Chapter 1

Optimal control theory and the linear Bellman Equation

Hilbert J. Kappen[1]

1.1 Introduction

Optimizing a sequence of actions to attain some future goal is the general topic of control theory Stengel (1993); Fleming and Soner (1992). It views an agent as an automaton that seeks to maximize expected reward (or minimize cost) over some future time period. Two typical examples that illustrate this are motor control and foraging for food. As an example of a motor control task, consider a human throwing a spear to kill an animal. Throwing a spear requires the execution of a motor program that is such that at the moment that the spear releases the hand, it has the correct speed and direction such that it will hit the desired target. A motor program is a sequence of actions, and this sequence can be assigned a cost that consists generally of two terms: a path cost, that specifies the energy consumption to contract the muscles in order to execute the motor program; and an end cost, that specifies whether the spear will kill the animal, just hurt it, or misses it altogether. The optimal control solution is a sequence of motor commands that results in killing the animal by throwing the spear with minimal physical effort. If x denotes the state space (the positions and velocities of the muscles), the optimal control solution is a function $u(x,t)$ that depends both on the actual state of the system at each time and also depends explicitly on time.

When an animal forages for food, it explores the environment with the objective to find as much food as possible in a short time window. At each time t, the animal considers the food it expects to encounter in the period $[t, t+T]$. Unlike the motor control example, the time horizon recedes into the future with the current time and the cost consists now only of a path contribution and no end-cost. Therefore, at each time the animal faces the same task, but possibly from a different location of the animal in the environment. The optimal control solution $u(x)$ is now time-independent and specifies for each location in the environment x the direction u in which the animal should move.

The general stochastic control problem is intractable to solve and requires an exponential amount of memory and computation time. The reason is that the state space needs to be discretized and thus becomes exponentially large in the number of dimensions. Computing the expectation values means that all states need to be visited and requires the summation of exponentially large sums. The same intractabilities are encountered in reinforcement learning.

[1]Donders' Institute for Neuroscience, Radboud University, 6525 EZ Nijmegen, the Netherlands

In this tutorial, we aim to give a pedagogical introduction to control theory. For simplicity, we will first consider in section 1.2 the case of discrete time and discuss the dynamic programming solution. This gives us the basic intuition about the Bellman equations in continuous time that are considered later on. In section 1.3 we consider continuous time control problems. In this case, the optimal control problem can be solved in two ways: using the Hamilton-Jacobi-Bellman (HJB) equation which is a partial differential equation Bellman and Kalaba (1964) and is the continuous time analogue of the dynamic programming method, or using the Pontryagin Minimum Principle (PMP) Pontryagin et al. (1962) which is a variational argument and results in a pair of *ordinary* differential equations. We illustrate the methods with the example of a mass on a spring.

In section 1.4 we generalize the control formulation to the stochastic dynamics. In the presence of noise, the PMP formalism has no obvious generalization (see however Yong and Zhou (1999)). In contrast, the inclusion of noise in the HJB framework is mathematically quite straight-forward. However, the numerical solution of either the deterministic or stochastic HJB equation is in general difficult due to the curse of dimensionality.

There are some stochastic control problems that can be solved efficiently. When the system dynamics is linear and the cost is quadratic (LQ control), the solution is given in terms of a number of coupled ordinary differential (Ricatti) equations that can be solved efficiently Stengel (1993). LQ control is useful to maintain a system such as for instance a chemical plant, operated around a desired point in state space and is therefore widely applied in engineering. However, it is a linear theory and too restricted to model the complexities of intelligent behavior encountered in agents or robots.

The simplest control formulation assumes that all model components (the dynamics, the environment, the costs) are known and that the state is fully observed. Often, this is not the case. Formulated in a Bayesian way, one may only know a probability distribution of the current state, or over the parameters of the dynamics or the costs. This leads us to the problem of partial observability or the problem of joint inference and control. We discuss two different approaches to learning: adaptive control and dual control. Whereas in the adaptive control approach the learning dynamics is exterior to the control problem, in the dual control approach it is recognized that learning and control are interrelated and the optimal solution for combined learning and control problem is computed. We illustrate the complexity of joint inference and control with a simple example. We discuss the concept of certainty equivalence, which states that for certain linear quadratic control problems the inference and control problems disentangle and can be solved separately. We will discuss these issues in section 1.5.

Recently, we have discovered a class of continuous non-linear stochastic control problems that can be solved more efficiently than the general case Kappen (2005a,b). These are control problems with a finite time horizon, where the control acts additive on the dynamics and is in some sense proportional to the noise. The cost of the control is quadratic in the control variables. Otherwise, the path cost and end cost and the intrinsic dynamics of the system are arbitrary. These control problems can have both time-dependent and time-independent solutions of the type that we encountered in the examples above. For these problems, the Bellman equation becomes a linear equation in the exponentiated cost-to-go (value function). The solution is formally written as a path integral. We discuss the path integral control method in section 1.6.

The path integral can be interpreted as a free energy, or as the normalization

of a probabilistic time-series model (Kalman filter, Hidden Markov Model). One can therefore consider various well-known methods from the machine learning community to approximate this path integral, such as the Laplace approximation and Monte Carlo sampling Kappen (2005b), variational approximations or belief propagation Broek et al. (2008a). In section 1.7.2 we show an example of an n joint arm where we compute the optimal control using the variational approximation for large n.

Non-linear stochastic control problems display features not shared by deterministic control problems nor by linear stochastic control. In deterministic control, only the globally optimal solution is relevant. In stochastic control, the optimal solution is typically a weighted mixture of suboptimal solutions. The weighting depends in a non-trivial way on the features of the problem, such as the noise and the horizon time and on the cost of each solution. This multi-modality leads to surprising behavior in stochastic optimal control. For instance, the optimal control can be qualitatively very different for high and low noise levels Kappen (2005a), where it was shown that in a stochastic environment, the optimal timing of the choice to move to one of two targets should be delayed in time. The decision is formally accomplished by a dynamical symmetry breaking of the cost-to-go function.

1.2 Discrete time control

We start by discussing the most simple control case, which is the finite horizon discrete time deterministic control problem. In this case the optimal control explicitly depends on time. See also Weber (2006) for further discussion.

Consider the control of a discrete time dynamical system:

$$x_{t+1} = x_t + f(t, x_t, u_t), \quad t = 0, 1, \ldots, T-1 \tag{1.1}$$

x_t is an n-dimensional vector describing the *state* of the system and u_t is an m-dimensional vector that specifies the *control* or *action* at time t. Note, that Eq. 1.1 describes a noiseless dynamics. If we specify x at $t = 0$ as x_0 and we specify a sequence of controls $u_{0:T-1} = u_0, u_1, \ldots, u_{T-1}$, we can compute future states of the system $x_{1:T}$ recursively from Eq.1.1.

Define a cost function that assigns a cost to each sequence of controls:

$$C(x_0, u_{0:T-1}) = \phi(x_T) + \sum_{t=0}^{T-1} R(t, x_t, u_t) \tag{1.2}$$

$R(t, x, u)$ is the cost that is associated with taking action u at time t in state x, and $\phi(x_T)$ is the cost associated with ending up in state x_T at time T. The problem of optimal control is to find the sequence $u_{0:T-1}$ that minimizes $C(x_0, u_{0:T-1})$.

The problem has a standard solution, which is known as dynamic programming. Introduce the *optimal cost to go*:

$$J(t, x_t) = \min_{u_{t:T-1}} \left(\phi(x_T) + \sum_{s=t}^{T-1} R(s, x_s, u_s) \right) \tag{1.3}$$

which solves the optimal control problem from an intermediate time t until the fixed end time T, starting at an arbitrary location x_t. The minimum of Eq. 1.2 is given by $J(0, x_0)$.

One can recursively compute $J(t,x)$ from $J(t+1,x)$ for all x in the following way:

$$\begin{aligned} J(T,x) &= \phi(x) \\ J(t,x_t) &= \min_{u_{t:T-1}} \left(\phi(x_T) + \sum_{s=t}^{T-1} R(s,x_s,u_s) \right) \\ &= \min_{u_t} \left(R(t,x_t,u_t) + \min_{u_{t+1:T-1}} \left[\phi(x_T) + \sum_{s=t+1}^{T-1} R(s,x_s,u_s) \right] \right) \\ &= \min_{u_t} \left(R(t,x_t,u_t) + J(t+1,x_{t+1}) \right) \\ &= \min_{u_t} \left(R(t,x_t,u_t) + J(t+1,x_t + f(t,x_t,u_t)) \right) \end{aligned} \tag{1.4}$$

Note, that the minimization over the whole path $u_{0:T-1}$ has reduced to a sequence of minimizations over u_t. This simplification is due to the Markovian nature of the problem: the future depends on the past and vise versa only through the present. Also note, that in the last line the minimization is done for each x_t separately.

The algorithm to compute the optimal control $u^*_{0:T-1}$, the optimal trajectory $x^*_{1:T}$ and the optimal cost is given by

1. Initialization: $J(T,x) = \phi(x)$
2. Backwards: For $t = T-1, \ldots, 0$ and for all x compute

$$\begin{aligned} u^*_t(x) &= \arg\min_u \{ R(t,x,u) + J(t+1, x + f(t,x,u)) \} \\ J(t,x) &= R(t,x,u^*_t) + J(t+1, x + f(t,x,u^*_t)) \end{aligned}$$

3. Forwards: For $t = 0, \ldots, T-1$ compute

$$x^*_{t+1} = x^*_t + f(t, x^*_t, u^*_t(x^*_t))$$

The execution of the dynamic programming algorithm is linear in the horizon time T and linear in the size of the state and action spaces.

1.3 Continuous time control

In the absence of noise, the optimal control problem in continuous time can be solved in two ways: using the Pontryagin Minimum Principle (PMP) Pontryagin et al. (1962) which is a pair of *ordinary* differential equations or the Hamilton-Jacobi-Bellman (HJB) equation which is a *partial* differential equation Bellman and Kalaba (1964). The latter is very similar to the dynamic programming approach that we have treated above. The HJB approach also allows for a straightforward extension to the noisy case. We will first treat the HJB description and subsequently the PMP description. For further reading see Stengel (1993); Jönsson et al. (2002).

1.3.1 The HJB equation

Consider the dynamical system Eq. 1.1 where we take the time increments to zero, ie. we replace $t+1$ by $t+dt$ with $dt \to 0$:

$$x_{t+dt} = x_t + f(x_t, u_t, t)dt \tag{1.5}$$

In the continuous limit we will write $x_t = x(t)$. The initial state is fixed: $x(0) = x_0$ and the final state is free. The problem is to find a control signal $u(t), 0 < t < T$, which we denote as $u(0 \to T)$, such that

$$C(x_0, u(0 \to T)) = \phi(x_T)) + \int_0^T d\tau R(x(\tau), u(\tau), \tau) \tag{1.6}$$

is minimal. C consists of an end cost $\phi(x)$ that gives the cost of ending in a configuration x, and a path cost that is an integral over time that depends on the trajectories $x(0 \to T)$ and $u(0 \to T)$.

