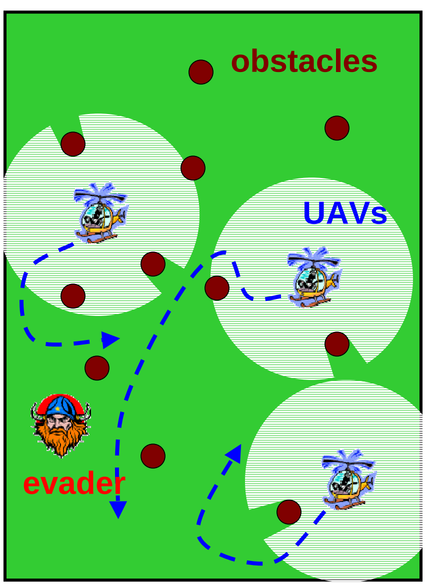
Approximate Solutions for Pursuit-Evasion Games with Unmanned Air Vehicles

João P. Hespanha
Univ. Southern California

In collaboration with J. Kim, M. Prandini, and S. Sastry

The "rules" of the game



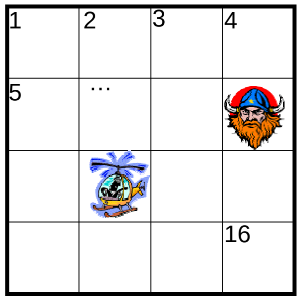
Terrain:
* with **fixed obstacles**
* not accurately mapped

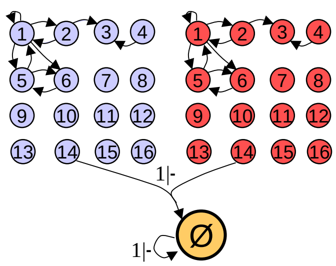
UAVs (pursuers) capable of:
* **flying** between obstacles
* **seeing** a region around them (limited by the occlusions)

Evader capable of:
* **moving** between obstacles (possibly actively avoiding detection)

Objective:
find the evader in minimum time

Competitive, Hidden MDPs





follower's optimal policy μ^*

$$\mu^* = \text{argmin}_{\mu} \, J_{\mu,\delta^*}$$

- deterministic (at least one of the optimal)

$$\mu^*(u, Y, U) \in \{0,1\}$$

- non Markov, but

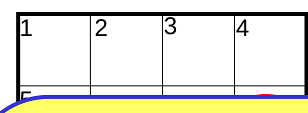
$$\mu^*(u, Y, U) = f\big(u, I(Y,U)\big)$$

where

$$I(Y,U) := \underbrace{\left\{ \mathrm{P}\left(\begin{array}{c} x_t = x, \boldsymbol{Z}_t = Z, \boldsymbol{D}_{t-1} = D \\ | \, \boldsymbol{Y}_t = Y, \boldsymbol{U}_{t-1} = U \end{array} \right) : \begin{array}{c} x \in \mathcal{X}, \\ Z \in \mathcal{Z}^*, \\ D \in \mathcal{D}^* \end{array} \right\}}_{\textbf{information state}}$$

information state
(size grows with t)

Competitive, Hidden MDPs



- follower's optimal policy is finite dimensional only if $\boldsymbol{\delta^*}$ uses ***finite memory***
- even when $\boldsymbol{\delta^*}$ is memoryless, the size of the information state tends to be prohibitively large, e.g.,

$$2^N \cdot N^{(m+1)} \sim 2.6\times 10^{69}$$

number of cells (200 with uniformly distributed obstacles)

number of pursuers (3)

follower's optimal policy μ^*

$$\mu^* = \text{argmin}_{\mu} \, J_{\mu,\delta^*}$$

- …ministic (at least one of the optimal)

$$\mu^*(u, Y, U) \in \{0,1\}$$

- …Markov, but

$$\mu^*(u, Y, U) = f\big(u, I(Y,U)\big)$$

$$\ldots) := \left\{ \mathrm{P}\left(\begin{array}{c} x_t = x, \boldsymbol{Z}_t = Z, \boldsymbol{D}_{t-1} = D \\ | \, \boldsymbol{Y}_t = Y, \boldsymbol{U}_{t-1} = U \end{array} \right) : \begin{array}{c} x \in \mathcal{X}, \\ Z \in \mathcal{Z}^*, \\ D \in \mathcal{D}^* \end{array} \right\}$$

information state
(size grows with t)

Problem formulation

At each time t:

