

Safe Learning of Adaptive Control Policies for Remote Patient Monitoring

Siddharth Chandak

Joint work with Ramanan Tamizholi, Isha Thapa, Nicholas Bambos, David Scheinker

Outline

- 1 Background and Motivation
- 2 Tiered Remote Patient Monitoring
- 3 Learning the Optimal Policy Safely
- 4 Guarantees and Simulations

Remote Patient Monitoring

Technology-enabled healthcare, in patients' daily lives



**Continuous Glucose
Monitors**



Smartwatches



Implantable Devices

The Tradeoffs



\$

Ordinary Monitoring

less support, lower cost and burden



\$\$\$

Intensive Monitoring

better support but more costly

The Tradeoffs

Remote Monitoring



\$



\$\$\$



Critical State



\$\$\$\$\$\$\$\$\$\$\$\$

What is the optimal level of monitoring?

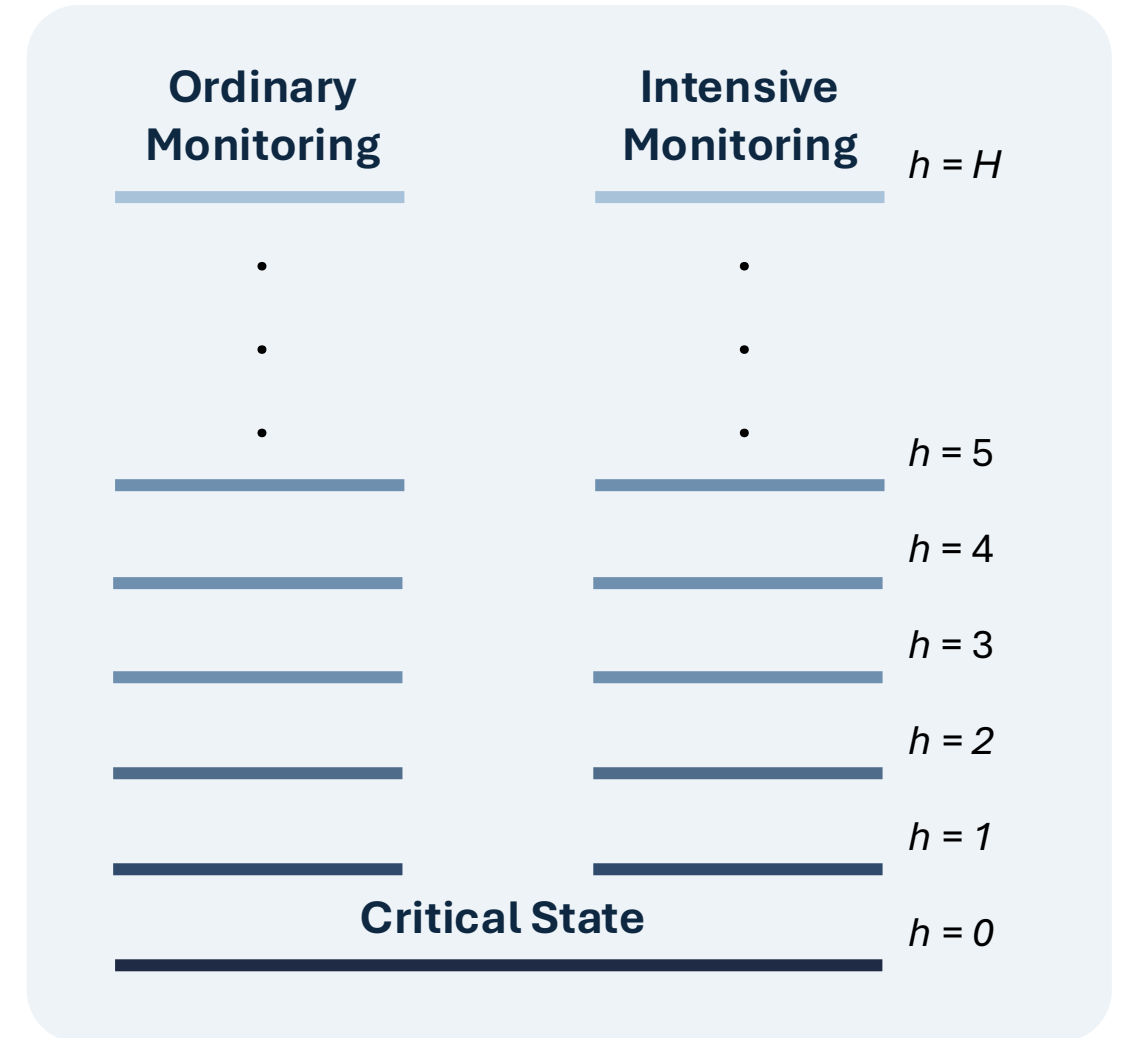
The Model

Time: $t \in \{0, 1, 2, \dots\}$

Patient State:

