Dynamic Patient Priority Assignment for Emergency Medical Services

Soovin Yoon

Department of Industrial Engineering, University of Wisconsin-Madison

Introduction

Efficient emergency medical service design requires rationing limited medical resources through patient triage. However, traditional triage systems are static and do not depend on the changes in available resources.

A spatial emergency medical system with multiple resource types and multiple demand types is considered. We aim to

- dispatch the appropriate type of server to an arriving emergency call
- reach the patient fast to meet time standard(9 min/13 min)

Therefore we prioritize based on the resource availability of the system. The resource availability keep changes due to call arrivals and service completions, which naturally results in dynamic reprioritization.

An Intuitive Example

An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

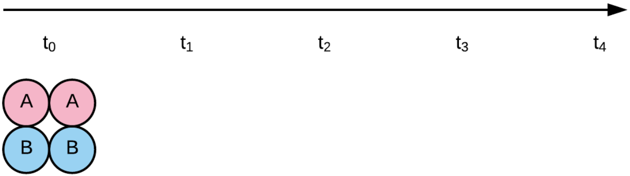
- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

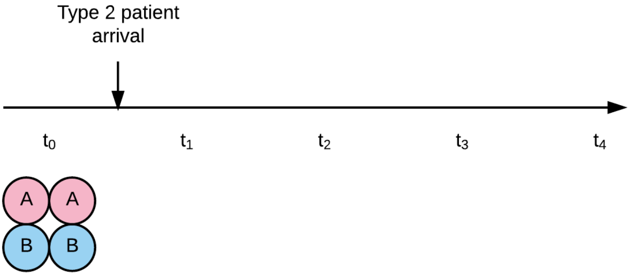


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

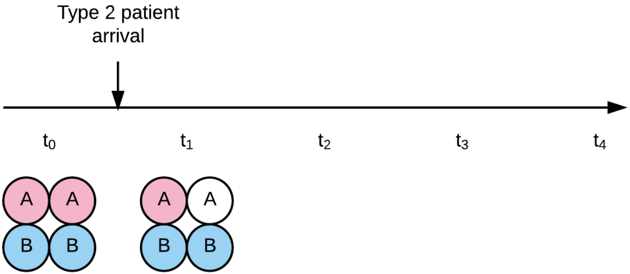


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

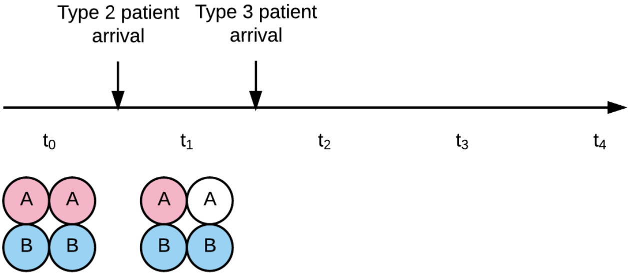


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

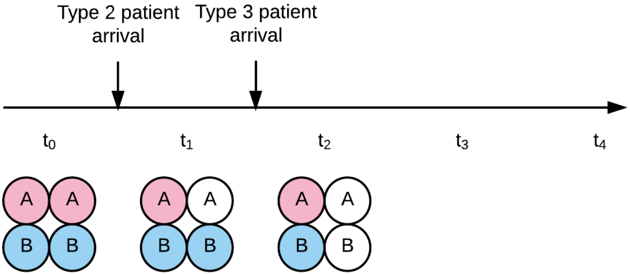


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

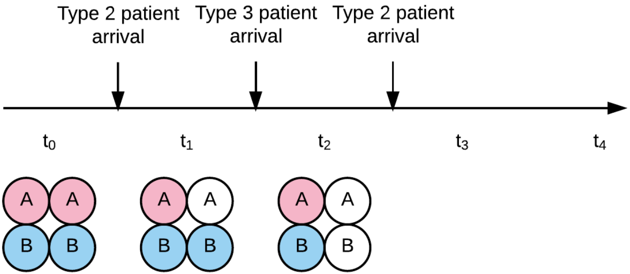


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

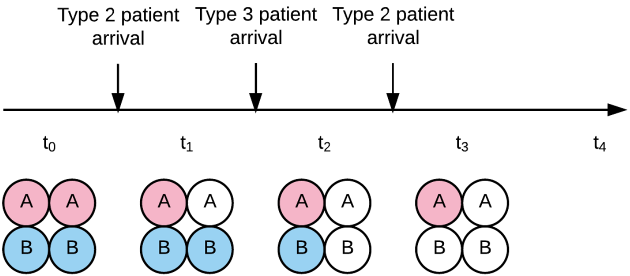


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

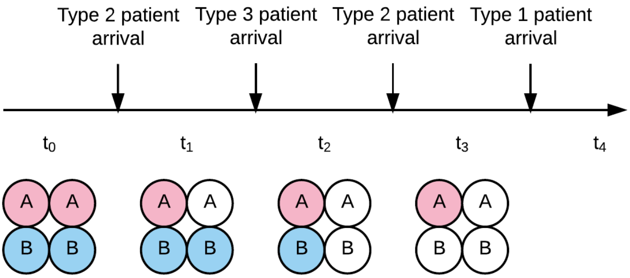


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B

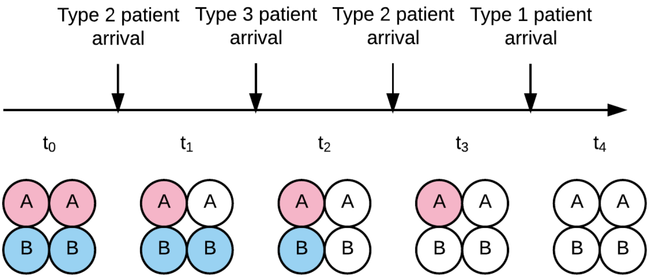


An Intuitive Example

2 types of servers are available: Advanced(A) and Basic(B)

3 types of patients arrive: 1, 2 and 3

- type 1: need A
- type 2: need A with probability of α
- type 3: can be served by either A or B



Setting

Setting

Resources

Ambulances are differentiated by the types of treatment they can provide.

- Advanced Life Support(ALS)
