Optimal Control in Large Stochastic Multi-Agent Systems

Bart van den Broek, Wim Wiegerinck, and Bert Kappen

SNN, Radboud University Nijmegen, Geert Grooteplein 21, Nijmegen,
The Netherlands
{b.vandenbroek, w.wiegerinck, b.kappen}@science.ru.nl

Abstract. We study optimal control in large stochastic multi-agent systems in continuous space and time. We consider multi-agent systems where agents have independent dynamics with additive noise and control. The goal is to minimize the joint cost, which consists of a state dependent term and a term quadratic in the control. The system is described by a mathematical model, and an explicit solution is given. We focus on large systems where agents have to distribute themselves over a number of targets with minimal cost. In such a setting the optimal control problem is equivalent to a graphical model inference problem. Exact inference will be intractable, and we use the mean field approximation to compute accurate approximations of the optimal controls. We conclude that near to optimal control in large stochastic multi-agent systems is possible with this approach.

1 Introduction

A collaborative multi-agent system is a group of agents in which each member behaves autonomously to reach the common goal of the group. Some examples are teams of robots or unmanned vehicles, and networks of automated resource allocation. An issue typically appearing in multi-agent systems is decentralized coordination; the communication between agents may be restricted, there may be no time to receive all the demands for a certain resource, or an unmanned vehicle may be unsure about how to anticipate another vehicles movement and avoid a collision.

In this paper we focus on the issue of optimal control in large multi-agent systems where the agents dynamics are continuous in space and time. In particular we look at cases where the agents have to distribute themselves in admissible ways over a number of targets. Due to the noise in the dynamics, a configuration that initially seems attainable with little effort may become harder to reach later on.

Common approaches to derive a coordination rule are based on discretizations of space and time. These often suffer from the curse of dimensionality, as the complexity increases exponentially in the number of agents. Some successfull ideas, however, have recently been put forward, which are based on structures that are assumed to be present [1, 2].

Here we rather model the system in continuous space and time, following the approach of Wiegerinck et al. [3]. The agents satisfy dynamics with additive control and noise, and the joint behaviour of the agents is valued by a joint cost function that is quadratic in the control. The stochastic optimization problem may then be transformed into a linear partial differential equation, which can be solved using generic path integral methods [4, 5]. The dynamics of the agents are assumed to factorize over the agents, such that the agents are coupled by their joint task only.

The optimal control problem is equivalent to a graphical model inference problem [3]. In large and sparsely coupled multi-agent systems the optimal control can be computed using the junction tree algorithm. Exact inference, however, will break down when the system is both large and densely coupled. Here we explore the use of graphical model approximate inference methods in optimal control of large stochastic multi-agent systems. We apply the mean field approximation to show that optimal control is possible with accuracy in systems where exact inference breaks down.

2 Stochastic Optimal Control of a Multi-Agent System

We consider n agents in a k-dimensional space $\mathbb{R}^k$, the state of each agent a is given by a vector x_a in this space, satisfying stochastic dynamics

$$dx_a(t) = b_a(x_a(t), t)dt + Bu_a(t)dt + \sigma dw(t), \tag{1}$$

where u_a is the control of agent a, b_a is an arbitrary function representing autonomous dynamics, w is a Wiener process, and B and σ are $k \times k$ matrices.

The agents have to reach a goal at the end time T, they will receive a reward $\phi(x(T))$ at the end time depending on their joint end state $x(T) = (x_1(T), \ldots, x_n(T))$, but to reach this goal they will have to make an effort which depends on the agents controls and states over time. At any time $t < T$, the expected cost-to-go is

$$C(x, t, u(t \to T)) = \left\langle \phi(x(T)) + \int_t^T d\theta\, V(x(\theta), \theta) + \sum_{a=1}^{n} \int_t^T d\theta\, \frac{1}{2} u_a(\theta)^\top R u_a(\theta) \right\rangle, \tag{2}$$

given the agents initial state x, and the joint control over time $u(t \to T)$. R is a symmetric $k \times k$ matrix with positive eigenvalues, such that $u_a(\theta)^\top R u_a(\theta)$ is always a non-negative number, $V(x(\theta), \theta)$ is the cost for the agents to be in a joint state $x(\theta)$ at time θ. The issue is to find the optimal control which minimizes the expected cost-to-go.

