Journal of Artificial Intelligence Research 32 (2008) 95-122 Submitted 10/07; published 05/08

Graphical Model Inference in Optimal Control of Stochastic Multi-Agent Systems

Bart van den Broek B.vandenBroek@science.ru.nl
Wim Wiegerinck W.Wiegerinck@science.ru.nl
Bert Kappen B.Kappen@science.ru.nl
SNN, Radboud University Nijmegen, Geert Grooteplein 21,
Nijmegen, The Netherlands

Abstract

In this article we consider the issue of optimal control in collaborative multi-agent systems with stochastic dynamics. The agents have a joint task in which they have to reach a number of target states. The dynamics of the agents contains additive control and additive noise, and the autonomous part factorizes over the agents. Full observation of the global state is assumed. The goal is to minimize the accumulated joint cost, which consists of integrated instantaneous costs and a joint end cost. The joint end cost expresses the joint task of the agents. The instantaneous costs are quadratic in the control and factorize over the agents. The optimal control is given as a weighted linear combination of single-agent to single-target controls. The single-agent to single-target controls are expressed in terms of diffusion processes. These controls, when not closed form expressions, are formulated in terms of path integrals, which are calculated approximately by Metropolis-Hastings sampling. The weights in the control are interpreted as marginals of a joint distribution over agent to target assignments. The structure of the latter is represented by a graphical model, and the marginals are obtained by graphical model inference. Exact inference of the graphical model will break down in large systems, and so approximate inference methods are needed. We use naive mean field approximation and belief propagation to approximate the optimal control in systems with linear dynamics. We compare the approximate inference methods with the exact solution, and we show that they can accurately compute the optimal control. Finally, we demonstrate the control method in multi-agent systems with nonlinear dynamics consisting of up to 80 agents that have to reach an equal number of target states.

1. Introduction

The topic of control in multi-agent systems is characterized by many issues, originating from various sources, including a wide variety of possible execution plans, uncertainties in the interaction with the environment, limited operation time and supporting resources, and a demand for robustness of the joint performance of the agents. Such issues are encountered in, for example, air traffic management (Tomlin, Pappas, & Sastry, 1998; van Leeuwen, Hesseling, & Rohling, 2002), formation flight (Ribichini & Frazzoli, 2003; Hu, Prandini, & Tomlin, 2007), radar avoidance for unmanned air vehicles or fighter aircraft (Pachter & Pachter, 2001; Kamal, Gu, & Postlethwaite, 2005; Larson, Pachter, & Mears, 2005; Shi, Wang, Liu, Wang, & Zu, 2007), and persistent area denial (Subramanian & Cruz, 2003; Liu, Cruz, & Schumacher, 2007; Castanon, Pachter, & Chandler, 2004).

In many control approaches in multi-agent systems, stochastic influences in the dynamics of the agents are not taken into account or assumed negligible, and the dynamics are modeled deterministically. If the system is truly deterministic, then the agents can be optimally controlled by open loop controls. However, when the stochastic influences in the dynamics are too large to be ignored, open loop controls become far from optimal, and the multi-agent system should no longer be modeled deterministically. The usual approach to control in multi-agent systems with stochastic dynamics is to model the system by a Markov Decision Processes (MDP) (Boutilier, 1996; Sadati & Elhamifar, 2006). In principle, these are solved in discrete space and time by backward dynamic programming. However, the discretization will make the joint state space of the multi-agent system increase exponentially in the number of agents, and a basic dynamic programming approach will generally be infeasible (Boutilier, 1996). An attempt to overcome this is to exploit structures in the problem and describe the system by a factored MDP. In general these structures will not be conserved in the value functions, and exact computations remain exponential in the system size. Guestrin, Koller, and Parr (2002a) and Guestrin, Venkataraman, and Koller (2002b) assumed a predefined approximate structure of the value functions, and thereby provided an efficient approximate MDP model for multi-agent systems. A similar approach was taken by Becker, Zilberstein, Lesser, and Goldman (2003, 2004), assuming independent collaboration of the agents with a global reward function, resulting in transition-independent decentralized MDPs.

In this paper we concentrate on multi-agent systems where the agents have a joint task in which they have to reach a number of target states. We model the multi-agent system in continuous space and time, following the approach of Wiegerinck, van den Broek, and Kappen (2006). We make the following assumptions. The agents are assumed to have complete and accurate knowledge of the global state of the system (assumption 1). The dynamics of each agent is additive in the control and disturbed by additive Wiener noise (assumption 2). The performance of the agents is valued by a global cost function, which is an integral of instantaneous costs plus an end cost. The joint task of the agents is modeled by the end cost. The instantaneous costs are assumed to be quadratic in the control (assumption 3). The noise level in the dynamics of the agents is inversely proportional to the control cost (assumption 4). Finally, we assume that both the autonomous dynamics and the instantaneous costs factorize over the agents (assumption 5).

Under the assumptions 1 and 2, the optimal control problem is partially solved by finding the optimal expected cost-to-go, which satisfies the so-called stochastic Hamilton-Jacobi-Bellman (SHJB) equation. Once the optimal expected cost-to-go is given, the optimal control is provided as the gradient of the optimal expected cost-to-go by adopting assumption 3. The SHJB equation is a nonlinear partial differential equation (PDE), and this nonlinearity makes it difficult to solve. A common approach to solving the SHJB equation is to assume, in addition to assumption 3, that the instantaneous costs and the end cost in the cost function are quadratic in the state, and that the dynamics are linear in the state as well—this is known as linear-quadratic control. The optimal expected cost-to-go then is quadratic in the state with time-varying coefficients, and the problem reduces to solving the Riccati equations that these coefficients satisfy (Stengel, 1993; Øksendal, 1998). Otherwise, approximation methods are needed. An approximate approach is given by the iterative linear-quadratic Gaussian method (Todorov & Li, 2005); this yields a locally optimal feedback control, and is valid in case there is little noise. We instead follow the approach of Fleming (1978) and adopt assumption 4. Under this assumption the SHJB equation can be transformed into a linear PDE by performing a logarithmic transformation. Its solution equals the expectation value of a stochastic integral of a diffusion process. In general, this is not a closed form expression. In this paper we will estimate this expression by formulating it as a path integral (Kappen, 2005a, 2005b), and we estimate the latter using Metropolis-Hastings sampling. There are several other ways to estimate the path integral, such as Hamilton Monte Carlo sampling and the Laplace approximation, but these are not covered in this paper.

The structure of the optimal expected cost-to-go will generally be very complex due to the dynamic couplings between the agents. By adopting assumption 5, the agents will only be coupled through the joint end cost, which then solely determines the structure of the optimal expected cost-to-go. This will result in state transition probabilities that factorize over the agents. It follows that the optimal control becomes a weighted combination of single-agent to single-target controls. The weights are given by a joint distribution over agent to target assignments. The joint distribution has the same structure as the joint end cost. The structure of the joint distribution is representable by a factor graph, and the optimal control problem becomes a graphical model inference problem (Wiegerinck et al., 2006). The complexity of the graphical model inference is exponential in the tree width of the factor graph. Exact inference will be possible by using the junction tree algorithm, given that the graph is sufficiently sparse and the number of agents is not too large. In more complex situations approximate inference methods are necessary, and we show that the optimal control can accurately be approximated in polynomial time, using naive mean field (MF) approximation or belief propagation (BP). This makes distributed coordination possible in multi-agent systems that are much larger than those that could be treated with exact inference.

The paper is organized as follows. In Sections 2 and 3, we provide a review of both the single and the multi-agent stochastic optimal control framework, developed by Kappen (2005a, 2005b) and Wiegerinck et al. (2006). As an example, we will rederive linear quadratic control. The general solution is given in terms of a path integral, and we explain how it can be approximated with Metropolis-Hastings sampling.

In Section 4, we give a factor graph representation of the end cost function. We discuss two graphical model approximate inference methods: naive mean field approximation and belief propagation. We show that the approximation of the optimal control in both methods is obtained by replacing the exact weights in the controls with their respective approximations.

In Section 5, we present numerical results. We make a comparison of the approximate optimal controls, infered by the naive mean field approximation, belief propagation and a greedy method, with the exact optimal control; this we do in a multi-agent system of 18 agents with linear dynamics in a two-dimensional state space, and with two target states. Furthermore, we present results from control in multi-agent systems with nonlinear dynamics and a four-dimensional state space, in which agents control their forward velocity and driving direction. The controls are approximated by a combination of Metropolis-Hastings sampling, to infer the path integrals, and naive mean field approximation, to infer the agent to target assignments. This allowed us to control systems of up to 80 agents with 80 target states. These results regarding nonlinear dynamics have only an illustrative purpose.

2. Stochastic Optimal Control of a Single Agent

We consider an agent in a k-dimensional continuous state space $\mathbb{R}^k$, its state $x(t)$ evolving over time according to the controlled stochastic differential equation

$$dx(t) = b(x(t), t)dt + u(x(t), t)dt + \sigma dw(t), \tag{1}$$

in accordance with assumptions 1 and 2 in the introduction. The control of the agent is the $\mathbb{R}^k$-valued function u of $x(t)$ and t. The noise in the dynamics is modeled by the Wiener process $w(t)$, i.e., a normally distributed k-dimensional stochastic process in continuous time with mean 0 and variance t, and the $k \times k$ matrix σ which represents the variance of the noise. Any autonomous dynamics are modeled by b, which is a $\mathbb{R}^k$-valued function of $x(t)$ and t. The state change $dx(t)$ is the sum of the noisy control and the autonomous dynamics.

The behavior of the agent is valued by a cost function. Given the agent's state $x(t) = x$ at the present time t, and a control u, there is an expected future cost for the agent:

$$C^u(x, t) = \mathbb{E}^u_{x,t}\left[\phi(x(T)) + \int_t^T d\theta \left(\frac{1}{2}\|Ru(x(\theta), \theta)\|^2 + V(x(\theta), \theta)\right)\right]. \tag{2}$$

The expectation $\mathbb{E}^u_{x,t}$ is taken with respect to the probability measure under which $x(t)$ is the solution to (1) given the control law u and the condition $x(t) = x$. The cost is a combination of the end cost $\phi(x(T))$, which is a function of the end state $x(T)$, and an integral of instantaneous costs. The instantaneous cost is a sum of a state and a control dependent term. The state dependent term $V(x(\theta), \theta)$ is the cost of being in state $x(\theta)$ at time θ. The function V is arbitrary, and represents the environment of the agent. The control dependent term $\frac{1}{2}\|Ru(x(\theta), \theta)\|^2$ is the cost of the control in state $x(\theta)$ at time θ, where $\|z\|^2 = z^\top z$ is the Euclidean norm, and R is a full rank $k \times k$ matrix. It is quadratic in the control, in accordance with assumption 3 in the introduction, and by assumption 4, R is related to the variance of the noise in the control via the relation

$$\sigma\sigma^\top = \lambda(R^\top R)^{-1}, \tag{3}$$

where λ is a scalar.