Eq. 1.4 becomes

$$\begin{aligned} J(t,x) &= \min_u \left(R(t,x,u)dt + J(t+dt, x+f(x,u,t)dt)\right) \\ &\approx \min_u \left(R(t,x,u)dt + J(t,x) + \partial_t J(t,x)dt + \partial_x J(t,x) f(x,u,t)dt\right) \\ -\partial_t J(t,x) &= \min_u \left(R(t,x,u) + f(x,u,t)\partial_x J(x,t)\right) \end{aligned} \tag{1.7}$$

where in the second line we have used the Taylor expansion of $J(t+dt, x+dx)$ around x, t to first order in dt and dx and in the third line have taken the limit $dt \to 0$. We use the shorthand notation $\partial_x J = \frac{\partial J}{\partial x}$. Eq. 1.7 is a partial differential equation, known as the *Hamilton-Jacobi-Bellman (HJB) equation*, that describes the evolution of J as a function of x and t and must be solved with boundary condition $J(x,T) = \phi(x)$. ∂_t and ∂_x denote partial derivatives with respect to t and x, respectively.

The optimal control at the current x, t is given by

$$u(x,t) = \arg\min_u \left(R(x,u,t) + \partial_x J(t,x) f(x,u,t)\right) \tag{1.8}$$

Note, that in order to compute the optimal control at the current state $x(0)$ at $= 0$ one must compute $J(x,t)$ for all values of x and t.

1.3.2 Example: Mass on a spring

To illustrate the optimal control principle consider a mass on a spring. The spring is at rest at $z = 0$ and exerts a force proportional to $F = -z$ towards the rest position. Using Newton's Law $F = ma$ with $a = \ddot{z}$ the acceleration and $m = 1$ the mass of the spring, the equation of motion is given by.

$$\ddot{z} = -z + u$$

with u a unspecified control signal with $-1 < u < 1$. We want to solve the control problem: Given initial position and velocity z_i and $\dot{z}_i$ at time 0, find the control path $u(0 \to T)$ such that $z(T)$ is maximal.

Introduce $x_1 = z, x_2 = \dot{z}$, then

$$\dot{x} = Ax + Bu, \qquad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and $x = (x_1, x_2)^T$. The problem is of the above type, with $\phi(x) = C^T x$, $C^T = (-1, 0)$, $R(x,u,t) = 0$ and $f(x,u,t) = Ax + Bu$. Eq. 1.7 takes the form

$$-J_t = (\partial_x J)^T Ax - |(\partial_x J)^T B|$$

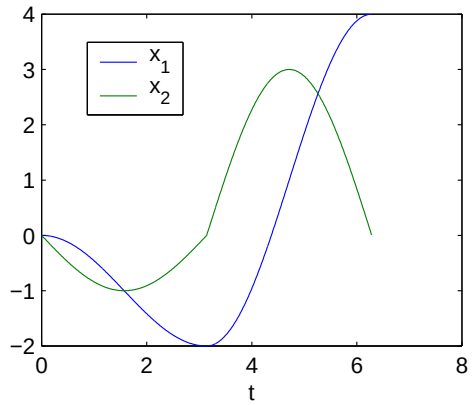


Figure 1.1: Optimal control of mass on a spring such that at $t = 2\pi$ the amplitude is maximal. x_1 is position of the spring, x_2 is velocity of the spring.

We try $J(t,x) = \psi(t)^T x + \alpha(t)$. The HJBE reduces to two ordinary differential equations

$$\begin{aligned} \dot{\psi} &= -A^T\psi \\ \dot{\alpha} &= |\psi^T B| \end{aligned}$$

These equations must be solved for all t, with final boundary conditions $\psi(T) = C$ and $\alpha(T) = 0$. Note, that the optimal control in Eq. 1.8 only requires $\partial_x J(x,t)$, which in this case is $\psi(t)$ and thus we do not need to solve α. The solution for ψ is

$$\begin{aligned} \psi_1(t) &= -\cos(t-T) \\ \psi_2(t) &= \sin(t-T) \end{aligned}$$

for $0 < t < T$. The optimal control is

$$u(x,t) = -\text{sign}(\psi_2(t)) = -\text{sign}(\sin(t-T))$$

As an example consider $x_1(0) = x_2(0) = 0$, $T = 2\pi$. Then, the optimal control is

$$\begin{aligned} u &= -1, \quad 0 < t < \pi \\ u &= 1, \quad \pi < t < 2\pi \end{aligned}$$

The optimal trajectories are for $0 < t < \pi$

$$x_1(t) = \cos(t) - 1, \quad x_2(t) = -\sin(t)$$

and for $\pi < t < 2\pi$

$$x_1(t) = 3\cos(t) + 1, \quad x_2(t) = -3\sin(t)$$

The solution is drawn in fig. 1.1. We see that in order to excite the spring to its maximal height at T, the optimal control is to first push the spring down for $0 < t < \pi$ and then to push the spring up between $\pi < t < 2\pi$, taking maximally advantage of the intrinsic dynamics of the spring.

Note, that since there is no cost associated with the control u and u is hard limited between -1 and 1, the optimal control is always either -1 or 1. This is known as bang-bang control.

1.3.3 Pontryagin minimum principle

In the last section, we solved the optimal control problem as a partial differential equation, with a boundary condition at the end time. The numerical solution requires a discretization of space and time and is computationally expensive. The solution is an optimal cost-to-go function $J(x,t)$ for all x and t. From this we compute the optimal control sequence Eq. 1.8 and the optimal trajectory.

An alternative to the HJB approach is a variational approach that directly finds the optimal trajectory and optimal control and bypasses the expensive computation of the cost-to-go. This approach is known as the Pontryagin Minimum Principle. We can write the optimal control problem as a constrained optimization problem with independent variables $u(0 \rightarrow T)$ and $x(0 \rightarrow T)$. We wish to minimize

$$\min_{u(0\rightarrow T), x(0\rightarrow T)} \phi(x(T)) + \int_0^T dt R(x(t), u(t), t)$$

subject to the constraint that $u(0 \rightarrow T)$ and $x(0 \rightarrow T)$ are compatible with the dynamics

$$\dot{x} = f(x, u, t) \tag{1.9}$$

and the boundary condition $x(0) = x_0$. $\dot{x}$ denotes the time derivative dx/dt.

We can solve the constraint optimization problem by introducing the Lagrange multiplier function $\lambda(t)$ that ensures the constraint Eq. 1.9 for all t:

$$\begin{aligned}
\mathcal{C} &= \phi(x(T)) + \int_0^T dt \left[R(t, x(t), u(t)) - \lambda(t)(f(t, x(t), u(t)) - \dot{x}(t))\right] \\
&= \phi(x(T)) + \int_0^T dt[-H(t, x(t), u(t), \lambda(t)) + \lambda(t)\dot{x}(t))] \\
-H(t, x, u, \lambda) &= R(t, x, u) - \lambda f(t, x, u)
\end{aligned} \tag{1.10}$$

The solution is found by extremizing $\mathcal{C}$. If we vary the action wrt to the trajectory x, the control u and the Lagrange multiplier λ, we get:

$$\begin{aligned}
\delta\mathcal{C} &= \phi_x(x(T))\delta x(T) \\
&+ \int_0^T dt[-H_x\delta x(t) - H_u\delta u(t) + (-H_\lambda + \dot{x}(t))\delta\lambda(t) + \lambda(t)\delta\dot{x}(t)] \\
&= (\phi_x(x(T)) + \lambda(T))\, \delta x(T) \\
&+ \int_0^T dt \left[(-H_x - \dot{\lambda}(t))\delta x(t) - H_u\delta u(t) + (-H_\lambda + \dot{x}(t))\delta\lambda(t)\right]
\end{aligned}$$

where the subscripts x, u, λ denote partial derivatives. For instance, $H_x = \frac{\partial H(t,x(t),u(t),\lambda(t))}{\partial x(t)}$. In the second line above we have used partial integration:

$$\int_0^T dt\lambda(t)\delta\dot{x}(t) = \int_0^T dt\lambda(t)\frac{d}{dt}\delta x(t) = -\int_0^T dt\frac{d}{dt}\lambda(t)\delta x(t) + \lambda(T)\delta x(T) - \lambda(0)\delta x(0)$$

and $\delta x(0) = 0$.

The stationary solution ($\delta\mathcal{C} = 0$) is obtained when the coefficients of the independent variations $\delta x(t), \delta u(t), \delta\lambda(t)$ and $\delta x(T)$ are zero. Thus,

$$\begin{aligned}
\dot{\lambda} &= -H_x(t, x(t), u(t), \lambda(t)) \\
0 &= H_u((t, x(t), u(t), \lambda(t)) \\
\dot{x} &= H_\lambda(t, x, u, \lambda) \\
\lambda(T) &= -\phi_x(x(T))
\end{aligned} \tag{1.11}$$

We can solve Eq. 1.11 for u and denote the solution as $u^*(t, x, \lambda)$. This solution is unique if H is convex in u. The remaining equations are

$$\begin{aligned} \dot{x} &= H^*_\lambda(t, x, \lambda) \\ \dot{\lambda} &= -H^*_x(t, x, \lambda) \end{aligned} \tag{1.12}$$

where we have defined $H^*(t, x, \lambda) = H(t, x, u^*(t, x, \lambda), \lambda)$ and with boundary conditions

$$x(0) = x_0 \qquad \lambda(T) = -\phi_x(x(T)) \tag{1.13}$$

The solution provided by Eqs. 1.12 with boundary conditions Eq. 1.13 are coupled ordinary differential equations that describe the dynamics of x and λ over time with a boundary condition for x at the initial time and for λ at the final time. Compared to the HJB equation, the complexity of solving these equations is low since only time discretization and no space discretization is required. However, due to the mixed boundary conditions, finding a solution that satisfies these equations is not trivial and requires sophisticated numerical methods. The most common method for solving the PMP equations is called (multiple) shooting Fraser-Andrews (1999); Heath (2002).

The Eqs. 1.12 are also known as the so-called Hamilton equations of motion that arise in classical mechanics, but then with initial conditions for both x and λ Goldstein (1980). In fact, one can view control theory as a generalization of classical mechanics.

In classical mechanics H is called the Hamiltonian. Consider the time evolution of H:

$$\dot{H} = H_t + H_u \dot{u} + H_x \dot{x} + H_\lambda \dot{\lambda} = H_t \tag{1.14}$$

where we have used the dynamical equations Eqs. 1.12 and Eq. 1.11. In particular, when f and R in Eq. 1.10 do not explicitly depend on time, neither does H and $H_t = 0$. In this case we see that H is a constant of motion: the control problem finds a solution such that $H(t = 0) = H(t = T)$.