The optimal controls are given by the gradient

$$u_a(x, t) = -R^{-1} B^\top \partial_{x_a} J(x, t), \tag{3}$$

where $J(x,t)$ the optimal expected cost-to-go, i.e. the cost (2) minimized over all possible controls; a brief derivation is contained in the appendix. An important implication of equation (3) is that at any moment in time, each agent can compute its own optimal control if it knows its own state and that of the other agents: there is no need to discuss possible strategies! This is because the agents always perform the control that is optimal, and the optimal control is unique.

To compute the optimal controls, however, we first need to find the optimal expected cost-to-go J. The latter may be expressed in terms of a forward diffusion process:

$$J(x,t) = -\lambda \log \int dy\, \rho(y,T|x,t) e^{-\phi(y)/\lambda}, \tag{4}$$

$\rho(y,T|x,t)$ being the transition probability for the system to go from a state x at time t to a state y at the end time T. The constant λ is determined by the relation $\sigma\sigma^\top = \lambda B R^{-1} B^\top$, equation (14) in the appendix. The density $\rho(y,\theta|x,t)$, $t < \theta \leq T$, satisfies the forward Fokker-Planck equation,

$$\partial_\theta \rho = -\frac{V}{\lambda} - \sum_{a=1}^n \partial_{y_a}^\top b_a \rho + \sum_{a=1}^n \frac{1}{2} \mathrm{Tr}\left(\sigma\sigma^\top \partial_{y_a}^2 \rho\right). \tag{5}$$

The solution to this equation may generally be estimated using path integral methods [4, 5], in a few special cases a solution exists in closed form:

Example 1. Consider a multi-agent system in one dimension in which there is noise and control in the velocities of the agents, according to the set of equations

$$\begin{cases} dx_a(t) = \dot{x}_a(t) dt \\ d\dot{x}_a(t) = u_a(t) dt + \sigma dw(t). \end{cases}$$

Note that this set of equations can be merged into a single equation of the form (1) by a concatenation of x_a and $\dot{x}_a$ into a single vector. We choose the potential $V = 0$. Under the task where each agent a has to reach a target with location μ_a at the end time T, and arrive with speed $\dot{\mu}_a$, the end cost function ϕ can be given in terms of a product of delta functions, that is

$$e^{-\phi(x,\dot{x})/\lambda} = \prod_{a=1}^n \delta(x_a - \mu_a)\delta(\dot{x}_a - \dot{\mu}_a),$$

and the system decouples into n independent single-agent systems. The dynamics of each agent a is given by a transition probability

$$\rho_a(y_a, \dot{y}_a, T | x_a, \dot{x}_a, t) = \frac{1}{\sqrt{\det(2\pi c)}} \exp\left(-\frac{1}{2}\left\| c^{-1/2} \begin{pmatrix} y_a - x_a - (T-t)\dot{x}_a \\ \dot{y}_a - \dot{x}_a \end{pmatrix} \right\|^2\right), \tag{6}$$

where

$$c = \frac{1}{6} \begin{pmatrix} 2(T-t)^3 & 3(T-t)^2 \\ 3(T-t)^2 & 6(T-t) \end{pmatrix} \sigma^2.$$

The optimal control follows from equations (3) and (4) and reads

$$u_a(x_a, \dot{x}_a, t) = \frac{6(\mu_a - x_a - (T-t)\dot{x}_a) - 2(T-t)(\dot{\mu}_a - \dot{x}_a)}{(T-t)^2}. \tag{7}$$

The first term in the control will steer the agent towards the target μ_a in a straight line, but since this may happen with a speed that differs from $\dot{\mu}_a$ with which the agent should arrive, there is a second term that initially 'exaggerates' the speed for going in a straight line, so that in the end there is time to adjust the speed to the end speed $\dot{\mu}_a$.

2.1 A Joint Task: Distribution over Targets

We consider the situation where agents have to distribute themselves over a number of targets $s = 1, \ldots, m$. In general, there will be m^n possible combinations of assigning the n agents to the targets—note, in example 1 we considered only one assignment. We can describe this by letting the end cost function ϕ be given in terms of a positive linear combination of functions

$$\Phi(y_1, \ldots, y_n, s_1, \ldots, s_n) = \prod_{a=1}^{n} \Phi_a(y_a, s_a)$$

that are peaked around the location $(\mu_{s_1}, \ldots, \mu_{s_n})$ of a joint target $(s_1, \ldots, s_n)$, that is

$$e^{-\phi(y)/\lambda} = \sum_{s_1, \ldots, s_n} w(s_1, \ldots, s_n) \prod_{a=1}^{n} \Phi_a(y_a, s_a),$$