  - staffed by paramedics to serve urgent calls, service rate μ_A
- Basic Life Support(BLS)
  - staffed by EMTs to serve less serious calls, service rate μ_B

Setting

Resources

Ambulances are differentiated by the types of treatment they can provide.

- Advanced Life Support(ALS)
  - staffed by paramedics to serve urgent calls, service rate μ_A
- Basic Life Support(BLS)
  - staffed by EMTs to serve less serious calls, service rate μ_B

Patients

Calls arriving at rate $\lambda = \sum_i \lambda_{ip}$ have

- types $i \in \{1, \ldots, m\}$: Any available information that is correlated with the urgency of a call can be used as a type information.
  - Incident type
  - Geographic information
- pre-assigned priorities $p \in \{1, 2, 3\}$

Setting

Resources

Ambulances are differentiated by the types of treatment they can provide.

- Advanced Life Support(ALS)
  - staffed by paramedics to serve urgent calls, service rate μ_A
- Basic Life Support(BLS)
  - staffed by EMTs to serve less serious calls, service rate μ_B

Patients

Calls arriving at rate $\lambda = \sum_i \lambda_{ip}$ have

- types $i \in \{1, \ldots, m\}$: Any available information that is correlated with the urgency of a call can be used as a type information.
  - Incident type
  - Geographic information
- pre-assigned priorities $p \in \{1, 2, 3\}$

Signal

The true urgency of an arriving call with priority 2 is known only probabilistically based on its call type, with parameter $\alpha^i = P(\text{urgent}|\text{call type}=i,\text{priority}=2)$. For priority 1 calls $\alpha^i = 1$, for priority 3 calls $\alpha^i = 0$ for any call type i.

MDP Model

MDP Model

<u>**Time**</u> $t \in \{1, \cdots, T\}$
By uniformization with factor Λ, we get discrete time epochs.

MDP Model

Time $t \in \{1, \cdots, T\}$
By uniformization with factor Λ, we get discrete time epochs.

State $s_t = (s^A, s^B)$
where $s^A(s^B)$ is the number of available ALS(BLS) servers.

MDP Model

Time $t \in \{1, \cdots, T\}$
By uniformization with factor Λ, we get discrete time epochs.

State $s_t = (s^A, s^B)$
where $s^A(s^B)$ is the number of available ALS(BLS) servers.