1.3.4 Again mass on a spring

We consider again the example of the mass on a spring that we introduced in section 1.3.2 where we had

$$\begin{aligned} \dot{x_1} &= x_2, \qquad \dot{x_2} = -x_1 + u \\ R(x, u, t) &= 0 \qquad \phi(x) = -x_1 \end{aligned}$$

The Hamiltonian Eq. 1.10 becomes

$$H(t, x, u, \lambda) = \lambda_1 x_2 + \lambda_2(-x_1 + u)$$

Using Eq. 1.11 we obtain $u^* = -\text{sign}(\lambda_2)$ and

$$H^*(t, x, \lambda) = \lambda_1 x_2 - \lambda_2 x_1 - |\lambda_2|$$

The Hamilton equations

$$\begin{aligned} \dot{x} = \frac{\partial H^*}{\partial \lambda} \qquad &\Rightarrow \qquad \dot{x_1} = x_2, \qquad \dot{x_2} = -x_1 - \text{sign}(\lambda_2) \\ \dot{\lambda} = -\frac{\partial H^*}{\partial x} \qquad &\Rightarrow \qquad \dot{\lambda_1} = -\lambda_2, \qquad \dot{\lambda_2} = \lambda_1 \end{aligned}$$

with $x(t = 0) = x_0$ and $\lambda(t = T) = 1$.

1.3.5 Comments

The HJB method gives a sufficient (and often necessary) condition for optimality. The solution of the PDE is expensive. The PMP method provides a necessary condition for optimal control. This means that it provides candidate solutions for optimality.

The PMP method is computationally less complicated than the HJB method because it does not require discretization of the state space. The PMP method can be used when dynamic programming fails due to lack of smoothness of the optimal cost-to-go function.

The subject of optimal control theory in continuous space and time has been well studied in the mathematical literature and contains many complications related to the existence, uniqueness and smoothness of the solution, particular in the absence of noise. See Jönsson et al. (2002) for a clear discussion and further references. In the presence of noise and in particular in the path integral framework, as we will discuss below, it seems that many of these intricacies disappear.

1.4 Stochastic optimal control

In this section, we consider the extension of the continuous control problem to the case that the dynamics is subject to noise and is given by a stochastic differential equation. First, we give a very brief introduction to stochastic differential equations.

1.4.1 Stochastic differential equations

Consider the random walk on the line:

$$x_{t+1} = x_t + \xi_t \qquad \xi_t = \pm\sqrt{\nu}$$

with $x_0 = 0$. The increments ξ_t are iid random variables with mean zero, $\langle \xi_t^2 \rangle = \nu$ and ν is a constant. We can write the solution for x_t in closed form as

$$x_t = \sum_{i=1}^{t} \xi_i$$

Since x_t is a sum of random variables, x_t becomes Gaussian in the limit of large t. We can compute the evolution of the mean and covariance:

$$\begin{aligned} \langle x_t \rangle &= \sum_{i=1}^{t} \langle \xi_i \rangle = 0 \\ \langle x_t^2 \rangle &= \sum_{i,j=1}^{t} \langle \xi_i \xi_j \rangle = \sum_{i=1}^{t} \langle \xi_i^2 \rangle + \sum_{i,j=1, j\neq i}^{t} \langle \xi_i \rangle \langle \xi_j \rangle = \nu t \end{aligned}$$

Note, that the fluctuations $\sigma_t = \sqrt{\langle x_t^2 \rangle} = \sqrt{\nu t}$ increase with the square root of t. This is a characteristic property of a diffusion process, such as for instance the diffusion of ink in water or warm air in a room.

In the continuous time limit we define

$$dx_t = x_{t+dt} - x_t = d\xi \tag{1.15}$$

with $d\xi$ an infinitesimal mean zero Gaussian variable. In order to get the right scaling with t we must choose $\left\langle d\xi^2 \right\rangle = \nu dt$. Then in the limit of $dt \to 0$ we obtain

$$\begin{aligned}
\frac{d}{dt}\left\langle x \right\rangle &= \lim_{dt\to 0}\left\langle \frac{x_{t+dt}-x_t}{dt} \right\rangle = \lim_{dt\to 0}\frac{\left\langle d\xi \right\rangle}{dt} = 0, \quad \Rightarrow \quad \left\langle x \right\rangle(t) = 0 \\
\frac{d}{dt}\left\langle x^2 \right\rangle &= \nu, \quad \Rightarrow \quad \left\langle x^2 \right\rangle(t) = \nu t
\end{aligned}$$

The conditional probability distribution of x at time t given x_0 at time 0 is Gaussian and specified by its mean and variance. Thus

$$\rho(x,t|x_0,0) = \frac{1}{\sqrt{2\pi\nu t}}\exp\left(-\frac{(x-x_0)^2}{2\nu t}\right)$$

The process Eq. 1.15 is called a Wiener process.

1.4.2 Stochastic optimal control theory

Consider the stochastic differential equation which is a generalization of Eq. 1.5:

$$dx = f(x(t),u(t),t)dt + d\xi. \tag{1.16}$$

$d\xi$ is a Wiener processes with $\langle d\xi_i d\xi_j \rangle = \nu_{ij}(t,x,u)dt$ and ν is a symmetric positive definite matrix.

Because the dynamics is stochastic, it is no longer the case that when x at time t and the full control path $u(t \to T)$ are given, we know the future path $x(t \to T)$. Therefore, we cannot minimize Eq. 1.6, but can only hope to be able to minimize its expectation value over all possible future realizations of the Wiener process:

$$C(x_0, u(0\to T)) = \left\langle \phi(x(T)) + \int_0^T dt R(x(t),u(t),t) \right\rangle_{x_0} \tag{1.17}$$

The subscript x_0 on the expectation value is to remind us that the expectation is over all stochastic trajectories that start in x_0.

The solution of the control problem proceeds very similar as in the deterministic case, with the only difference that we must add the expectation value over trajectories. Eq. 1.4 becomes

$$J(t,x_t) = \min_{u_t} R(t,x_t,u_t)dt + \langle J(t+dt, x_{t+dt})\rangle_{x_t}$$

We must again make a Taylor expansion of J in dt and dx. However, since $\left\langle dx^2 \right\rangle$ is of order dt because of the Wiener process, we must expand up to order dx^2:

$$\begin{aligned}
\langle J(t+dt,x_{t+dt})\rangle &= \int dx_{t+dt}\mathcal{N}(x_{t+dt}|x_t,\nu dt)J(t+dt,x_{t+dt}) \\
&= J(t,x_t) + dt\partial_t J(t,x_t) + \langle dx\rangle \partial_x J(t,x_t) + \frac{1}{2}\left\langle dx^2 \right\rangle \partial_x^2 J(t,x_t) \\
\langle dx\rangle &= f(x,u,t)dt \\
\left\langle dx^2 \right\rangle &= \nu(t,x,u)dt
\end{aligned}$$

Thus, we obtain

$$-\partial_t J(t,x) = \min_u\left(R(t,x,u) + f(x,u,t)\partial_x J(x,t) + \frac{1}{2}\nu(t,x,u)\partial_x^2 J(x,t)\right) \tag{1.18}$$

which is the *Stochastic Hamilton-Jacobi-Bellman Equation* with boundary condition $J(x,T) = \phi(x)$. Eq. 1.18 reduces to the deterministic HJB equation Eq.1.7 in the limit $\nu \to 0$.

1.4.3 Linear quadratic control

In the case that the dynamics is linear and the cost is quadratic one can show that the optimal cost to go J is also a quadratic form and one can solve the stochastic HJB equation in terms of 'sufficient statistics' that describe J.

x is n-dimensional and u is p dimensional. The dynamics is linear

$$dx = [A(t)x + B(t)u + b(t)]dt + \sum_{j=1}^{m}(C_j(t)x + D_j(t)u + \sigma_j(t))d\xi_j \tag{1.19}$$

with dimensions: $A = n \times n, B = n \times p, b = n \times 1, C_j = n \times n, D_j = n \times p, \sigma_j = n \times 1$ and $\langle d\xi_j d\xi_{j'} \rangle = \delta_{jj'} dt$. The cost function is quadratic

$$\phi(x) = \frac{1}{2}x^T G x \tag{1.20}$$

$$f_0(x,u,t) = \frac{1}{2}x^T Q(t)x + u^T S(t)x + \frac{1}{2}u^T R(t)u \tag{1.21}$$

with $G = n \times n, Q = n \times n, S = p \times n, R = p \times p$.

We parametrize the optimal cost to go function as

$$J(t,x) = \frac{1}{2}x^T P(t)x + \alpha^T(t)x + \beta(t) \tag{1.22}$$

which should satisfy the stochastic HJB equation eq. 1.18 with $P(T) = G$ and $\alpha(T) = \beta(T) = 0$. $P(t)$ is an $n \times n$ matrix, $\alpha(t)$ is an n-dimensional vector and $\beta(t)$ is a scalar. Substituting this form of J in Eq. 1.18, this equation contains terms quadratic, linear and constant in x and u. We can thus do the minimization with respect to u exactly and the result is

$$u(t) = -\Psi(t)x(t) - \psi(t)$$

with

$$\begin{aligned}
\hat{R} &= R + \sum_{j=1}^{m} D_j^T P D_j, \quad (p \times p) \\
\hat{S} &= B^T P + S + \sum_{j=1}^{m} D_j^T P C_j, \quad (p \times n) \\
\Psi &= \hat{R}^{-1}\hat{S}, \quad (p \times n) \\
\psi &= \hat{R}^{-1}(B^T\alpha + \sum_{j=1}^{m} D_j^T P \sigma_j), \quad (p \times 1)
\end{aligned}$$

The stochastic HJB equation then decouples as three ordinary differential equations

$$-\dot{P} = PA + A^T P + \sum_{j=1}^{m} C_j^T P C_j + Q - \hat{S}^T \hat{R}^{-1}\hat{S} \tag{1.23}$$

$$-\dot{\alpha} = [A - B\hat{R}^{-1}\hat{S}]^T\alpha + \sum_{j=1}^{m}[C_j - D_j\hat{R}^{-1}\hat{S}]^T P\sigma_j + Pb \tag{1.24}$$

$$\dot{\beta} = \frac{1}{2}\left|\sqrt{\hat{R}}\psi\right|^2 - \alpha^T b - \frac{1}{2}\sum_{j=1}^{m}\sigma_j^T P\sigma_j \tag{1.25}$$

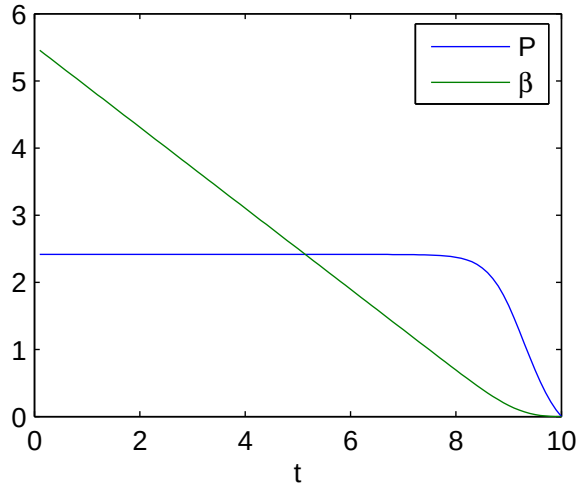


Figure 1.2: Stochastic optimal control in the case of a linear system with quadratic cost. $T = 10$, time discretization $dt = 0.1$, $\nu = 0.05$. The optimal control is to steer towards the origin with $-P(t)x$, where P is roughly constant until $T \approx 8$. Afterward control weakens because the expected diffusion is proportional to the time-to-go.