where the $w(s_1, \ldots, s_n)$ are positive weights. We will refer to these weights as coupling factors, since they introduce dependencies between the agents. The optimal control of a single agent is obtained using equations (3) and (4), and is a weighted combination of single-target controls,

$$u_a = \sum_{s=1}^{m} p_a(s) u_a(s) \tag{8}$$

(the explicit (x, t) dependence has been dropped in the notation). Here $u_a(s)$ is the control for agent a to go to target s,

$$u_a(s) = -R^{-1} B^{\top} \partial_{x_a} Z_a(s), \tag{9}$$

with $Z_a(s)$ defined by

$$Z_a(s_a) = \int dy_a \rho_a(y_a, T | x_a, t) \Phi_a(y_a, s_a).$$

The weights $p_a(s)$ are marginals of the joint distribution

$$p(s_1, \ldots, s_n) \propto w(s_1, \ldots, s_n) \prod_{a=1}^{n} Z_a(s_a). \tag{10}$$

p thus is a distribution over all possible assignments of agents to targets.

Example 2. Consider the multi-agent system of example 1, but with a different task: each of the agents $a = 1, \ldots, n$ has to reach a target $s = 1, \ldots, n$ with location μ_s at the end time T, and arrive with zero speed, but no two agents are allowed to arrive at the same target. We model this by choosing an end cost function $\phi(x, \dot{x})$ given by

$$e^{-\phi(x,\dot{x})/\lambda} = \sum_{s_1,\ldots,s_n} w(s_1, \ldots, s_n) \prod_{a=1}^{n} \delta(y_a - \mu_a)\delta(\dot{y}_a)$$

with coupling factors

$$w(s_1, \ldots, s_n) = \prod_{a,a'=1}^{n} \exp\left(\frac{c}{\lambda n} \delta_{s_a, s_{a'}}\right).$$

For any agent a, the optimal control under this task is a weighted average of single target controls (7),

$$u_a(x_a, \dot{x}_a, t) = \frac{6(\langle \mu_a \rangle - x_a - (T-t)\dot{x}_a) + 2(T-t)\dot{x}_a}{(T-t)^2}, \tag{11}$$

where $\langle \mu_a \rangle$ the averaged target for agent a,

$$\langle \mu_a \rangle = \sum_{s=1}^{n} p_a(s)\mu_s.$$

The average is taken with respect to the marginal p_a of the joint distribution

$$p(s_1, \ldots, s_n) \propto w(s_1, \ldots, s_n) \prod_{a=1}^{n} \rho_a(\mu_{s_a}, 0, T | x_a, \dot{x}_a, t),$$

the densities ρ_a given by (6).

In general, and in example 2 in particular, the optimal control of an agent will not only depend on the state of this agent alone, but also on the states of other agents. Since the controls are computed anew at each instant in time, the agents are able to continuously adapt to the behaviour of the other agents, adjusting their control to the new states of all the agents.

2.2 Factored End Costs

The additional computational effort in multi-agent control compared to single-agent control lies in the computation of the marginals of the joint distribution p, which involves a sum of at most m^n terms. For small systems this is feasible, for large systems this will only be feasible if the summation can be performed efficiently. Whether an efficient way of computing the marginals exists, depends on the joint task of the agents. In the most complex case, to fulfil the task each agent will have to take the joint state of the entire system into account. In less complicated cases, an agent will only consider the states of a few agents in the system, in other words, the coupling factors will have a nontrivial factorized form:

$$w(s_1, \ldots, s_n) = \prod_A w_A(s_A),$$

where the A are subsets of agents. In such cases we may represent the couplings, and thus the joint distribution, by a factor graph; see Figure 1 for an example.

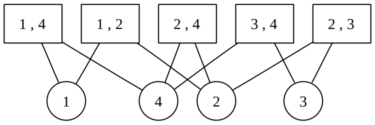


Fig. 1. Example of a factor graph for a multi-agent system of four agents. The couplings are represented by the factors A, with $A = \{1, 4\}, \{1, 2\}, \{2, 4\}, \{3, 4\}, \{2, 3\}$.

2.3 Graphical Model Inference

In the previous paragraph we observed that the joint distribution may be represented by a factor graph. This implies that the issue of assigning agents to targets is equivalent to a graphical model inference problem. Both exact methods (junction tree algorithm [6]) and approximate methods (mean field approximation [7], belief propagation [8]) can be used to compute the marginals in (8). In this paper we will use the mean field approximation to tackle optimal control in large multi-agent systems.