- the pursuers take ***measurements*** y_t (positions of enemies and obstacles seen)
- ***control actions*** u_t are sent to the pursuers (motion control)

A ***pursuit policy*** is a function ν that maps past measurements to the next control action:

$$u_{t+1} = \nu(y_1, y_2, \ldots, y_t)$$

Goal: design pursuit policies that "maximize" probability of finding the evader

Optimality criteria

- ***Probability of finding*** the evader in the next T time steps (to be maximized)

$$J_T = \text{Prob(find an evader from time 1 to time } T\text{)}$$

- ***Expected time*** needed to find the evader (to be minimized)

$$J^* = E[T^*] \qquad T^* \equiv \text{time at which the evader is found}$$

Optimality criteria

- ***Probability of finding*** the evader in the next T time steps (to be maximized)

$$J_T = 1 - \prod_{t=1}^{T} \left(1 - p(t)\right)$$

- ***Expected time*** needed to find the evader (to be minimized)

$$J^* = \sum_{t=1}^{\infty} t\, p(t) \left(\prod_{\tau=1}^{t-1} \left(1 - p(t)\right) \right)$$

where $p(t) \equiv$ Prob(finding evader at time t, given that none was found up to t)

Greedy policy

greedy policy $\equiv$ pursuit policy that, at each instant, maximizes the probability of finding the evader at the next time instant

computationally very attractive because it scales well with the size of the region, number of obstacles, etc.

Greedy policy

greedy policy $\equiv$ pursuit policy that, at each instant, maximizes the probability of finding the evader at the next time instant

- **probability of finding** the evader in finite time

$$\mathrm{P}(\mathbf{T}^* < \infty) = ?$$

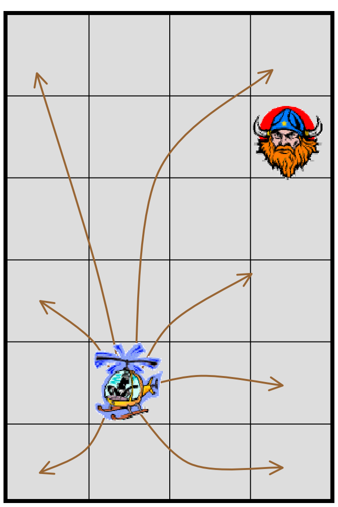
where $\mathbf{T}^* \equiv$ time at which the evader is found

- **expected time** needed to find the evader

$$\mathrm{E}[\mathbf{T}^*] = ?$$

Unconstrained motion for pursuers

Assume the whole area is reachable by a single pursuer in ***one time step***



At every time *t*, there must exist at least one cell for which

$$\mathrm{P}(\text{evader at that cell}) \geq \frac{1}{\text{n. cells}}$$

Therefore a greedy policy can always make

$$p(t) \geq \frac{1}{\text{n. cells}}$$

where

$p(t) \equiv$ P(finding evader at time t, given that none was found up to t)

Unconstrained motion for pursuers

Assume the whole area is reachable by each pursuer in ***one time step***

There must exist at least m cells for which

$$\mathrm{P}(\text{evader at one of the } m \text{ cells}) \geq \frac{\mathrm{m}}{\mathrm{n.\ cells}}$$

Therefore a greedy policy with m pursuers can always make

$$p(t) \geq \frac{\mathrm{m}}{\mathrm{n.\ cells}}$$

where

$p(t) \equiv$ P(finding evader at time t, given that none was found up to t)

Persistent policies

Definition: A policy is ***persistent*** if for every time t

$$p(t) \geq \varepsilon$$

for some $\varepsilon > 0$

$p(t) \equiv$ P(finding evader at time t, given that none was found up to t)

Theorem: For a persistent policy

$$P(\mathbf{T}^* < \infty) = 1$$

$$P(\mathbf{T}^* \leq T) \geq 1 - (1-\varepsilon)^T$$

$$\mathrm{E}[\mathbf{T}^*] \leq \frac{1}{\varepsilon}$$

where $\mathbf{T}^* \equiv$ time at which the evader is found

Persistent policies

Definition: A policy is ***persistent*** if for every time t

$$p(t) \geq \varepsilon$$

$p(t) \equiv$ P(finding evader at time t, given that none was found up to t)

for some $\varepsilon > 0$

Theorem: For a persistent policy

$$P(\mathbf{T}^* < \infty) = 1$$

$$P(\mathbf{T}^* \leq T) = 1 - \prod_{t=1}^{T} \big(1 - p(t)\big) \geq 1 - (1-\varepsilon)^T$$