The way to solve these equations is to first solve eq. 1.23 for $P(t)$ with end condition $P(T) = G$. Use this solution in eq. 1.24 to compute the solution for $\alpha(t)$ with end condition $\alpha(T) = 0$. Finally,

$$\beta(s) = -\int_s^T dt\dot{\beta}dt$$

can be computed from eq. 1.25.

1.4.4 Example of LQ control

Find the optimal control for the dynamics

$$dx = (x+u)dt + d\xi, \qquad \left\langle d\xi^2 \right\rangle = \nu dt$$

with end cost $\phi(x) = 0$ and path cost $R(x,u) = \frac{1}{2}(x^2+u^2)$.

The Ricatti equations reduce to

$$\begin{aligned} -\dot{P} &= 2P + 1 - P^2 \\ -\dot{\alpha} &= 0 \\ \dot{\beta} &= -\frac{1}{2}\nu P \end{aligned}$$

with $P(T) = \alpha(T) = \beta(T) = 0$ and

$$u(x,t) = -P(t)x$$

We compute the solution for P and β by numerical integration. The result is shown in figure 1.2. The optimal control is to steer towards the origin with $-P(t)x$, where P is roughly constant until $T \approx 8$. Afterward control weakens because the total future state cost reduces to zero when t approaches the end time.

Note, that in this example the optimal control is independent of ν. It can be verified from the Ricatti equations that this is true in general for 'non-multiplicative' noise ($C_j = D_j = 0$).

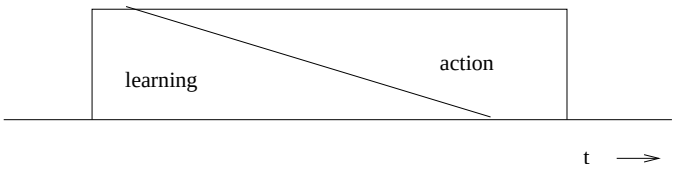


Figure 1.3: When life is finite and is executed only one time, we should first learn and then act.

1.5 Learning

So far, we have assumed that all aspects that define the control problem are known. But in many instances this is not the case. What happens if (part of) the state is not observed? For instance, as a result of measurement error we do not know x_t but only know a probability distribution $p(x_t|y_{0:t})$ given some previous observations $y_{0:t}$. Or, we observe x_t, but do not know the parameters of the dynamical equation Eq. 1.16. Or, we do not know the cost/rewards functions that appear in Eq. 1.17.

Using a Bayesian point of view, the agent can represent the uncertainty as beliefs, ie. probability distributions over the hidden states, parameters or rewards. The optimal behaviour is then a trade-off between two objectives: choosing actions that optimize the expected future reward given the current beliefs and choosing actions that improve the accuracy of the beliefs. In other words, the agent faces the problem of finding the right compromise between learning and control, a problem which is known in control theory as dual control and was originally introduced by Feldbaum (1960) (see Filatov and Unbehauen (2004) for a recent review). In addition to the observed state variables x, there are an additional number of variables θ that specify the belief distributions. The dual control solution is the ordinary control solution in this extended (x, θ) state space. The value function becomes a function of the extended state space and the Bellman equation describes the evolution in this extended state space. Some approaches to partially observed MDPs (POMDPs) Sondik (1971); Kaelbling et al. (1998); Poupart and Vlassis (2008) are an example of dual control problems.

A typical solution to the dual control problem for a finite time horizon problem is a control strategy that first chooses actions that explore the state space in order to learn a good model and use it at later times to maximize reward. In other words, the dual control problem solves the exploration exploitation problem by making explicit assumptions about the belief distributions. This is very reminiscent of our own life. Our life is finite and we have only one life. Our aim is to maximize accumulated reward during our lifetime, but in order to do so we have to allocate some resources to learning as well. It requires that we plan our learning and the learning problem becomes an integral part of the control problem. At $t = 0$, there is no knowledge of the world and thus making optimal control actions is impossible. $t = T$, learning has become useless, because we will have no more opportunity to make use of it. So we should learn early in life and act later in life, as is schematically shown in fig. 1.3. See Bertsekas (2000) for a further discussion. We discuss a simple example in section 1.5.1.

Note, that reinforcement learning is typically defined as an adaptive control problem rather than a dual control problem. These approaches use beliefs that are specified in terms of hyper parameters θ, but the optimal cost to go is still a function of the original state x only. The Bellman equation is an evolution equation for $J(x)$ where unobserved quantities are given in terms of their expected values

that depend on θ. This control problem is then in priciple solved for fixed θ (although in reinforcement learning often a sample based approach is taken and no strict convergence is enforced). θ is adapted as a result of the samples that are collected. In this formulation, the exploration exploitation dilemma arises since the control computation will propose actions that are only directed towards exploitation assuming the wrong θ (its optimal value still needs to be found). As a result, the state space is not fully explored and the updates for θ thus obtained are biased. The common heuristic to improve the learning is to mix these actions with 'exploratory actions' that explore the state space in directions that are not dictated by exploitation. Well-known examples of this appraoch are Bayesian reinforcement learning Dearden et al. (1999) and some older methods that are reviewd in Thrun (1992). Nevertheless, the principled solution is to explore all space, for instance by using a dedicated exploration strategy such as proposed in the E^3 algorithm Kearns and Singh (2002).

In the case of finite time control problems the difference between the dual control formulation and the adaptive control formulation become particularly clear. The dual control formulation requires only one trial of the problem. It starts at $t = 0$ with its initial belief θ_0 and initial state x_0 and computes the optimal solution by solving the Bellman equation in the extended state space for all intermediate times until the horizon time T. The result is a single trajectory $(x_{1:T}, \theta_{1:T})$. The adaptive control formulation requires many trials. In each trial i, the control solution is computed by solving the Bellman equation in the ordinary state space where the beliefs are given by θ^i. The result is a trajectory $(x_{1:T}, \theta^i)$. Between trials, θ is updated using samples from the previous trial(s). Thus, in the adaptive control approach the learning problem is not solved as part of the control problem but rather in an 'outer loop'. The time scale for learning is unrelated to the horizon time T. In the dual control formulation, learning must take place in a single trial and is thus tightly related to T.

Needless to say, the dual control formulation is more attactive than the adaptive control formulation, but is computationally significantly more costly.

1.5.1 Inference and control

As an example Florentin (1962); Kumar (1983), consider the simple LQ control problem

$$dx = \alpha u dt + d\xi \tag{1.26}$$

with α *unobserved* and x observed. Path cost $R(x, u, t)$ and end cost $\phi(x)$ and noise variance ν are given.

Although α is unobserved, we have a means to observe α indirectly through the sequence $x_t, u_t, t = 0, \ldots$. Each time step we observe dx and u and we can thus update our belief about α using the Bayes formula:

$$p_{t+dt}(\alpha|dx, u) \propto p(dx|\alpha, u)p_t(\alpha) \tag{1.27}$$

with $p(dx|\alpha, u)$ a Normal distribution in dx with variance νdt and $p_t(\alpha)$ a probability distribution that expresses our belief at time t about the values of α. The problem is that the future information that we receive about α depends on u: if we use a large u, the term $\alpha u dt$ is larger than the noise term $d\xi$ and we will get reliable information about α. However, large u values are more costly and also may drive us away from our target state $x(T)$. Thus, the optimal control is a balance between optimal inference and minimal control cost.

The solution is to augment the state space with parameters θ_t (sufficient statistics) that describe $p_t(\alpha) = p(\alpha|\theta_t)$ and θ_0 known, which describes our initial belief in the possible values of α. The cost that must be minimized is

$$C(x_0, \theta_0, u(0 \to T)) = \left\langle \phi(x(T)) + \int_0^T dt R(x, u, t) \right\rangle \tag{1.28}$$

where the average is with respect to the noise $d\xi$ as well as the uncertainty in α.

For simplicity, consider the example that α attains only two values $\alpha = \pm 1$. Then $p_t(\alpha|\theta) = \sigma(\alpha\theta)$, with the sigmoid function $\sigma(x) = \frac{1}{2}(1 + \tanh(x))$. The update equation Eq. 1.27 implies a dynamics for θ:

$$d\theta = \frac{u}{\nu} dx = \frac{u}{\nu}(\alpha u dt + d\xi) \tag{1.29}$$

With $z_t = (x_t, \theta_t)$ we obtain a standard HJB Eq. 1.18:

$$-\partial_t J(t,z)dt = \min_u \left(R(t,z,u)dt + \langle dz \rangle_z \partial_z J(z,t) + \frac{1}{2} \left\langle dz^2 \right\rangle_z \partial_z^2 J(z,t) \right)$$

with boundary condition $J(z,T) = \phi(x)$. The expectation values appearing in this equation are conditioned on (x_t, θ_t) and are averages over $p(\alpha|\theta_t)$ and the Gaussian noise. We compute $\langle dx \rangle_{x,\theta} = \bar{\alpha} u dt$, $\langle d\theta \rangle_{x,\theta} = \frac{\bar{\alpha} u^2}{\nu} dt$, $\left\langle dx^2 \right\rangle_{x,\theta} = \nu dt$, $\left\langle d\theta^2 \right\rangle_{x,\theta} = \frac{u^2}{\nu} dt$, $\langle dx d\theta \rangle = u dt$, with $\bar{\alpha} = \tanh(\theta)$ the expected value of α for a given value θ. The result is

$$-\partial_t J = \min_u \left(f_0(x,u,t) + \bar{\alpha} u \partial_x J + \frac{u^2 \bar{\alpha}}{\nu} \partial_\theta J + \frac{1}{2} \nu \partial_x^2 J + \frac{1}{2} \frac{u^2}{\nu} \partial_\theta^2 J + u \partial_x \partial_\theta J \right)$$

with boundary conditions $J(x, \theta, T) = \phi(x)$.

