APPENDIX
Learning Optimal Control for Remote Patient Monitoring Systems

APPENDIX I

A. Dynamic Programming Equation

The dynamic programming equations satisfied by the optimal control $V^*(\cdot,\cdot)$ are as follows.

(i) At the critical health state $h = 0$,

$$V^*(i,0) = V^*(o,0) = C_c,$$

(ii) For health states $1 \le h \le H-1$,

$$\begin{aligned} V^*(i,h) &= V^*(o,h) \\ &= \min\Bigg\{ C_i + \gamma\Big[\lambda_i V^*(i,h+1) + \mu_i V^*(i,h-1)\Big], \\ &\qquad C_o + \gamma\Big[\lambda_o V^*(o,h+1) + \mu_o V^*(o,h-1)\Big]\Bigg\} \end{aligned}$$

(iii) At health state H,

$$\begin{aligned} V^*(o,H) &= V^*(i,H) \\ &= \min\Bigg\{ C_i + \gamma\Big[\lambda_i V^*(i,H) + \mu_i V^*(i,H-1)\Big], \\ &\qquad C_o + \gamma\Big[\lambda_o V^*(o,H) + \mu_o V^*(o,H-1)\Big]\Bigg\} \end{aligned}$$

B. Proof for Theorem 1

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

$$\gamma(\lambda_i - \lambda_o) \le C_i/c_c. \tag{1}$$

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

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

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

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

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

C. Experiment Details

In this section, we present complete experimental details for the plots presented in the main paper.

a) **Figure 2** - *Evolution of the computed policy for a single run (optimal threshold = 2)*
- Each episode comprised of 10 patients.
- 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 $\alpha = 0,\ h_{\text{peak}} = H-3,\ \sigma = \frac{H}{4}$.

b) **Figure 3a** - *Average time to reach critical health for different initial thresholds*
- Each episode comprised of 5 patients.
- 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 3b** - *Percentage of converged runs after each episode for different prior strengths*
- Each episode comprised of 5 patients.
- 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 3c** - *Percentage of converged runs after each episode for different optimism levels*
- Each episode comprised of 10 patients.
- 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 3d** - *Average discounted cost for different optimism levels*

- Each episode comprised of 10 patients.
- 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 $\alpha \in \{10, 40, 60, 100, 1000\}$, $h_{\text{peak}} = H - 3$, $\sigma = \frac{H}{4}$.

D. Health-State-Dependent Parameters

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.