In the mean field approximation we minimize the mean field free energy, a function of single agent marginals q_a defined by

$$F_{\mathrm{MF}}(\{q_a\}) = -\langle \lambda \log w \rangle_q - \sum_a \langle \lambda \log Z_a \rangle_{q_a} - \lambda \sum_a H(q_a),$$

where $q(s) = q_1(s_1) \cdots q_n(s_n)$. Here the $H(q_a)$ are the entropies of the distributions q_a,

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

The minimum

$$J_{\mathrm{MF}} = \min_{\{q_a\}} F_{\mathrm{MF}}(\{q_a\})$$

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

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

with additional constraints for normalization of the probability vectors q_a. Solutions to this set of equations are implicitely given by the *mean field equations*

$$q_a(s_a) = \frac{Z_a(s_a)\exp\left(\langle \log w | s_a \rangle\right)}{\sum_{s'_a=1}^{n} Z_a(s'_a)\exp\left(\langle \log w | s'_a \rangle\right)} \tag{12}$$

where $\langle \log w | s_a \rangle$ the conditional expectation of $\log w$ given s_a,

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

The mean field equations are solved by means of iteration, and the solutions are the local minima of the mean field free energy. Thus the mean field free energy minimized over all solutions to the mean field equations equals the minimum J_{MF}.

The mean field approximation of the optimal control is found by taking the gradient of the minimum J_{MF} of the mean field free energy, similar to the exact case where the optimal control is the gradient of the optimal expected cost-to-go, equation (3):

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

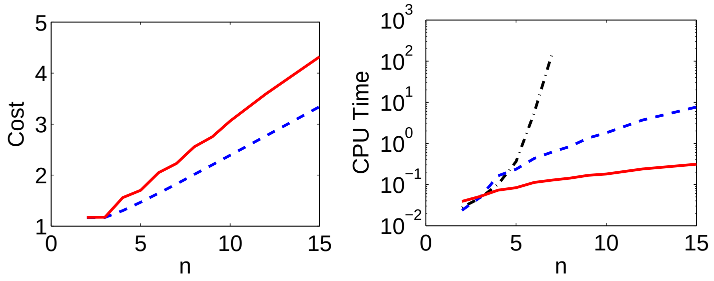
Similar to the exact case, it is an average of single-agent single-target optimal controls $u_a(x_a, t; s_a)$, the controls $u_a(x_a, t; s_a)$ given by equation (9), where the average is taken with respect to the mean field approximate marginal $q_a(s_a)$ of agent a.

3 Control of Large Multi-Agent Systems

Exact inference of multi-agent optimal control is intractable in large and densely coupled systems. In this section we present numerical results from approximate inference in optimal control of a large multi-agent system. We focus on the system presented in example 2. A group of n agents have to distribute themselves over an equal number of targets, each target should be reached by precisely one agent. The agents all start in the same location at $t = 0$, and the time they reach the targets lies at $T = 1$, as illustrated in figure 3. The variance of the noise equals 0.1 and the control cost parameter R equals 1, both are the same for each agent. The coupling strength c in the coupling factors equals -10. For implementation, time had to be discretized: each time step equaled 0.05 times the time-to-go $T - t$.

We considered two approximate inference methods for obtaining the marginals in (8), the mean field approximation described in section 2.3, and an approximation which at each moment in time assigns each agent to precisely one target. In the latter method the agent that is nearest to any of the targets is assigned first to its nearest target, then, removing this pair of agent and target, this is repeated for the remaining agents and targets, until there are no more remaining agents and targets. We will refer to this method as the *sort distances* (SD) method.

For several sizes of the system we computed the control cost and the required CPU time to calculate the controls. This we did under both control methods. Figures 2(a) and (b) show the control cost and the required CPU time as a function of the system size n; each value is an average obtained from 100 simulations. For comparison, in figure 2(b) we included the required CPU time under exact inference. Both under the SD method and the MF method the required CPU time appears to increase polynomially with n, the SD method requiring less computation time than the MF method. In contrast, the required CPU time under exact inference increases exponentially with n, as we may have expected. Though the SD method is faster than the MF method, it also is more costly: the control cost under the SD method is significantly higher than under the MF method. The MF method thus better approximates the optimal control.