The expected cost-to-go at time t minimized over all controls u defines the optimal expected cost-to-go

$$J(x, t) = \min_u C^u(x, t). \tag{4}$$

In Appendix A, it is explained that due to the linear-quadratic form of the optimization problem—the dynamics (1) is linear in the action u, the cost function (2) is quadratic in the action—the minimization can be performed explicitly, yielding a nonlinear partial differential equation in J, the so-called stochastic Hamilton-Jacobi-Bellman (SHJB) equation. The minimum is attained in

$$u(x, t) = -(R^\top R)^{-1}\partial_x J(x, t). \tag{5}$$

This is the optimal control. Note that it explicitely depends on the state x of the agent at time t, making it a feedback control.

The optimal expected cost-to-go can be re-expressed in terms of a diffusion process (for a derivation, we again refer to Appendix A):

$$J(x,t) = -\lambda \log Z(x,t) \tag{6}$$

where $Z(x,t)$ is the expectation value

$$Z(x,t) = \mathbb{E}_{x,t}\left[\exp\left(-\frac{1}{\lambda}\phi(y(T)) - \frac{1}{\lambda}\int_t^T d\theta\, V(y(\theta),\theta)\right)\right] \tag{7}$$

and $y(\theta)$ is a diffusion process with $y(t) = x$ and satisfying uncontrolled dynamics:

$$dy(\theta) = b(y(\theta),\theta)d\theta + \sigma dw(\theta). \tag{8}$$

Substituting relations (3) and (6) in (5), we find the optimal control in terms of $Z(x,t)$:

$$u(x,t) = \sigma\sigma^\top \partial_x \log Z(x,t). \tag{9}$$

Example 1. *Consider an agent in one dimension with a state $x(t)$ described by the dynamical equation* (1) *without autonomous dynamics ($b = 0$). The instantaneous cost V is zero, and the end cost ϕ is a quadratic function around a target state μ:*

$$\phi(y) = \frac{\alpha}{2}|y-\mu|^2.$$

The diffusion process $y(\theta)$ that satisfies the uncontrolled dynamics (8) *is normally distributed around the agents state $x = y(t)$ at time t and with a variance $\sigma^2(\theta - t)$, hence the state transition probability for the agent to go from (x,t) to (y,T) in space-time is given by the Gaussian density*

$$\rho(y,T|x,t) = \frac{1}{\sqrt{2\pi\sigma^2(T-t)}}\exp\left(-\frac{|y-x|^2}{2\sigma^2(T-t)}\right).$$

The expectation value (7) *is given by the integral*

$$Z(x,t) = \int dy \rho(y,T|x,t)e^{-\frac{1}{\lambda}\phi(y)} = \sqrt{\frac{R^2/\alpha}{T-t+R^2/\alpha}}\exp\left(-\frac{|x-\mu|^2}{2\sigma^2(T-t+R^2/\alpha)}\right),$$

where relation (3) *is used. The optimal control follows from* (6) *and* (9) *and reads*

$$u(x,t) = \frac{\mu - x}{T-t+R^2/\alpha}. \tag{10}$$

This result is well known (Stengel, 1993)*.*

2.1 A Path Integral Formulation

Example 1 shows that for a simple system with no autonomous dynamics ($b = 0$) or costs due to the environment ($V = 0$), we can write down the control explicitly. This is because the uncontrolled dynamics is normally distributed, and consequently the expectation value (7) with quadratic end cost has a closed form expression. In the general situation where b and V are arbitrary, there no longer exists an explicit expression for the expectation value, and the optimal control can only be obtained by approximation. We will now discuss how this is done by taking a path integral approach (Kleinert, 2006). A detailed derivation of the expressions presented here is given in Appendix B.

In the path integral approach, we write the expectation value (7) as a path integral:

$$Z(x,t) = \lim_{\varepsilon \downarrow 0} Z_\varepsilon(x(t_0), t_0) \tag{11}$$

where $x(t_0) = x$, $t_0 = t$ and

$$Z_\varepsilon(x(t_0), t_0) = \frac{1}{\sqrt{\det(2\pi\varepsilon\sigma^2)^N}} \int dx(t_1) \dots \int dx(t_N)\, e^{-\frac{1}{\lambda} S_\varepsilon(x(t_0),\dots,x(t_N),t_0)}.$$

It is an integral over paths $(x(t_0), \dots, x(t_N))$ in discrete time, the start $x(t_0)$ kept fixed and $\varepsilon N = T - t$, taken in a continuous time limit of sending the length of the time steps $\varepsilon = t_{i+1} - t_i$ to zero. Note that in this limit N goes to infinity and the paths become infinite dimensional objects. The function in the exponent is the cost of the path:

$$S_\varepsilon(x(t_0), \dots, x(t_N), t_0) = \\ \phi(x(T)) + \sum_{i=0}^{N-1} \varepsilon\, V(x(t_i), t_i) + \sum_{i=0}^{N-1} \varepsilon \frac{1}{2} \left\| R \left(\frac{x(t_{i+1}) - x(t_i)}{\varepsilon} - b(x(t_i), t_i) \right) \right\|^2,$$

The optimal control becomes a weighted average over controls that are derived from a single path:

$$u(x(t_0), t_0) = \lim_{\varepsilon \downarrow 0} \int dx(t_1) \dots \int dx(t_N)\, p(x(t_0), \dots, x(t_N), t_0)\, u(x(t_0), \dots, x(t_N), t_0). \tag{12}$$

The weights are given by

$$p(x(t_0), \dots, x(t_N), t_0) = \frac{e^{-\frac{1}{\lambda} S_\varepsilon(x(t_0),\dots,x(t_N),t_0)}}{\sqrt{\det(2\pi\varepsilon\sigma^2)^N} Z_\varepsilon(x(t_0), t_0)}.$$

The control derived from a path $(x(t_0), \dots, x(t_N))$ reads

$$u(x(t_0), \dots, x(t_N), t_0) = \frac{x(t_1) - x(t_0)}{\varepsilon} - b(x(t_0), t_0). \tag{13}$$

Note that it only depends on the first two entries $x(t_0)$ and $x(t_1)$ in the path.

2.2 Path Integration by Metropolis-Hastings Sampling

The path integral formulation (12) of the optimal control can generally not be computed, because it is an integral over uncountably many paths, but there exist several ways to approximate it. A natural approach goes by stochastic sampling of paths. Several methods of stochastic sampling exist, the one we will use here is known as Metropolis-Hastings sampling (Hastings, 1970). In its implementation time will be discretized: we do not take the limit in (12) of ε decreasing to zero, but instead keep ε at a fixed value. A sample path will be a sequence $(x^s(t_0), \ldots, x^s(t_N))$ of vectors in the state space $\mathbb{R}^k$, with $x(t_0) = x$ the current state of the agent at the current time $t_0 = t$. According to equation (13), we only need $x^s(t_0)$ and $x^s(t_1)$ to derive the control from a sample path $(x(t_0), \ldots, x(t_N))$. The Metropolis-Hastings sampling ensures that different paths are properly weighted, hence the optimal control is approximated as follows:

$$u(x(t_0), t_0) \approx \frac{\langle x(t_1) \rangle - x(t_0)}{t_1 - t_0} - b(x(t_0), t_0), \tag{14}$$

where $\langle x(t_1) \rangle$ is the mean value of $x^s(t_1)$ taken over the sample paths. Pseudo-code for the algorithm is given in Algorithm 1.

Algorithm 1: Metropolis-Hastings sampling

```
Input: initial path $(x(t_0), \ldots, x(t_N))$
$s = 1$
repeat $M$ times:
    define Gaussian proposal distribution centered around $(x(t_1), \ldots, x(t_N))$
    with variance equal to the noise
    draw sample path $(x'(t_1), \ldots, x'(t_N))$ from proposal distribution
    $a = \exp\left(\frac{1}{\lambda} S_\varepsilon(x(t_0), x(t_1), \ldots, x(t_N), t_0) - \frac{1}{\lambda} S_\varepsilon(x(t_0), x'(t_1), \ldots, x'(t_N), t_0)\right)$
    if $a \geq 1$
        set $(x(t_1), \ldots, x(t_N)) = (x'(t_1), \ldots, x'(t_N))$
    else
        set $(x(t_1), \ldots, x(t_N)) = (x'(t_1), \ldots, x'(t_N))$ with probability $a$
    end if
    $(x^s(t_0), \ldots, x^s(t_N)) = (x(t_0), \ldots, x(t_N))$
    $s = s + 1$
end repeat
compute approximate control with equation (14)
```

3. Stochastic Optimal Control of a Multi-Agent System

We now turn to the issue of optimally controlling a multi-agent system of n agents. In principle, the theory developed for a single agent straightforwardly generalizes to the multi-agent situation. Each agent a has a k-dimensional state x_a that satisfies a dynamics similar to (1):

$$dx_a(t) = b_a(x_a(t), t)dt + u_a(x(t), t)dt + \sigma_a dw_a(t), \tag{15}$$

in accordance with assumptions 1, 2 and 5 in the introduction. Note that the control of each agent not only depends on its own state x_a, but on the joint state $x = (x_1, \ldots, x_n)$

of the system. The system has a joint cost function similar to (2), depending on the joint state x and joint control $u = (u_1, \ldots, u_n)$ of the system:

$$C^u(x,t) = \mathbb{E}^u_{x,t}\left[\phi(x(T)) + \sum_{a=1}^{n} \int_t^T d\theta \, \left(\frac{1}{2}\|R_a u_a(x(\theta),\theta)\|^2 + V(x_a(\theta),\theta)\right)\right].$$