$$\mathrm{E}[\mathbf{T}^*] = \sum_{t=1}^{\infty} t\, p(t) \left(\prod_{\tau=1}^{t-1} \big(1 - p(\tau)\big) \right) \leq \frac{1}{\varepsilon}$$

where $\mathbf{T}^* \equiv$ time at which the evader is found

Persistent policies

Definition: A policy is ***persistent***

$$p(t) \geq \varepsilon$$

for some $\varepsilon > 0$

The case of a greedy policy when the motion of the pursuers is unconstrained

Theorem: For a persistent policy

$$P(\mathbf{T}^* < \infty) = 1$$

$$P(\mathbf{T}^* \leq T) \geq 1 - (1-\varepsilon)^T$$

$$\mathrm{E}[\mathbf{T}^*] \leq \frac{1}{\varepsilon}$$

where $\mathbf{T}^* \equiv$ time at which the evader is found

Greedy policy

greedy policy ≡ pursuit policy that, at each instant, maximizes the probability of finding the evader at the next time instant

- **probability of finding** the evader in finite time is one and, e.g.,

$$P(\mathbf{T}^* \leq T) \geq 1 - \left(1 - \frac{\text{n. pursuers}}{\text{n. cells}}\right)^T$$

when each pursuer can check any single unit of area at each instant of time (unconstrained motion for pursuers)

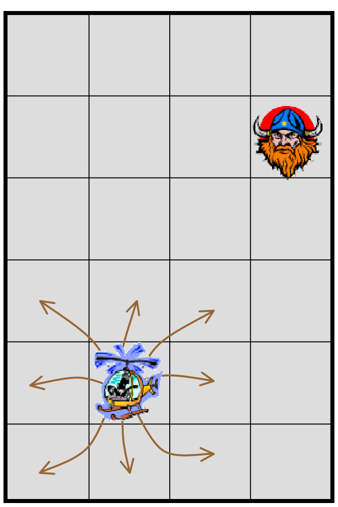
- **expected time** needed to find the evader is finite, e.g.,

$$E[\mathbf{T}^*] \leq \frac{\text{area}}{\text{n. pursuers}}$$

under same assumptions

Constrained motion for pursuers

Assume the whole area is reachable by each pursuer in ***T time steps***



With constrained motion for the pursuers it is not possible to guarantee

$$p(t) \geq \varepsilon$$

for some $\varepsilon > 0$, where

$p(t)$ ≡ P(finding evader at time t, given that none was found up to t)

But we can guarantee that this will hold ***over intervals of length T***

Persistent on-the-average policies

Definition: A policy is ***persistent on-the-average*** if there is an integer T such that for every time t

$$p_T(t) \geq \varepsilon \qquad p_T(t) \equiv \text{P(finding evader from } t \text{ to } t+T\text{-1, given that none was found up to } t)$$

for some $\varepsilon > 0$

Theorem: For a persistent on-the-average policy:

$$P(\mathbf{T}^* < \infty) = 1$$

$$P(\mathbf{T}^* \leq t) \geq 1 - (1-\varepsilon)^{\left\lfloor \frac{t}{T} \right\rfloor}$$

$$\mathrm{E}[\mathbf{T}^*] \leq \frac{T}{\varepsilon}$$

where $\mathbf{T}^* \equiv$ time at which the evader is found

Persistent on-the-average policies

Definition: A policy is ***persistent on-the-average*** if there is an integer T such that for every time t

The case of a greedy policy over an horizon T

$$p_T(t) \geq \varepsilon \qquad p_T(t) \equiv \text{P(finding evader from } t \text{ to } t+T\text{-1, given that none was found up to } t)$$

for some $\varepsilon > 0$

Theorem: For a persistent on-the-average policy:

$$P(\mathbf{T}^* < \infty) = 1$$

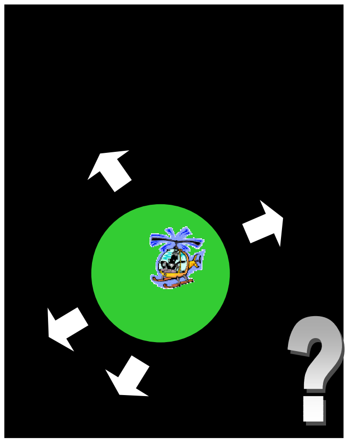
$$P(\mathbf{T}^* \leq t) \geq 1 - (1-\varepsilon)^{\left\lfloor \frac{t}{T} \right\rfloor}$$

