$$= \epsilon_1 + \frac{\epsilon_2 N \gamma^{\tau_{\min}} \|V_{t+1}\|}{1 - \gamma} + \gamma^{\tau_{\min}} \|V_{t+1} - V_t\|, \quad (17)$$

where $N$ is the size of the finite state-space $\mathcal{S}$, and $(1^*)$

follows from $|\hat{p}_t^{ij} - p^{ij}| \leq \epsilon_2$ and statement (i) of Lemma 1. Substituting (15) and (17) in (14), we get:

$$\mathcal{Z}_2 \leq \frac{\gamma^{\tau_{\min}}}{1-\gamma^{\tau_{\min}}}\|V_{t+1} - V_t\| + f(\|V_{t+1}\|), \qquad (18)$$

where $f(\|V_{t+1}\|) := \frac{1}{1-\gamma^{\tau_{\min}}}\left(\epsilon_1 + \frac{\epsilon_2 N \gamma^{\tau_{\min}}\|V_{t+1}\|}{1-\gamma}\right)$ is bounded for bounded $\|V_{t+1}\|$ and $f(\|V_{t+1}\|) \mapsto 0$, when $\epsilon_1, \epsilon_2 \mapsto 0$. Substituting (13) and (18) in (10), we get

$$\mathcal{V}_t^* \leq \frac{2\gamma^{\tau_{\min}}}{1-\gamma^{\tau_{\min}}}\|V_{t+1} - V_t\| + f(\|V_{t+1}\|). \qquad (19)$$

Recall that for any $\epsilon_3 > 0$, there exists $\tilde{t}_1 \geq \tilde{t_0}$ such that $|V_{t+1} - V_t\| \leq \epsilon_3$, for $t \geq \tilde{t}_1$. Choosing $\epsilon_1, \epsilon_2 \mapsto 0$, and $\epsilon_3 = \frac{\epsilon(1-\gamma^{\tau_{\min}})}{2\gamma^{\tau_{\min}}}$, we get that there exists $\tilde{t}_1$, such that $\|V_{t+1} - V_t\| \leq \frac{\epsilon(1-\gamma^{\tau_{\min}})}{2\gamma^{\tau_{\min}}}$, and $\mathcal{V}_t^* = \|V_t^* - V^*\| < \epsilon$, for any $t \geq \tilde{t}_1$.

Recall that $\mathcal{V}_t^*$ only considers states $s \in B_t$. However, since each state is updated infinitely often, there exists $\tilde{t}$ such that $\|V_t^* - V^*\| < \epsilon$, for any $t \geq \tilde{t}$. Furthermore, for $t \geq \tilde{t}$, $\|V_{t+1}^* - V_t^*\| \leq \|V_{t+1}^* - V^*\| + \|V_t^* - V^*\| \leq 2\epsilon$. ■

*Proof of Theorem 1*: For any state $s \in \mathcal{S}$, define a sequence $\{t_i^s\}_{i=1}^{\infty}$ of times at which state $s$ is updated by the asynchronous VI, and consider the updates after time $\tilde{t}$, i.e., consider the sequence $\{t_i^s\}_{i=k}^{\infty}$ such that $t_k^s \geq \tilde{t}$. Let $\mathcal{V}_t(s) := |V_{t+1}(s) - V_t^*(s)|$. Therefore, using statement (ii) of Lemma 1, $\mathcal{V}_{t_{i+1}^s} = |V_{t_{i+1}^s+1}(s) - V_{t_{i+1}^s}^*(s)| \leq \gamma^{\tau_{\min}}\|V_{t_{i+1}^s} - V_{t_{i+1}^s}^*\|$, and therefore upper-bounded by:

$$\begin{aligned}\mathcal{V}_{t_{i+1}^s}(s) &\leq \gamma^{\tau_{\min}}(\|V_{t_{i+1}^s} - V_{t_i^s}^*\| + \|V_{t_i^s}^* - V_{t_{i+1}^s}^*\|) \\ &\overset{(1^*)}{\leq} \gamma^{\tau_{\min}}\left(\|\mathcal{V}_{t_i^s}\| + 2\epsilon\right), \qquad (20)\end{aligned}$$

for $i \geq k$, where $(1^*)$ follows from statement (ii) of Lemma 2. From (20), we get the following recursion:

$$\|\mathcal{V}_{t_{i+1}^s}\| \leq \gamma^{\tau_{\min}}\left(\|\mathcal{V}_{t_i^s}\| + 2\epsilon\right), \qquad (21)$$

for $i \geq k$.