The expectation $\mathbb{E}^u_{x,t}$ is taken with respect to the probability measure under which $x(t)$ is the solution to (15) given the control law u and the condition that $x(t) = x$. The cost is a combination of the joint end cost $\phi(x(T))$, which is a function of the joint end state $x(T)$, and an integral of instantaneous costs. The instantaneous cost factorizes over the agents, in accordance with assumption 5 in the introduction. For each agent, it is a sum of a state dependent term $V(x_a(\theta),\theta)$ and a control dependent term $\frac{1}{2}\|R_a u_a(x_a(\theta),\theta)\|^2$, similar to the single agent case. In accordance with assumption 4 in the introduction, the control cost of each agent is related to the noise in the agent's dynamics via the relation

$$\sigma_a \sigma_a^\top = \lambda (R_a^\top R_a)^{-1},$$

where λ is the same for each agent. The joint cost function is minimized over the joint control, yielding the optimal expected cost-to-go J. The optimal expected cost-to-go is expressed in terms of a diffusion process via the relation

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

where $Z(x,t)$ is the joint expectation value

$$Z(x,t) = \mathbb{E}_{x,t}\left[\exp\left(-\frac{1}{\lambda}\phi(y(T)) - \frac{1}{\lambda}\sum_{a=1}^{n}\int_t^T d\theta \, V(y_a(\theta),\theta)\right)\right] \tag{16}$$

and the $y_1(t), \ldots, y_n(t)$ are diffusion processes, with $y = (y_1, \ldots, y_n)$ and $y(t) = x$, satisfying uncontrolled dynamics

$$dy_a(\theta) = b_a(y_a(\theta),\theta)d\theta + \sigma_a dw_a(\theta), \qquad a = 1, \ldots, n. \tag{17}$$

The multi-agent equivalent of the optimal control (9) reads

$$u_a(x,t) = \sigma_a {\sigma_a}^\top \partial_{x_a} \log Z(x,t). \tag{18}$$

We will now show that the optimal control of an agent can be understood as an expected control, that is, an integral over target states y_a of a transition probability to the target times the optimal control to that target. To this end, we write the expectation (16) as an integral over the end state:

$$Z(x,t) = \int dy e^{-\frac{1}{\lambda}\phi(y)} \prod_{a=1}^{n} Z_a(y_a, T; x_a, t), \tag{19}$$

where the $Z_a(y_a, T; x_a, t)$ are implicitly defined by

$$\int dy_a \, Z_a(y_a, T; x_a, t) f(y_a) = \mathbb{E}_{x_a,t}\left[f(y_a(T)) \exp\left(-\frac{1}{\lambda}\int_t^T d\theta \, V(y_a(\theta),\theta)\right)\right]$$

for arbitrary functions f. Substituting (19) into (18) yields

$$u_a(x,t) = \int dy_a \, p_a(y_a|x,t) \, u_a(y_a;x_a,t) \tag{20}$$

where

$$u_a(y_a;x_a,t) = \sigma_a {\sigma_a}^\top \partial_{x_a} \log Z_a(y_a,T;x_a,t) \tag{21}$$

is the optimal control for agent a to go from state x_a at the current time t to state y_a at the end time T, and $p_a(y_a|x,t)$ is a marginal of

$$p(y|x,t) = \frac{1}{Z(x,t)} e^{-\frac{1}{\lambda}\phi(y)} \prod_{a=1}^{n} Z_a(y_a,T;x_a,t).$$

3.1 Discrete End States

The agents have to fulfill a task of arriving at a number of target states at the end time according to an initially specified way: for example, they should all arrive at the same target, or they should all arrive at different targets. The targets are considered regions $G_1,\ldots,G_m$ in the state space, and the end cost ϕ is modeled as follows:

$$e^{-\frac{1}{\lambda}\phi(y)} = \sum_s w(s) \prod_{a=1}^{n} w_a(y_a;s_a), \qquad w_a(y_a;s_a) = e^{-\frac{1}{\lambda}\phi_a(y_a;s_a)}, \tag{22}$$

where the sum runs over assignments $s = (s_1,\ldots,s_n)$ of agents a to regions G_{s_a}. $\phi_a(y_a;s_a)$ is a cost function associated to region G_{s_a}, returning a low cost if the end state y_a of agent a lies in the region G_{s_a} and a high cost otherwise. $w(s)$ is a weight, grading the assignments s and thereby specifying the joint task of the agents. Assignments that result in a better fulfillment of the task have a higher weight. In a situation where all agents have to go to the same target, for example, a vector s that assigns each agent to a different target will have a low weight $w(s)$.

With this choice of end cost, equation (19) factorizes as

$$Z(x,t) = \sum_s w(s) \prod_{a=1}^{n} Z_a(s_a;x_a,t)$$

where

$$Z_a(s_a;x_a,t) = \int dy_a Z_a(y_a,T;x_a,t) w_a(y_a;s_a). \tag{23}$$

The interpretation of $Z_a(s_a;x_a,t)$ is that $-\lambda \log Z_a(s_a;x_a,t)$ is the expected cost for agent a to move from x_a to target s_a. The optimal control (20) of a single agent a becomes

$$u_a(x,t) = \sum_{s_a=1}^{m} p(s_a|x,t) u_a(s_a;x_a,t), \tag{24}$$

where

$$u_a(s_a;x_a,t) = \sigma_a {\sigma_a}^\top \partial_{x_a} \log Z_a(s_a;x_a,t) \tag{25}$$

is the control for agent a to go to target s_a, and the weights $p(s_a|x,t)$ are the single-agent marginals

$$p(s_a|x,t) = \sum_{s \setminus s_a} p(s|x,t) \tag{26}$$

of the joint distribution

$$p(s|x,t) = \frac{1}{Z(x,t)} w(s) \prod_{a=1}^{n} Z_a(s_a; x_a, t). \tag{27}$$

The weight $p(s|x,t)$ equals the ratio $\exp\left(-\frac{1}{\lambda}J(s;x,t)\right) / \exp\left(-\frac{1}{\lambda}J(x,t)\right)$, where $J(s;x,t) = -\lambda \log w(s) - \sum_{a=1}^{n} \lambda \log Z_a(s_a; x, t)$ is the optimal expected cost-to-go in case the agents have predetermined targets that are specified by the assignment s; an assignment of agents to targets that has a low expected cost $J(s;x,t)$ will yield a high weight $p(s|x,t)$, and the associated single-agent to single-target controls $u_a(s_a; x_a, t)$ will be predominant in the optimal controls $u_a(x,t)$.

3.2 Metropolis-Hastings Sampling in Multi-Agent Systems

In general, both the controls $u_a(s_a; x_a, t)$ and the marginals $p(s_a|x,t)$ in the optimal control (24) do not have a closed form solution, but have to be inferred approximately. The controls $u_a(s_a; x_a, t)$ can be approximated by the Metropolis-Hastings sampling discussed in Section 2.2. Inference of the marginals involves the inference of the path integral formulations of the $Z_a(s_a; x_a, t)$:

$$Z_a(s_a; x_a, t) = \lim_{\varepsilon \downarrow 0} \frac{1}{\sqrt{\det(2\pi\varepsilon\sigma^2)^N}} \int dx_a(t_1) \ldots \int dx_a(t_N) e^{-\frac{1}{\lambda} S_\varepsilon(x_a(t_0),\ldots,x_a(t_N),t_0;s_a)}$$

with $x_a(t_0) = x_a$, $t_0 = t$ and

$$\begin{aligned} S(x_a(t_0), \ldots, x_a(t_N), t_0; s_a) &= \phi_a(x_a(T); s_a) \\ &\quad + \sum_{i=0}^{N-1} \varepsilon\, V(x_a(t_i), t_i) + \sum_{i=0}^{N-1} \varepsilon\, \frac{1}{2} \left\| R_a \left(\frac{x_a(t_{i+1}) - x_a(t_i)}{\varepsilon} - b_a(x_a(t_i), t_i) \right) \right\|^2 . \end{aligned}$$

The value of $Z_a(s_a; x_a, t)$ is generally hard to determine (MacKay, 2003). Possible approximations include the maximum a posteriori (MAP) estimate and the inclusion of the variance in the sample paths. A third approximation is to take the average of the path costs as an estimate of $\log Z_a(s_a; x_a, t)$; this means that the entropy of the distribution in the path integral is neglected.

4. Graphical Model Inference

The additional computational effort in multi-agent control compared to single-agent control lies in the computation of the marginals $p(s_a|x,t)$ of the joint distribution $p(s|x,t)$, which involves a sum over all m^n assignments s. For small systems this is feasible, but for large systems this is only so if the summation can be performed efficiently. An efficient approach is provided by graphical model inference, which relies on a factor graph representation of the joint distribution.

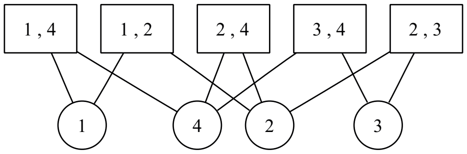


Figure 1: Example of a factor graph for a multi-agent system of four agents. The couplings are represented by the factors A, with $A = \{1, 4\}, \{1, 2\}, \{2, 4\}, \{3, 4\}, \{2, 3\}$.

4.1 A Factor Graph Representation of the Joint Distribution

The complexity of the joint distribution is in part determined by the weights $w(s)$ in the end cost function (22). These weights determine how the agents consider the states of the other agents. In the most complex case, the way one agent takes the state of another agent into account will depend on the states of all the other agents. The situation is less complicated when an agent considers the states of some agents independently of the states of the others. This means that the joint end cost has a factorized form:

$$w(s) = \prod_A w_A(s_A), \tag{28}$$

the A being subsets of agents. This structure is represented graphically by a so-called factor graph (Kschischang, Frey, & Loeliger, 2001). See Figure 1 for an example. The agents a and the factors A are nodes in the factor graph, represented by circles and squares respectively, and there is an edge between an agent a and a factor A when a is a member of subset A, that is, when w_A in the factorization of w depends on s_a. From (27) it is immediate that the joint distribution $p(s|x, t)$ factorizes according to the same factor graph.