- Monitoring State: $m_t \in M = \{o, i\}$
- Health State: $h_t \in \{0, 1, 2, \dots, H\}$
- Overall State: $s_t := (m_t, h_t)$

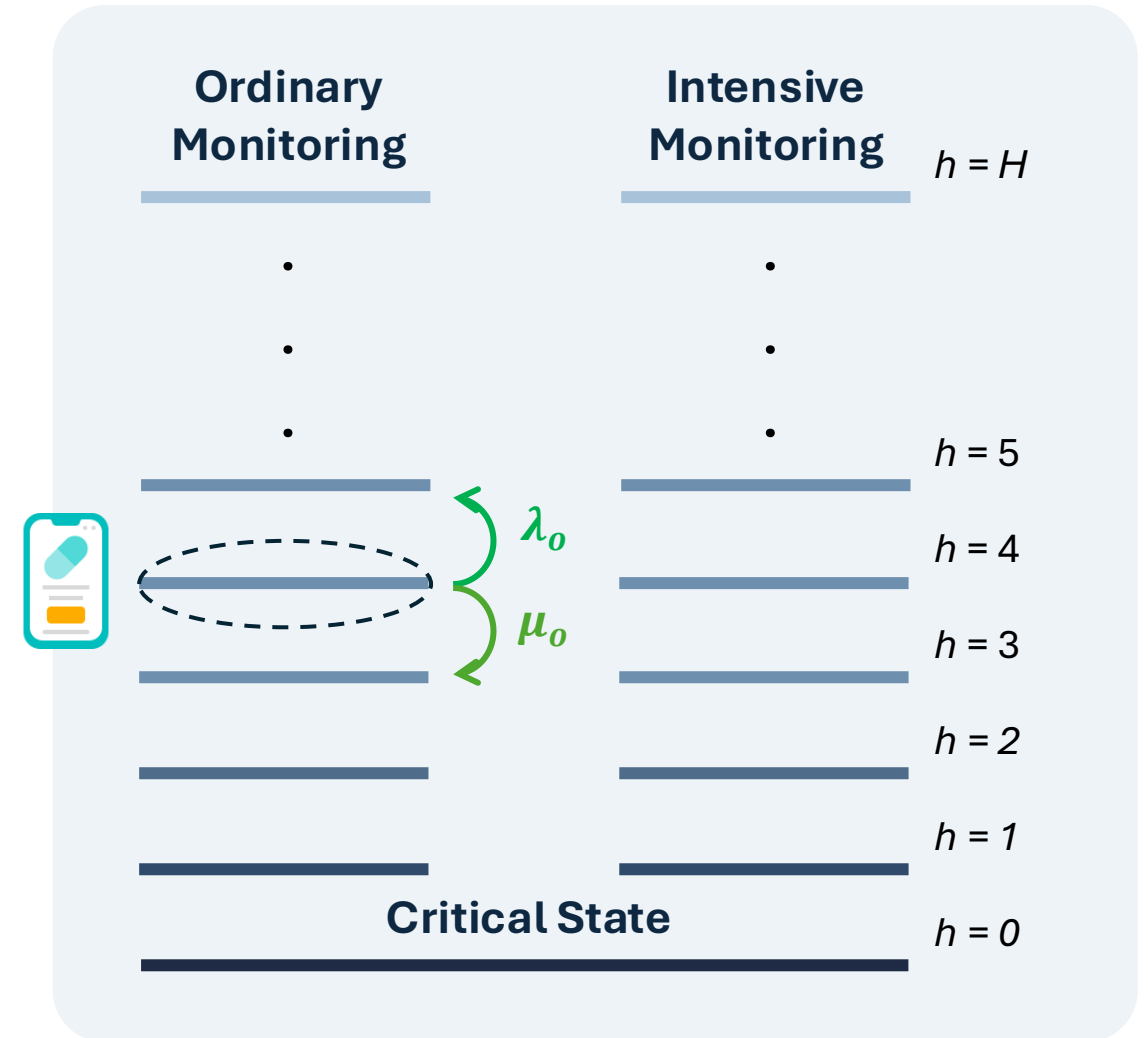