Thus, the dual control problem (joint inference on α and control problem on x) has become an ordinary control problem in x, θ. Quoting Florentin (1962): "It seems that any systematic formulation of the adaptive control problem leads to a meta-problem which is not adaptive". Note also, that dynamics for θ is non-linear (due to the u^2 term) although the original dynamics for dx was linear. The solution to this non-linear stochastic control problem requires the solution of this PDE and was studied in Kappen and Tonk (2010). An example of the optimal control solution $u(x, \theta, t)$ for $x = 2$ and different θ and t is given in fig. 1.4. Note, the 'probing' solution with u much larger when α is uncertain (θ small) then when α is certain $\theta = \pm\infty$. This exploration strategy is optimal in the dual control formulation. In Kappen and Tonk (2010) we further demonstrate that exploration is achieved through symmetry breaking in the Bellman equation; that optimal actions can be discontinuous in the beliefs (as in fig. 1.4); and that the optimal value function is typically non-differentiable. This poses a challenge for the design of value function approximations for POMDPs, which typically assumes a smooth class of functions.

1.5.2 Certainty equivalence

Although in general adaptive control is much more complex than non-adaptive control, there exists an exception for a large class of linear quadratic problems, such as the Kalman filter Theil (1957). Consider the dynamics

$$\begin{aligned} dx &= (x + u)dt + d\xi \\ y &= x + \eta \end{aligned}$$

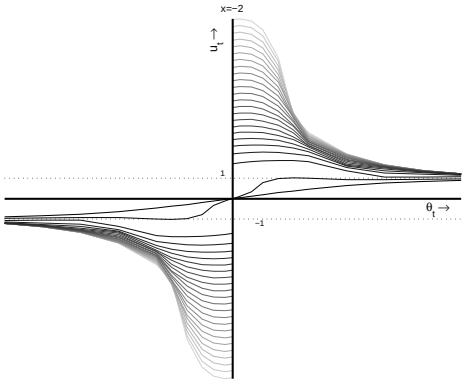


Figure 1.4: Dual control solution with end cost $\phi(x) = x^2$ and path cost $\int_t^{t_f} dt' \frac{1}{2} u(t')^2$ and $\nu = 0.5$. Plot shows the deviation of the control from the certain case: $u_t(x,\theta)/u_t(x,\theta = \pm\infty)$ as a function of θ for different values of t. The curves with the larger values are for larger times-to-go.

where now x is not observed, but y is observed and all other model parameters are known.

When x is observed, we can compute the quadratic cost, which we assume of the form

$$C(x_t, t, u_{t:T}) = \left\langle \sum_{\tau=t}^{T} \frac{1}{2}(x_\tau^2 + u_\tau^2) \right\rangle$$

We denote the optimal control solution by $u(x,t)$.

When x_t is not observed, we can compute $p(x_t|y_{0:t})$ using Kalman filtering and the optimal control minimizes

$$C_{\mathrm{KF}}(y_{0:t}, t, u_{t:T}) = \int dx_t p(x_t|y_{0:t}) C(x_t, t, u_{t:T})$$

with C as above.

Since $p(x_t|y_{0:t}) = \mathcal{N}(x_t|\mu_t, \sigma_t^2)$ is Gaussian and

$$\begin{aligned}
C_{\mathrm{KF}}(y_{0:t}, t, u_{t:T}) &= \int dx_t C(x_t, t, u_{t:T}) \mathcal{N}(x_t|\mu_t, \sigma_t^2) = \sum_{\tau=t}^{T} \frac{1}{2} u_\tau^2 + \sum_{\tau=t}^{T} \left\langle x_\tau^2 \right\rangle_{\mu_t, \sigma_t} \\
&= \sum_{\tau=t}^{T} \frac{1}{2} u_\tau^2 + \frac{1}{2}(\mu_t^2 + \sigma_t^2) + \frac{1}{2} \int dx_t \left\langle x_{t+dt}^2 \right\rangle_{x_t, \nu dt} \mathcal{N}(x_t|\mu_t, \sigma_t^2) + \cdots \\
&= \sum_{\tau=t}^{T} \frac{1}{2} u_\tau^2 + \frac{1}{2}(\mu_t^2 + \sigma_t^2) + \frac{1}{2} \left\langle x_{t+dt}^2 \right\rangle_{\mu_t, \nu dt} + \frac{1}{2}\sigma_t^2 + \cdots \\
&= C(\mu_t, t, u_{t:T}) + \frac{1}{2}(T-t)\sigma_t^2
\end{aligned}$$

The first term is identical to the observed case with $x_t \to \mu_t$. The second term does not depend on u and thus does not affect the optimal control. Thus, the optimal control for the Kalman filter $u_{\mathrm{KF}}(y_{0:t}, t)$ computed from C_{KF} is identical to the optimal control function $u(x,t)$ that is computed for the observed case C, with x_t replaced by μ_t:

$$u_{\mathrm{KF}}(y_{0:t}, t) = u(\mu_t, t)$$

This property is known as Certainty Equivalence Theil (1957), and implies that for these systems the control computation and the inference computation can be done separately, without loss of optimality.

1.6 Path integral control

1.6.1 Introduction

As we have seen, the solution of the general stochastic optimal control problem requires the solution of a partial differential equation. This is for may realistic applications not an attractive option. The alternative considered often, is to approximate the problem somehow by a linear quadratic problem which can then be solved efficiently using the Ricatti equations.

In this section, we discuss the special class of non-linear, non-quadratic control problems for which some progress can be made Kappen (2005a,b). For this class of problems, the non-linear Hamilton-Jacobi-Bellman equation can be transformed into a linear equation by a log transformation of the cost-to-go. The transformation stems back to the early days of quantum mechanics and was first used by Schrödinger to relate the Hamilton-Jacobi formalism to the Schrödinger equation (a linear diffusion-like equation). The log transform was first used in the context of control theory by Fleming (1978) (see also Fleming and Soner (1992)).

Due to the linear description, the usual backward integration in time of the HJB equation can be replaced by computing expectation values under a forward diffusion process. The computation of the expectation value requires a stochastic integration over trajectories that can be described by a path integral. This is an integral over all trajectories starting at x, t, weighted by $\exp(-S/\lambda)$, where S is the cost of the path (also know as the Action) and λ is a constant that is proportional to the noise.

The path integral formulation is well-known in statistical physics and quantum mechanics, and several methods exist to compute path integrals approximately. The Laplace approximation approximates the integral by the path of minimal S. This approximation is exact in the limit of $\nu \to 0$, and the deterministic control law is recovered.

In general, the Laplace approximation may not be sufficiently accurate. A very generic and powerful alternative is Monte Carlo (MC) sampling. The theory naturally suggests a naive sampling procedure, but is also possible to devise more efficient samplers, such as importance sampling.

We illustrate the control method on two tasks: a temporal decision task, where the agent must choose between two targets at some future time; and a simple n joint arm. The decision task illustrates the issue of spontaneous symmetry breaking and how optimal behavior is qualitatively different for high and low noise. The n joint arm illustrates how the efficient approximate inference methods (the variational approximation in this case) can be used to compute optimal controls in very high dimensional problems.

1.6.2 Path integral control

Consider the special case of Eqs. 1.16 and 1.17 where the dynamic is linear in u and the cost is quadratic in u:

$$dx_i = f_i(x,t)dt + \sum_{j=1}^{p} g_{ij}(x,t)(u_j dt + d\xi_j) \tag{1.30}$$

$$R(x,u,t) = V(x,t) + \frac{1}{2}u^T R u \tag{1.31}$$

with R a non-negative matrix. $f_i(x,t), g_{ij}(x,t)$ and $V(x,t)$ are arbitrary functions of x and t, and $\langle d\xi_j d\xi_{j'}\rangle = \nu_{jj'}dt$. In other words, the system to be controlled can be arbitrary complex and subject to arbitrary complex costs. The control instead, is restricted to the simple linear-quadratic form when $g_{ij} = 1$ and in general must act in the same subspace as the noise. We will suppress all component notation from now on. Quantities such as f, u, x, dx are vectors and R, g, ν are matrices.

The stochastic HJB equation 1.18 becomes

$$-\partial_t J = \min_u \left(\frac{1}{2} u^T R u + V + (\nabla J)^T (f + gu) + \frac{1}{2} \mathrm{Tr} \nu g^T \nabla^2 J g \right)$$

Due to the linear-quadratic appearance of u, we can minimize with respect to u explicitly which yields:

$$u = -R^{-1} g^T \nabla J \tag{1.32}$$

which defines the optimal control u for each x, t. The HJB equation becomes

$$-\partial_t J = V + (\nabla J)^T f + \frac{1}{2} \mathrm{Tr} \left(-g R^{-1} g^T (\nabla J)(\nabla J)^T + g \nu g^T \nabla^2 J \right)$$

Note, that after performing the minimization with respect to u, the HJB equation has become non-linear in J. We can, however, remove the non-linearity and this will turn out to greatly help us to solve the HJB equation. Define $\psi(x,t)$ through $J(x,t) = -\lambda \log \psi(x,t)$. We further assume that there exists a constant λ such that the matrices R and ν satisfy[2]:

$$\lambda R^{-1} = \nu \tag{1.33}$$

This relation basically says that directions in which control is expensive should have low noise variance. It can also be interpreted as saying that all noise directions are controllable (in the correct proportion). Then the HJB becomes

$$-\partial_t \psi(x,t) = \left(-\frac{V}{\lambda} + f^T \nabla + \frac{1}{2} \mathrm{Tr} \left(g \nu g^T \nabla^2 \right) \right) \psi \tag{1.34}$$

Eq. 1.34 must be solved backwards in time with $\psi(x,T) = \exp(-\phi(x)/\lambda)$.

The linearity allows us to reverse the direction of computation, replacing it by a diffusion process, in the following way. Let $\rho(y,\tau|x,t)$ describe a diffusion process for $\tau > t$ defined by the Fokker-Planck equation

$$\partial_\tau \rho = -\frac{V}{\lambda}\rho - \nabla^T (f\rho) + \frac{1}{2} \mathrm{Tr} \left(\nabla^2 (g \nu g^T \rho) \right) \tag{1.35}$$

with initial condition $\rho(y,t|x,t) = \delta(y-x)$. Note, that when $V = 0$, Eq. 1.35 describes the evolution of diffusion process Eq. 1.30 with $u = 0$.

Define $A(x,t) = \int dy \rho(y,\tau|x,t)\psi(y,\tau)$. It is easy to see by using the equations of motion Eq. 1.34 and 1.35 that $A(x,t)$ is independent of τ. Evaluating $A(x,t)$ for $\tau = t$ yields $A(x,t) = \psi(x,t)$. Evaluating $A(x,t)$ for $\tau = T$ yields $A(x,t) = \int dy \rho(y,T|x,t)\psi(x,T)$. Thus,

$$\psi(x,t) = \int dy \rho(y,T|x,t) \exp(-\phi(y)/\lambda) \tag{1.36}$$

[2]Strictly, the weaker condition $\lambda g(x,t) R^{-1} g^T(x,t) = g(x,t)\nu g^T(x,t)$ should hold.

