APPENDIX
Learning Adaptive Control Policies for Remote Patient Monitoring

APPENDIX I

A. Proof for Theorem 1

Proof. Theorem 1 from [1] states that when $H \to \infty$, the optimal policy is π_o when

$$\gamma(\lambda_i - \lambda_o) \leq C_i/C_c. \tag{1}$$

Here π_o is the policy where the patient always stays in ordinary monitoring, i.e., threshold-based policy $\pi_{t,\bar{h}}$ with $\bar{h} = 0$. Theorem 2 from [1] states that the optimal policy is $\pi_{t,\bar{h}}$ with some $\bar{h}$ when the following are satisfied -

$$\gamma(\lambda_i - \lambda_o) > C_i/C_c, \tag{2}$$

and

$$\frac{\gamma\mu_o(1+\gamma\mu_0)}{1-\gamma^2\lambda_o\mu_o} \leq 1. \tag{3}$$

Note that (2) is the complement of (1) and hence one of them is always true. Hence we only need to verify that equation (3) is satisfied. Condition (3) is satisfied when

$$\gamma \leq \frac{-1+\sqrt{1+4/\mu_o}}{2},$$

which is satisfied for sufficiently large $1/\gamma$. This completes the proof for Theorem 1. □

B. Experiment Details

In this section, we present complete experimental details for the plots presented in the main paper. We set $\sigma = H/4$, $h_{\text{peak}} = H - 3$ across all experiments in this section.

We set σ as $H/4$ because with $\sigma = 1$ the weight decays too quickly across health levels, leaving little exploration away from h_{peak}. We choose $h_{\text{peak}} = H - 3$ to minimize exploration in low-h states for safety while encouraging it in the upper–mid range.

a) **Figure 2a** - *Average time to reach critical health for different initial thresholds*
- Each epoch comprised of 75 timesteps.
- True Parameters: $H = 10, \lambda_o = 0.2, \lambda_i = 0.5, C_o = 0, C_i = 15, C_c = 500, \gamma = 0.9$. The optimal threshold is 4.
- Priors: $\lambda_{o,p} = 0.2, \lambda_{i,p} = 0.5, C_{o,p} = 0, C_{i,p} \in \{4, 9, 18, 31\}$. The initial thresholds are $8, 6, 4, 2$ for $C_{i,p} = 4, 9, 18, 31$, respectively.
- Prior Strength: $n_0 = 1000$
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha = 10,\ h_{\text{peak}} = H - 3,\ \sigma = \frac{H}{4}$.

b) **Figure 2b** - *Average discounted cost for different initial thresholds*
- Each epoch comprised of 75 timesteps.
- True Parameters: $H = 10, \lambda_o = 0.2, \lambda_i = 0.5, C_o = 0, C_i = 15, C_c = 500, \gamma = 0.9$. The optimal threshold is 4.
- Priors: $\lambda_{o,p} = 0.2, \lambda_{i,p} = 0.5, C_{o,p} = 0, C_{i,p} \in \{4, 9, 18, 31\}$. The initial thresholds are $8, 6, 4, 2$ for $C_{i,p} = 4, 9, 18, 31$, respectively.
- Prior Strength: $n_0 = 1000$
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha = 10,\ h_{\text{peak}} = H - 3,\ \sigma = \frac{H}{4}$.

c) **Figure 2c** - *Percentage of converged runs after each epoch for different prior strengths*
- Each epoch comprised of 75 timesteps.
- True Parameters: $H = 10, \lambda_o = 0.2, \lambda_i = 0.5, C_o = 0, C_i = 30, C_c = 500, \gamma = 0.9$. The optimal threshold is 2.
- Priors: $\lambda_{o,p} = 0.1, \lambda_{i,p} = 0.2, C_{o,p} = 0, C_{i,p} = 3$. The initial threshold is 9.
- Prior Strength: $n_0 \in \{10, 100, 1000, 10000\}$
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha = 1000,\ h_{\text{peak}} = H - 3,\ \sigma = \frac{H}{4}$.

d) **Figure 3a** - *Percentage of converged runs after each epoch for different optimism levels*
- Each epoch comprised of 150 timesteps.
- True Parameters: $H = 10, \lambda_o = 0.2, \lambda_i = 0.5, C_o = 0, C_i = 1, C_c = 500, \gamma = 0.9$. The optimal threshold is 9.
- Priors: $\lambda_{o,p} = 0.1, \lambda_{i,p} = 0.2, C_{o,p} = 0, C_{i,p} = 20$. The initial threshold is 0.
- Prior Strength: $n_0 = 150$
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha \in \{10, 40, 60, 100, 1000\},\ h_{\text{peak}} = H - 3,\ \sigma = \frac{H}{4}$.

