Optimal control of stochastic multi-agent systems in continuous space and time

Wim Wiegerinck, Bart van den Broek, Bert Kappen

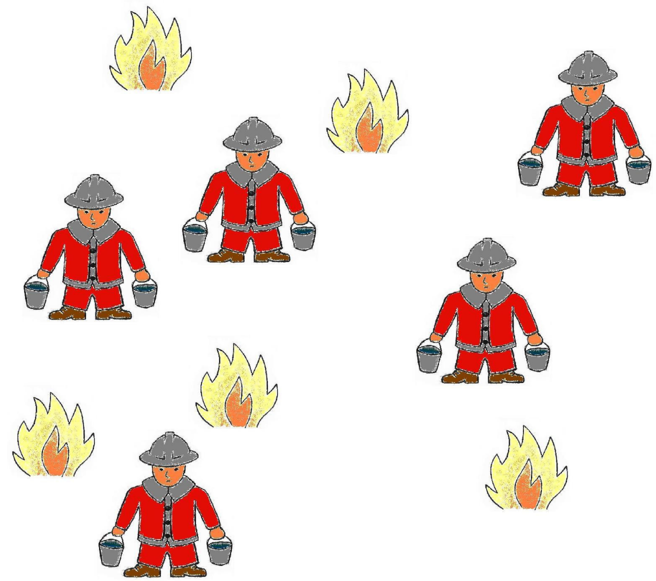
SNN, Radboud University Nijmegen

Presented at UAI 2006

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant BSIK03024.

Stochastic multi-agent systems and optimal control

- Multi-agent systems (e.g. firemen - see figure) that have to distribute themselves over a number of targets (e.g. fires)
  - noisy, non-linear dynamics in continuous space-time
  - additive control of the dynamics
- Optimal control:
  - minimize total joint cost
    (= effort cost + end cost)
  - to which fire should a fireman go?
  - when to decide?



Contents

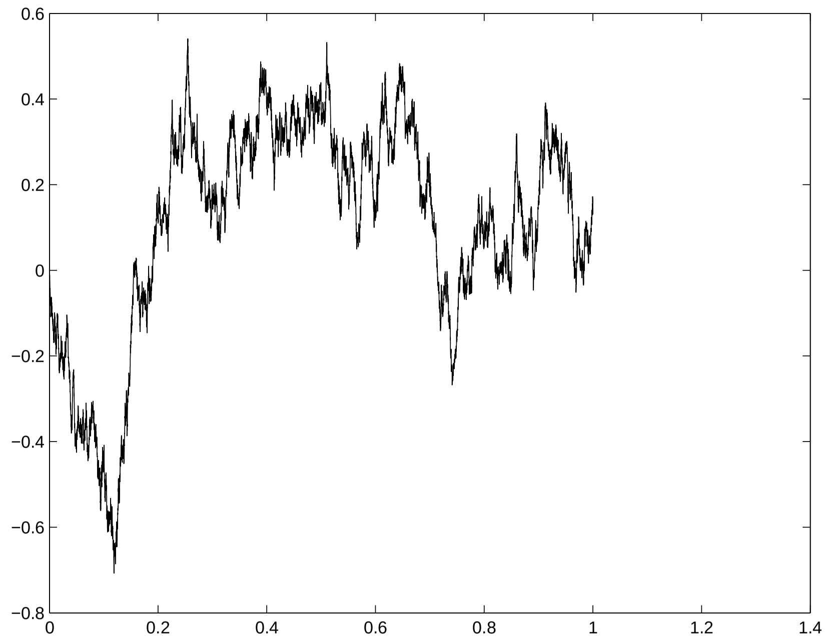
- Stochastic optimal control in continuous space time, Hamilton-Jacobi-Bellman equation
- Transform into a linear PDE
- From single-agent single-target systems to multi-agent multi-target systems
- MAS control and graphical models
- Simulations
- Summary, discussion

Stochastic dynamics

We consider a system in continuous space and time. Its state $\boldsymbol{x} \in R^k$ obeys the stochastic dynamics

$$d\boldsymbol{x} = (\boldsymbol{b}(\boldsymbol{x}, t) + \boldsymbol{u})dt + d\boldsymbol{\xi}$$

- $\boldsymbol{b}$: drift term modeling the dynamics due to the environment,
- $\boldsymbol{u}$: control to influence the dynamics,
- $d\boldsymbol{\xi}$ a Wiener process (i.e. noise) with $\langle d\xi_i d\xi_j \rangle = \boldsymbol{\nu}_{ij} dt$.