We arrive at the important conclusion that the optimal cost-to-go $J(x,t) = -\lambda \log \psi(x,t)$ can be computed either by backward integration using Eq. 1.34 or by forward integration of a diffusion process given by Eq. 1.35. The optimal control is given by Eq. 1.32.

Both Eq. 1.34 and 1.35 are partial differential equations and, although being linear, still suffer from the curse of dimensionality. However, the great advantage of the forward diffusion process is that it can be simulated using standard sampling methods which can efficiently approximate these computations. In addition, as is discussed in Kappen (2005b), the forward diffusion process $\rho(y,T|x,t)$ can be written as a path integral and in fact Eq. 1.36 becomes a path integral. This path integral can then be approximated using standard methods, such as the Laplace approximation.

Example: linear quadratic case

The class of control problems contains both additive and multiplicative cases. We give an example of both. Consider the control problem Eqs. 1.30 and 1.31 for the simplest case of controlled free diffusion:

$$V(x,t) = 0, \quad f(x,t) = 0, \quad \phi(x) = \frac{1}{2}\alpha x^2$$

In this case, the forward diffusion described by Eqs. 1.35 can be solved in closed form and is given by a Gaussian with variance $\sigma^2 = \nu(T-t)$:

$$\rho(y,T|x,t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \tag{1.37}$$

Since the end cost is quadratic, the optimal cost-to-go Eq. 1.36 can be computed exactly as well. The result is

$$J(x,t) \quad = \quad \nu R \log\left(\frac{\sigma}{\sigma_1}\right) + \frac{1}{2}\frac{\sigma_1^2}{\sigma^2}\alpha x^2 \tag{1.38}$$

with $1/\sigma_1^2 = 1/\sigma^2 + \alpha/\nu R$. The optimal control is computed from Eq. 1.32:

$$u = -R^{-1}\partial_x J = -R^{-1}\frac{\sigma_1^2}{\sigma^2}\alpha x = -\frac{\alpha x}{R + \alpha(T-t)}$$

We see that the control attracts x to the origin with a force that increases with t getting closer to T. Note, that the optimal control is independent of the noise ν as we also saw in the previous LQ example in section 1.4.4.

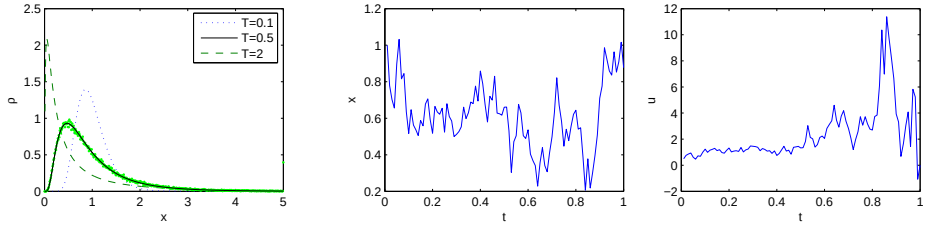
Example: multiplicative case

Consider as a simple example of a multiplicative case, $f = 0$, $g = x$, $V = 0$ in one dimension and $R = 1$. Then the forward diffusion process reduces to

$$dx = x(udt + d\xi) \tag{1.39}$$

and $x(t_i) = x_0$. If we define $y = \log x$ then

$$dy = \frac{dy}{dx}dx + \frac{1}{2}\frac{d^2y}{dx^2}dx^2 = udt + d\xi - \frac{\nu}{2}dt$$



(a) $\rho(x', t_f|x = 1, t_i = 0)$ vs. x' for $t_f = 0.1, 0.5, 2$ (b) Optimal trajectory $x(t)$ (c) Optimal control $u(t)$

Figure 1.5: Optimal control for the one-dimensional multiplicative process Eq. 1.39 with quadratic control cost $\int_{t_1}^{t_f} dt \frac{1}{2} u(t)^2$ to reach a fixed target $x' = 1$, starting from an initial position $x = 1$. Figure a) shows the forward diffusion solution in the absence of control Eq. 1.40 which is used to compute the optimal control solution Eq. 1.41.

with $y(t_i) = \log x_0$ and the solution in terms of y is simply a Gaussian distribution

$$\tilde{\rho}(y', t|y, t_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-(y' - y - (u - \nu/2)(t - t_i))^2/2\sigma^2)$$

with $\sigma^2 = (t - t_i)\nu$. In terms of x the solution becomes:

$$\rho(x', t|x, t_i) = \frac{1}{x'} \tilde{\rho}(\log x', t| \log x, t_i) \tag{1.40}$$

The solution is shown in fig. 1.5a for $u = 0$ and $t_f = 0.1, 0.5$ and $t_f = 2$. For $t_f = 0, 5$ the solution is compared with forward simulation of Eq. 1.39. Note, that the diffusion drifts towards the origin, which is caused by the state dependent noise. The noise is proportional to x and therefore the conditional probability $p(x_{\text{small}}|x_{\text{large}})$ is greater than the reverse probability $p(x_{\text{large}}|x_{\text{small}})$. This results in a netto drift towards small x.

From Eq. 1.40, we can compute the optimal control. Consider the control task to steer to a fixed end point x' from an arbitrary initial point x. Then,

$$\begin{aligned}
\psi(x, t) &= \rho(x', t_f|x, t) = \frac{1}{\sqrt{2\pi\nu T}} \frac{1}{x'} \exp\left(-(\log(x') - \log(x) + \nu T/2)^2/2\nu T\right) \\
J(x, t) &= -\nu \log \psi(x, t) = \nu \log \sqrt{2\pi\nu T} + \nu \log x' + (\log(x') - \log(x) + \nu T/2)^2/2T \\
u(x, t) &= -x \frac{dJ(x, t)}{dx} = \frac{1}{T} \log\left(\frac{x'}{x}\right) + \nu/2
\end{aligned} \tag{1.41}$$

with $T = t_f - t$. The first term attracts x to x' with strength increasing in $1/T$ as usual. The second term is a constant positive drift, to counter the tendency of the uncontrolled process to drift towards the origin. An example of the solution for a task to steer from $x = 1$ at $t = 0$ to $x = 1$ at $t = 1$ is shown in fig. 1.5b,c.

1.6.3 The diffusion process as a path integral

The diffusion equation Eq. 1.35 contains three terms. The second and third terms describe drift $f(x, t)dt$ and diffusion $g(x, t)d\xi$ as in Eq. 1.30 with $u = 0$. The first term describes a process that kills a sample trajectory with a rate $V(x, t)dt/\lambda$. This term does not conserve the probability mass. Thus, the solution of Eq. 1.35 can be

obtained by sampling the following process

$$\begin{aligned} dx &= f(x,t)dt + g(x,t)d\xi \\ x &= x + dx, \quad \text{with probability } 1 - V(x,t)dt/\lambda \\ x_i &= \dagger, \quad \text{with probability } V(x,t)dt/\lambda \end{aligned} \tag{1.42}$$

We can thus obtain a sampling estimate of

$$\begin{aligned} \Psi(x,t) &= \int dy \rho(y,T|x,t) \exp(-\phi(y)/\lambda) \\ &\approx \frac{1}{N} \sum_{i \in \text{alive}} \exp(-\phi(x_i(T))/\lambda) \end{aligned} \tag{1.43}$$

by computing N trajectories $x_i(t \to T), i = 1, \ldots, N$. Each trajectory starts at the same value x and is sampled using the dynamics Eq. 1.43. 'Alive' denotes the subset of trajectories that do not get killed along the way by the † operation.

The diffusion process can formally be 'solved' as a path integral. We restrict ourselves to the simplest case $g_{ij}(x,t) = \delta_{ij}$. The general case can also be written as a path integral, but is somewhat more involved. The argument follows simply by splitting the time interval $[t,T]$ is a large number n of infinitesimal intervals $[t, t+dt]$. For each small interval, $\rho(y, t+dt|x,t)$ is a product of a Gaussian distribution due to the drift f and diffusion $gd\xi$, and the annihilation process $\exp(-V(x,t)dt/\lambda)$: $\rho(y, t+dt|x,t) = \mathcal{N}(y|x+f(x,t)dt, \nu)$. We can then compute $\rho(y,T|x,t)$ by multiplying all these infinitesimal transition probabilities and integrating the intermediate variables y. The result is

$$\begin{aligned} \rho(y,T|x,t) &= \int [dx]_x^y \exp\left(-\frac{1}{\lambda} S_{\text{path}}(x(t \to T))\right) \\ S_{\text{path}}(x(t \to T)) &= \int_t^T d\tau \frac{1}{2} (\dot{x}(\tau) - f(x(\tau),\tau))^T R (\dot{x}(\tau) - f(x(\tau),\tau)) \\ &+ \int_t^T d\tau V(x(\tau),\tau) \end{aligned} \tag{1.44}$$

Combining Eq. 1.44 and Eq. 1.36, we obtain the cost-to-go as

$$\begin{aligned} \Psi(x,t) &= \int [dx]_x \exp\left(-\frac{1}{\lambda} S(x(t \to T))\right) \\ S(x(t \to T)) &= S_{\text{path}}(x(t \to T)) + \phi(x(T)) \end{aligned} \tag{1.45}$$

Note, that Ψ has the general form of a partition sum. S is the energy of a path and λ the temperature. The corresponding probability distribution is

$$p(x(t \to T)|x,t) = \frac{1}{\Psi(x,t)} \exp\left(-\frac{1}{\nu} S(x(t \to T))\right)$$

$J = -\lambda \log \Psi$ can be interpreted as a free energy. See Kappen (2005b) for details.

Although we have solved the optimal control problem formally as a path integral, we are still left with the problem of computing the path integral. Here one can resort to various standard methods such as Monte Carlo sampling Kappen (2005b) of which the naive forward sampling Eq. 1.42 is an example. One can however, improve on this naive scheme using importance sampling where one changes the drift term such as to minimize the annihilation of the diffusion by the $-V(x,t)dt/\lambda$ term.

A particularly cheap approximation is the Laplace approximation, that finds the trajectory that minimizes S in Eq. 1.45. This approximation is exact in the limit of $\lambda \to 0$ which is the noiseless limit. The Laplace approximation gives the classical path. On particular effective forward importance sampling method is to use the classical path as a drift term. We will give an example of the naive and importance forward sampling scheme below for the double slit problem.

One can also use a variational approximation to approximate the path integral using the variational approach for diffusion processes Archambeau et al. (2008), or use the EP approximation Mensink et al. (2010). An illustration of the variational approximation to a particular simple n joint arm is presented in section 1.7.2.

1.7 Approximate inference methods for control

1.7.1 MC sampling

In this section, we illustrate the path integral control method for the simple example of a double slit. The example is sufficiently simple that we can compute the optimal control solution in closed form. We use this example to compare the Monte Carlo and Laplace approximations to the exact result.