Action $a_t(s^A, s^B) = (a_t^1, \cdots, a_t^i, \cdots, a_t^m)$
Based on type(i) and priority of an arriving call, assign either an ALS ($a^i = 0$) or BLS ($a^i = 1$), depending on the system congestion (s^A, s^B).

MDP Model

Time $t \in \{1, \cdots, T\}$
By uniformization with factor Λ, we get discrete time epochs.

State $s_t = (s^A, s^B)$
where $s^A(s^B)$ is the number of available ALS(BLS) servers.

Action $a_t(s^A, s^B) = (a_t^1, \cdots, a_t^i, \cdots, a_t^m)$
Based on type(i) and priority of an arriving call, assign either an ALS ($a^i = 0$) or BLS ($a^i = 1$), depending on the system congestion (s^A, s^B).

Transition $P_t(j|s_t, a_t)$

- An urgent call arrives $(s^A, s^B) \to (s^A - 1, s^B)$
- A less urgent call arrives $(s^A, s^B) \to (s^A, s^B - 1)$
- An ALS server finishes service $(s^A, s^B) \to (s^A + 1, s^B)$
- A BLS server finishes service $(s^A, s^B) \to (s^A, s^B + 1)$
- A dummy event $(s^A, s^B) \to (s^A, s^B)$

MDP Model

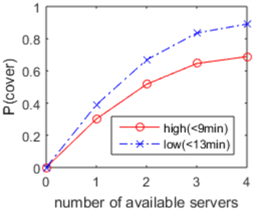
Reward $R_t(s_t, a_t)$

MDP Model

Reward $R_t(s_t, a_t)$

Since ambulances are spatially located, only a subset can respond fast. The probability that an arriving high(low) priority call is served in a timely fashion is modelled as a non-decreasing reachability function $f^H(s)(f^L(s))$.

	High	**Low**
send ALS	$f^H(s^A)$	$f^L(s^A)$
send BLS	$f^H(s^B)$	$f^L(s^B)$

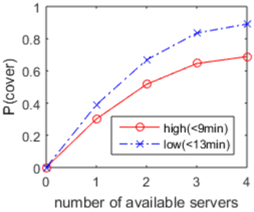


MDP Model

Reward $R_t(s_t, a_t)$

Since ambulances are spatially located, only a subset can respond fast. The probability that an arriving high(low) priority call is served in a timely fashion is modelled as a non-decreasing reachability function $f^H(s)(f^L(s))$.

	High	Low
send ALS	$f^H(s^A)$	$f^L(s^A)$
send BLS	$f^H(s^B)$	$f^L(s^B)$



We get different utility depending on the match between the server type and true priority of the call.

	High	Low
send ALS	U_{HA}	U_{LA}
send BLS	U_{HB}	U_{LB}

- $U_{HA} > U_{LA}$ and $U_{HA} > U_{LB}$: we get more benefit by serving a high priority call with a ALS than serving a low priority call
- $U_{HA} > U_{HB}$: Under-service of urgent call is penalized
- $U_{LA} = U_{LB}$: A low priority call can be served equally well by ALS or BLS

Solution Methodology

Finite-time discrete MDP is solved by backward induction to maximize total expected reward.

$$V_t(s_t) = \sup_a \{R_t(s_t, a) + \sum_j P_t(j|s_t, a) V_{t+1}(j)\}$$

The size of possible action set to be evaluated at each time epoch t grows exponentially with the number of type m. However, we don't really have to evaluate the whole set to find the optimal solution, due to the structural property of the problem that follows.

Type Independence of Optimal Action

Proposition 1

For any time epoch t and state s, the optimal action for the call type i is to send an ALS server if and only if the following equality is true:

$$\alpha^{i2} U_{LA} f^{H}(s^{A}) + (1-\alpha^{i2}) U_{LA} f^{L}(s^{A}) > \alpha^{i2} U_{HB} f^{H}(s^{B}) + (1-\alpha^{i2}) U_{LB} f^{L}(s^{B}))$$

which does not depend on a^{k} for all $k \in \{1, \cdots, m\}, k \neq i$.

Proposition 1 implies that the optimal decision of sending an ALS server or a BLS server to type i call can be made regardless of decision for other call types. Therefore, the number of action we have to evaluate at each time epoch t to solve by the backward induction can be reduced from exponential 2^{m} to linear m.