Control problem

Find the control $\boldsymbol{u}(.)$ that minimizes the expected cost-to-go

$$C(\boldsymbol{x}_i, t_i, \boldsymbol{u}(.)) = \left\langle \phi(\boldsymbol{x}(T)) + \int_{t_i}^{T} dt \Big(\frac{1}{2} \boldsymbol{u}(\boldsymbol{x}, t)^\top \boldsymbol{R} \boldsymbol{u}(\boldsymbol{x}, t) + V(\boldsymbol{x}(t), t) \Big) \right\rangle$$

in which

- $\boldsymbol{x}_i$: initial state,
- t_i: initial time,
- T: Fixed end-time,
- $\phi(\boldsymbol{x}(T))$: cost of being in state $\boldsymbol{x}$ at end-time T,
- $V(\boldsymbol{x}(t), t)dt$: cost of being in state $\boldsymbol{x}$ during time interval $[t, t + dt]$,
- $\boldsymbol{u}^\top \boldsymbol{R} \boldsymbol{u} dt$: cost of control during the same time interval,
- $\boldsymbol{R}$ is a constant $k \times k$ matrix (parametrizing the cost of control).

Hamilton-Jacobi-Bellman equation and optimal control

- Optimal (expected) cost-to-go

$$J(\boldsymbol{x},t) = \min_{\boldsymbol{u}(.)} C(\boldsymbol{x},t,\boldsymbol{u}(.)).$$

- It satisfies the stochastic Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{aligned} -\partial_t J &= \min_{\boldsymbol{u}(.)} \left(\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{R}\boldsymbol{u} + V + (\boldsymbol{b}+\boldsymbol{u})^\top \nabla J + \frac{1}{2}\mathrm{Tr}(\boldsymbol{\nu}\nabla^2 J) \right) \\ &= -\frac{1}{2}\nabla J^\top \boldsymbol{R}^{-1}\nabla J + V + \boldsymbol{b}^\top \nabla J + \frac{1}{2}\mathrm{Tr}(\boldsymbol{\nu}\nabla^2 J) \end{aligned}$$

  with boundary condition $J(\boldsymbol{x},T) = \phi(\boldsymbol{x})$.

- The minimization with respect to $\boldsymbol{u}$ yields

$$\boldsymbol{u} = -\boldsymbol{R}^{-1}\nabla J,$$

  which defines the optimal control.

Log transformation and linear evolution

Assume $\boldsymbol{\nu} = \lambda \boldsymbol{R}^{-1}$, then the non-linear PDE of J can transformed into a linear one by the log transform (W. Flemming, 1978). Set

$$J(\boldsymbol{x}, t) = -\lambda \log Z(\boldsymbol{x}, t)$$

with “partition function”

$$Z(\boldsymbol{x}, t) = \int d^k y \rho(\boldsymbol{y}, T|\boldsymbol{x}, t) \exp(-\phi(\boldsymbol{y})/\lambda)$$

in which the ‘probability density’ ρ satisfies the Fokker-Planck equation

$$\partial_\vartheta \rho(\boldsymbol{y}, \vartheta|\boldsymbol{x}, t) = -\frac{V}{\lambda}\rho - \nabla_y^\top b\rho + \frac{1}{2}\mathrm{Tr}\big(\boldsymbol{\nu}\nabla_y^2 \rho\big).$$

$\rho(\boldsymbol{y}, T|\boldsymbol{x}, t)$: probability of getting at $\boldsymbol{y}$ at time T given initial state $\boldsymbol{x}$ at time t,