$$\mathrm{E}[\mathbf{T}^*] \leq \frac{T}{\varepsilon}$$

where $\mathbf{T}^* \equiv$ time at which the evader is found

Greedy policies

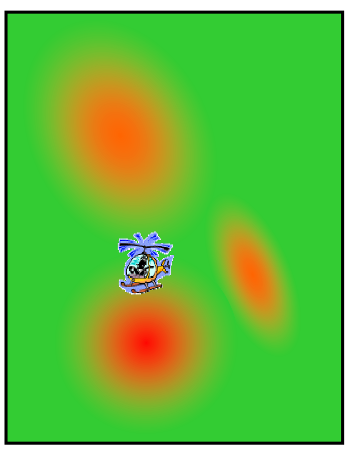
greedy policy ≡ pursuit policy that, at each instant, maximizes the probability of finding the evader at the next time instant



Q: How does one evaluate which control action will maximize the probability of finding the evader?

Greedy policies

greedy policy ≡ pursuit policy that, at each instant, maximizes the probability of finding the evader at the next time instant



Q: How does one evaluate which control action will maximize the probability of finding the evader?

A: By keeping track of the probability of the evader being at each possible position

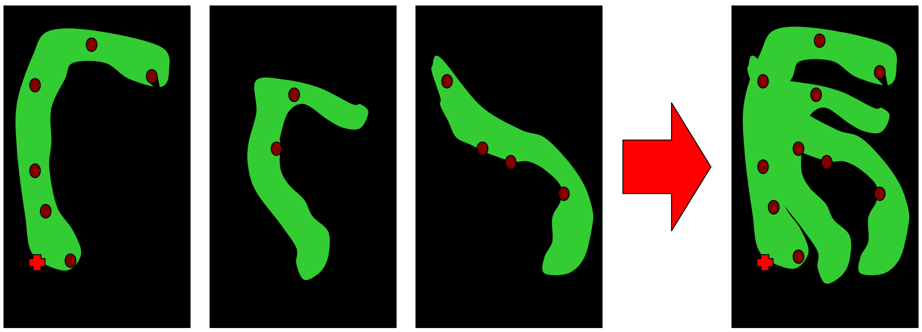
$$P(\text{evader}(t+1) = x \mid y_1, y_2, \ldots, y_t)$$

probabilistic map for the position of the evader

Map building

Problem:

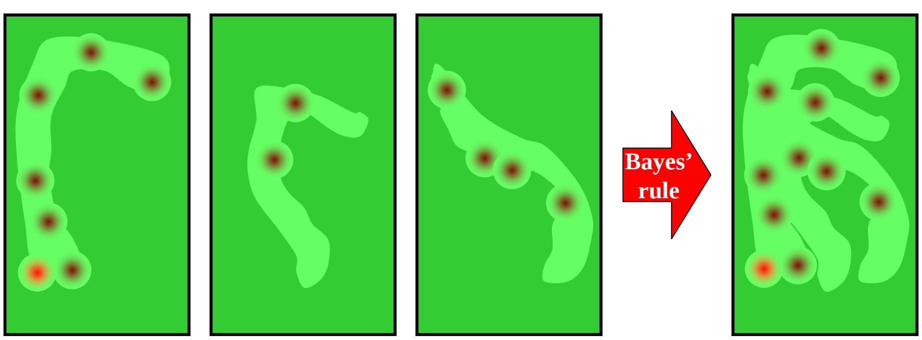
How to combine data sensed by one or several UAVs over a period of time to build a global probabilistic map?



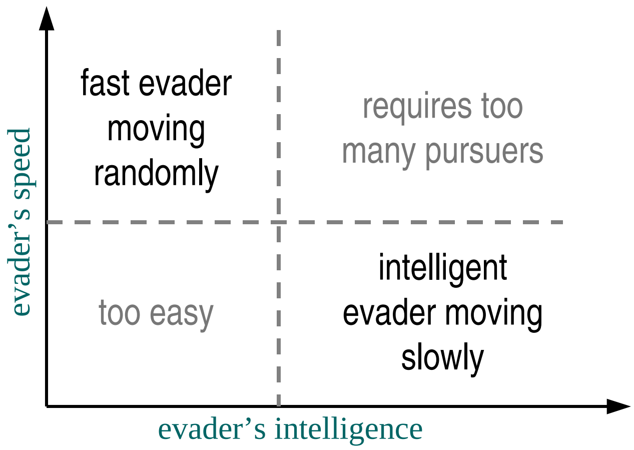
Probabilistic map building

Problem:

