ABSTRACT
We consider multi-agent systems with stochastic non-linear dynamics in continuous space-time. We focus on systems of agents that aim to visit a number of given target locations at given points in time at minimal control cost. The on-line optimization of which agent has to visit which target requires the solution of the Hamilton-Jacobi-Bellman (HJB) equation, which is a non-linear partial differential equation (PDE). Under some conditions, the log-transform can be applied to turn the HJB equation into a linear PDE. We then show that the optimal solution in the multi-agent scheduling problem can be expressed in closed form as a sum of single schedule solutions.

Categories and Subject Descriptors
G.1.6 [Numerical analysis]: Optimization—stochastic programming; I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search—control theory, dynamic programming, scheduling; I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—Multiagent systems; G.3 [Probability and statistics]: Stochastic processes

General Terms
Algorithms, Theory

Keywords
continuous-time, exact optimal controls, multialgent systems, nonlinear stochastic systems, optimal stochastic control, stochastic processes

1. INTRODUCTION
A collaborative multi-agent system (MAS) is a collection of agents that autonomously control their behavior to achieve a common goal or to maximize the performance of the group. Examples are teams of soccer-robots and teams of unmanned rescue vehicles in a hazardous disaster area. In practical applications, agents often have to deal with stochastic dynamics, e.g., due to noise from the environment, and with limitations of resources.

In this paper, we are interested in optimal control in multi-agent systems that have to operate in a space-like environment in continuous time. The typical example is a group of micro air vehicles (MAVs) whose mission is to visit a number of predefined targets, e.g., for surveillance purposes. The agents dynamics is subject to noise. In the example of MAVs, this may be due to local winds and turbulences in the air. A typical mission is that the agents start at some initial positions and distribute themselves such that each target is reached at the required times by at least one vehicle. Then the agents have to return to base at given end-time. The agents should continuously control themselves such that the mission is realized at minimal expected effort. The additional complexity is that due to the noise in the dynamics, a schedule that seems optimal from the initial positions may become suboptimal in a later stage. Thus an on-line schedule could be beneficial.

In this paper we consider optimal control in multi-agent Markov Decision Processes (MDPs). The system is fully observable, i.e. each agent knows the joint utility function as well as the current and past state of itself and of the other agents. In addition, the common knowledge that each agent follows the same optimization approach is assumed. State transitions, however are stochastic. This makes optimization of control non-trivial.

A common approach is to model such a system as an MDP in discrete space and time: the optimal actions in an MDP optimization problem are in principle solved by backward dynamic programming. In general, the joint action space and the joint state space of the agents will be large due discretization. Since the number of states will also increase exponentially in the number of agents, this approach will generally be infeasible [1]. Often, one can describe the system more compactly as a factored MDP. In such systems both the transition probabilities and reward functions have some structure. Unfortunately, this structure is not conserved in the value functions and exact computation remains exponential in the system size. Recently, a number of advanced and powerful approximate methods have been proposed. The common denominator of these approaches is that they basically assume some predefined approximate structure of the
value functions [4, 5].

In this paper, we remain in the continuous domain. We consider agents that evolve under a non-linear dynamics with additive Wiener noise. The optimal control follows from the solution of the Hamilton-Jacobi-Bellman equation, which is a non-linear partial differential equation (PDE) [9]. By assuming that control is additive to the dynamics and contributes quadratically to the cost, in such a way that control in directions with strong noise is cheap and in directions with weak noise expensive, we can apply the log-transform. This transform turns this nonlinear PDE into a linear PDE [2, 3, 6, 7].

We then follow the approach in [11] where a similar problem is considered: the optimal control of a system of agents that have to distribute themselves over a number of targets at a single given end-time. If the log-transform can be applied, the optimal control solution in the multi-agent multi-target system can be shown to be the sum of single-agent single-target optimal controls weighted by a factor that decreases with the expected costs to reach each target.

In this paper we further extend the work in [11] by considering the optimal control of agents that have to visit a number of targets at different points of time. The system has to not only to decide on which agent goes to which target, but also on the ordering of the visits. In the decision of which targets are to be visited first, the agents have to take into account possible consequences at later times. We call each ordering of visits a schedule. So, in addition to the stochastic optimization of the control to a given target, the system should decide on-line how it optimally realizes at least one of the schedules.