- following stochastic system dynamics in absence of control, i.e., $\boldsymbol{u} = 0$
- with annihilation probability $V(\boldsymbol{x}, t)dt/\lambda$

Linear theory

- Expected optimal cost to go

$$J(\boldsymbol{x}, t) = \lambda \log Z(\boldsymbol{x}, t)$$

- Z is expressed as

$$Z(\boldsymbol{x}, t) = \int d^k y \rho(\boldsymbol{y}, T|\boldsymbol{x}, t) \exp(-\phi(\boldsymbol{y})/\lambda)$$

in which ρ satisfies the Fokker-Plank equation.

- The optimal control is given by

$$\boldsymbol{u}(\boldsymbol{x}, t) = \boldsymbol{\nu} \nabla \log Z(\boldsymbol{x}, t) \ .$$

Example: Quadratic end-cost

- $\boldsymbol{b} = 0, V = 0$
- R and ν scalars,
- Solution of the diffusion equation

$$\rho(\boldsymbol{y}, T|\boldsymbol{x}, t) = (2\pi\nu(T-t))^{k/2} \exp(-\frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{2\nu(T-t)})$$

- Assume end-cost: $\phi(\boldsymbol{x}) = \alpha|\boldsymbol{x} - \mu|^2$
- Then Z follows from convolution with $\exp(-\phi(\boldsymbol{y})/\lambda)$,

$$Z(\boldsymbol{x}, t) \propto \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{\mu}|^2}{2\nu(T-t+R/\alpha)}\right).$$

- The optimal control follows from $\boldsymbol{u} = \nu\nabla\log Z$,

$$u(\boldsymbol{x}, t) = \frac{\boldsymbol{\mu}-\boldsymbol{x}}{T-t+R/\alpha}.$$

Single target

- Back to arbitrary $\boldsymbol{b}$ and V.
- To enforce an end-state at target $\boldsymbol{\mu}$, we set

$$\exp(-\phi(\boldsymbol{y})/\lambda) \propto \delta(\boldsymbol{y} - \boldsymbol{\mu})$$

- This implies

$$Z(\boldsymbol{x}, t; \boldsymbol{\mu}) \propto \rho(\boldsymbol{\mu}, T | \boldsymbol{x}, t) \ ,$$

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{x}, t; \boldsymbol{\mu}) &= \boldsymbol{\nu} \nabla \log Z(\boldsymbol{x}, t; \boldsymbol{\mu}) \\ &= \boldsymbol{\nu} \nabla \log \rho(\boldsymbol{\mu}, T | \boldsymbol{x}, t) \end{aligned}$$

Running example

Assume $\boldsymbol{b} = 0$, $V = 0$, R and ν scalars.

$$\begin{aligned} Z(\boldsymbol{x}, t; \boldsymbol{\mu}) &\propto \exp\left[-\frac{|\boldsymbol{x} - \boldsymbol{\mu}|^2}{2\nu(T - t)}\right] , \\ \boldsymbol{u}(\boldsymbol{x}, t; \boldsymbol{\mu}) &= \frac{\boldsymbol{\mu} - \boldsymbol{x}}{T - t}. \end{aligned}$$

- For any $\boldsymbol{b}$ linear in $\boldsymbol{x}$, and $V = 0$, in the Fokker-Planck equation can be solved analytically. Its solution is a Gaussian. Z and u are of essentially the same form as in the running example.

Multiple targets

To enforce an end-state at one of m targets $\boldsymbol{\mu}_s$, with target preferences expressed by relative cost $E(s)$,

$$\exp(-\phi(\boldsymbol{y})/\lambda) \propto \sum_{s=1}^{m} \exp(-E(s)/\lambda)\delta(\boldsymbol{y} - \boldsymbol{\mu}_s) \ .$$