4.2 The Junction Tree Algorithm

Efficient inference of the distribution $p(s|x, t)$ by means of its factor graph representation is accomplished by using the junction tree algorithm (Lauritzen & Spiegelhalter, 1988). The complexity of this algorithm is exponential in the induced tree width of the graph. A small tree width can be expected in systems where the factor graph is sparse, which is the case when the agents take the states into account of a limited number of other agents. This implies that multi-agent systems with sparse graphs and a limited number of targets are tractable (Wiegerinck et al., 2006). The factor graph in Figure 1 is an example of a sparse graph. On the other hand, should each agent take the state of each other agent into account, then the junction tree algorithm does not really help: the underlying factor graph is fully connected and the tree width of the graph equals the number of agents in the system.

Exact computation of the optimal control will be intractable in large and complex multi-agent systems, since the junction tree algorithm requires memory exponential in the tree width of the factor graph. Instead we can use graphical model approximate inference methods to approximately infer the marginals (26). We will proceed with a discussion of two such methods: naive mean field (MF) approximation (Jordan, Ghahramani, Jaakkola, & Saul, 1999) and belief propagation (BP) (Kschischang et al., 2001; Yedidia, Freeman, & Weiss, 2001).

4.3 Naive Mean Field Approximation

Our starting point is to note that the optimal expected cost-to-go is a log partition sum, also known as a free energy. Consider the variational free energy

$$F(q) = -\langle \lambda \log w \rangle_q - \sum_a \langle \log Z_a \rangle_{q_a} - \lambda H(q),$$

where $\langle\ \rangle_q$ and $\langle\ \rangle_{q_a}$ denote expectation values with respect to distribution q and marginals q_a respectively, and $H(q)$ is the entropy of q:

$$H(q) = -\sum_s q(s) \log q(s).$$

The optimal expected cost-to-go equals the variational free energy minimized over all distributions q. In the naive mean field approximation one considers the variational free energy restricted to factorized distributions $q(s) = \prod_a q_a(s_a)$. The minimum

$$J_{\mathrm{MF}} = \min_{q=\prod_a q_a} F(q)$$

is an upper bound for the optimal expected cost-to-go J, it equals J in case the agents are uncoupled. F has zero gradient in its local minima, that is,

$$0 = \frac{\partial F(q_1(s_1) \cdots q_n(s_n))}{\partial q_a(s_a)} \qquad a = 1, \ldots, n, \tag{29}$$

with additional constraints for normalization of the distributions q_a. Solutions to this set of equations are implicitly given by the mean field equations

$$q_a(s_a) = \frac{Z_a(s_a)\langle w|s_a\rangle_q}{\sum_{s'_a=1}^{n} Z_a(s'_a)\langle w|s'_a\rangle_q} \tag{30}$$

where $\langle w|s_a\rangle_q$ is the conditional expectation of w under q given s_a:

$$\langle w|s_a\rangle_q = \sum_{s_1,\ldots,s_n \backslash s_a} \left(\prod_{a' \neq a} q_{a'}(s_{a'}) \right) w(s_1, \ldots, s_n).$$

The mean field equations are solved by means of iteration; this procedure results in a convergence to a local minimum of the free energy.

The mean field approximation of the optimal control is found by taking the gradient with respect to x of the minimum J_{MF} of the free energy. This is similar to the exact case where the optimal control is the gradient of the optimal expected cost-to-go, equation (18). Using (29), we find

$$u_a(x,t) = -\frac{1}{\lambda}\sigma_a {\sigma_a}^\top \partial_{x_a} J_{\mathrm{MF}}(x,t) = \sum_{s_a} q_a(s_a) u_a(x_a, t; s_a).$$

Similar to the exact case, it is an average of the single-agent to single-target optimal controls $u_a(x_a, t; s_a)$ given by equation (25), where the average is taken with respect to the mean field approximate marginal $q_a(s_a)$ of agent a.

4.4 Belief Propagation

In belief propagation, we approximate the free energy by the Bethe free energy, and we minimize the latter. The Bethe free energy is defined by

$$F_{\text{Bethe}}(\{q_a, q_A\}) = -\sum_A \langle \lambda \log w_A \rangle_{q_A} - \sum_a \langle \lambda \log Z_a \rangle_{q_a} - \lambda \sum_A H(q_A) + \lambda \sum_a (n_a - 1) H(q_a). \tag{31}$$

It is a function of 'beliefs' $q_a(s_a)$ and $q_A(s_A)$, which are non-negative normalized functions that satisfy consistency relations:

$$\forall a \; \forall A \ni a : \sum_{s_{A \backslash a}} q_A(s_A) = q_a(s_a).$$

The $H(q_a)$ and $H(q_A)$ are the entropies of the beliefs q_a and q_A, n_a denotes the number of neighbors of node a in the factor graph.

Belief propagation is an algorithm that computes the beliefs (Kschischang et al., 2001). In case the joint distribution p has a factor graph representation that is a tree, belief propagation will converge to beliefs that are the exact marginals of p, and the Bethe free energy of these beliefs equals the optimal expected cost-to-go J. If the factor graph representation of p contains cycles, we may still apply belief propagation. Yedidia et al. (2001) showed that the fixed points of the algorithm correspond to local extrema of the Bethe free energy. In particular, more advanced variations on the algorithm (Heskes, Albers, & Kappen, 2003; Teh & Welling, 2001; Yuille, 2002) are guaranteed to converge to local minima of the Bethe free energy (Heskes, 2003).

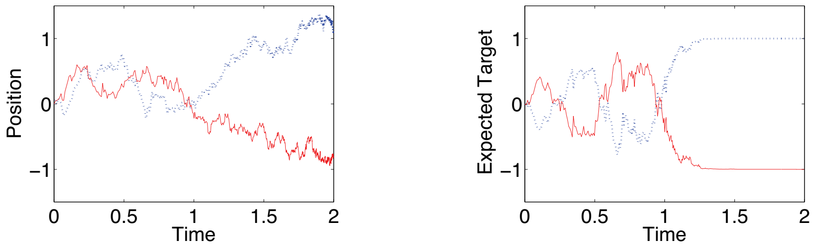
We find the BP approximation of the optimal control by taking the gradient of the minimum J_{Bethe} of the Bethe free energy:

$$u_a(x,t) = -\frac{1}{\lambda} \sigma_a {\sigma_a}^\top \partial_{x_a} J_{\text{Bethe}}(x,t) = \sum_{s_a} q_a(s_a) u_a(x_a, t; s_a),$$

with the $u_a(x_a, t; s_a)$ given by equation (25). Similar to the exact case and the mean field approximation, the BP approximation of the optimal control is an average of single-agent single-target optimal controls, where the average is taken with respect to the belief $q_a(s_a)$.

5. Numerical Results

In this section, we present numerical results of simulations of optimal control in multi-agent systems. The problem of computing the optimal controls (24) consists of two parts: the inference of the single-agent to single-target controls (25), and the inference of the marginals (26) of the global distribution over agent to target assignments. When the dynamics are linear, and the instantaneous costs V are zero, the single-agent to single-target controls can be given in closed form. Such multi-agent systems therefore only know the issue of infering marginal distributions. In Section 5.1 we will consider multi-agent systems of this kind. Section 5.2 deals with the general problem of infering the optimal controls when the dynamics are nonlinear and the instantaneous costs V are nonzero. In both sections



(a) Positions

(b) Expected Targets

Figure 2: Two agents, with noise and control in their positions, need to reach target locations at -1 and 1 at end time $t = 2$, each agent at a different target location. The positions (a) and expected targets (b) over time.

the joint end cost is given by equation (22), with

$$w(s) = \prod_{a,b}^{n} w_{a,b}(s_a, s_b), \qquad w_{a,b}(s_a, s_b) = \exp\left(-\frac{c}{n\lambda}\delta_{s_a,s_b}\right), \tag{32}$$

$$w_a(y_a; s_a) = \exp\left(-\frac{1}{\lambda}\phi_a(y_a; s_a)\right), \qquad \phi_a(y_a; s_a) = \frac{\alpha}{2}|y_a - \mu_{s_a}|^2, \tag{33}$$

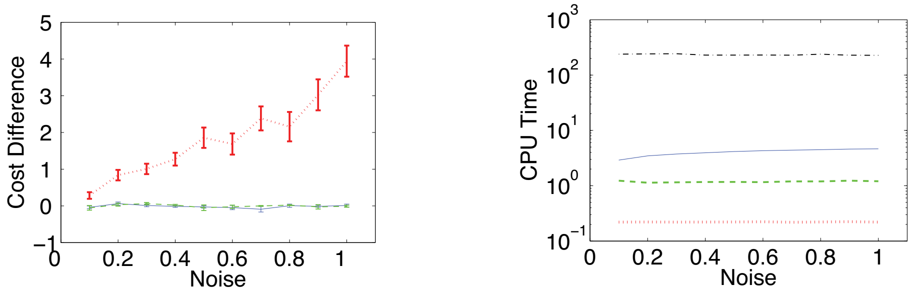
where c determines the coupling strength between the agents, and the μ_{s_a} are the target states.

5.1 Linear Dynamics

We begin with an illustration of optimal control by showing a simulation of an exactly solvable stochastic multi-agent system. In this system of two agents in one dimension, the agents satisfy dynamics (15) with b_a equal to zero. There are two target states, $x = \mu_1 = -1$ and $x = \mu_2 = 1$. The task of the agents is for each one to go to a different target. The instantaneous costs V in the cost function are zero, and the end cost function is given by equations (22), (32) and (33) with $\alpha = 20$ and $c = -4$. The negative sign of the coupling strength c implies a repulsion between the agents. The control cost parameter R equals 1, the noise level σ^2 lies at 0.5. The agents start at $x = 0$ at time $t = 0$, the end time lies at $T = 2$. To prevent overshooting the targets, udt should be small compared to the distance to the target states. This is done by choosing $dt = 0.05(T - t + 0.05)$.

Figure 2 shows the agents' positions and expected targets $\sum_{s_a=1,2} p(s_a|x,t)\mu_{s_a}$ over time. We see that up to time $t = 1$, the agents have not decided to which target each of them will go, and they remain between the two targets. Then, after $t = 1$, a final decision seems to have been made. This delayed choice is due to a symmetry breaking in the cost-to-go as time increases. Before the symmetry breaking, it is better to keep options open, and see what the effect of the noise is. After the symmetry breaking, time is too short to wait longer and a choice has to be made. This phenomenon is typical for multi-modal problems.

We proceed with a quantitative comparison of the different control methods that arise from the exact or approximate inferences of the marginals of the joint distribution (27).