To deal with this problem, we generalize the traditional stochastic optimal control problem by replacing the end-cost that is usually included in the stochastic cost function by a term that also depends on the states of the system at earlier times, i.e., a cost potential that depends on the path of the system rather than only the end-state. In section 2, we review the traditional framework of optimal control in continuous stochastic systems [9] and the solution with the log transform, as described in [6, 7] and closely following [11]. In section 3 we generalize the framework by including a path dependent cost potential and we provide, using the log transform, an optimal control solution in the extended framework. In section 4, we first apply the theory to an agent that has to control itself to a single target, discussed earlier in [11]. Then we consider a schedule, i.e. a sequence of targets that have to be visited after each other at prescheduled times. Since this is just a concatenation of single-target-single time problems, it is not suprisingly that the control is just control to the next target in line. However, the optimal expected cost-to-go will be affected by the targets later in the schedule. In section 5, we consider online scheduling, i.e. where the control should be such that at end-time one of the schedules is realized at minimal expected cost. We show that in the log-transform framework, the optimal control can be expressed as a weighted combination of single schedule controls, weighted by a factor that decreases with the expected cost-to-go for each schedule. This result is a direct generalization of the result in [11] mentioned earlier. We illustrate the result by a numerical example where an agent has to hit a moving target before a given end-time (rather than exactly at the given end-time). Optimal control can be calculated by modeling this problem with schedules. In section 6, we show that the result generalizes straightforwardly to on-line scheduling in multi-agent systems. This is analogous to the multi-agent control to targets at end-time discussed in [11]. Numerical examples are provided of a two agent system, that have to visit a number of targets at a number of times. Finally, we end with some conclusions in section 7.

2. A REVIEW OF STOCHASTIC OPTIMAL CONTROL THEORY

In this section, we follow the review in [11] of the stochastic control framework with the log-transform as developed and described in [2, 6, 7].

We consider an agent moving in $\mathbb{R}^k$. Its position $x$ obeys the stochastic dynamics

$$dx = (b(x,t) + u)dt + d\xi,$$  \hspace{1cm} (1)

with $d\xi$ a Wiener process with $\langle d\xi, d\xi \rangle = \nu dt$, $b(x,t)$ an arbitrary function of $x$ and $t$, modeling the dynamics due to the environment. The dynamics can be influenced the additive control $u$.

Given $x$ at initial time $t_0$, the problem is to find a control policy $u(\cdot)$ from $t_0$ to the end-time $t_n$ such that the expected cost-to-go

$$C(u(t_0 \rightarrow t_n), x, t_0, \cdot) = \left< \phi(x(t_n)) \right. \left. + \int_{t_0}^{t_n} \left( \frac{1}{2}u(t)^T Ru(t) + V(x(t), t) \right) dt \right>$$ \hspace{1cm} (2)

is minimal. The expectation is taken over all noise realizations, resulting in different trajectories $x(t)$ in state space that start in $x(t_0) = x$. $\phi(x(t_n))$ is the end cost, depending only on the end state $x(t_n)$. $V(x(t), t)dt$ is the cost of being at position $x(t)$ during the time interval $[t, t+dt]$. $u(t)^T Ru(t) dt$ is the cost of the control during the same time interval. $R$ is a constant $k \times k$ matrix.

The expected cost-to-go at time $t$ needs to be minimized over all strategies $u(t \rightarrow t_n)$, this yields the optimal (expected) cost-to-go

$$J(x, t) = \min_{u(t \rightarrow t_n)} C(u(t \rightarrow t_n), x, t).$$ \hspace{1cm} (3)

In the appendix, it is briefly explained that due to the linear-quadratic form of the optimization problem—the dynamics (1) is linear in the action $u$, the cost (2) is quadratic in the action—the minimization with respect to $u(\cdot)$ can be performed explicitly, yielding a non-linear partial differential equation in $J$, the well-known Hamilton-Jacobi-Bellman equation [9]. If, in addition, the matrices $\nu$ and $R$ can be linked via a scalar $\lambda$ such that $\nu = \lambda R^{-1}$, the log-transform can be applied, and the optimal cost-to-go is re-expressed as the log of an ordinary integral,

$$J(x, t) = -\lambda \log Z(x, t)$$ \hspace{1cm} (4)

with “partition function”

$$Z(x, t) = \int \rho(y, t_n|x, t) \exp \left( -\frac{\phi(y)}{\lambda} \right) dy$$ \hspace{1cm} (5)

in which $\rho(y, t_n|x, t)$ is the probability of arriving in state $y$ at time $t_n$, when starting in $x$ at time $t$, under the dynamics (55) in the appendix. This dynamics (55) is a diffu-
sion process, it equals the stochastic system dynamics without control, i.e., \( u = 0 \), with in addition the probability \((V(x,t)/\lambda)dt\) of being removed from the system between \( t \) and \( t + dt \) and thus not arriving in state \( y \).