Consider a stochastic particle that moves with constant velocity from t to T in the horizontal direction and where there is deflecting noise in the x direction:

$$dx = udt + d\xi \tag{1.46}$$

The cost is given by Eq. 1.31 with $\phi(x) = \frac{1}{2}x^2$ and $V(x, t_1)$ implements a slit at an intermediate time t_1, $t < t_1 < T$:

$$\begin{aligned} V(x,t_1) &= 0, \quad a < x < b, \quad c < x < d \\ &= \infty, \quad \text{else} \end{aligned}$$

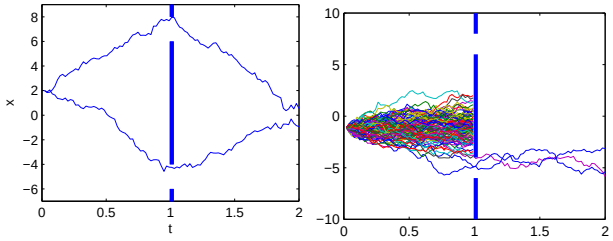
The problem is illustrated in Fig. 1.6a where the constant motion is in the t (horizontal) direction and the noise and control is in the x (vertical) direction.

The cost to go can be solved in closed form. The result for $t > t_1$ is a simple linear quadratic control problem for which the solution is given by Eq. 1.38 and for $t < t_1$ is Kappen (2005b):

$$\begin{aligned} J(x,t) &= \nu R \log\left(\frac{\sigma}{\sigma_1}\right) + \frac{1}{2}\frac{\sigma_1^2}{\sigma^2}x^2 \\ &- \nu R \log\frac{1}{2}\left(F(b,x) - F(a,x) + F(d,x) - F(c,x)\right) \\ F(x_0,x) &= \text{Erf}\left(\sqrt{\frac{A}{2\nu}}(x_0 - \frac{B(x)}{A})\right) \\ A &= \frac{1}{t_1 - t} + \frac{1}{R + T - t_1} \qquad B(x) = \frac{x}{t_1 - t} \end{aligned} \tag{1.47}$$

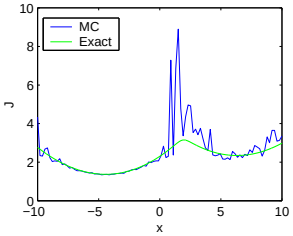
The solution Eq. 1.47 is shown for $t = 0$ in fig. 1.6b. We can compute the optimal control from Eq. 1.32.

We assess the quality of the naive MC sampling scheme, as given by Eqs. 1.42 and 1.43 in fig. 1.6b,c. Fig. 1.6b shows the sampling trajectories of the naive MC sampling procedure for one particular value of x. Note, the inefficiency of the sampler because most of the trajectories are killed at the infinite potential at $t = t_1$. Fig. 1.6c shows the accuracy of the naive MC sampling estimate of $J(x, 0)$ for all x between -10 and 10 using $N = 100000$ trajectories. We note, that the number

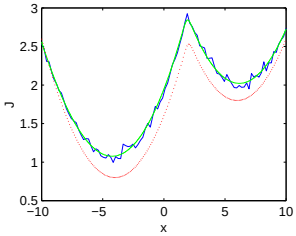


(a) Controlled trajectories

(b) Naive MC sampling trajectories



(c) Naive MC sampling results



(d) Laplace approximation and importance sampling result

Figure 1.6: Double slit experiment. (a) Set-up of the experiment. Particles travel from $t = 0$ to $t = 2$ under dynamics Eq. 1.46. A slit is placed at time $t = t_1$, blocking all particles by annihilation. Two trajectories are shown under optimal control. (b) Naive Monte Carlo sampling trajectories to compute $J(x = -1, t = 0)$ through Eq. 1.43. Only trajectories that pass through a slit contribute to the estimate. (c) Comparison of naive MC estimates with $N = 100000$ trajectories and exact result for $J(x, t = 0)$ for all x. (d) Comparison of Laplace approximation (dotted line) and Monte Carlo importance sampling (solid jagged line) of $J(x, t = 0)$ with exact result Eq. 1.47 (solid smooth line). The importance sampler used $N = 100$ trajectories for each x.

of trajectories that are required to obtain accurate results, strongly depends on the initial value of x due to the annihilation at $t = t_1$. As a result, low values of the cost-to-go are more easy to sample accurately than high values.

In addition, the efficiency of the sampling procedures depends strongly on the noise level. For small noise, the trajectories spread less by themselves and it is harder to generate trajectories that do not get annihilated. In other words, sampling becomes more accurate for high noise, which is a well-known general feature of sampling.

The sampling is of course particularly difficult in this example because of the infinite potential that annihilates most of the trajectories. However, similar effects will be observed in general due to the multi-modality of the Action.

We can improve the sampling procedure using the importance sampling procedure using the Laplace approximation (see Kappen (2005b)). The Laplace approximation in this case are the two piece-wise linear trajectories that pass through one of the slits to the goal. The Laplace approximation and the results of the importance sampler are given in fig. 1.6d. We see that the Laplace approximation is quite good for this example, in particular when one takes into account that a constant shift in J does not affect the optimal control. The MC importance sampler dramatically improves over the naive MC results in fig. 1.6, in particular since 1000 times less samples are used and is also significantly better than the Laplace approximation.

1.7.2 The variational method

In this example we illustrate the use of the variational approximation for optimal control computation. We consider a particularly simple realization of an n joint arm in two dimensions. We will demonstrate how this approximation will be useful even for large n.

Consider an arm consisting of n joints of length 1. The location of the ith joint in the 2d plane is

$$x_i = \sum_{j=1}^{i} \cos\theta_i$$
$$y_i = \sum_{j=1}^{i} \sin\theta_i$$

with $i = 1, \ldots, n$. Each of the joint angles is controlled by a variable u_i. The dynamics of each joint is

$$d\theta_i = u_i dt + d\xi_i, \quad i = 1, \ldots, n$$

with $d\xi_i$ independent Gaussian noise with $\left\langle d\xi_i^2 \right\rangle = \nu dt$. Denote by $\vec{\theta}$ the vector of joint angles, and $\vec{u}$ the vector of controls. The expected cost for the control path $\vec{u}_{t:T}$ is

$$\begin{aligned} C(\vec{\theta}, t, \vec{u}_{t:T}) &= \left\langle \phi(\theta(T)) + \int_t^T \frac{1}{2}\vec{u}^T(t)\vec{u}(t) \right\rangle \\ \phi(\vec{\theta}) &= \frac{\alpha}{2}\left((x_n(\vec{\theta}) - x_{\text{target}})^2 + (y_n(\vec{\theta}) - y_{\text{target}})^2 \right) \end{aligned}$$

with $x_{\text{target}}, y_{\text{target}}$ the target coordinates of the end joint.

Because the state dependent path cost V and the intrinsic dynamics of f are zero, the solution to the diffusion process Eq. 1.42 that starts with the arm in the

configuration $\vec{\theta^0}$ is a Gaussian so that Eq. 1.36 becomes [3]

$$\Psi(\vec{\theta^0}, t) = \int d\vec{\theta} \left(\frac{1}{\sqrt{2\pi\nu(T-t)}} \right)^n \exp\left(-\sum_{i=1}^{n} (\theta_i - \theta_i^0)^2 / 2\nu(T-t) - \phi(\vec{\theta})/\nu \right)$$

The control at time t for all components i is computed from Eq. 1.32 and is given by

$$u_i = \frac{1}{T-t} \left(\langle \theta_i \rangle - \theta_i^0 \right) \tag{1.48}$$

where $\langle \theta_i \rangle$ is the expectation value of θ_i computed wrt the probability distribution

$$p(\vec{\theta}) = \frac{1}{\Psi(\vec{\theta^0}, t)} \exp\left(-\sum_{i=1}^{n} (\theta_i - \theta_i^0)^2 / 2\nu(T-t) - \phi(\vec{\theta})/\nu \right) \tag{1.49}$$

Thus, the stochastic optimal control problem reduces the inference problem to compute $\langle \theta_i \rangle$. There are several ways to compute this. One can use a simple importance sampling scheme, where the proposal distribution is the n dimensional Gaussian centered on $\vec{\theta^0}$ (first term in Eq. 1.49) and where samples are weighted with $\exp(-\phi(\vec{\theta})/\nu)$. I tried this, but that does not work very well (results not shown). One can also use a Metropolis Hastings methods with a Gaussian proposal distribution. This works quite well (results not shown). One can also use a very simple variational method which we will now discuss.

We compute the expectations $\left\langle \vec{\theta} \right\rangle$ by introducing a factorized Gaussian variational distribution $q(\vec{\theta}) = \prod_{i=1}^{n} \mathcal{N}((\theta_i | \mu_i, \sigma_i)$ that will serve as an approximation to $p(\vec{\theta})$ in Eq. 1.49. We compute μ_i and σ_i by by minimizing the KL divergence between $q(\vec{\theta})$ and $p(\vec{\theta})$:

$$KL = \int d\theta q(\theta) \log \frac{q(\theta)}{p(\theta)}$$

$$= -\sum_{i=1}^{n} \log \sqrt{2\pi\sigma_i^2} + \log \Psi(\vec{\theta^0}, t) + \frac{1}{2\nu(T-t)} \sum_{i=1}^{n} \left(\sigma_i^2 + (\mu_i - \theta_i^0)^2 \right) + \frac{1}{\nu} \left\langle \phi(\vec{\theta}) \right\rangle_q$$

where we omit irrelevant constants. Because ϕ is quadratic in x_n and y_n and these are defined in terms of sines and cosines, the $\left\langle \phi(\vec{\theta}) \right\rangle$ can be computed in closed form. The computation of the variational equations result from setting the derivative of the KL with respect to μ_i and σ_i^2 equal to zero. The result is

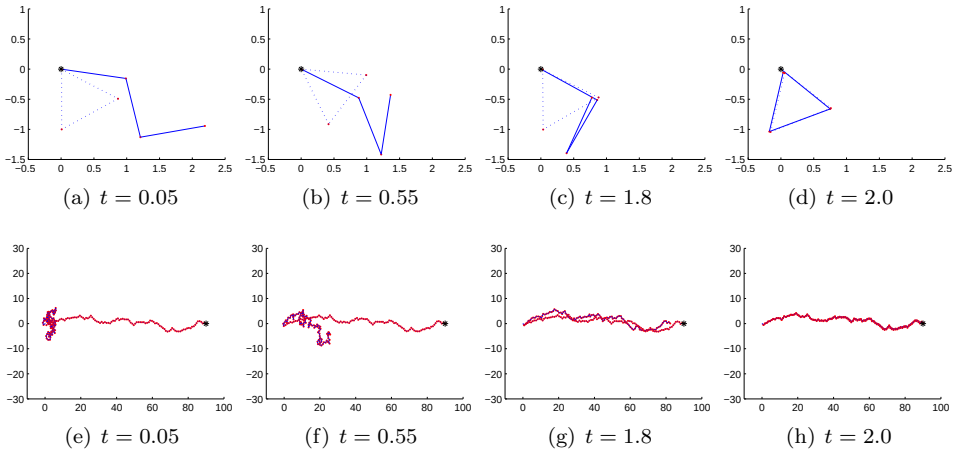
$$\mu_i \leftarrow \theta_i^0 + \alpha(T-t) \left(\sin \mu_i e^{-\sigma_i^2/2} (\langle x_n \rangle - x_{\text{target}}) - \cos \mu_i e^{-\sigma_i^2/2} (\langle y_n \rangle - y_{\text{target}}) \right)$$

$$\frac{1}{\sigma_i^2} \leftarrow \frac{1}{\nu} \left(\frac{1}{(T-t)} + \alpha e^{-\sigma_i^2} - \alpha \left(\langle x_n \rangle - x_{\text{target}} \right) \cos \mu_i e^{-\sigma_i^2/2} - \alpha \left(\langle y_n \rangle - y_{\text{target}} \right) \sin \mu_i e^{-\sigma_i^2/2} \right)$$

After convergence the estimate for $\langle \theta_i \rangle = \mu_i$.