(a) Costs

(b) CPU Time

Figure 3: The deviation from the optimal cost (a) and the required CPU Time in seconds (b) as functions of the noise. The lines represent exact $(\cdot - \cdot)$, Greedy $(\cdots)$, MF (—) and BP (−−) control.

The example we consider is a multi-agent system of $n = 18$ agents in a two-dimensional state space with zero instantaneous costs ($V = 0$) and no autonomous dynamics ($b_a = 0$). The end cost function is given by equations (22), (32) and (33). The two targets are located at $\mu_1 = (-1, 0)$ and $\mu_2 = (1, 0)$. $\alpha = 20$ and $c = -0.5$. The control cost matrix R equals the identity matrix. The agents start in $(0, 0)$ at time $t = 0$, the end time lies at $T = 2$, and time steps are of size $dt = 0.05(T - t + 0.05)$.

The approximations are naive mean field approximation and belief propagation, as described in Section 4, and greedy control. By greedy control we mean that at each time step each agent chooses to go to its nearest target. We include this approximation because it is simple and requires little computation time, and for those reasons it is an obvious choice for a naive approximation. Because a greedy control policy neglects the choices of the other agents, we expect that it will give an inferior performance.

For each approximation, Figure 3(a) shows the cost under the approximate (optimal) control minus the cost under exact (optimal) control, averaged over 100 simulations, and for different noise levels. The same noise samples were used for the approximate and the exact control. We see that both naive mean field approximation and belief propagation yield costs that on average coincide with the cost under exact control: the average cost difference under both methods does not significantly differ from zero. Greedy control, on the other hand, yields costs that are significantly higher than the costs under exact control; only in the deterministic limit does it converge to the cost under exact control, when both controls coincide. Figure 3(b) shows the CPU time required for the calculation of the controls under the different control methods. This is the average CPU time of an entire simulation. Each simulation consists of 73 time steps, and at each time step the control is calculated for each agent. We observe that greedy control is at least 10 times faster than the other methods, and exact control is nearly 100 times more time consuming than the other methods. Belief propagation gives a performance that for all considered noise levels is a bit quicker than the naive mean field approximation, but this may be the result of implementation details. We have also done simulations with attractive coupling $c = 0.5$; this returned results similar to the ones with repulsive coupling $c = -0.5$ that we presented here.

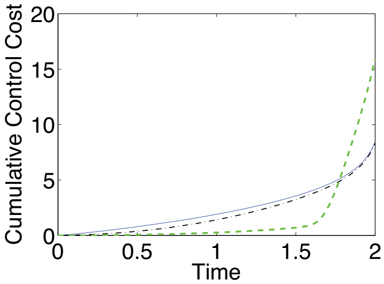


Figure 4: The cumulative control cost over time, in case of a strong repulsive coupling $c = -2$ and a low noise level $\sigma^2 = 0.1$. The curves represent exact ($\cdot - \cdot$), MF (—), and BP control (– –).

Although Figure 3 suggests that belief propagation and naive mean field approximation perform equally well, this is not always the case, since for certain combinations of the noise level and the coupling strength the BP control is more costly than MF control and exact control. The origin of this difference lies in the symmetry breaking, which tends to occur later under BP and earlier under MF when compared to exact control. We observe this in Figure 4, which shows the cumulative cost over time for the control methods in the multi-agent system, now with a coupling strength $c = -2$ and a fixed noise level $\sigma^2 = 0.1$. The cumulative costs are averages over 100 simulations. The cost under MF control lies a bit higher than the cost under exact control, whereas the cost under BP control initially is lower than the cost under the other control methods, but at $t = 1.7$ it starts to increase much faster and eventually ends up higher. Including the end costs, we found total costs 26.13 ± 0.12 under exact control, 26.19 ± 0.12 under MF control, and 35.5 ± 0.4 under BP control. This suggests that it is better to have an early symmetry breaking than a late symmetry breaking.

The time required for computing the control under the various methods depends on the number of agents in the multi-agent system. Figure 5 shows the required CPU time as a function of the number of agents n in the two-dimensional multi-agent system considered above. We see that the exact method requires a CPU time that increases exponentially with the number of agents. This is what may be expected from the theory, because the exact method uses the junction tree algorithm which has a complexity that is exponential in the tree width of the underlying graph, i.e., exponential in n. For the greedy method, the CPU time increases linearly with the number of agents, which is in agreement with the fact that under greedy control there is no coupling between the agents. The required CPU time increases polynomially for both the mean field approximation and belief propagation.

5.2 Nonlinear Dynamics

We now turn to multi-agent systems with nonlinear dynamics. To control these systems, we must approximate both the graphical model inference as well as the single-agent to single-target control problem (12). We consider a multi-agent system in which the agents move in

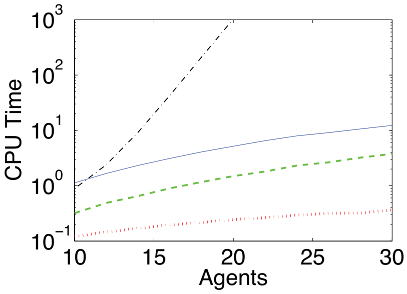


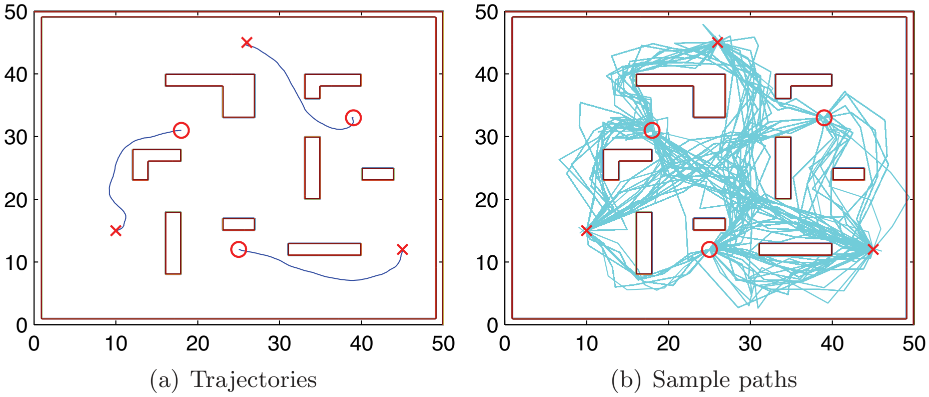
Figure 5: The required CPU time in seconds for the calculation of the controls at a different number of agents. Exact ($\cdot - \cdot$), greedy ($\cdots$), MF (—), and BP control (– –).

two dimensions and have a four-dimensional state that is specified by the agents location (x_a, y_a), its forward velocity v_a, and its driving direction φ_a. The dynamics of each agent is given by the equations

$$\begin{aligned} dx_a &= v_a \cos\varphi_a \, dt \\ dy_a &= v_a \sin\varphi_a \, dt \\ dv_a &= u_a dt + \sigma_a dw_a \\ d\varphi_a &= \omega_a dt + \nu_a d\xi_a. \end{aligned}$$

The first two equations model the kinematics of the agent's position for a given forward velocity and driving direction. The last two equations describe the control of the speed and the driving direction by application of a forward acceleration u_a and an angular velocity ω_a. The noise in the control is modeled by the standard normal Wiener processes w_a and ξ_a and the noise level parameters σ_a and ν_a. Note that the noise does not act in dimensions other than those of the control. Although the control space counts less dimensions than the state space, the example does fit in the general framework: we refer to Appendix C for details.

We look at two different tasks. The first task is that of obstacle avoidance in a multi-agent system of three agents. The agents each have to reach one of three target locations and avoid any obstacles in the environment. Each target location should be reached by precisely one agent; we model this with an end cost function, given by equations (22), (32) and (33), with $\alpha = \infty$ and $c = -0.5$. The targets are located at $(10, 15)$, $(45, 12)$ and $(26, 45)$, and the agents should arrive with zero velocity. The control cost matrix R is the identity matrix. $\lambda = 0.1$. The instantaneous cost V equaled 1000 at the locations of the obstacles, and zero otherwise. The agents start at time $t = 0$, the end time lies at $T = 20$, and time steps dt are of size 0.2. The starting locations of the agents are $(18, 31)$, $(25, 12)$ and $(39, 33)$, and the agents start with zero velocity. The sample paths are discrete time paths in the two-dimensional space of the forward velocity v and the driving direction φ. They are specified by their values at times $t_i = t + i\,\varepsilon$, $i = 0, \ldots, N-1$, with $\varepsilon = \frac{T-t}{N-1}$ and $N = 7$, the value at time t_0 equals the current state of one of the agents, and the value at time t_N equals one of the target end states. The control for each agent to one of the targets is computed



(a) Trajectories (b) Sample paths

Figure 6: Three agents, with noise and control in their forward velocities and driving directions, have to reach three targets (marked by 'X') in an environment containing also a number of walls. Each agent starts at a different location (marked by 'O') and with zero forward velocity, and each agent should arrive at a different target with zero velocity without hitting the walls. (a) The trajectories that the agents followed to reach the targets. (b) Sample paths.

with a Metropolis-Hastings sampling of paths, according to Subsection 3.2. The proposal distribution is a $2N$-dimensional Gaussian, centered around the agent's current planned path, and with a variance equal to the noise level in the agent's dynamics. The expectation values $Z_a(s_a; x_a, t)$ are estimated by the average costs of the sample paths. We have also tried MAP estimation of $Z_a(s_a; x_a, t)$ and an inclusion of the variance in the sample paths, but the former did not show a significant difference, and the latter returned estimates that fluctuated heavily. Figure 6(a) shows the environment and the trajectories of the agents from their starting locations to the targets. Each agent manages to avoid the obstacles and arrive at one of the targets with zero velocity, such that each target is reached by a different agent.

The second task is that of coordination in the multi-agent system as shown in Figure 7(a). In this system there are no instantaneous costs ($V = 0$). The agents have to move from their initial positions to a number of target locations. They should arrive at these locations with zero velocity and horizontal driving direction. There is an equal number of agents and target locations, and each agent has to reach a different target. The initial locations are aligned vertically, and so are the target locations, but there is a vertical displacement between the two. Thus the agents have to coordinate their movements in order to reach the targets in a satisfactory way.