The main advantage of the log-transform is that the HJB equations which is a nonlinear PDE in \( \rho \) is turned into a linear PDE in \( \rho \), which is often much easier to solve—at least approximately (although still difficult, see [6, 7]). In fact, for any \( b(x) \) linear in \( x \), and \( V = 0 \), the PDE, which is well-known as the Fokker-Planck equation, can be solved analytically, and its solution is known to be a Gaussian, see e.g. [10]. In particular, with \( b = 0 \), \( V = 0 \) and \( \nu > 0 \) a scalar, the solution is

\[
\rho(y,t_n | x,t) = (2\pi \nu (t_n - t))^{-k/2} \exp \left[ -\frac{|y-x|^2}{2\nu (t_n - t)} \right].
\]

(6)

Exact solutions of Fokker-Planck equations with nonlinear drift terms \( b \) have been studied in e.g. [8].

The optimal control of the agent is directly obtained from the optimal cost-to-go, by taking its gradient (equation (52) in the appendix), which implies the following relation between control and partition function,

\[
u (x,t) = \nu \partial_x \log Z(x,t).
\]

(7)

Finally we remark the optimal control problem at time \( t \) with end-cost \( \phi(x) \) is equivalent to the case with optimal cost-to-go \( J(x,t') \) at intermediate time \( t < t' < t_n \), i.e.,

\[
\int \rho(y,t_n | x,t) \exp \left( -\frac{\phi(y)}{\lambda} \right) dy = \int \rho(y',t' | x,t) \exp \left( -\frac{J(y,t')}{\lambda} \right) dy'.
\]

(8)

This implies the following relation for the partition function

\[
Z(x,t) = \int \rho(y',t' | x,t) Z(y',t') dy'.
\]

(9)

3. PATH DEPENDENT COST POTENTIAL

In this section we extend the framework outlined in the previous section by replacing the end-cost \( \phi \) in the cost-to-go (2) by a potential that not only depends on the state at end-time \( t_n \) but also on the states at a number of predefined earlier times \( t_1, \ldots, t_n \). Under the same dynamics as previously, the objective is now to find a control that minimizes

\[
C(u(t_0 \rightarrow t_n), x(t_n), t_0) = \left\langle \phi(x(t_1), \ldots, x(t_n)) \right\rangle
\]

\[
+ \int_{t_0}^{t_n} \left( \frac{1}{2} u(t)^T R u(t) + V(x(t),t) \right) dt.
\]

(10)

For an intermediate time \( t \), with \( t_k < t < t_{k+1} \), the expected cost-to-go is

\[
C(u(t \rightarrow t_n), x(t_1), \ldots, x(t_k), x(t), t) = \left\langle \phi(x(t_1), \ldots, x(t_k)) \right\rangle
\]

\[
+ \int_{t}^{t_k} \left( \frac{1}{2} u(t')^T R u(t') + V(x(t'),t') \right) dt'.
\]

(11)

which depends on the current state \( x(t) \) but also on the past states \( x(t_1), \ldots, x(t_k) \). The optimal expected cost-to-go is

\[
J(x(t_1), \ldots, x(t_k), x(t), t) = \min_{u(t \rightarrow t_n)} C(u(t \rightarrow t_n), x(t_1), \ldots, x(t_k), x(t), t).
\]

(12)

The \( J \) in (12) can be understood as an conditional optimal cost-to-go \( J_k \), conditioned on the states visited at the past times \( t_1, \ldots, t_k \).

\[
J_k(x(t), t \mid x(t_1), \ldots, x(t_k)) = J(x(t_1), \ldots, x(t_k), x(t), t).
\]

(13)

The conditional optimal cost-to-go functions \( J_k \) are connected via the boundary conditions

\[
J_k(x(t_{k+1}), x_{k+1} | x(t_1), \ldots, x_k) = J_{k+1}(x(t_{k+1}), x_{k+1} | x(t_1), \ldots, x_{k+1}).
\]

(14)

To proceed, let us assume that the time \( t \) is in the interval \( t_k \leq t < t_{k+1} \), that the previous states \( x(t_1) = x_1, \ldots, x(t_k) = x_k \) are given and that we know the optimal cost to go \( J_k(x_{k+1}, t_{k+1} | x_1, \ldots, x_k) \) for all possible states \( x_{k+1} \) at the end-time of the interval \( t_{k+1} \). Then we can apply for this time interval the theory of the previous section. The conditional optimal cost-to-go is re-expressed by applying the log-transformation

\[
J_k(x, t | x_1, \ldots, x_k) = -\lambda \log Z_k(x, t | x_1, \ldots, x_k)
\]

in which the conditional partition function satisfies (cf 9)

\[
Z_k(x, t | x_1, \ldots, x_k) = \int \rho(x_{k+1}, t_{k+1} | x, t) \times Z_k(x_{k+1}, t_{k+1} | x_1, \ldots, x_k) dx_{k+1}
\]