The problem is illustrated in fig. 1.7 Note, that the computation of $\langle \theta_i \rangle$ solves the coordination problem between the different joints. Once $\langle \theta_i \rangle$ is known, each θ_i is steered independently to its target value $\langle \theta_i \rangle$ using the control law Eq. 1.48. The computation of $\langle \theta_i \rangle$ in the variational approximation is very efficient and can be used to control arms with hundreds of joints.

[3] This is not exactly correct because θ is a periodic variable. One should use the solution to diffusion on a circle instead. We can ignore this as long as $\sqrt{\nu(T-t)}$ is small compared to 2π.



(a) $t = 0.05$ (b) $t = 0.55$ (c) $t = 1.8$ (d) $t = 2.0$

(e) $t = 0.05$ (f) $t = 0.55$ (g) $t = 1.8$ (h) $t = 2.0$

Figure 1.7: (a-d) Path integral control of a $n = 3$ joint arm. The objective is that the end joint reaches a target location at the end time $T = 2$. Solid line: current joint configuration in Cartesian coordinates $(\vec{x}, \vec{y})$ corresponding to the angle state $\vec{\theta}_0$ at time t. Dashed: expected joint configuration computed at the horizon time $T = 2$ corresponding to the expected angle state $\left\langle \vec{\theta} \right\rangle$ from Eq. 1.49 with $\vec{\theta}^0$ the current joint position. Target location of the end effector is at the origin, resulting in a triangular configuration for the arm. As time t increases, each joint moves to its expected target location due to the control law Eq. 1.48. At the same time the expected configuration is recomputed, resulting in a different triangular arm configuration. (e-h). Same, with $n = 100$.

1.8 Discussion

In this paper, we have given a basic introduction to some notions in optimal deterministic and stochastic control theory and have discussed recent work on the path integral methods for stochastic optimal control. We would like to mention a few additional issues.

One can extend the path integral control formalism to multiple agents that jointly solve a task. In this case the agents need to coordinate their actions not only through time, but also among each other to maximize a common reward function. The approach is very similar to the n-joint problem that we studied in the last section. The problem can be mapped on a graphical model inference problem and the solution can be computed exactly using the junction tree algorithm Wiegerinck et al. (2006, 2007) or approximately Broek et al. (2008b,a).

There is a relation between the path integral approach discussed and the linear control formulation proposed in Todorov (2007). In that work the discrete space and time case is considered and it is shown, that if the immediate cost can be written as a KL divergence between the controlled dynamics and a passive dynamics, the Bellman equation becomes linear in a very similar way as we derived for the continuous case in Eq. 1.34. In Todorov (2008) it was further observed that the linear Bellman equation can be interpreted as a backward message passing equation in a HMM.

In Kappen et al. (2009) we have taken this analogy one step further. When the immediate cost is a KL divergence between transition probabilities for the controlled and passive dynamics, the total control cost is also a KL divergence between probability distributions describing controlled trajectories and passive trajectories. Therefore, the optimal control solution can be directly inferred as a Gibbs distribu-

tion. The optimal control computation reduces to the probabilistic inference of a marginal distribution on the first and second time slice. This problem can be solved using efficient approximate inference methods. We also show how the path integral control problem is obtained as a special case of this KL control formulation.

The path integral approach has recently been applied to the control of character animation da Silva et al. (2009). In this work the linearity of the Bellman equation Eq. 1.34 and its solution Eq. 1.36 is exploited by noting that if ψ_1 and ψ_2 are solutions for end costs ϕ_1 and ϕ_2, then $\psi_1 + \psi_2$ is a solution to the control problem with end cost $-\lambda \log \left(\exp(-\phi_1/\lambda) + \exp(-\phi_2/\lambda)\right)$. Thus, by computing the control solution to a limited number of archetypal tasks, one can efficiently obtain solutions for arbitrary combinations of these tasks.

In robotics, Theodorou et al. (2009, 2010); Evangelos A. Theodorou (2010) has shown the the path integral method has great potential for application in robotics. They have compared the path integral method with some state-of-the-art reinforcement learning methods, showing very significant improvements. In addition, they have successful implemented the path integral control method to a walking robot dog.

Acknowledgments

This work is supported in part by the Dutch Technology Foundation and the BSIK/ICIS project.

Bibliography

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. (2008). Variational inference for diffusion processes. In Koller, D. and Singer, Y., editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA. MIT Press.

Bellman, R. and Kalaba, R. (1964). *Selected papers on mathematical trends in control theory*. Dover.

Bertsekas, D. (2000). *Dynamic Programming and optimal control*. Athena Scientific, Belmont, Massachusetts. Second edition.

Broek, B., W., W., and Kappen, H. (2008a). Graphical model inference in optimal control of stochastic multi-agent systems. *Journal of AI Research*, 32:95–122.

Broek, B. v. d., Wiegerinck, W., and Kappen, H. (2008b). Optimal control in large stochastic multi-agent systems. In *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning*, volume 4865/2008, pages 15–26, Berlin / Heidelberg. Springer.

da Silva, M., Durand, F., and Popović, J. (2009). Linear bellman combination for control of character animation. In *SIGGRAPH '09: ACM SIGGRAPH 2009 papers*, pages 1–10, New York, NY, USA. ACM.

Dearden, R., Friedman, N., and Andre, D. (1999). Model based bayesian exploration. In *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–159.

Evangelos A. Theodorou, Jonas Buchli, S. S. (2010). reinforcement learning of motor skills in high dimensions: a path integral approach. In *international conference of robotics and automation (icra 2010) - accepted*.

Feldbaum, A. (1960). Dual control theory. I-IV. *Automation remote control*, 21-22:874–880, 1033–1039, 1–12, 109–121.

Filatov, N. and Unbehauen, H. (2004). *Adaptvie dual control*. Springer Verlag.

Fleming, W. (1978). Exit probabilties and optimal stochastic control. *Applied Math. Optim.*, 4:329–346.

Fleming, W. and Soner, H. (1992). *Controlled Markov Processes and Viscosity solutions*. Springer Verlag.

Florentin, J. (1962). Optimal, probing, adaptive control of a simple bayesian system. *International Journal of Electronics*, 13:165–177.

Fraser-Andrews, G. (1999). A multiple-shooting technique for optimal control. *Journal of Optimization Theory and Applications*, 102:299–313.

Goldstein, H. (1980). *Classical mechanics*. Addison Wesley.

Heath, M. (2002). *Scientific Computing: An Introductory Survey*. McGraw-Hill, New York. 2nd Edition.

Jönsson, U., Trygger, C., and Ögren, P. (2002). Lectures on optimal control. Unpublished.

Kaelbling, L., Littman, M., and Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *ARtificail Intelligence*, 101:99–134.

Kappen, H. (2005a). A linear theory for control of non-linear stochastic systems. *Physical Review Letters*, 95:200201.

Kappen, H. (2005b). Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and Experiment*, page P11011.

Kappen, H., Gómez, V., and Opper, M. (2009). Optimal control as a graphical model inference problem. *Journal of Machine Learning Research*. Submitted,`http://arxiv.org/abs/0901.0633`.

Kappen, H. and Tonk, S. (2010). Optimal exploration as a symmetry breaking phenomenon. In *Advances in Neural Information Processing Systems*. Rejected.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, pages 209–232.

Kumar, P. R. (1983). Optimal adaptive control of linear-quadratic-gaussian systems. *SIAM Journal on Control and Optimization*, 21(2):163–178.

Mensink, T., Verbeek, J., and Kappen, H. (2010). Ep for efficient stochastic control with obstacles. In *Advances in Neural Information Processing Systems*. Rejected.

Pontryagin, L., Boltyanskii, V., Gamkrelidze, R., and Mishchenko, E. (1962). *The mathematical theory of optimal processes*. Interscience.

Poupart, P. and Vlassis, N. (2008). Model-based bayesian reinforcement learning in partially observable domains. In *Proceedings International Symposium on Artificial Intelligence and Mathematics (ISAIM)*.

Sondik, E. (1971). *The optimal control of partially observable Markov processes*. PhD thesis, Stanford University.

Stengel, R. (1993). *Optimal control and estimation*. Dover publications, New York.

Theil, H. (1957). A note on certainty equivalence in dynamic planning. *Econometrica*, 25:346–349.

Theodorou, E., Buchli, J., and Schaal, S. (2010). Learning policy improvements with path integrals. In *international conference on artificial intelligence and statistics (aistats 2010) - accepted*.

Theodorou, E. A., Buchli, J., and Schaal, S. (2009). path integral-based stochastic optimal control for rigid body dynamics. In *adaptive dynamic programming and reinforcement learning, 2009. adprl '09. ieee symposium on*, pages 219–225.

Thrun, S. B. (1992). The role of exploration in learning control. In White, D. and Sofge, D., editors, *Handbook of intelligent control*. Multiscience Press.

Todorov, E. (2007). Linearly-solvable markov decision problems. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1369–1376. MIT Press, Cambridge, MA.

Todorov, E. (2008). General duality between optimal control and estimation. In *47th IEEE Conference on Decision and Control*, pages 4286–4292.

Weber, R. (2006). Lecture notes on optimization and control. Lecture notes of a course given autumn 2006.

Wiegerinck, W., Broek, B. v. d., and Kappen, H. (2006). Stochastic optimal control in continuous space-time multi-agent systems. In *Uncertainty in Artificial Intelligence. Proceedings of the 22th conference*, pages 528–535. Association for UAI.

Wiegerinck, W., Broek, B. v. d., and Kappen, H. (2007). Optimal on-line scheduling in stochastic multi-agent systems in continuous space and time. In *Proceedings AAMAS*, page 8 pages.

Yong, J. and Zhou, X. (1999). *Stochastic controls. Hamiltonian Systems and HJB Equations*. Springer.