Partition function and optimal control can be expressed as sum of single-target quantities,

$$\begin{aligned} Z(\boldsymbol{x}, t) &\propto \sum_{s=1}^{m} \exp(-E(s)/\lambda) Z(\boldsymbol{x}, t; \boldsymbol{\mu}_s) \ , \\ \boldsymbol{u}(\boldsymbol{x}, t) &= \sum_{s=1}^{m} p(s|\boldsymbol{x}, t) \boldsymbol{u}(\boldsymbol{x}, t; \boldsymbol{\mu}_s) \ , \end{aligned}$$

in which $p(s|\boldsymbol{x}, t)$ is the probability

$$p(s|\boldsymbol{x}, t) = \frac{\exp(-E(s)/\lambda) Z(\boldsymbol{x}, t; \boldsymbol{\mu}_s)}{\sum_{s'=1}^{m} \exp(-E(s')/\lambda) Z(\boldsymbol{x}, t; \boldsymbol{\mu}_{s'})} .$$

Example: Multiple targets

In the running example, optimal control with multiple targets is

$$\boldsymbol{u}(\boldsymbol{x},t) = \sum_s p(s|\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t;\boldsymbol{\mu}_s) = \sum_s p(s|\boldsymbol{x},t)\left(\frac{\boldsymbol{\mu}_s-\boldsymbol{x}}{T-t}\right) = \frac{\bar{\boldsymbol{\mu}}-\boldsymbol{x}}{T-t}$$

with $\bar{\boldsymbol{\mu}}$ the ‘expected target’

$$\bar{\boldsymbol{\mu}} = \sum_{s=1}^{m} p(s|\boldsymbol{x},t)\boldsymbol{\mu}_s$$

which is the expected value of the target according to the probability

$$p(s|\boldsymbol{x},t) = \frac{\exp(-E(s))\exp\left[-\frac{|\boldsymbol{x}-\boldsymbol{\mu}_s|^2}{2\nu(T-t)}\right]}{\sum_{s=1}^{m}\exp(-E(s))\exp\left[-\frac{|\boldsymbol{x}-\boldsymbol{\mu}_s|^2}{2\nu(T-t)}\right]}$$

Multi-agents, multiple targets

- MAS state $\vec{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, single agent state $\boldsymbol{x}_a$.
- Dynamics non-interactive:
  - $\boldsymbol{b}_a(\vec{\boldsymbol{x}}, t) = \boldsymbol{b}_a(\boldsymbol{x}_a, t)$
  - $V(\vec{\boldsymbol{x}}, t) = \sum_a V_a(\boldsymbol{x}_a, t)$.
  - Furthermore, $\boldsymbol{\nu} = \lambda \boldsymbol{R}^{-1}$ globally.
  - Therefore ρ factorizes,

$$\rho(\vec{\boldsymbol{y}}, T|\vec{\boldsymbol{x}}, t) = \prod_a \rho_a(\boldsymbol{y}_a, T|\boldsymbol{x}_a, t) \ .$$

- Coupling via joint task: distribute over targets

$$\exp(-\phi(\vec{\boldsymbol{y}})/\lambda) = \sum_s \exp(-E(\vec{s})/\lambda) \prod_a \delta(\boldsymbol{y}_a - \boldsymbol{\mu}_{s_a}) \ .$$

  - s_a label of target reached by agent a.
  - $E(\vec{s}) = E(s_1, \ldots, s_n)$ is the cost when agent 1 reaches $\boldsymbol{\mu}_{s_1}$, agent 2 reaches $\boldsymbol{\mu}_{s_2}$ etc.

Multi-agents, multiple targets (cont)

- Control for agent a (depends on joint state)

$$\boldsymbol{u}_a(\vec{\boldsymbol{x}}, t) = \sum_{s_a=1}^{m} p(s_a|\vec{\boldsymbol{x}}, t)\boldsymbol{u}_a(\vec{\boldsymbol{x}}_a, t; \boldsymbol{\mu}_{s_a}) \ .$$

Control involves marginal distribution for agent a

$$p(s_a|\vec{\boldsymbol{x}}, t) = \frac{\sum_{\vec{s}\setminus s_a} \exp(-E(\vec{s})/\lambda) \prod_b Z_b(\vec{\boldsymbol{x}}_b, t; s_b)}{\sum_{\vec{s}} \exp(-E(\vec{s})/\lambda) \prod_c Z_c(\vec{\boldsymbol{x}}_c, t; s_b)} \ ,$$