(16)

in which the integral is over all possible states \( x_{k+1} \) at the time \( t_{k+1} \). The control is given by

\[
u_k(x, t | x_1, \ldots, x_k) = \nu \partial_x Z_k(x, t | x_1, \ldots, x_k).
\]

(17)

We can solve \( Z_k \) recursively with the use of boundary conditions (14), definition (13), and the end-condition

\[
J(x(t_1), \ldots, x(t_n), t_n) = \phi(x(t_1), \ldots, x(t_n)).
\]

(18)

This yields for \( t_k \leq t < t_{k+1} \) the following expression for \( Z_k \),

\[
Z_k(x, t | x_1, \ldots, x_k) = \int \exp \left( -\frac{\phi(x_1, \ldots, x_k)}{\lambda} \right)
\]

\[	imes \rho(x_1, t, x_{n-1}, t_{n-1}) \ldots \rho(x_{k+1}, t_{k+1} | x, t) dx_{k+1} \ldots dx_n.
\]

(19)

In the remainder of the paper, we drop the conditioning on earlier states \( x_1, \ldots, x_k \) in the notation. In addition, we drop the subindex \( k \). Furthermore, we redefine without loss of generality \( t_{k+1} \) as \( t_1 \), and (19) reduces to

\[
Z(x, t) = \int \exp \left( -\frac{\phi(x_1, \ldots, x_n)}{\lambda} \right)
\]

\[	imes \rho(x, t, x_{n-1}, t_{n-1}) \ldots \rho(x_1, t_1 | x, t) dx_1 \ldots dx_n.
\]

(20)

which contains (5) as a special case. The expressions for the cost-to-go (15) reduces to (4), and the expression for the control (17) reduces to (7), where it should be understood that the partition function is as in (20).
4. CONTROL WITH A SCHEDULE

In this section, we apply the framework to the case where a single agent has to visit a number of targets. We start with the simplest case, where the agent has only to visit one target at given time. Then we consider schedules. In a schedule, given targets should be visited at given times. We assume that dynamics $b$, noise $v$ and environment costs $V$ are given such that the diffusion process (55) results in a (given) distribution $p(x’, t’|x, t)$. The matrix $R$ is assumed to be such that $v = \lambda R^{-1}$ holds.

4.1 Control to a single target

First we consider the case where the agent in state $x$ at time $t$ has to visit a given target $\mu_1$ at a given time $t_1 > t$. To enforce that $x(t_1)$ is close to target $\mu_1$, we choose the following cost potential, where we assume $\epsilon$ to be small,

$$\phi(x(t_1)) = \begin{cases} -c & \text{if } D(\mu_1, x(t_1)) < \epsilon \\ \infty & \text{otherwise} \end{cases} \quad (21)$$

in which $D(\mu, x)$ is a distance measure between $\mu$ and $x$, e.g., $D(\mu, x) = |\mu - x|$. With $-c$ being a cost, $c$ is the reward for visiting the target. We compute the partition function according to (20) with end cost (21), resulting in

$$Z(x, t) = \exp(c/\lambda)V_{\epsilon} \int_{D(\mu_1, x(t_1)) < \epsilon} \rho(x_1, t_1|x, t)dx_1 . \quad (22)$$

Since $\epsilon$ is small (and $\rho$ is assumed to be smooth) we can approximate this by

$$Z(x, t) = \rho(\mu_1, t_1|x, t) \exp(c/\lambda)V_{\epsilon} , \quad (23)$$

where we defined the volume of points within distance $\epsilon$ around $\mu_1$ as $V_{\epsilon} = \int_{D<\epsilon} dx$.

The control follows from (7),

$$u(x, t; \mu_1, t_1) = \nu \partial_x \log Z(x, t) = \nu \partial_x \log \rho(\mu_1, t_1|x, t) . \quad (24)$$

which is in this case independent of the values of $\exp(c/\lambda)$ and $V_{\epsilon}$. The fact that it is independent of $c$ was to be expected: the agent has to visit the target regardless the reward, since the cost of missing the target is infinite. The independence of $V_{\epsilon}$ is due to the approximation $\epsilon \to 0$.

In the example where $b = 0$, $V = 0$ and $\nu > 0$ a scalar, so that $\rho$ given by (6), the control reduces to

$$u(x, t; \mu_1, t_1) = \frac{\mu_1 - x}{t_1 - t} , \quad (25)$$

which is well known from standard linear-quadratic control theory [9].

