Motivation and main challenges in mobile health: the importance of continuing task adaptation of RLSVI is useful for tackling two of the main challenges in mobile health: the importance of efficient learning and online learning as a continuing task. Tableaux simulate the task of balancing reward and burden.

**Problem Formulation**

**Motivation**

- Want to model a simplified mobile health problem of determining when to ping the user.
- Current memory window affects the user’s immediate reward.
- Waiting decreases short-term burden.

**Problem Formulation (cont.)**

At time step $t$:

- $s$ is $(x_0, s_0) \in \{0, \ldots, d(t+1)\} \times \{0, \ldots, T\}$
- $a_t \in \{0, 1\}$ where 1 corresponds to pinging user and 0 corresponds to waiting.
- Transition function is $T(a_t, s_t) \rightarrow s_{t+1}$ such that:
  
  $T(s_t, a_t) = s_{t+1}$ with probability $1 - \epsilon$, and
  
  $T(s_t, a_t)$ otherwise, with $p_{s_t}$ being the number of pings in last 7 time steps.
  
  - Reward observed from taking action $a_t$ at state $s_t$ is:
    
    $r_t = R(T(s_t, a_t)) + \epsilon_t$
    
    where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is random noise and our reward function is:
    
    $R(s^0, s^1) = \sum_{1 \leq t \leq T} r_t$
    
    where $\epsilon_t \geq 0$, $r_t \geq 1$, and $\epsilon$ are fixed constants.

**Algorithm**

**Algorithm 1: Continuation task RLSVI**

Inputs:

- $\Phi$, $\rho$, $\epsilon_t$, $r_T > 0$, $\lambda > 0$, $N > 0$, $H > 0$

Output:

- $\hat{s}_t$, $\hat{a}_t$, $R_t$, $r_t > 0$, $H > 1$ for $t = H - 1, \ldots, 1$	

Generate regression problem $A \in \mathbb{R}^{T N} 

\theta_t = \max_{x_t} \Phi(x_t, a_t)$ if $H = 1$ or

 otherwise

Bayesian linear regression for the value function $\hat{s}_{t+1} = \mathbb{E}[a_t | A_t]$

Sample $\hat{s}_{t+1} \sim \mathcal{N}(\hat{s}_t, \Sigma_t)$ from Gaussian posterior

**Algorithm 2: Policy learning procedure for continuing tasks**

Inputs:

- Features $\Phi$; $\rho > 0$, $\lambda > 0$, $H$

Output:

- $\hat{s}_t$, $\hat{a}_t$, $R_t$, $r_t > 0$, $H > 1$ for $t = H - 1$

Use random $\theta_0$, $\theta_{H-1} \sim \mathcal{N}(0, 1)$/Algorithm 1

Obtain $\hat{s}_t$, $\hat{a}_t$ from Algorithm 1

Sample $s_0 | \arg \max_{s_0} \mathbb{E}[\Phi(x_t, a_t)]$

Observe $r_t$ and $s_{t+1}$

**Results**

**Baseline**

- Compare with Least-Squares Value Iteration (LSVI) with $\epsilon$-greedy exploration as our baseline in these experiments

**Evaluation Metrics**

- Final reward: the average total reward obtained in last 100 time steps
- Burden size
- Maximum final reward
- Minimum MSE

**Conclusions**

- Overall, these simulations show that our proposed continuing task extension of RLSVI is more effective than LSVI for the purpose of mobile health problems: it efficiently learns a near-optimal policy that achieves higher rewards, lower MSE, and better robustness over varying hyperparameters than LSVI.

- However, although the testbed was made to reflect a simplified mobile health problem, it was unsuccessful at directly testing the need to explore efficiently – this made it so that success was limited and generalized poorly to a larger state space.

- Thus our next step is to test this algorithm on a testbed that will reward efficient exploration, since our ultimate goal is to apply this algorithm to mobile health problems where efficient exploration is necessary.

**Future Work**

- Since we would like to further investigate our algorithm’s ability to explore efficiently, it may be useful to revert to a model-up approach starting with simpler testbeds and sparse rewards, such as the modified chain setting from [1], and increasing in complexity.

- Explicitly trying an approach that requires deep exploration approach such as a modified chain setting from [1], and increasing in complexity.

**References**