e) **Figure 3b** - *Average discounted cost for different optimism levels*
- Each epoch comprised of 150 timesteps.
- True Parameters: $H = 10, \lambda_o = 0.2, \lambda_i = 0.5, C_o = 0, C_i = 1, C_c = 500, \gamma = 0.9$. The optimal threshold is 9.
- Priors: $\lambda_{o,p} = 0.1, \lambda_{i,p} = 0.2, C_{o,p} = 0, C_{i,p} = 20$.

The initial threshold is 0.

- Prior Strength: $n_0 = 150$
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha \in \{10, 40, 60, 100, 1000\}$, $h_{\text{peak}} = H - 3$, $\sigma = \frac{H}{4}$.

f) **A Representative Run - Figure 4** - *Evolution of the computed policy for a single run (optimal threshold = 2)*

- Each epoch comprised of 150 timesteps.
- True Parameters: $H = 10, \lambda_o = 0.2, \lambda_i = 0.5, C_o = 0, C_i = 30, C_c = 500, \gamma = 0.9$. The optimal threshold is 2.
- Priors: $\lambda_{o,p} = 0.1, \lambda_{i,p} = 0.2, C_{o,p} = 0, C_{i,p} = 3$. The initial threshold is 9.
- Prior Strength: $n_0 = 1000$
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha = 0$, $h_{\text{peak}} = H - 3$, $\sigma = \frac{H}{4}$.

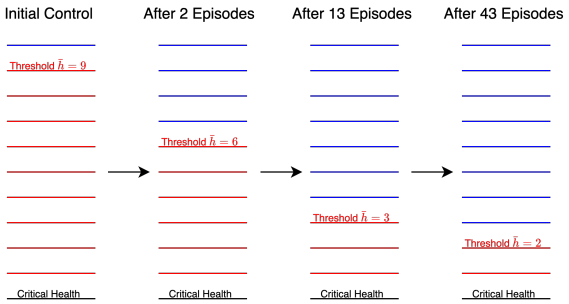

Fig. 4: Evolution of the computed policy for a single run (optimal threshold = 2).

C. Health-State-Dependent Parameters

In this section, we detail state-dependent parameterization and the experiments we conducted. Here, the MDP parameters are defined per state rather than globally. Specifically, for each nonterminal health level $h \in \{1, \ldots, H\}$ and action $a \in \{\text{o}, \text{i}\}$, the probability of moving to $h + 1$ is $\lambda_{a,h}$ and of moving to $h - 1$ is $1 - \lambda_{a,h}$. At $h = H$ the chain either remains at H or moves to $H-1$ w.p. $\lambda_{a,H}$ and $1 - \lambda_{a,H}$, respectively. As before, the state $h = 0$ is absorbing. Costs are likewise state-action specific, $C_{a,h}$, and a terminal penalty C_c is incurred upon entering $h = 0$.

With this statewise model in place, we estimate parameters online from per-(h, a) aggregates. Specifically, for each $h \in \{1, \ldots, H\}$ and $a \in \{\text{o}, \text{i}\}$, we maintain a visit count $N_{a,h}$, a count of upwards health transitions $N^+_{a,h}$, and a running cost sum $S_{a,h}$. The transition probability and cost are then estimated statewise using the same rules as in the main algorithm but indexed by (h, a).

All other components of the algorithm remain as defined for the state-independent case, planning is discounted with factor $\gamma \in (0, 1)$, UCB-style optimism is applied with a state-dependent exploration weight α_h that retains the same Gaussian shape used previously, and value iteration

Algorithm 1 Online Model-Based RL with Optimism (State-Dependent

Input: Priors $\lambda_{o,p}, \lambda_{i,p}, C_{o,p}, C_{i,p}$, prior strength n_0, exploration schedule α_h for $h \in \mathcal{H}$ and critical cost C_c
Initialization (for all $h \in \mathcal{H}$ and $a \in \mathcal{A}$): Counts $N_{a,h} \leftarrow n_0$, $N^+_{a,h} \leftarrow \lambda_{a,p} n_0$, cost sums $S_{a,h} \leftarrow C_{a,p} n_0$ for $a \in \{i, o\}$
for epoch $k = 1, \ldots,$ **do**
  (1) Parameter Estimation (statewise):
  for $a \in \mathcal{A}$ and $h \in \mathcal{H}$ compute estimates

$$\widehat{\lambda}_{a,h} \leftarrow N^+_{a,h}/N_{a,h} \quad \text{and} \quad \widehat{C}_{a,h} \leftarrow S_{a,h}/N_{a,h}$$