4.2 A single schedule

We define a schedule $\sigma$ as a sequence of given targets, $\mu_1, \ldots, \mu_n$ which are to be visited at a sequence of given (later) times $t_1, \ldots, t_n$. We can model the schedule with the following cost-potential

$$\phi(x(t_1), \ldots, x(t_n); \sigma) = \begin{cases} -c(\sigma) & \text{if } D(\mu_i, x(t_i)) < \epsilon \forall i \\ \infty & \text{otherwise} \end{cases} \quad (26)$$

The agent obtains a reward $c(\sigma)$ if the schedule is realized, i.e. if all the states $x(t_i)$ are within $\epsilon$ distance of the targets. We compute the partition function according to (5) with end cost (26). In the approximation of small $\epsilon$, we obtain

$$Z(x, t; \sigma) = \rho(\mu_1, t_1|x, t) \times \prod_{i=1}^{n-1} \rho(\mu_{i+1}, t_{i+1} | \mu_i, t_i) \exp \left( \frac{c(\sigma)}{\lambda} \right) V_{\epsilon_i} . \quad (27)$$

The control follows from (17),

$$u(x, t; \sigma) = \nu \partial_x \log Z(x, t; \sigma) = \nu \partial_x \log \rho(\mu_1, t_1|x, t) = u(x, t; \mu_1, t_1) , \quad (28)$$

which is in this case independent of the values of $\exp(c/\lambda)$ and $V_{\epsilon}$ and the targets at later times $t_i > t_1$. The latter is due to the fact that in the schedule $\sigma$ the agent has visited the first target at time $t_1$ regardless the positions and times of the targets later in the schedule. These latter targets do, however, of course contribute in the optimal cost-to-go.

5. ON-LINE SCHEDULING

Now we consider the case where the agent can choose a schedule from a given set of schedules $\sigma \in \Sigma$. Each schedule consists of a sequence of targets $\mu_i(\sigma)$, $i = 1, \ldots, n(\sigma)$ that are to be visited at times $t_i(\sigma)$ and a reward $c(\sigma)$ when the targets are visited. The schedule can be chosen on-line. In other words, optimal control is such that at end-time at least one schedule is realized, with minimal expected total cost.

We model the optimal control problem by a cost potential

$$\phi(x(t_1), \ldots, x(t_n)) = \min_{\sigma} \phi(x(t_1), \ldots, x(t_n); \sigma) , \quad (29)$$

where $t_1, \ldots, t_n$ is the union of times $t_i(\sigma)$ in all schedules. We assume that the schedules are different, i.e. each two schedules differ at least in one pair $(\mu_i, t_i)$, since only one (the one with lowestest cost) will contribute to the cost-potential (29). For the same reason we assume that if a schedule $\sigma'$ is a superset of another schedule $\sigma$, i.e. that $\sigma'$ contains all $(\mu_i, t_i)$ of $\sigma$ plus some other targets at other times, then $c(\sigma') > c(\sigma)$. We compute the partition function according to (20), which contains a term exp$\left(-\phi/\lambda\right)$. For the cost (29), this term equals

$$\exp \left( -\frac{\phi(x(t_1), \ldots, x(t_n))}{\lambda} \right) = \max_{\sigma} \exp \left( -\frac{\phi(x(t_1), \ldots, x(t_n); \sigma)}{\lambda} \right) . \quad (30)$$

Next we do the small $\epsilon$ approximation. First we consider the case that the target times $t_i$ of each of the schedules are identical. Then, if $\epsilon$ is small enough, a sequence of states $x(t_i)$ can at most realize one of the schedules, i.e. the schedules do not ‘overlap’. Since $\exp(\phi(x(t_1), \ldots, x(t_n); \sigma)) = 0$ if the schedule $\sigma$ is not realized, we may conclude that

$$\max_{\sigma} \exp \left( -\frac{\phi(x(t_1), \ldots, x(t_n); \sigma)}{\lambda} \right) = \sum_{\sigma} \exp \left( -\frac{\phi(x(t_1), \ldots, x(t_n); \sigma)}{\lambda} \right) , \quad (31)$$