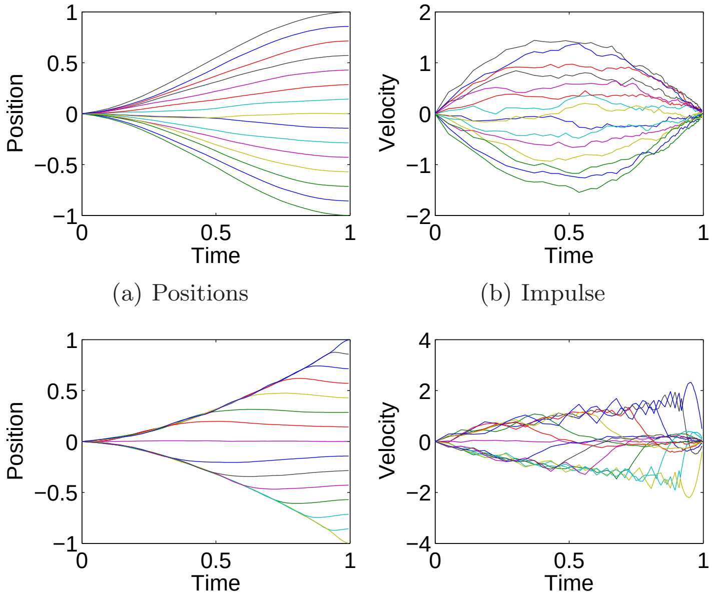


(a) Cost (b) CPU Time

Fig. 2. The control cost (a) and the required CPU Time (b) under the exact method $(\cdot - \cdot)$, the MF method $(--)$, and the SD method (—).

Figure 3 shows the positions and the velocities of the agents over time, both under the control obtained using the MF approximation and under the control obtained with the SD method. We observe that under MF control, the agents determine their targets early, between $t = 0$ and $t = 0.5$, and the agents velocities gradually increase from zero to a maximum value at $t = 0.5$ to again gradually decrease to zero, as required. This is not very surprising, since the MF approximation is known to show an early symmetry breaking. In contrast, under the SD method the decision making process of the agents choosing their targets takes place over almost the entire time interval, and the velocities of the agents are subject to frequent changes; in particular, as time increases the agents who have not yet chosen a target seem to exchange targets in a frequent manner. This may be understood by realising that under the SD method agents always perform a control to their nearest target only, instead of a weighted combination of controls to different targets which is the situation under MF control.

Further more, compared with the velocities under the MF method the velocities under the SD method take on higher maximum values. This may account for the relatively high control costs under SD control.



(a) Positions

(b) Impulse

(c) Positions

(d) Impulse

Fig. 3. A multi-agent system of 15 agents. The positions (a) and the velocities (b) over time under MF control, and the positions (c) and the velocities (d) over time under SD control.

4 Discussion

In this paper we studied optimal control in large stochastic multi-agent systems in continuous space and time, focussing on systems where agents have a task to distribute themselves over a number of targets. We followed the approach of Wiegerinck et al. [3]: we modeled the system in continuous space and time, resulting in an adaptive control policy where agents continuously adjust their controls to the environment. We considered the task of assigning agents to targets as a graphical model inference problem. We showed that in large and densely coupled systems, in which exact inference would break down, the mean field approximation manages to compute accurate approximations of the optimal controls of the agents.

We considered the performances of the mean field approximation and an alternative method, referred to as the sort distances method, on an example system in which a number of agents have to distribute themselves over an equal number of targets, such that each target is reached by precisely one agent. In the sort distances method each agent performs a control to a single nearby target, in such a way that no two agents head to the same target at the same time. This method has an advantage of being fast, but it results in relatively high control costs. Because each agent performs a control to a single target, agents switch targets frequently during the control process. In the mean field approximation each agent performs a control which is a weighted sum of controls to single targets. This requires more computation time than the sort distances method, but involves significantly lower control costs and therefore is a better approximation to the optimal control.

An obvious choice for a graphical model inference method not considered in the present paper would be belief propagation. Results of numeric simulations with this method in the context of multi-agent control, and comparisons with the mean field approximation and the exact junction tree algorithm will be published elsewhere.

There are many possible model extensions worthwhile exploring in future research. Examples are non-zero potentials V in case of a non-empty environment, penalties for collisions in the context of robotics, non-fixed end times, or bounded state spaces in the context of a production process. Typically, such model extensions will not allow for a solution in closed form, and approximate numerical methods will be required. Some suggestions are given by Kappen [4, 5]. In the setting that we considered the model which describes the behaviour of the agents was given. It would be worthwhile, however, to consider cases of stochastic optimal control of multi-agent systems in continuous space and time where the model first needs to be learned.

Acknowledgments

We thank Joris Mooij for making available useful software. This research is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant BSIK03024.