Optimality of Threshold-Type Policy

Proposition 2

For any time epoch t and state s, a threshold value $\bar{\alpha}_t(s)$ can be specified such that it is optimal to send ALS server to type i call if and only if $\alpha^i > \bar{\alpha}_t(s)$, if

$$U_{HA} f^H(s^A) - U_{HB} f^H(s^B) - U_{LA} f^L(s^A) + U_{LB} f^L(s^B) > 0.$$

and the threshold value is

$$\bar{\alpha}_t(s) = \frac{U_{LA} f^L(s^A) - V_{LB} f^L(s^B) - V_{t+1}(s^A+1, s^B) + V_{t+1}(s^A, s^B+1)}{U_{HA} f^H(s^A) - U_{HB} f^H(s^B) - U_{LA} f^L(s^A) + U_{LB} f^L(s^B)}$$

The condition is satisfied independent of the state if U_{HA} is significantly larger than U_{HB}.

Optimality of Monotone Policy

Proposition 3

For each call type i, the optimal action a_t^i is

1. *nonincreasing in s^A if the value function $V_t(s^A, s^B)$ is concave in s^A*
2. *nondecreasing in s^B if the value function $V_t(s^A, s^B)$ is concave in s^B,*

and the value function $V_t(s^A, s^B)$ is supermodular in (s^A, s^B).

Corollary 1

Value function $V_t(s^A, s^B)$ is

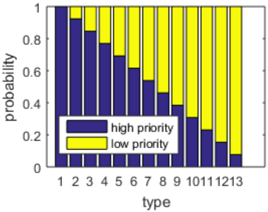
1. *Monotone nondecreasing in s^A and s^B.*
2. *Convex(Concave) in s^A and s^B, if $f^H(s)$ and $f^L(s)$ is convex(concave) in s.*
3. *Modular.*

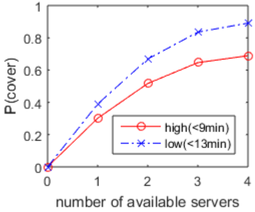
Computational Setup

Resources
The EMS has 2 ALS servers and 4 BLS servers with service rate normalized to $\mu_A = \mu_B = 1$.

Demands
Emergency call datasets from Hanover County, Virginia (June 2009∼December 2011) is used to create arrival rates.

Information and Reachability
The type information, reachability function and utility are created as

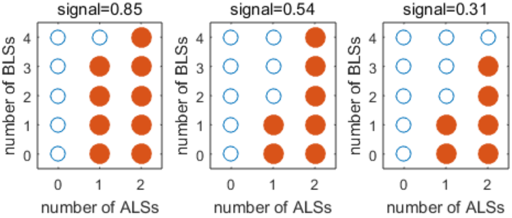




Utilities are set as $U_{HA} = 1$, $U_{HB} = 0.4$, $U_{LA} = U_{LB} = 0.3$.
It is assumed that (1) calls arriving when there is no server available is served by external system (2) calls are always served otherwise.

Result

A snapshot of optimal policies at $t = 217 = T/2$:



The dynamic solution created from solving the MDP is compared to two other static policies, case 1 (always send BLS to priority 2 calls) and case 2 (always send ALS to priority 2 calls).

	dynamic	case 1	case 2
Value	11.0427	10.6036	10.8117

From the use of dynamic policy, the expected coverage is improved by 4.14% compared to the case 1 policy and 2.14% compared to the case 2 policy.

Discussions

- In this research, we examine the potential of dynamic priority assignment to increase the emergency medical service coverage under resource limitations.
- We show conditions under which we have optimal threshold-type, monotone policies. Our computational study provides that the dynamic policy achieves significant improvement in coverage over existing static policies.
- Future work might concentrate on the extension of the model to geographical call types so that the model can explicitly consider spatial features of the system.

References

- Alanis R, Ingolfsson A, Kolfal B (2013) A Markov Chain Model for an EMS System with Repositioning. *Production and Operations Management* 22(1):216:231.
- McLay LA, Mayorga ME (2013) A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions* 45(1):1:24.
- Uzun Jacobson E, Argon NT, Ziya S (2012) Priority assignment in emergency response. *Operations Research* 60(4):813:832.

Acknowledgement

This work was funded by the National Science Foundation [Award 1444219]. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.