Action: $a_t \in A = \{o, i\}$



The Model

Transition Probabilities

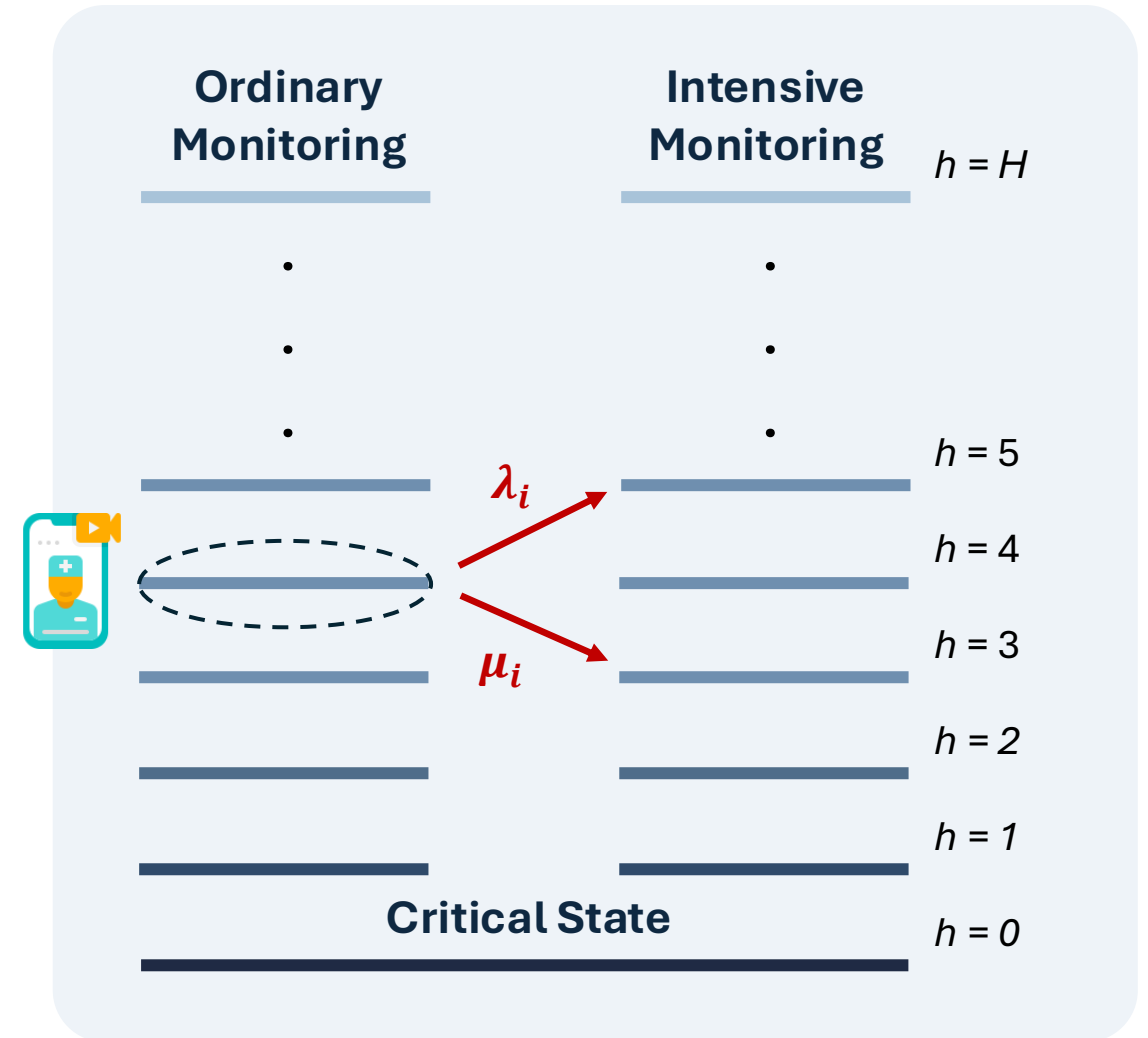
- **Ordinary monitoring:** λ_o , $\mu_o = 1 - \lambda_o$