Recursively performing (21) to obtain upper-bounds on $\|\mathcal{V}_{t_j^s}\|$, for $j = k, \ldots i$, and substituting in (20), we get:

$$\mathcal{V}_{t_{i+1}^s}(s) \leq \gamma^{(i+1)\tau_{\min}}\|\mathcal{V}_{t_k^s}\| + \frac{2\gamma^{\tau_{\min}}(1-\gamma^{(i+1)\tau_{\min}})}{1-\gamma^{\tau_{\min}}}\epsilon,$$

In the limit $i \to \infty$, $\mathcal{V}_{t_{i+1}^s} = |V_{t_{i+1}^s+1}(s) - V_{t_{i+1}^s}^*(s)| \leq \epsilon_4$, where $\epsilon_4 := \frac{2\gamma^{\tau_{\min}}}{1-\gamma^{\tau_{\min}}}\epsilon$, and $\epsilon_4 \mapsto 0$ for $\epsilon \mapsto 0$. Since for each $s \in \mathcal{S}$, $V_{t_{i+1}^s}^*(s)$ converges to $V^*(s)$ (Lemma 2), and $\epsilon$ is arbitrary, $V_t(s)$ converges to $V^*(s)$ for any $s$. ■

### *B. Proof of Theorem 2*

We prove Theorem 2 using the following Theorem 4.

*Theorem 4:* $T_r$ is a contraction mapping, and hence, there exists a unique fixed point satisfying, $T_r(V) = V$.

*Proof:* The proof follows similar to the case of robust MDPs [15]. ■

*Proof of Theorem 2*: Since $T_r$ is a contraction mapping (Theorem 4), and each state is updated infinitely often, the robust adaptive asynchronous VI converges to a fixed point $T_r(V) = V$. Furthermore, bounded $\mathcal{P}^a$ implies that the value function at any time $t$ remains bounded. Once $\mathcal{P}^a$ converges to the singleton estimate $\mathbb{P}$, the robust adaptive asynchronous VI reduces to the adaptive asynchronous VI. Hence, the proof follows using Theorem 1. ■

## REFERENCES


[1] I. R. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human-robot teaming for search and rescue," *IEEE Pervasive Computing*, vol. 4, no. 1, pp. 72–78, 2005.

[2] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, 2018.

[3] L. F. Bertuccelli and M. L. Cummings, "Operator choice modeling for collaborative uav visual search tasks," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 5, pp. 1088–1099, 2012.

[4] V. Srivastava, P. Holmes, and P. Simen, "Explicit moments of decision times for single- and double-threshold drift-diffusion processes," *Journal of Mathematical Psychology*, vol. 75, no. 2016, pp. 96–109, 2016, special Issue in Honor of R. Duncan Luce.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[6] P. Gupta and V. Srivastava, "Optimal fidelity selection for human-in-the-loop queues using semi-Markov decision processes," in *American Control Conference*, Philadelphia, PA, Jul. 2019, pp. 5266–5271.

[7] C. C. White III and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 42, no. 4, pp. 739–749, 1994.

[8] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, "Bias and variance approximation in value function estimates," *Management Science*, vol. 53, no. 2, pp. 308–322, 2007.

[9] V. Borkar and R. Jain, "Risk-constrained Markov decision processes," *IEEE Transactions on Automatic Control*, vol. 59, no. 9, pp. 2574–2579, 2014.

[10] E. Delage and S. Mannor, "Percentile optimization for Markov decision processes with parameter uncertainty," *Operations Research*, vol. 58, no. 1, pp. 203–213, 2010.

[11] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust Markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[12] L. F. Bertuccelli, A. Wu, and J. P. How, "Robust adaptive Markov decision processes: Planning with model uncertainty," *IEEE Control Systems Magazine*, vol. 32, no. 5, pp. 96–109, 2012.

[13] S. Stidham Jr and R. R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Operations Research*, vol. 37, no. 4, pp. 611–625, 1989.

[14] L. I. Sennott, "Average cost semi-Markov decision processes and the control of queueing systems," *Probability in the Engineering and Informational Sciences*, vol. 3, no. 2, pp. 247–272, 1989.

[15] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[16] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[17] V. Gullapalli and A. G. Barto, "Convergence of indirect adaptive asynchronous value iteration algorithms," in *Advances in Neural Information Processing Systems*, 1994, pp. 695–702.

[18] S. C. Kramer and H. W. Sorenson, "Bayesian parameter estimation," *IEEE Transactions on Automatic Control*, vol. 33, no. 2, pp. 217–222, 1988.

[19] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[20] R. P. Rao, *Brain-Computer Interfacing: An Introduction*. Cambridge University Press, 2013.

[21] K. P. Murphy, "Conjugate bayesian analysis of the Gaussian distribution," University of British Columbia, BC, Tech. Rep., Oct 2007.

[22] L. Li and M. Littman, "Prioritized sweeping converges to the optimal value function," Rutgers University, NJ, Tech. Rep. DCS-TR-631, June 2008.

[23] S. Chib, "Markov chain Monte Carlo methods: computation and inference," *Handbook of Econometrics*, vol. 5, pp. 3569–3649, 2001.

[24] A. Gosavi, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, 2nd ed. Springer, 2015.