References

1. Guestrin, C., Koller, D., Parr, R.: Multiagent planning with factored MDPs. In: Proceedings of NIPS. Volume 14. (2002) 1523–1530
2. Guestrin, C., Venkataraman, S., Koller, D.: Context-specific multiagent coordination and planning with factored MDPs. In: Proceedings of AAAI. Volume 18. (2002) 253–259
3. Wiegerinck, W., van den Broek, B., Kappen, B.: Stochastic optimal control in continuous space-time multi-agent systems. In: UAI'06. (2006)
4. Kappen, H.J.: Path integrals and symmetry breaking for optimal control theory. Journal of statistical mechanics: theory and experiment (2005) P11011
5. Kappen, H.J.: Linear theory for control of nonlinear stochastic systems. Physical Review Letters **95**(20) (2005) 200201

6. Lauritzen, S., Spiegelhalter, D.: Local computations with probabilities on graphical structures and their application to expert systems (with discussion). J. Royal Statistical Society Series B **50** (1988) 157–224
7. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. In M.I., J., ed.: Learning in Graphical Models. MIT Press (1999)
8. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Trans. Info. Theory **47** (2001) 498–519

A Stochastic Optimal Control

In this appendix we give a brief derivation of equations (3), (4) and (5), starting from (2). Details can be found in [4, 5].

The optimal expected cost-to-go J, by definition the expected cost-to-go (2) minimized over all controls, satisfies the stochastic Hamilton-Jacobi-Bellman (HJB) equation

$$-\partial_t J = \min_u \sum_{a=1}^{n} \left(\frac{1}{2} u_a^\top R u_a + (b_a + B u_a)^\top \partial_{x_a} J + \frac{1}{2} \mathrm{Tr}\left(\sigma \sigma^\top \partial^2_{x_a} J \right) \right) + V,$$

with boundary condition $J(x, T) = \phi(x)$. The minimization with respect to u yields equation (3), which specifies the optimal control for each agent. Substituting these controls in the HJB equation gives a non-linear equation for J. We can remove the non-linearity by using a log transformation: if we introduce a constant λ, and define $Z(x, t)$ through

$$J(x, t) = -\lambda \log Z(x, t), \tag{13}$$

then

$$\begin{aligned} \frac{1}{2} u_a^\top R u_a + (B u_a)^\top \partial_{x_a} J &= -\frac{1}{2} \lambda^2 Z^{-2} (\partial_{x_a} Z)^\top B R^{-1} B^\top \partial_{x_a} Z, \\ \frac{1}{2} \mathrm{Tr}\left(\sigma \sigma^\top \partial^2_{x_a} J \right) &= \frac{1}{2} \lambda Z^{-2} (\partial_{x_a} Z)^\top \sigma \sigma^\top \partial_{x_a} Z - \frac{1}{2} \lambda Z^{-1} \mathrm{Tr}\left(\sigma \sigma^\top \partial^2_{x_a} Z \right). \end{aligned}$$

The terms quadratic in Z vanish when $\sigma^\top \sigma$ and R are related via

$$\sigma \sigma^\top = \lambda B R^{-1} B^\top. \tag{14}$$

In the one dimensional case a constant λ can always be found such that equation (14) is satisfied, in the higher dimensional case the equation puts restrictions on the matrices σ and R, because in general $\sigma \sigma^\top$ and $B R^{-1} B^\top$ will not be proportional.

When equation (14) is satisfied, the HJB equation becomes

$$\begin{aligned} \partial_t Z &= \left(\frac{V}{\lambda} - \sum_{a=1}^{n} b_a^\top \partial_{x_a} - \sum_{a=1}^{n} \frac{1}{2} \mathrm{Tr}\big(\sigma \sigma^\top \partial^2_{x_a} \big) \right) Z \\ &= -H Z, \end{aligned} \tag{15}$$

where H a linear operator acting on the function Z. Equation (15) is solved backwards in time with $Z(x,T) = e^{-\phi(x)/\lambda}$. However, the linearity allows us to reverse the direction of computation, replacing it by a diffusion process, as we will now explain.

The solution to equation (15) is given by

$$Z(x,t) = \int dy \rho(y,T|x,t) e^{-\phi(y)/\lambda}, \tag{16}$$

the density $\rho(y,\vartheta|x,t)$ $(t < \vartheta \leq T)$ satisfying the forward Fokker-Planck equation (5). Combining the equations (13) and (16) yields the expression (4) for the optimal expected cost-to-go.