The Model

Transition Probabilities

- Ordinary monitoring: $\lambda_o, \mu_o = 1 - \lambda_o$
- **Intensive monitoring: $\lambda_i, \mu_i = 1 - \lambda_i$**



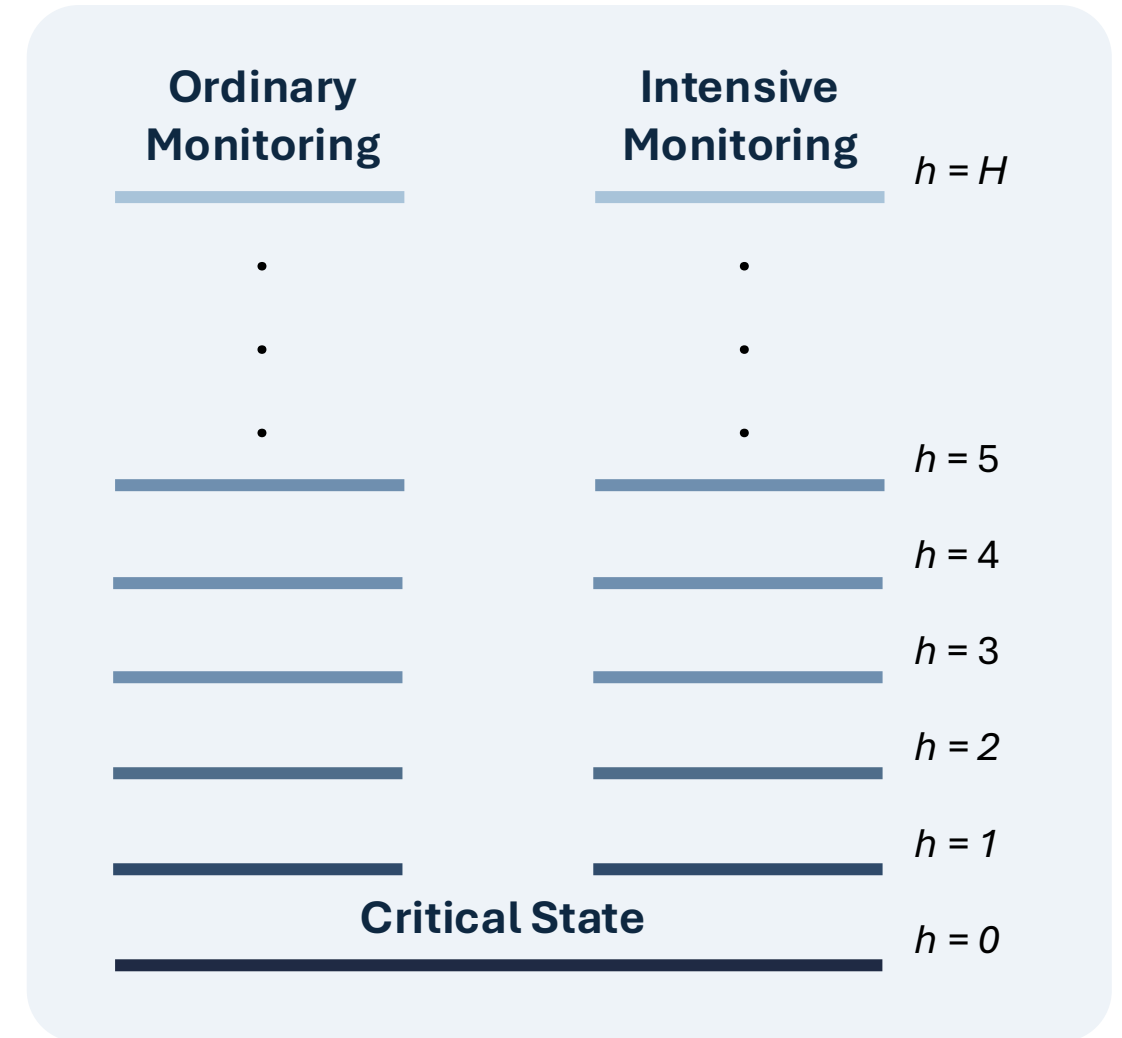
The Model

Transition Probabilities

- Ordinary monitoring: $\lambda_o, \mu_o = 1 - \lambda_o$
- Intensive monitoring: $\lambda_i, \mu_i = 1 - \lambda_i$
- Satisfies $\lambda_i \geq \lambda_o$