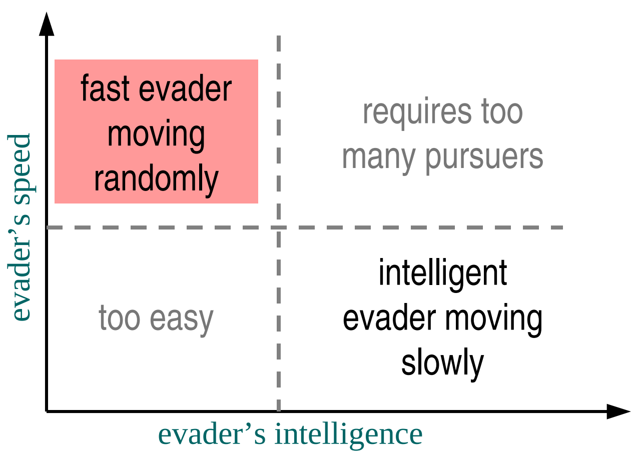
How to combine data sensed by one or several UAVs over a period of time to build a global map?



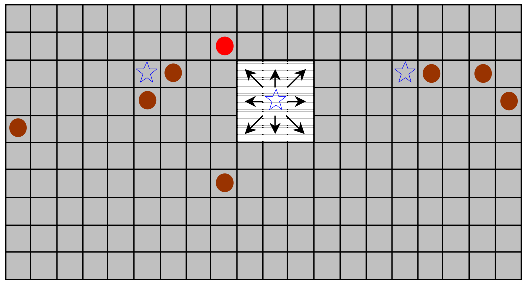
Scenarios: intelligence v.s. speed



Scenarios: intelligence v.s. speed



Scenarios: fast evader moving randomly

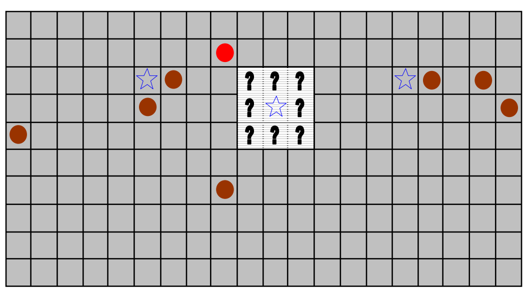


- ***region***: 2-dimensional (200 cells)
- ***obstacles***: uniformly distributed
- ***agents***: 3 pursuers & 1 evader

} Markov chain with 2.6×10^{69} states

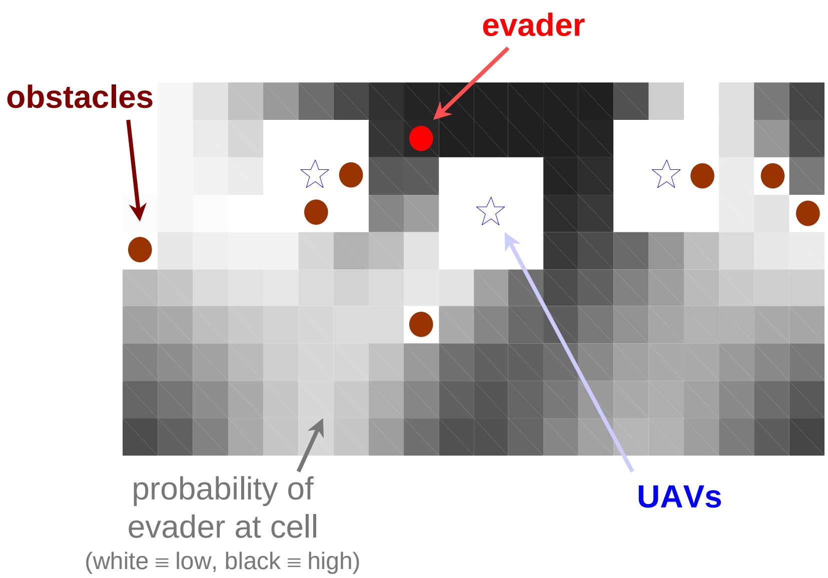
- ***actions***: agents can move to any empty adjacent cell w.p.1 by selecting appropriate action