and

$$Z(x, t) = \sum_{\sigma} Z(x, t; \sigma) . \quad (32)$$

In the following we will argue that for schedules with arbitrary times $t_i(\sigma)$, expression (32) is also valid, except in
some degenerate cases. If the times $t_i$ are not equal, there can be overlap between the schedules, i.e., a single sequence of states can realize more than one schedule. For example consider the case of two schedules $\sigma = 1$ and $\sigma = 2$, each with times $t_i(\sigma) \in T_\sigma$. If the set of times that are in both schedules $T_{12} = T_1 \cap T_2$ is non-empty, and if the targets at these times are the same in both schedules, then there is overlap. The overlap is realized by the sequences of states that visit the targets of $\sigma = 1$ at times in $T_1$ and the targets of $\sigma = 2$ at times in $T_2 \setminus T_1$. In such a case, (31) is only approximately true. If we assume that the rewards satisfy $c(1) \geq c(2)$, the righthand side of (31) is over-counted by $\exp((1/2(x(t_1), \ldots , x(t_n); \sigma = 2))$ for sequences of states $x(t_1), \ldots , x(t_n)$ that realize both schedules. To correct the partition function (32) we should replace $Z(x,t;2)$ by $(1-f(x,t))Z(x,t;2)$ where $f(x,t)$ is the ratio of the number of particles that realize both schedules to the number of particles that realize schedule $\sigma = 2$ under the (uncontrolled) diffusion process implied by $\rho$. In the 'degenerate case' where schedules are automatically realized by this dynamics, $f$ is a significantly larger than zero. In the typical case, however, $f \approx V_{\rho}^{-n(2)}$, where $n(2)$ is the length of schedule $\sigma = 2$ and $n$ the total number of different targets in both schedules. From $c(1) \geq c(2)$ we can conclude $n(2) < n$: if $n(2) = n$, the schedule $\sigma = 2$ would be equal to or a superset of schedule $\sigma = 1$. With $c(2) \leq c(1)$ this case has been excluded earlier. From $V_\rho \approx 0$ follows $f \approx 0$. So, in the generic case, even with overlapping schedules, (32) is a valid approximation.

The expected cost-to-go $J$ follows from (4). Note that the expected cost-to-go $J(x,t)$ is smaller than the cost-to-go $J(x,t;\sigma)$ for any $\sigma \in \chi$, since

$$-\lambda \log \sum_{\sigma' \in \chi} Z(x,t;\sigma') < -\lambda \log Z(x,t;\sigma).$$

The control $u$ follows from (7). It can be expressed as a weighted combination of single schedule controls,

$$u(x,t) = p(\sigma|x,t)u(x,t;\sigma),$$

with $u(x,t;\sigma)$ as in (28), and the probability over $\sigma$,

$$p(\sigma|x,t) = \frac{Z(x,t;\sigma)}{\sum_{\sigma' \in \chi} Z(x,t;\sigma')}.$$  

We remarked earlier that each single-schedule control is actually a single-target control (cf. (24) and (28)). Therefore the weighted combination (34) is actually a linear combination of single controls $u(x,t;\mu_1, t_1)$. The weight factor is given by the sum of the weight factors of schedules for which $\mu_1$ is the first target and $t_1$ the first time. Denoting this set of schedules as

$$\chi(\mu_1, t_1) = \{ \sigma : \mu_1(\sigma) = \mu_1, t_1(\sigma) = t_1 \},$$

then the weight factor is given by the marginal probability (35),

$$p(\mu_1, t_1|x,t) = \sum_{\sigma \in \chi(\mu_1, t_1)} p(\sigma|x,t),$$

and the control can be written as

$$u(x,t) = \sum_{\mu_1, t_1} p(\mu_1, t_1|x,t)u(x,t;\mu_1, t_1).$$

Assuming that $\rho$ and its derivatives are $O(1)$, and for simplicity that $\chi(\mu_1, t_1)$ contains at most one schedule, we see from (32) and (27) that contributions of a schedule to the partition function is of the order of $\exp(c(\sigma)/\lambda)V_{\rho}^{n(\sigma)}$ where $n(\sigma)$ is the number of targets to be visited. In other words, the expected total reward of a schedule is of order

$$E(\sigma) \equiv c(\sigma) - \lambda n(\sigma) \log V_{\rho}.$$  

So if several schedules are to be considered by the agent, the rewards $c(\sigma)$ should be such that $E(\sigma)$ is about the same for all $\sigma$. Schedules with significantly larger $c(\sigma)$ will dominate the control of the agent, while schedules with significantly smaller $c(\sigma)$ will practically just be ignored. So, in a way from an economists viewpoint, $\lambda \log(V_{\rho}^{-1})$ can be considered as a reasonable price for visiting one additional target in a schedule.

In the case where $\chi(\mu_1, t_1)$ contain more schedules for some $\mu_1, t_1$, we see from (37) that schedules that share their initial targets become more dominant. This makes sense since control to such targets is more save: it keeps options open for the agent to choose between schedules depending on the unknown future noise.