Costs

- Ordinary monitoring: C_o
- Intensive monitoring: C_i
- Critical State: C_c
- Satisfies $0 \leq C_o \leq C_i \ll C_c$



Optimal Control

Find a policy π (states $\rightarrow$ actions) minimizing the expected discounted cost

$$V_{\pi}(s) = \mathbb{E} \left[\sum_{t=0}^{T-1} \gamma^t c(s_t, \pi(s_t)) + \gamma^T C_c \mid s_0 = s \right]$$

T

time the patient reaches
the critical state (random)

$\gamma \in (0,1)$ discount factor

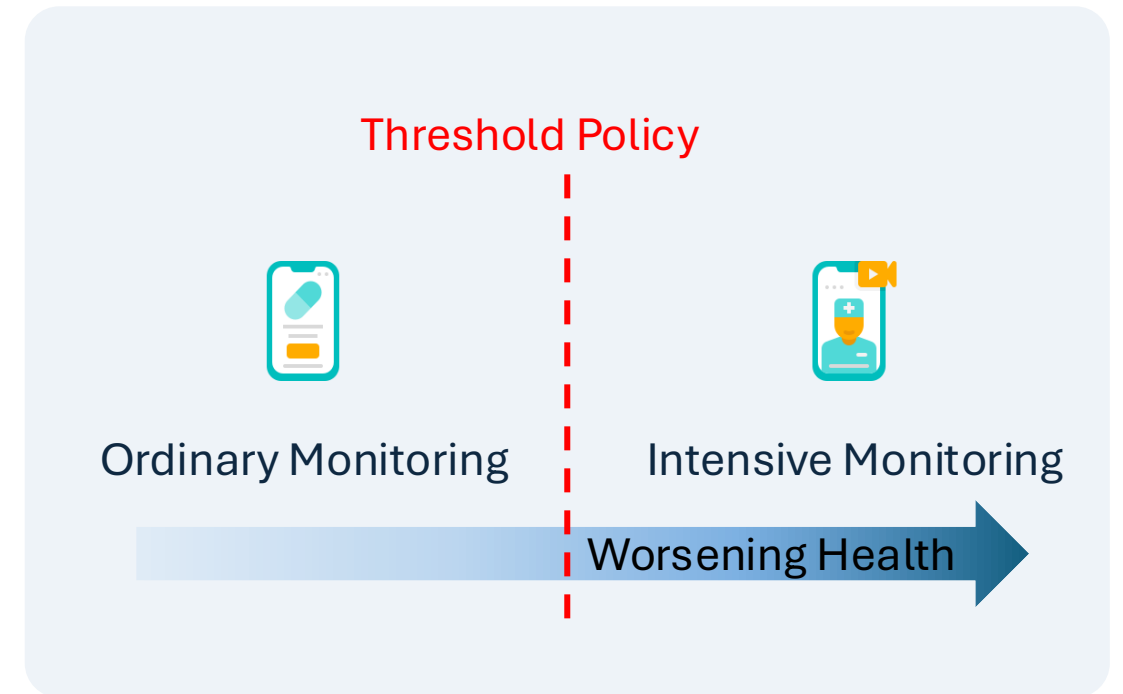
Optimal Policy: Threshold Structure

Known parameters → dynamic programming gives the optimal policy

Theorem

The optimal policy is a threshold policy:

intensive monitoring $\Leftrightarrow h \leq h^*$



But the parameters $\lambda_o, \lambda_i, C_o, C_i$ are unknown...

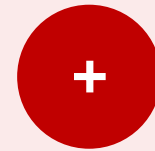
How do we learn the optimal threshold?

What Should the Learning Algorithm Do?



Efficient

- Learn the parameters
 $\hat{\lambda}_a, \hat{C}_a \rightarrow \lambda_a, C_a$
- Converge to the optimal policy
 $\tilde{h}_k \rightarrow h^*$
- Experiment on as few patients as possible