The agents start at time 0, the end time lies at 100, and they make time steps of size $dt = \frac{T-t}{2(N-1)}$, with $N = 7$, until $dt < 0.01$. At each time step the controls are computed by a Metropolis-Hastings sampling of paths and a naive mean field approximation to infer the marginals $p_a(s_a|x, t)$ that weigh the single-agent to single-target controls, equations (24) and (26). The sample paths were discretized into seven equidistant time points from the present time to the end time. The proposal distribution was taken a Gaussian, which was centered around the agent's current planned path and with a variance equal to the noise level in the agent's dynamics. Figure 7(a) shows an example of the trajectories of a system of 10 agents. It was obtained with 10 sample paths per agent-target combination. We observe that the agents reach the targets, and that each target is reached by precisely one agent, as required. Due to the noise in the second order dynamics of the agents, it takes the agents less effort to approach a target than to remain there, since the former allows exploitation of the noise while the latter requires a constant correction of the state changes caused by the noise. The result is that the trajectories of the agents are more curved and elongated than what would be expected in the situation without noise. The simulation was carried out as well for a larger number of agents. Figure 7(b) shows the required CPU time as a function of the number of agents, both under exact and MF inference of the marginals of the agents. Note that the complexity of the graphical model inference problem scales as n^n, with n the number of agents. Exact inference using the junction tree algorithm was only feasible for $n < 10$.

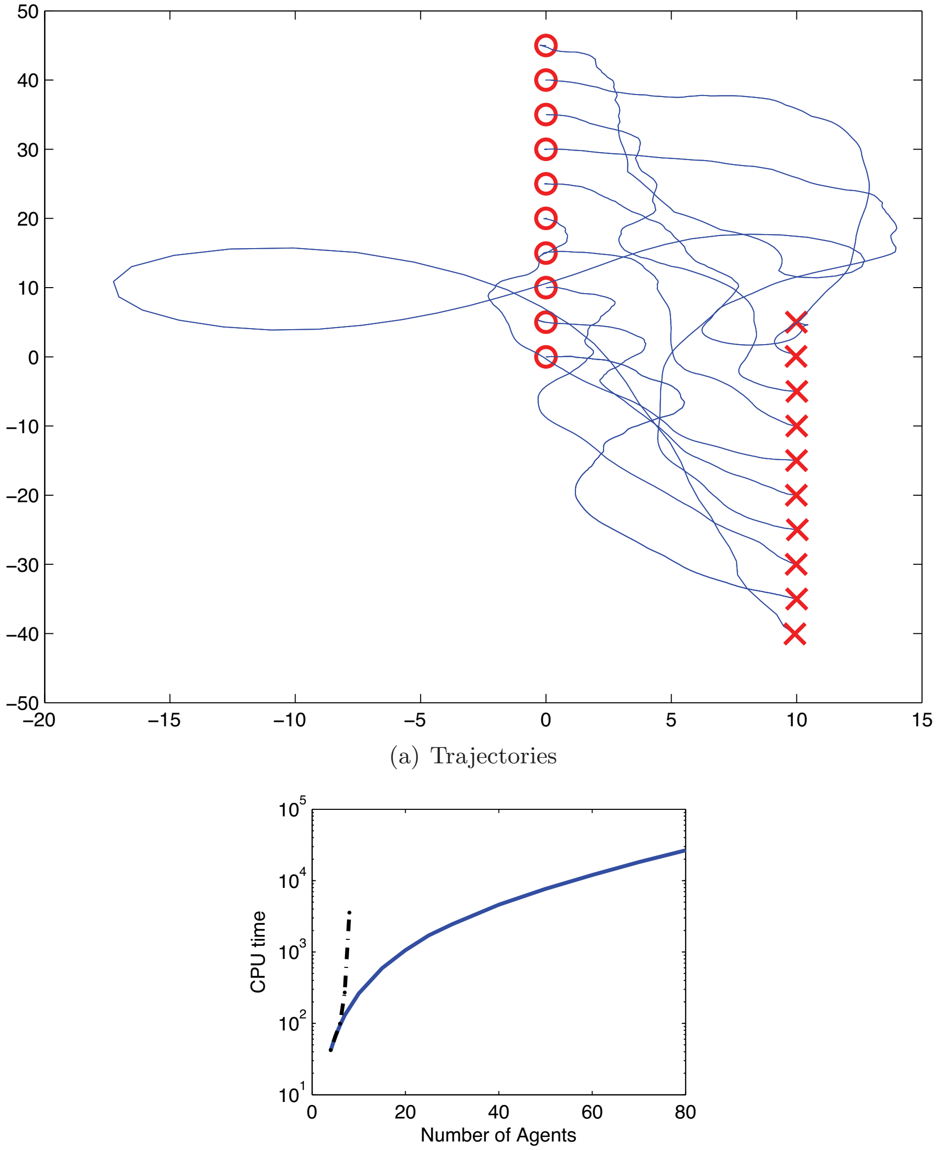
6. Discussion

We studied the use of graphical model inference methods in optimal control of stochastic multi-agent systems in continuous space and time where the agents have a joint task to reach a number of target states. Rather than discretizing, as is commonly done and typically makes large systems intractable due to the curse of dimensionality, we followed the approach developed by Wiegerinck et al. (2006), modeling the system in continuous space and time. Under certain assumptions on the dynamics and the cost function, the solution can be given in terms of a path integral.

The path integral can be computed in closed form in a few special cases, such as the linear-quadratic case, but in general it has to be approximated. This can be done by a variety of methods. The method we considered in this paper is MCMC sampling. The dimension of the sample paths was kept low ($N = 7$) to limit the curvature of the sample paths. The gain of limiting the curvature is that the variance in the samples is reduced and less samples are needed. By limiting the curvature, however, we introduce a bias. In addition, in the presence of obstacles insufficient curvature would make the sampler return sample paths that run through the obstacles. We believe that more advanced MCMC methods such as Hybrid MC sampling (Duane, Kennedy, Pendleton, & Roweth, 1987) and overrelaxation (Neal, 1998) can improve the inference of the path integrals.

Apart from MCMC sampling, there are other approximation methods that one could consider, such as the Laplace approximation or a variational approximation. The Laplace approximation becomes exact in the noiseless limit and could be useful in low noise regimes as well. The variational approximation approximates the path integral (11) by a Gaussian process (Archambeau, Opper, Shen, Cornford, & Shawe-Taylor, 2007), and could be particularly useful in the high noise regime. A drawback of the variational approach, however, is that it cannot be straightforwardly applied to situations with infinite instantaneous costs, like hard obstacles in the environment that we considered here.

Wiegerinck et al. (2006) showed that for systems that are sufficiently sparse and in which the single-agent to single-target controls can be determined in closed form, e.g. linear-quadratic control with time-independent coefficients, exact inference can be achieved using



(a) Trajectories

(b) CPU time

Figure 7: (a) The trajectories of 10 agents from starting locations ‘O’ to 10 targets ‘X’. (b) The required CPU time in seconds as a function of the number of agents, with the number of targets equal to the number of agents. The lines represent exact $(\cdot - \cdot)$ and MF (—) inference of the marginals.

the junction tree algorithm. Van den Broek, Wiegerinck, and Kappen (2007) considered a multi-agent system with second-order dynamics, linear autonomous dynamics and zero instantaneous costs, and showed that graphical model inference by naive mean field approximation significantly outperformed a greedy inference. Here we showed that a close to optimal result can be achieved as well in dense systems, using graphical model approximate inference methods. The approximation methods that we considered were naive mean field approximation and belief propagation. We demonstrated their performances in an example system where exact inference is significantly more time consuming. Mean field approximation showed to work very well, returning costs for control equal to the optimal ones, belief propagation performed similarly. Below a certain value for the ratio of coupling strength to the noise level, the symmetry breaking in the control process takes place earlier under mean field approximation when compared to exact inference, and later under belief propagation. An early symmetry breaking does not increase the costs for coordination much, however, a late symmetry breaking does, making the performance under belief propagation suboptimal.

Some variations on the considered case are also possible within the general framework. Wiegerinck, van den Broek, and Kappen (2007) discuss situations where agents sequentially visit a number of targets, and where the end time is not fixed. It focusses on prefered trajectories in state space over time, instead of prefered states at the end time; this is achieved by modeling the path cost in a way similar to how we have modeled the end cost. The problem where agents have to intercept a moving target with noisy dynamics is also covered there.

The control formalism developed by Kappen (2005a, 2005b) and applied to multi-agent coordination by Wiegerinck et al. (2006) and in this article, demands that the noise and the control act in the same dimensions. One way to satisfy this constraint is to assume that the agents are identical. In addition, the single agent dynamics should be such that the noise and the control act in the same dimensions. We saw that for the two-dimensional second order system in Section 5.2 this condition was satisfied in a natural way. However, in general one can think of examples of control problems where equation (3) is violated. An interesting future direction of research is to investigate to what extend the path integral approach can be used as an approximation in such cases.

The paper assumes that the joint state space of the agents is observable to all agents. For large multi-agent systems, however, it will be more realistic that an agent only observes its own state and the states of agents that are physically nearby. Our approach does not directly apply to such situations. Depending on the joint task of the agents, it may be a valid approximation to do optimal control in the sub-system consisting of those agents that one agent does observe. If the task of the agents is to avoid collisions, then it will be sufficient to consider only the states of agents that are nearby, but if the task is to all go to the same target then it will be crucial to have information about the states of all other agents. A natural alternative to deal with partial observability is to describe the multi-agent system by a decentralized POMDP (Seuken & Zilberstein, 2008). It is not clear however, how such an approach would combine with the path integral formalism.

The topic of learning has not been addressed in this paper, but clearly is of great interest. However, one could argue that a sampling procedure to compute the path integral corresponds to a learning of the environment. A discussion on this line of thought can be found in (Kappen, 2007).

There are many more possible model extensions worthwhile exploring in future research. Obvious examples are bounded controls, or a limited observation of the global state of the system; these issues are already of interest to study in the single agent situation. Others apply typically to the multi-agent situation. In the context of physical agents, introducing penalties for collisions between agents would become relevant. Typically, these types of model extensions will not have a solution in closed form, and will require additional approximate numerical methods. Some suggestions are given by Kappen (2005a, 2005b).

Acknowledgments

We thank the reviewers for their useful comments. We thank Joris Mooij for making available useful software (`www.mbfys.ru.nl/~jorism/libDAI/`). This research is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant BSIK03024.