5.1 Example: A moving target

As an example, we consider the case where there is a moving target with given position $\mu(s)$ at increasing times $t(s) = s\Delta t$ with $s = 1, \ldots , S$. The agent has to hit the target at one of the times, gets the reward $c(s)$, then the game is over. This is modelled by considering $s$ as a schedule. Applying the results of previous subsections, we find the partition function

$$Z(x,t) = \sum_{s:t(s) > t} Z(x,t;s),$$

with

$$Z(x,t;s) = \rho(\mu(s), t(s)|x,t) \exp \left( \frac{c(s)}{\lambda} \right) V_{\rho}.$$

The control is

$$u(x,t) = \sum_{s:t(s) > t} p(s|x,t)u(x,t;\mu(s), t(s))$$

with

$$p(s|x,t) = \frac{\exp \left( \frac{c(s)}{\lambda} \right) \rho(\mu(s), t(s)|x,t)}{\sum_{s':t(s') > t} \exp \left( \frac{c(s')}{\lambda} \right) \rho(\mu(s'), t(s')|x,t)}.$$  

In figure 1, we show a the result of an 1-d simulation of an agent trying to visit a target moving according $\mu(t) = t^2$. The constraint is that it should visit the target between $t = 0$ and $t = 1$. The dynamics and environment is given by $\rho = 0$, $V = 0 \nu = 1$, and $R = 1$. So $\rho$ is a Gaussian (0), and $u$ is given by (25). This simulation result can be compared with the case where the agent has to visit the target at exactly $t = 0.8$, using the control (25), see plots in figure 2. As can be seen from the figure, this control is much costlier, because the agent has to try hard to make the visit at the precise required time. On the other hand, it does not exploit the opportunities to visit if it is accidentally close to the target.
appropriate components of the gradient

control follows from (44). So no additional coordination

is required since each agent has full information about the

control by observing the joint state at that time from which

= hold globally, so that we can apply the log-

factorizes into single agent partition functions

The trivial case where the schedule is fixed beforehand,

i.e. \( \chi = \{ \sigma \} \). Then each agent \( a \) has just to visit the targets

according to his own part of the schedule \( \sigma^a \). Control by the

agents is independent of each other. The partition function

factorizes into single agent partition functions

\begin{align*}
Z(x, t; \sigma) &= \exp \left( \frac{c(\sigma)}{\lambda} \right) Z_0(x, t; \sigma) \\
&= \exp \left( \frac{c(\sigma)}{\lambda} \right) \prod_a Z_0(x^a, t; \sigma^a),
\end{align*}

in which \( Z_0(x, t; \sigma) \) is defined as the partition function for

schedule \( \sigma \) with substitution of cost \( c(\sigma) = 0 \), i.e. \( Z \equiv 
\exp(c(\sigma)/\lambda) Z_0 \). In the case where the MAS can choose the

schedule on-line, we have a similar expression as in the single

agent case,

\begin{equation}
Z(x, t) = \sum_{\sigma \in \chi} Z(x, t; \sigma),
\end{equation}

with the difference that the single-schedule partition functions

are now products of single-agent single-schedule partition functions,
as expressed in (46).
The expected cost-to-go $J$ follows again from (4). The control of agent $a$ is obtained using (44). In analogy to (34), the control can be expressed as a weighted combination of the single schedule controls of each agent,

$$u^a(x, t) = \sum_{\sigma^a} p(\sigma^a|x, t) u^a(x^a, t; \sigma^a), \quad (48)$$

with $u^a(x^a, t; \sigma^a)$ the control of agent $a$ with schedule $\sigma^a$ as in (28). The weight factors are the marginal probabilities,

$$p(\sigma^a|x, t) = \sum_{(a,b)\neq a} p(\sigma|x, t)$$

$$= \sum_{(a,b)\neq a} \frac{\exp\left(\frac{c(\sigma)}{\lambda}\right)\prod_b Z_b(x^b, t; \sigma^b)}{\sum_{\sigma'} \exp\left(\frac{c(\sigma')}{\lambda}\right)\prod_c Z_c(x^c, t; \sigma^c)} \quad (49)$$

As in the single-agent case, the control of agent $a$ can be expressed as a combination of single agent single-target controls

$$u^a(x, t) = \sum_{\mu_1, \tau_1} p(\mu_1^a, \tau_1^a|x, t) u^a(x^a, t; \mu_1^a, \tau_1^a), \quad (50)$$

with the marginal

$$p(\mu_1^a, \tau_1^a|x, t) = \sum_{\sigma^a\in \{\mu_1^a, \tau_1^a\}} p(\sigma^a|x, t), \quad (51)$$

defined as in (37). The expression of the control of an agent in the multi-agent case is the same as in the single agent case, with the difference that (1) the weight-factors $p(\sigma^a|x, t)$ are actually marginal distributions of a joint MAS distribution $p(\sigma|x, t)$, and (2) they do not only depend on the state of a single agent, but rather on the joint MAS state.