Example: MAS, multiple targets

In the running example, optimal control with multiple targets is

$$\boldsymbol{u}_a(\vec{\boldsymbol{x}}, t) \quad = \quad \frac{\bar{\boldsymbol{\mu}}_a - \boldsymbol{x}_a}{T - t}$$

with $\bar{\boldsymbol{\mu}}$ the 'expected target' for agent a

$$\bar{\boldsymbol{\mu}}_a = \sum_{s=1}^{m} p(s_a|\vec{\boldsymbol{x}}, t)\boldsymbol{\mu}_{s_a}$$

which is the expected value of the target according to the probability

$$p(s_a|\vec{\boldsymbol{x}}, t) \propto w(s_a|\vec{\boldsymbol{x}}_{\setminus a}, t) \exp\left[-\frac{|\boldsymbol{x}_a - \boldsymbol{\mu}_{s_a}|^2}{2\nu(T-t)}\right]$$

with

$$w(s_a|\vec{\boldsymbol{x}}_{\setminus a}, t) = \sum_{\vec{s}\setminus s_a} \exp(-E(\vec{s})) \exp\left[-\frac{\sum_{b\neq a} |\boldsymbol{x}_b - \boldsymbol{\mu}_{s_b}|^2}{2\nu(T-t)}\right]$$

MAS control and graphical models

- MAS control requires inference of $p(s_a|\vec{\boldsymbol{x}}, t)$.
- Graphical model methods can be exploited if $p(\vec{s})$ is a *factor graph*, i.e. if the cost can be written

$$E(\vec{s}) = \sum_{\alpha} E_\alpha(s_\alpha)$$

with α groups of agents. E.g. pairwise costs

$$E_{ab}(s_a, s_b) = -c_{ab}\delta_{s_a s_b} \ .$$

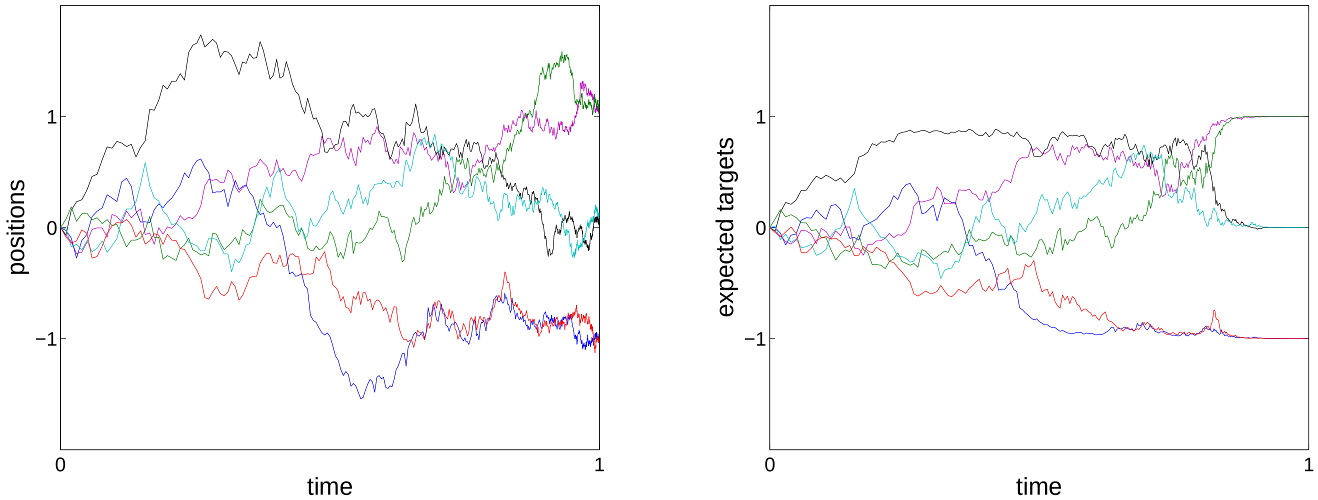
- Boltzmann machine analogy
  - $E_{a,b}(s_a, s_b)$ plays role of couplings in a Boltzmann machine, constant in the system
  - $Z_a(\boldsymbol{x}_a, t; \mu_{s_a})$, (i.e., $\rho(\mu_{s_a}, T|\boldsymbol{x}_a, t)$) plays role of bias in a BM, changes over time
- In general: graphical structure is preserved over time (unlike discrete time factored MDPs).
- Sparse graphs $\Rightarrow$ junction tree algorithm.