Appendix A. Stochastic Optimal Control

In this appendix we give a derivation of (5), (6) and (7), starting from (1), (2), (3) and (4). Detailed discussions can be found in many works on stochastic optimal control, for example that of Kushner (1967), Fleming and Rishel (1975), Fleming (1978), Øksendal (1998), Stengel (1993), and Kappen (2005a, 2005b).

The optimal expected cost-to-go J in a state x at time t is defined as

$$J(x,t) = \min_u C^u(x,t), \tag{34}$$

where

$$C^u(x,t) = \mathbb{E}^u_{x,t}\left[\phi(x(T)) + \int_t^T d\theta \left(\frac{1}{2}\|Ru(x(\theta),\theta)\|^2 + V(x(\theta),\theta)\right)\right] \tag{35}$$

is the expected cost given the control law u. These are the equations (4) and (2) in the main text. We first show that J satisfies the stochastic Hamilton-Jacobi-Bellman (SHJB) equation

$$-\partial_t J = \min_u \left(\frac{1}{2}\|Ru\|^2 + (b+u)^\top \partial_x J + \frac{1}{2}\mathrm{Tr}\left(\sigma\sigma^\top \partial_x^2 J\right) + V\right), \tag{36}$$

with boundary condition $J(x,T) = \phi(x)$. This equation is derived in the following way. For any moment in time θ between t and T it holds that

$$\begin{aligned} J(x,t) &= \min_u \mathbb{E}^u_{x,t}\left[C^u(x(\theta),\theta) + \int_t^\theta ds \left(\frac{1}{2}\|Ru(x(s),s)\|^2 + V(x(s),s)\right)\right] \\ &= \min_u \mathbb{E}^u_{x,t}\left[J(x(\theta),\theta) + \int_t^\theta ds \left(\frac{1}{2}\|Ru(x(s),s)\|^2 + V(x(s),s)\right)\right]. \end{aligned}$$

The first line follows from dividing the integral from t to T into two integrals, one from t to θ and one from θ to T, and using the definition of the cost function C, the second line follows from the definition of J. A rewriting yields

$$0 = \min_u \mathbb{E}^u_{x,t}\left[\frac{J(x(\theta),\theta) - J(x,t)}{\theta - t} + \frac{1}{\theta - t}\int_t^\theta ds\left(\frac{1}{2}\|Ru(x(s),s)\|^2 + V(x(s),s)\right)\right].$$

Taking the limit $\theta \to t$ we obtain

$$0 = \min_u \mathbb{E}^u_{x,t}\left[\frac{dJ(x(t),t)}{dt} + \frac{1}{2}\|Ru(x(t),t)\|^2 + V(x(t),t)\right]. \tag{37}$$

Subsequently, we apply to $dJ(x(t),t)$ the well known chain rule for diffusion processes:

$$dJ(x(t),t) = \sum_i \frac{\partial J(x(t),t)}{\partial x_i}dx_i(t) + \frac{\partial J(x(t),t)}{\partial t}dt + \frac{1}{2}\sum_{i,j}\frac{\partial^2 J(x(t),t)}{\partial x_i \partial x_j}dx_i(t)dx_j(t). \tag{38}$$

It differs from the chain rule for deterministic processes in that it also contains a term quadratic in dx. This extra term does not vanish, because the Wiener process appearing in the dynamics (1) has quadratic variation that increases linear in time:

$$\mathbb{E}^u_{x,t}\left[dw_i(t)dw_j(t)\right] = \delta_{ij}dt. \tag{39}$$

It follows that in expectation $dx_i(t)dx_j(t)$ is equal to $(\sigma\sigma^\top)_{ij}dt$. By substituting the dynamics (1) in (38), taking expectation values, and using (39), we obtain

$$\mathbb{E}^u_{x,t}\left[dJ(x(t),t)\right] = \frac{\partial J(x,t)}{\partial t}dt + (b(x,t)+u(x,t))^\top \frac{\partial J(x,t)}{\partial x}dt + \mathrm{Tr}\left(\sigma\sigma^\top \frac{\partial^2 J(x,t)}{\partial x \partial x}\right)dt.$$

Substitution into equation (37) then yields equation (36).

The minimum of the right-hand side of equation (36) is given by

$$u = -(R^\top R)^{-1}\partial_x J.$$

This is the optimal control.

The minimization in (36) is removed by inserting the optimal control. This yields a nonlinear equation for J. We can remove the nonlinearity by using a logarithmic transformation: if we introduce a constant λ, and define $Z(x,t)$ through $J(x,t) = -\lambda \log Z(x,t)$, then

$$\begin{aligned}\frac{1}{2}u^\top R^\top R u + u^\top \partial_x J &= -\frac{1}{2}\lambda^2 Z^{-2}(\partial_x Z)^\top (R^\top R)^{-1}\partial_x Z,\\ \frac{1}{2}\mathrm{Tr}\left(\sigma\sigma^\top \partial_x^2 J\right) &= \frac{1}{2}\lambda Z^{-2}(\partial_x Z)^\top \sigma\sigma^\top \partial_x Z - \frac{1}{2}\lambda Z^{-1}\mathrm{Tr}\left(\sigma\sigma^\top \partial_x^2 Z\right).\end{aligned}$$

The terms quadratic in $\partial_x Z$ vanish when $\sigma\sigma^\top$ and R are related via equation (3),

$$\sigma\sigma^\top = \lambda(R^\top R)^{-1}.$$

When this relation is satisfied, the SHJB equation becomes

$$\begin{aligned}\partial_t Z &= \left(\frac{V}{\lambda} - b^\top \partial_x - \frac{1}{2}\mathrm{Tr}\big(\sigma\sigma^\top \partial_x^2\big)\right)Z\\ &= -HZ,\end{aligned} \tag{40}$$

where H a linear operator acting on the function Z.

Equation (40) must be solved backwards in time with boundary condition $Z(x,T) = e^{-\frac{1}{\lambda}\phi(x)}$. We present a solution in terms of a forward diffusion process. It is a common approach in the theory of stochastic processes to give solutions to partial differential equations in terms of diffusion processes. The solution to equation (40) is the expectation value

$$Z(x,t) = \mathbb{E}_{x,t}\left[\exp\left(-\frac{1}{\lambda}\phi(y(T)) - \frac{1}{\lambda}\int_t^T d\theta\, V(y(\theta),\theta)\right)\right], \tag{41}$$

where $y(\theta)$ is a process that satisfies the uncontrolled dynamics

$$dy(\theta) = b(y(\theta),\theta)d\theta + \sigma dw(\theta),$$

and $y(t) = x$. The expectation $\mathbb{E}_{x,t}$ is taken with respect to the probability measure under which $y(\theta)$ satisfies the uncontrolled dynamics with condition $y(t) = x$. It is clear that (41) matches the boundary condition. To verify that it satisfies equation (40), we let

$$I(t) = \exp\left(-\frac{1}{\lambda}\int_t^T d\theta\, V(y(\theta),\theta)\right).$$

We see that

$$dI(t) = \frac{1}{\lambda}V(y(t),t)I(t)dt.$$

Let f be the function $f(y) = \exp\left(-\frac{1}{\lambda}\phi(y)\right)$. We again use the chain rule for stochastic processes and apply it to $f(y(T-\theta))$ to find

$$\begin{aligned} df(y(T-\theta)) &= \sum_{i=1}^{k}\frac{\partial f(y(T-\theta))}{\partial y_i}dy_i(T-\theta) + \frac{1}{2}\sum_{i,j=1}^{k}\frac{\partial^2 f(y(T-\theta))}{\partial y_i \partial y_j}dy_i(T-\theta)dy_j(T-\theta) \\ &= \left(\frac{\partial f(y(T-\theta))}{\partial y}\right)^\top(-b(y(T-\theta),T-\theta)d\theta + \sigma dw(T-\theta)) \\ &\quad -\frac{1}{2}\mathrm{Tr}\left(\sigma\sigma^\top\frac{\partial^2 f(y(T-\theta))}{\partial y \partial y}\right)d\theta. \end{aligned}$$

We then choose $\theta = 0$ and $d\theta = dt$ and combine this identity with the previous one to obtain

$$\begin{aligned} df(y(T))I(t) &= f(y(T))dI(t) + I(t)df(y(T)) \\ &= -Hf(y(T))I(t)dt + \partial_y f(y(T))I(t)\sigma dw(T). \end{aligned}$$

Taking the expectation value on both sides makes the term $\partial_y f(y(T))I(t)\sigma dw(T)$ disappear, and the remaining part,

$$d\mathbb{E}\left[f(y(T))I(t)\right] = -H\mathbb{E}\left[f(y(T))I(t)\right]dt,$$

is just equation (40).

Appendix B. The Path Integral Formulation

We are going to write the expectation value (7) as a path integral. Partitioning the time interval from t to T into N intervals of equal length ε, $t = t_0 < t_1 < \ldots < t_N = T$, the expectation value can be written as follows:

$$Z(x,t) = \int dx_1 \ldots \int dx_N \, e^{-\frac{1}{\lambda}\phi(x_N)} \prod_{i=0}^{N-1} Z(x_{i+1}, t_{i+1}; x_i, t_i) \tag{42}$$

where $x_0 = x$ and the $Z(x_{i+1}, t_{i+1}; x_i, t_i)$ are implicitly defined by

$$\int dx_{i+1} \, Z(x_{i+1}, t_{i+1}; x_i, t_i) f(x_{i+1}) = \mathbb{E}\left[f(x_{i+1}) \exp\left(-\frac{1}{\lambda} \int_{t_i}^{t_{i+1}} d\theta \, V(y(\theta), \theta) \right) \middle| y(t_i) = x_i \right]$$

for arbitrary functions f. In the limit of infinitesimal ε, the $Z(x_{i+1}, t_{i+1}; x_i, t_i)$ satisfy

$$Z(x_{i+1}, t_{i+1}; x_i, t_i) = \rho(x_{i+1}, t_{i+1} | x_i, t_i) \exp\left(-\frac{1}{\lambda} V(x_i, t_i) \varepsilon \right), \tag{43}$$

where $\rho(x_{i+1}, t_{i+1} | x_i, t_i)$ is the transition probability of the uncontrolled dynamics (8) to go from (x_i, t_i) to (x_{i+1}, t_{i+1}) in space-time. The transition probability is given by

$$\rho(x_{i+1}, t_{i+1} | x_i, t_i) = \frac{1}{\sqrt{\det(2\pi\varepsilon\sigma^2)}} \exp\left(-\frac{\|\sigma^{-1}(x_{i+1} - x_i - b(x_i, t_i)\varepsilon)\|^2}{2\varepsilon} \right).$$