Scenarios: fast evader moving randomly

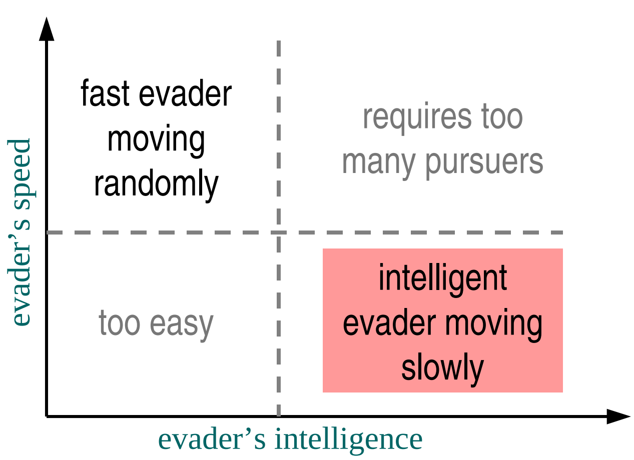


- ***observations***:
  - *pursuers* sense adjacent cells for obstacles and evader (1% of false positives & 1% of false negatives) } ***greedy policy***
  - *evader* is "blind" } ***random policy*** (unif. dist. over actions)

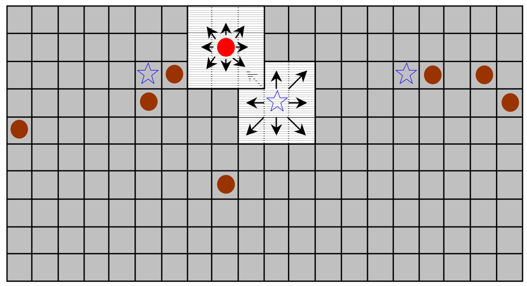
Simulation



Scenarios: intelligence v.s. speed



Scenarios: intelligent evader moving slowly



- ***region***: 2-dimensional (200 cells)
- ***obstacles***: uniformly distributed
- ***agents***: 3 pursuers & 1 evader

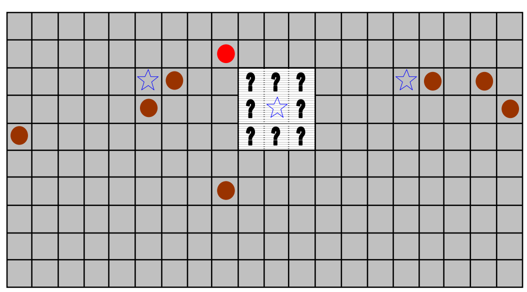
speed = 1 cell/time unit

- ***actions***:

  average speed = v cells/time unit

  – *pursuers* can move to any adjacent cell w.p.1
  – *evader* moves to adjacent cells w.p. $v < 1$ ($v = .5$)

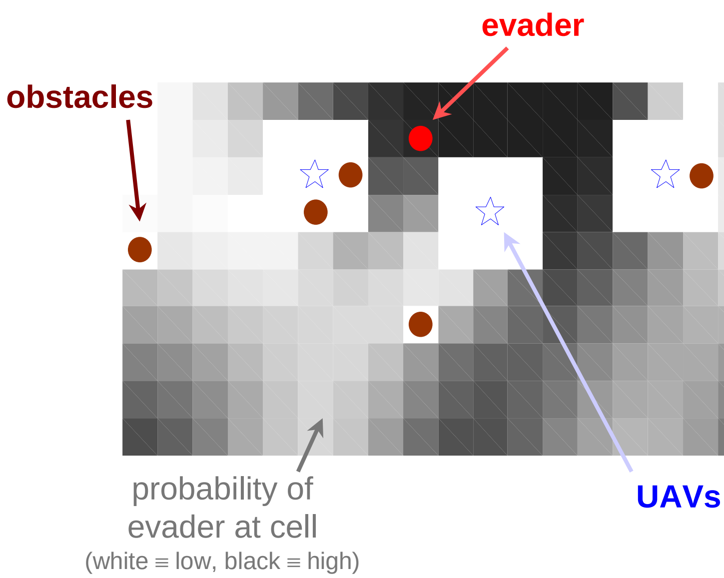
Scenarios: intelligent evader moving slowly



- ***observations***:
  – *pursuers* sense adjacent cells for obstacles and evader (1% of false pos. and 1% of false neg.) ***greedy policy***
  – *evader* has full information (positions and internal state of agents)

1-step Nash policy maximizing Euclidean distance to pursuers

Simulation



Conclusions

- Finding optimal policies for competitive, hidden, MDPs is ***computationally difficult***
- There are computationally efficient ***sub-optimal policies*** with guaranteed performance
- It is worth exploring techniques from reinforcement learning and neuro-dynamic programming