Simulations

- $b = 0$, $V = 0$, pairwise relations c_{ab}.
- 1-d 'Positions' x_a.
  - for plotting purposes. k-dimensional would be feasible as well.
- 'Expected targets' $\overline{\mu}_a = \sum_{s_a} p_a(s_a|x)\mu_{s_a}$.
- *Only for illustration:* $u_a = \frac{\overline{\mu}_a - x_a}{\nu(T-t)}$ *is mathematically optimal!*

Firemen problem

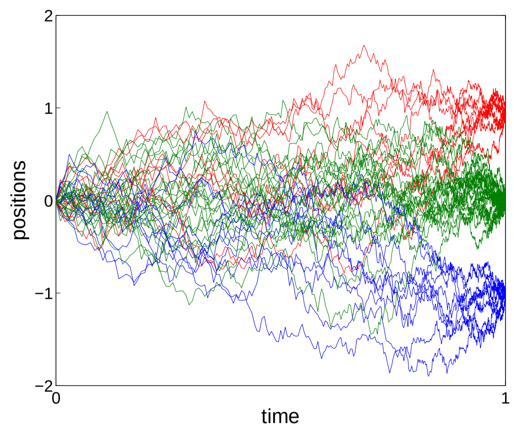
- 6 agents (firemen), three targets (fires).
- Fully connected graph with $c_{ab} = c$ negative $\Rightarrow$ aim to distribute evenly.

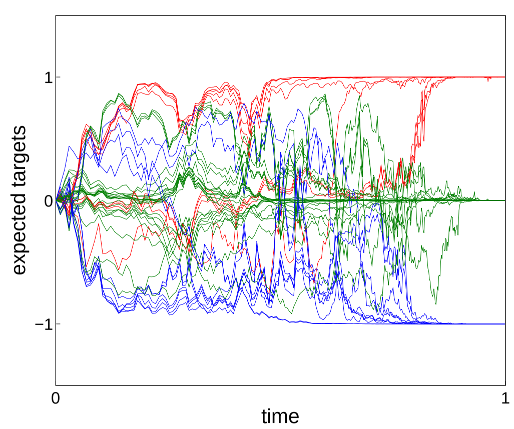


- Symmetry breaking as delayed choice (Kappen, 2005)

Holiday resort problem

- 42 agents, 3 targets (resorts).
- E represented by sparse graph: each agent has pairwise relations, $c_{ab} = \pm 1$, with three other agents. Agent only cares for related agents whether or not to have holiday in the same resort.
- Joint task is to optimally distribute MAS over the resorts.
- Clique-size = 7





Summary and discussion

We studied optimal control in stochastic MAS in continuous space-time

- Optimal control can be derived from the solution of Hamilton-Jacobi-Bellman equations
- Under some conditions, the log-transformation transforms the non-linear HJB equations (a non-linear PDE) into a linear PDE
- under these conditions, a superposition principle holds. This enables us to generate MAS multi-target solutions from single agent single-target solutions
- Additional computational cost for MAS involves probabilistic inference. which is tractable in sparsely connected systems.
- In dense MASs, exact inference is intractable; a natural approach would be approximate inference using message passing algorithms. (current study).
- In linear models, with $V = 0$ and no agent interactions, optimal control in MAS multi-target systems is solved analytically.
- In general, however, even the single agent problem requires computational intensive approximations (Kappen, 2005).