This follows from the dynamics

$$x_{i+1} - x_i = b(x_i, t_i)\varepsilon + \sigma w$$

over the infinitesimal time interval and the observation that the Wiener process w is normally distributed around zero with variance ε. Using equation (3), we may rewrite the transition probability as

$$\rho(x_{i+1}, t_{i+1} | x_i, t_i) = \frac{1}{\sqrt{\det(2\pi\varepsilon\sigma^2)}} \exp\left(-\frac{1}{2\lambda} \left\| R\left(\frac{x_{i+1} - x_i}{\varepsilon} - b(x_i, t_i) \right) \right\|^2 \varepsilon \right). \tag{44}$$

We obtain the path integral representation of $Z(x,t)$ by combining equations (42), (43) and (44) in the limit of ε going to zero:

$$Z(x,t) = \lim_{\varepsilon \downarrow 0} Z_\varepsilon(x_0, t_0) \tag{45}$$

with $x_0 = x$, $t_0 = t$,

$$Z_\varepsilon(x_0, t_0) = \frac{1}{\sqrt{\det(2\pi\varepsilon\sigma^2)^N}} \int dx_1 \ldots \int dx_N \, e^{-\frac{1}{\lambda} S_\varepsilon(x_0, \ldots, x_N, t_0)}$$

and

$$S_\varepsilon(x_0, \ldots, x_N, t_0) = \phi(x_N) + \sum_{i=0}^{N-1} \varepsilon \, V(x_i, t_i) + \sum_{i=0}^{N-1} \varepsilon \, \frac{1}{2} \left\| R\left(\frac{x_{i+1} - x_i}{\varepsilon} - b(x_i, t_i) \right) \right\|^2.$$

The optimal control is given by equation (9) and is proportional to the gradient of $\log Z(x,t)$. Substituting the path integral representation (45) of $Z(x,t)$, we find that

$$\begin{aligned}
u(x_0,t_0) &= \lim_{\varepsilon\downarrow 0}\int dx_1\ldots\int dx_N \frac{e^{-\frac{1}{\lambda}S_\varepsilon(x_0,\ldots,x_N,t_0)}}{\sqrt{\det(2\pi\varepsilon\sigma^2)^N}Z_\varepsilon(x,t_0)}\sigma\sigma^\top\partial_{x_0}\left(-\frac{1}{\lambda}S_\varepsilon(x_0,\ldots,x_N,t_0)\right)\\
&= \lim_{\varepsilon\downarrow 0}\int dx_1\ldots\int dx_N\, p(x_0,\ldots,x_N,t_0)u(x_0,\ldots,x_N,t_0)
\end{aligned}$$

where

$$u(x_0,\ldots,x_N,t_0) = \frac{x_1-x_0}{\varepsilon} - b(x_0,t_0)$$

and

$$p(x_0,\ldots,x_N,t_0) = \frac{e^{-\frac{1}{\lambda}S_\varepsilon(x_0,\ldots,x_N,t_0)}}{\sqrt{\det(2\pi\varepsilon\sigma^2)^N}Z_\varepsilon(x_0,t_0)}.$$

Note that the control $u(x_0,\ldots,x_N,t_0)$ that results from a path $(x_0,\ldots,x_N)$ only depends on the first two entries x_0 and x_1 of the path.

Appendix C. Dimension Reduction

The derivation of the path integral in Appendix B was given for the case that both the state and the control are k-dimensional. The particular case that only some dimensions of the state are controlled can be deduced by taking the limit of infinite control cost along the dimensions without control. The control along the latter dimensions then becomes zero, as can be seen from equation (5). The noise in these dimensions is equal to zero in accordance with relation (3). In the path integral formalism the transition probabilities (44) then reduce to delta functions along the dimensions without control. The implications for the MCMC sampling are that the dimension of the space in which to sample is also reduced, since sampling has only to be performed in the dimensions where there is noise.

References

Archambeau, C., Opper, M., Shen, Y., Cornford, D., & Shawe-Taylor, J. (2007). Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*.

Becker, R., Zilberstein, S., Lesser, V., & Goldman, C. V. (2003). Transition-independent decentralized Markov decision processes. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 41–48.

Becker, R., Zilberstein, S., Lesser, V., & Goldman, C. V. (2004). Solving transition independent decentralized Markov decision processes. *Journal of Artificial Intelligence Research*, *22*, 423–455.

Boutilier, C. (1996). Planning, learning and coordination in multiagent decision processes. In *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 195–210.

Castanon, D. A., Pachter, M., & Chandler, P. R. (2004). A game of deception. In *Proceedings of the 43rd IEEE Conference on Decision and Control*, pp. 3364–3369.

Duane, S., Kennedy, A., Pendleton, B., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222.

Fleming, W. H. (1978). Exit probabilities and optimal stochastic control. *Applied Mathematics and Optimization*, *4*, 329–346.

Fleming, W. H., & Rishel, R. W. (1975). *Deterministic and Stochastic Optimal Control*. Springer-Verlag, New York.

Guestrin, C., Koller, D., & Parr, R. (2002a). Multiagent planning with factored MDPs. In *Advances in Neural Information Processing Systems*, Vol. 14, pp. 1523–1530.

Guestrin, C., Venkataraman, S., & Koller, D. (2002b). Context-specific multiagent coordination and planning with factored MDPs. In *Eighteenth National Conference on Artificial Intelligence*, pp. 253–259.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Heskes, T. (2003). Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Advances in Neural Information Processing Systems*, Vol. 15, pp. 343–350.

Heskes, T., Albers, K., & Kappen, B. (2003). Approximate inference and constrained optimization. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pp. 313–320.

Hu, J., Prandini, M., & Tomlin, C. (2007). Conjugate points in formation constrained optimal multi-agent coordination: A case study. *SIAM Journal on Control and Optimization*, *45*(6), 2119–2137.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). An introduction to variational methods for graphical models. In *Learning in Graphical Models*. MIT Press, Cambridge.

Kamal, W. A., Gu, D.-W., & Postlethwaite, I. (2005). Real time trajectory planning for UAVs using MILP. In *Proceedings of the 4th IEEE Conference on Decision and Control, and the European Control Conference 2005*, pp. 3381–3386.

Kappen, H. J. (2005a). Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, P11011.

Kappen, H. J. (2005b). Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, *95*(20), 200201.

Kappen, H. J. (2007). An introduction to stochastic control theory, path integrals and reinforcement learning. In *AIP conference proceedings*, Vol. 887, pp. 149–181.

Kleinert, H. (2006). *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*. World Scientific, Singapore.

Kschischang, F. R., Frey, B. J., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*(2), 498–519.

Kushner, H. J. (1967). *Stochastic Stability and Control*. Academic Press Inc., New York.

Larson, R. A., Pachter, M., & Mears, M. (2005). Path planning by unmanned air vehicles for engaging an integrated radar network. In *Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit*.

Lauritzen, S., & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Royal Statistical Society Series B*, *50*, 157–224.

Liu, Y., Cruz, J. B., & Schumacher, C. J. (2007). Pop-up threat models for persistent area denial. *IEEE Transactions on Aerospace and Electronic Systems*, *43*(2), 509–521.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Neal, R. M. (1998). *Learning in Graphical Models*, pp. 205–225. Kluwer Academic Publishers.

Øksendal, B. (1998). *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag.

Pachter, L., & Pachter, M. (2001). Optimal paths for avoiding a radiating source. In *Proceedings of the 40th IEEE Conference on Decision and Control*, pp. 3581–3586.

Ribichini, G., & Frazzoli, E. (2003). Efficient coordination of multiple-aircraft systems. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, Vol. 1, pp. 1035–1040.

Sadati, N., & Elhamifar, E. (2006). Semi-decentralized control of multi-agent systems based on redundant manipulator optimization methods. In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control*, pp. 278–283.

Seuken, S., & Zilberstein, S. (2008). Formal models and algorithms for decentralized decision making under uncertainty. *Journal of Autonomous Agents and Multi-Agent Systems*.

Shi, X., Wang, X., Liu, Y., Wang, C., & Zu, C. (2007). Optimization of fighter aircraft evasive trajectories for radar threats avoidance. In *Proceedings of the 2007 IEEE International Conference on Control and Automation*, pp. 303–307.

Stengel, R. (1993). *Optimal Control and Estimation*. Dover Publications, New York.

Subramanian, S. K., & Cruz, J. B. (2003). Adaptive models of pop-up threats for multi-agent persistent area denial. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, pp. 510–515.

Teh, Y., & Welling, M. (2001). The unified propagation and scaling algorithm. In *Advances in Neural Information Processing Systems*, Vol. 14, pp. 953–960.

Todorov, E., & Li, W. (2005). A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the American Control Conference*, pp. 300–306.

Tomlin, C., Pappas, G. J., & Sastry, S. (1998). Conflict resolution for air traffic management: A study in multiagent hybrid systems. *IEEE Transactions on Automatic Control*, *43*(4), 509–521.

van den Broek, B., Wiegerinck, W., & Kappen, B. (2007). Optimal control in large stochastic multi-agent systems. In *Proceedings of the Seventh Symposium on Adaptive Learning Agents and Multi-Agent Systems*, pp. 9–20.

van Leeuwen, P., Hesseling, H., & Rohling, J. (2002). Scheduling aircraft using constraint satisfaction. *Electronic Notes in Theoretical Computer Science*, *76*, 252–268.

Wiegerinck, W., van den Broek, B., & Kappen, B. (2006). Stochastic optimal control in continuous space-time multi-agent systems. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pp. 528–535.

Wiegerinck, W., van den Broek, B., & Kappen, B. (2007). Optimal on-line scheduling in stochastic multi-agent systems in continuous space-time. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 744–751.

Yedidia, J., Freeman, W., & Weiss, Y. (2001). Generalized belief propagation. In *Advances in Neural Information Processing Systems*, Vol. 13, pp. 689–695.

Yuille, A. (2002). CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, *14*(7), 1691–1722.