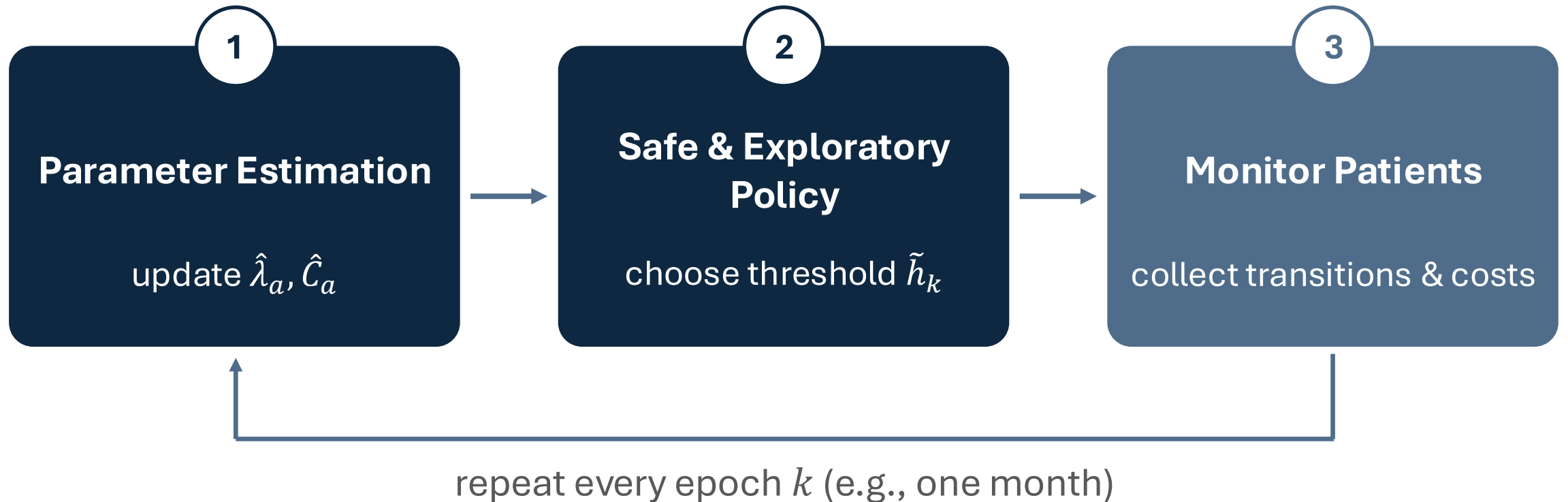


Safe

We are learning on real patients

Conservative threshold: $\tilde{h}_k \geq h^*$
at least as cautious as the optimal policy

The Algorithm

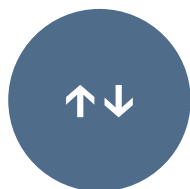


Parameter Estimation

Estimate

Policy

Monitor



Health transitions

Did health improve under monitoring a ?



$$\hat{\lambda}_a = \frac{N_a^+}{N_a}$$

fraction of steps where health improved



Monitoring burden

Patient feedback, e.g., surveys



$$\hat{C}_a = \frac{S_a}{N_a}$$

average reported cost per step

Initialized with priors $\lambda_{a,p}$, $C_{a,p}$ and prior strength n_0 (acts like n_0 pseudo-samples)

Safe & Exploratory Policy

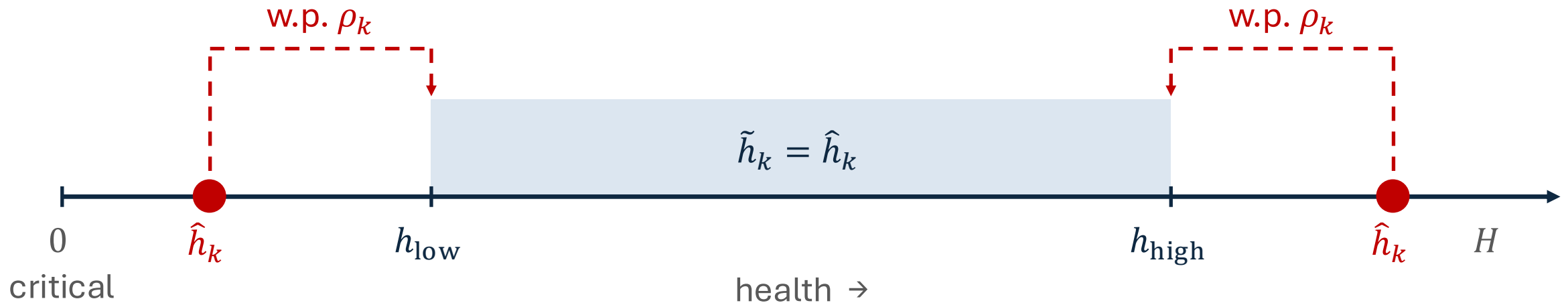
Estimate

Policy

Monitor

Step 1 $\hat{h}_k$ = optimal threshold for current estimates, via dynamic programming

Step 2 deploy $\tilde{h}_k$:



In $[h_{\text{low}}, h_{\text{high}}]$: deploy $\hat{h}_k$

Below h_{low} : raise to h_{low} w.p. ρ_k , else keep $\hat{h}_k$

Above h_{high} : lower to h_{high} w.p. ρ_k , else keep $\hat{h}_k$

$\rho_k \rightarrow 0$ slowly

e.g., $\rho_k \propto k^{-1/3}$

Why This Policy?