  (2) Optimistic Monitoring Control
  Compute optimistic cost for all $h \in \mathcal{H}, a \in \mathcal{A}$,

$$\widehat{C}^{opti}_{a,h} = \widehat{C}_{a,h} - \alpha_h \sqrt{1/N_{a,h}}$$

  Solve following for all $1 \le h \le H, a \in \{i, o\}$:

$$Q(h,a) = \widehat{C}^{opti}_{a,h} + \gamma\left(\widehat{\lambda}_{a,h} V(h^+) + \widehat{\mu}_{a,h} V(h^-)\right), \quad (4)$$

  where $h^+ = \min\{h + 1, H\}, h^- = h - 1$ and

$$V(h) = \begin{cases} \min_{a \in \{i,o\}} Q(h,a), & \text{if } 1 \le h \le H, \\ C_c, & \text{if } h = 0. \end{cases} \quad (5)$$

  Update control $\pi(h) = \arg\min_{a \in \{i,o\}} Q(h,a)$.
  (3) RPM Service and Data Collection:
  for each patient in system **do**
    for timesteps $t = 0, \ldots, T$ of epoch **do**
      Observe health state h_t
      Implement monitoring control $a = \pi(h_t)$
      Observe next state h_{t+1} and cost sample c_t
      Update:

$$N^+_{a,h} \leftarrow N^+_{a,h} + 1 \text{ if } h_{t+1} = h_t + 1$$

$$N_{a,h} \leftarrow N_{a,h} + 1, S_{a,h} \leftarrow S_{a,h} + c_t$$

    end for
  end for
end for
Output: Final control π and estimates $\hat{\lambda}_{a,h}, \hat{C}_{a,h}$.

action selection are unchanged except that they now use the statewise $\lambda_{a,h}$ and $C^{(a)}_h$.

In our experiment, we model the transition probabilities $\lambda_{a,h}$ as increasing in the health level h, and the per-step costs $C_{a,h}$ as decreasing in h. Clinically, patients at higher health levels are healthier and either action is more likely to maintain or improve status; thus the probability of an upward transition should be larger, i.e., $\lambda_{a,h}$ increases as h increases. Conversely, care at poorer health states typically requires more resources, making per-step action costs higher when h is small hence $C_{a,h}$ decreases as h increases. Furthermore, we use a linear, monotone dependence on health because

changes are gradual.

The parameters used for this experiment (**Figure 4** - *Average discounted cost with health-state-dependent dynamics*) are as follows: -

- Each epoch comprised of 150 timesteps.
- True parameters (state dependent): $H = 10,\ \gamma = 0.9,\ C_c = 500$. For each action $a \in \{\mathrm{o}, \mathrm{i}\}$ and health level $h \in \{0, \ldots, H\}$,

$$\lambda_{i,h} = 0.45 + (0.55 - 0.45)\frac{h}{H},$$
$$\lambda_{o,h} = 0.18 + (0.22 - 0.18)\frac{h}{H},$$
$$C_{i,h} = 0.5 + (1.5 - 0.5)\frac{H - h}{H},$$
$$C_{o,h} = 0.0 + (0.5 - 0.0)\frac{H - h}{H}.$$

- Priors (per state–action): The priors were identical for every health state h. $\lambda_{o,p} = 0.10, \lambda_{i,p} = 0.20$ cost priors $C_{o,p} = 0,\ C_{i,p} = 20$ with prior strength $n_0 = 1000$.
- Exploration schedule: $\alpha_h = \alpha \exp\left(-\frac{(h - h_{\text{peak}})^2}{2\sigma^2}\right)$ with $\alpha \in \{100, 500, 1000, 5000\},\ h_{\text{peak}} = H - 3,\ \sigma = \frac{H}{4}$.

REFERENCES

[1] S. Chandak, I. Thapa, N. Bambos, and D. Scheinker, "Tiered service architecture for remote patient monitoring," in *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, 2024, pp. 1–7.