### 6.1 Numerical Examples

In this paragraph we illustrate the theory of optimal control in stochastic MASs by some simulations. The environment is with $b = 0$, $V = 0$, and $\nu = 0.1$, so $\rho$ is Gaussian (6). Furthermore we took $R = 1$. Simulations are in 1-d, for plotting purposes. We simulated a group of agents starting at $t = 0$, $x = 0$ that have to visit a number of targets at times $t_s$ and located at $\mu_t$ where $(t_s, \mu_t) = \{(1, 0.1), (2, -0.1), (3.0.1), (4, -0.1)\}$. After their mission, the agents have to return to $x = 0$ at time $t = 5$.

In figure 3, we show a simulation where each target has to be visited by exactly one agent. So with 4 targets and 2 agents, there are $2^4 = 16$ schedules to be considered.

In figure 4, we show the case where agents are allowed to ignore targets, if the control to reach the target is too expensive. Each target is to be visited by agent 1, or agent 2, or none of the agents. So there are $3^4 = 81$ schedules to be considered.

### 7. DISCUSSION

We studied optimal control in collaborative multi-agent systems in continuous space-time. A straightforward approach to discretize the system in space and time would make the $n$ agent MAS intractable due to the exponential blow-up of the state-space. In this paper, we took the approach developed in [6, 7]. We generalized the result in [11], and showed that under given model assumptions, optimal distributed on-line scheduling and control can be solved analytically. The result in [11] is a special case were only at the end-time a joint decision is taken. The additional computational cost in on-line scheduling, compared to a schedule chosen before hand is the computation of the distribution $p(\sigma)$. If at each of the $n$ time points, each of the $A$ agents can choose one of $m$ targets, the state space of schedules in a multi-agent system grows in principle as $n^A$. In general this approach is therefore infeasible. In the examples in this paper we restricted state space by allowing only a limited number $M$ of possible schedules of the MAS, reducing the growth to $n^M$. Since the contributions of each schedule to the expected cost can be computed beforehand, one could reduce the space even more by pruning the schedules with the smallest contributions in $Z$. This will lead to a trade-off in expected cost/reward and computational complexity of for the on-line control in the system. In [11], sparse reward functions $c(\sigma)$ represented as a graphical models were considered. This allows the exploitation of efficient probabilistic graphical model inference methods. The results in this paper suggest the use of similar methods, extended to the temporal domain. This is currently under study.

There are many possible model extensions that need to be explored in future research. Obvious extensions are to consider systems with more realistic environments, such as allowing for obstacles are already of interest to study in the single agent situation. Others apply typically to the multi-agent situation, such as penalties for collisions between agents. Typically, these types of model extensions will prohibit an analytical solution of the control, and approximate numerical methods will be required. Some proposals can be found in [6, 7].

Finally we would like to stress, that although the model class is quite specific and maybe not generally applicable, we think that the study of this class is interesting because it is one of the few “exactly solvable” multi-agent systems, allowing the study of non-trivial collective optimal behaviour.
in distributed systems, both analytically as well as in simulations, and possibly providing insights that might help to develop efficient approximating methods for more general systems.

8. ACKNOWLEDGMENTS

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

9. REFERENCES


APPENDIX

A. STOCHASTIC OPTIMAL CONTROL

In this appendix we give a brief derivation of (4), (5) and (7), starting from (3). This appendix follows [11]. Details can be found in [6, 7].

The optimal cost-to-go $J$ in a state $x$ at time $t$ is found by minimizing $C(x,t,u(\cdot))$ over all control policies $u(\cdot)$:

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

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

$$-\partial_t J = \min_u \left( \frac{1}{2} u^T Ru + V + (b + u)^T \partial_x J + \frac{1}{2} \text{Tr}(\nu \partial_y^2 J) \right),$$

with boundary condition $J(x,t_n) = \phi(x)$. The minimization with respect to $u$ yields

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

which defines the optimal control. Substituting this control turns the HJB into a a non-linear partial differential equation for $J$. We can remove the non-linearity by using the log transformation: define $Z(x,t)$ through

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

with $\lambda$ a scalar such that $\nu = \lambda R^{-1}$ (this implies the assumption that the matrices $\nu$ and $R^{-1}$ are proportional to each other), and define

$$Z(x,t) = \int dy \rho(y,t_n|x,t) \exp(-\phi(y)/\lambda),$$

then the the density $\rho(y,t|x,t)$ ($t < \theta \leq t_n$) satisfies a forward Fokker-Planck equation

$$\partial_\theta \rho(y,\theta|x,t) = \frac{\nu^T}{\lambda} \rho - \partial_y^T b \rho + \frac{1}{2} \text{Tr}(\nu \partial_y^2 \rho),$$

which is a linear in $\rho$. If $\rho$ is solved, then $Z$ follows from (54), with end condition $Z(x,t_n) = \exp(-\phi(x)/\lambda)$, (since $\rho(y,t_n|x,t_n) = \delta(y-x)$). From $Z$ the solution of $J$ follows by applying the log-transform.