Explore

- $\hat{h}_k$ **too low** $\rightarrow$ few intensive-monitoring samples
- $\hat{h}_k$ **too high** $\rightarrow$ few ordinary-monitoring samples
- Occasional moves to h_{low} / h_{high} keep both actions sampled



Stay Safe

- Explore through the threshold, not arbitrary actions
- Aggressive thresholds are pushed **up** to h_{low} (fragile health)
- Conservative priors + large $n_0 \rightarrow \hat{h}_k \geq h^*$

Theoretical Guarantees

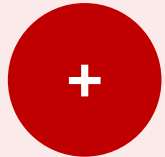
Under suitable assumptions:

Theorem (Convergence)



The deployed threshold $\tilde{h}_k$ converges almost surely to the optimal threshold h^*

Theorem (Safety)

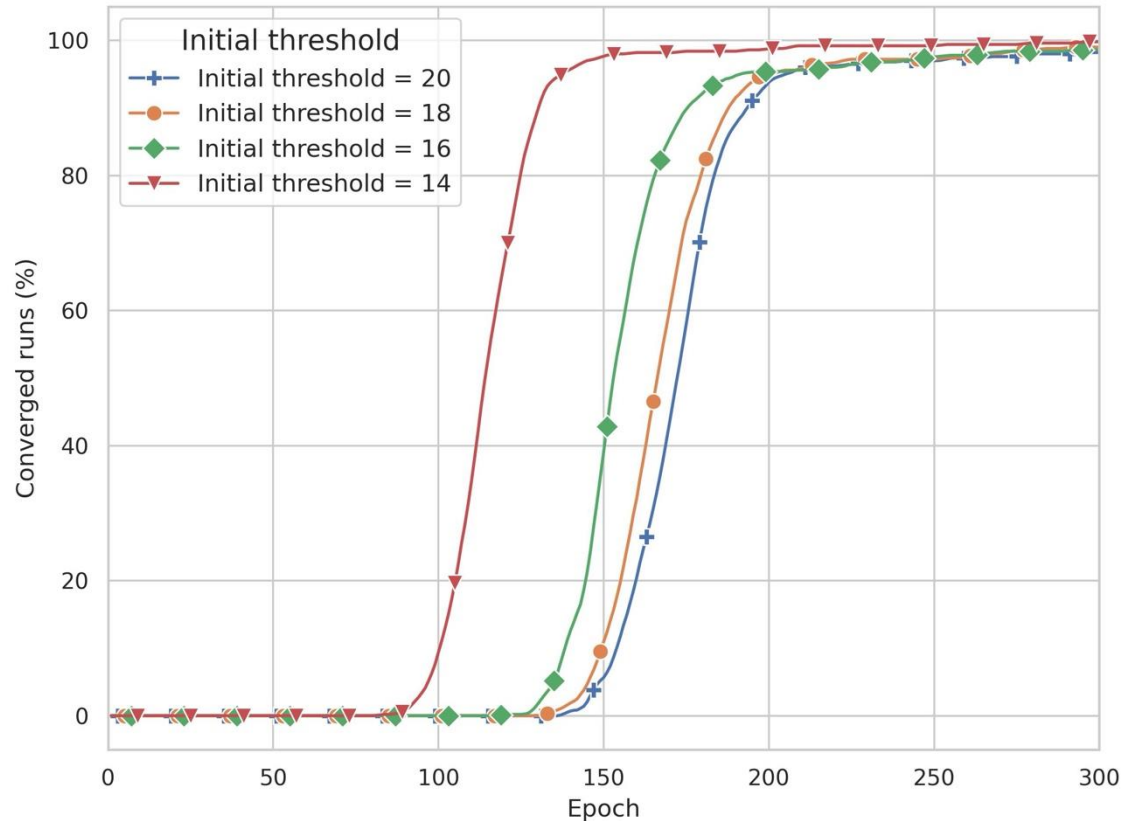


For any $\delta \in (0,1)$, if prior strength $n_0 \geq \Omega(\log(1/\delta))$, then with probability at least $1 - \delta$:

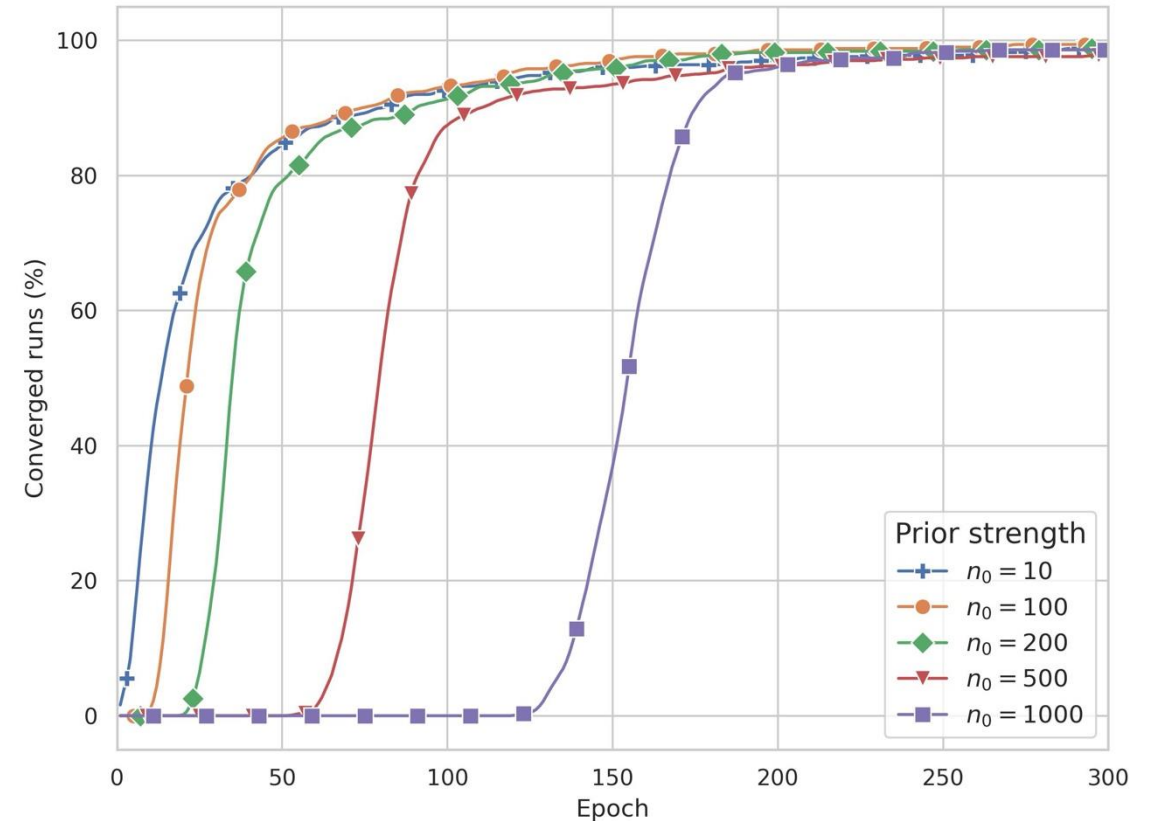
$$\tilde{h}_k \geq h^* \text{ for all epochs } k$$

Simulations: Convergence

Effect of initial threshold



Effect of prior strength n_0



Converged = all estimates within 5% of true values ($H = 20, h^* = 12$)

Safety of the Deployed Threshold

n_0	$\tilde{h} < h^*$	$\tilde{h} < h^* - 1$	$\tilde{h} < h^* - 2$	$\tilde{h} < h_{\text{low}}$
10	82	38	8	0
100	50	1	0	0
200	25	0	0	0
500	6	0	0	0
1000	0	0	0	0

runs (out of 100) where the deployed threshold ever fell below the cutoff ($h^* = 12$)

Key Takeaways

- 1 Learn RPM parameters online, while patients are being monitored
- 2 Explore through the threshold: interpretable and safe
- 3 A provably safe algorithm that converges to the optimal policy

Next: validation on real patient data

Thank You!